Understanding the effect of gene-smoking interactions on lung function and COPD risk utilising UK Biobank

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Carl A Melbourne BSc MSc

Department of Health Sciences

University of Leicester

February 2019

Abstract

Understanding the effect of gene-smoking interactions on lung function and COPD risk utilising UK Biobank

Carl A Melbourne

Chronic obstructive pulmonary disease (COPD), characterised by severe airflow obstruction, is a leading cause of mortality worldwide. Smoking is the biggest risk factor, however COPD and the lung function measures key to its diagnosis, have a strong genetic component. Known loci account for a small proportion of lung function heritability, and not all smokers develop COPD. This thesis questions whether smoking and genetic effects on lung function are independent, aiming to identify novel gene-smoking interactions, to aid in genetic risk prediction and treatment development.

The literature review undertaken here examines a range of approaches for interaction analyses. I applied two such methods using simulation studies to determine the power to detect interaction effects.

Interaction analyses were then undertaken utilising UK Biobank (n~500k). Firstly, as regions of the genome containing lung function associated loci might be more likely to contain SNPs producing interaction effects, a candidate region interaction analysis was undertaken. Two SNPs were identified, driven by stronger genetic effects in ever smokers. Secondly, I meta-analysed data from UK Biobank and the SpiroMeta consortium (n~400k) and determined that none of the 279 lung function and COPD associated signals reported to date had individually differing genetic effects between ever and never smokers. Thirdly, I undertook the largest genome-wide gene-smoking interaction analysis to date and identified 53 genetic loci that interact with smoking behaviour. Replication efforts were penalised by small effective sample sizes, but results provide direction for further research.

This work informs current estimates of relative and absolute risk for poor lung function and COPD due to genetic risk and smoking status. The 55 loci identified could have clinical importance by providing personalised risk prediction and treatment. To further understand the effect of gene-smoking interactions on lung function and COPD risk, larger replication sample sizes with better imputation quality are needed.

Acknowledgments

I would like to thank Professor Louise Wain and Professor Martin Tobin for their unwavering support and guidance during the completion of my PhD. I would also like to thank the University of Leicester Genetic Epidemiology group as a whole, who have made it a pleasure to study here.

I could not ask for better parents. They have been supportive throughout all of my academic achievements and I am very lucky to have them.

Finally, I would like to thank my partner Lauren, who has given me love, encouragement and reassurance whilst completing my PhD, and my step daughter Amelie, who is the smartest and funniest eight year old I know.

Publications related to this thesis

Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. - Nat Genet.2019 Mar;51(3):481-493.doi: 10.1038/s41588-018-0321-7.Epub 2019 Feb 25. (1546-1718 (Electronic); 1061-4036 (Linking)).

Table of Contents

List of Tables

List of Figures

Abbreviations and glossary

Term	Description
AAT	Alpha1-antitrypsin - A lung protecting protein
Additive model	A genetic model in which each copy of the effect allele linearly increases the genetic effect on the phenotype
A (Adenine)	One of the four nitrogenous bases in DNA that pairs with thymine
Allele	The name given to each of the possible bases at a SNP locus (i.e. A, C, G or T)
Amino acid	The individual units which join together to create the protein structure
Autosome	The name given to all chromosomes which are not the sex chromosomes (or allosome)
Base pair	A pair of nucleotides connected by a hydrogen bond
Broad-sense heritability	The observed variance of a trait attributable to genetic effect of all types (additive, dominant, interactions etc.)
Candidate gene study	A study in which analysis focusses on genes previously implicated by prior research or beliefs
Chromosome	A structure which contains a section of human DNA. Humans carry 23 pairs of chromosomes
CNV	Copy number variation - A type of mutation where sections of DNA are repeated or deleted
Codon	A group of three nucleotides that are part of the mRNA molecule and are responsible for a certain amino acid

Term	Description
COPD	Chronic obstructive pulmonary disease – A collective term for a number of lung diseases which cause progressive breathing problems e.g. emphysema
CPD	Cigarettes per day – the number of cigarettes smoked per day
(C) Cytosine	One of the four nitrogenous bases in DNA that pairs with guanine
DNA	Deoxyribonucleic acid - A molecule which carries our genetic code
Dominant model	A genetic model in which one copy of the effect allele carries the same risk as two copies of the effect allele
Exon / exonic	Part of a gene which encodes for a protein
FEV_1	Forced expiratory volume in 1 second – The volume of air forcibly exhaled after the first second during a spirometry test
FVC	Forced vital capacity – The maximum volume of air forcibly exhaled during a spirometry test
Gamete	A sex cell (sperm for male and ovum for female)
Gene	Fixed genomic positions which carry hereditary information and determine human characteristics
Genetic linkage	Variants close in proximity that are unaffected during meiosis and therefore inherited together
Genome	The complete genetic complement of a human
Genomic control	The act of scaling test statistics by a factor of λ to control for population stratification
Genotype	The two alleles taken from a consistent strand on a chromosomal pair at a SNP
Genotypic model	A genetic model in which the heterozygous and homozygous effect allele genotypes carry individual

Term	Description
	risks (non-linear per increase in effect allele in contrast to additive model)
Genotyping	The process of categorising individuals into genotype groups using allele probe intensities
GOLD	Global Initiative for Chronic Obstructive Lung Disease – The rules for categorising COPD severity
(G) Guanine	One of the four nitrogenous bases in DNA that pairs with cytosine
GWAS	Genome-wide association study – An analysis which tests variants across the genome for association with the phenotype
Haplotype	A group of SNPs or genes inherited together from a single parent
Hardy-Weinberg equilibrium	Assumes that genotype probabilities are stable from one generation to the next assuming no other influences
Heritability	The observed variance of a trait attributable to genetic effect
Heterozygosity	The proportion of genotypes which are heterozygous
Heterozygous genotype	A genotype constructed of the two differing alleles at a SNP locus
Homozygous genotype	A genotype constructed of two of the same alleles at a SNP locus
IBD	Identity by descent - The number or proportion of alleles shared between individuals because of a common ancestor
IBS	Identity by state - The number or proportion of alleles shared between individuals
Imputation	The process of predicting missing genotypes that were not directly genotyped using a reference panel

Term	Description
Indel	The insertion or deletion of a single base or small number of bases at a locus
Intergenic	A region of the genome that lies between genes
Intron / Intronic	The part of the gene removed during splicing
Inversion	A type of mutation where a section of DNA is reversed
Karyotype	The visual representation and number of chromosomes for an individual
λ (lambda)	A metric used to determine the inflation of observed test statistics compared with the expected test statistics in a genome-wide analysis
LD	Linkage Disequilibrium - The observed correlation (non-random association) between variants along the genome in a certain population
MAC	Minor allele count – The count of the less common allele at a SNP locus in the population being studied
MAF	Minor allele frequency – The frequency of the less common allele at a SNP locus
Mb (Megabase)	1 million base pairs (used to describe genomic distance)
Meiosis	A process of cell division used in the production of gametes
Missense	A mutation in which a nucleotide change replaces one codon for another and thus in turn produces a different amino acid
mRNA	Messenger Ribonucleic acid - A molecule created during splicing that encodes for a protein
Narrow-sense heritability	The observed variance of a trait attributable to additive genetic effect only

Term	Description
Nonsense	A mutation in which a nucleotide change replaces a codon for a stop codon
Non-synonymous	A type of SNP that modifies the amino acid sequence of a gene
Nucleotide	A nitrogenous base which forms the basis of DNA – Adenines (A), Cytosines (C), Guanines (G) and Thymines (T)
PEF	Peak expiratory flow – The maximum flow (volume per unit of time) achieved during a spirometry test
Percent predicted FEV_1 (%pred FEV_1)	The ratio of observed FEV_1 and predicted FEV_1 used to determine COPD severity
Phenotype	The outcome variable (the disease or trait being analysed)
Population stratification	When allele frequencies differ between populations which may cause result bias in genetic association analysis
Principal component analysis	Reducing multi-dimensional data into a number of uncorrelated linear variables
Protein	A molecule consisting of chains of amino acids which determine traits and functions in a human
Putative aneuploidy	A considered "abnormal" number of chromosomes e.g. XXY instead of XX or XY for a sex chromosome pairing
PY	Pack years – a trait calculated using both the number of cigarettes smoked per day and the number of years smoked
QC	Quality control
Recessive model	A genetic model in which two copies of the recessive allele are needed to increase risk

Term	Description
Recombination	A process in which blocks of DNA from each chromosomal pair are exchanged to create a new chromosomal pair during meiosis
RNA	Ribonucleic acid – A molecule created during DNA transcription in which the thymine (T) is substituted for a Uracil (U)
SC	Smoking cessation – whether an individual is a current or former smoker
Sex mismatch	Contradiction between recorded sex and sex inferred from genetic data
SI	Smoking initiation – whether an individual is an ever or never smoker
Silent substitution	A mutation in which a codon is replaced by another which infers the same amino acid
SNP or SNV	Single nucleotide polymorphism/variant – When a nucleotide is substituted for another at a genome locus
Spirometry	The act of measuring the volume of air forcibly exhaled by an individual
Splice site	A region where splicing occurs
Splicing	A process where introns are removed and the exons are joined together
Strand	One of the two nucleotide sequences (one side of the double helix DNA structure) in a single copy of a chromosome
Synonymous	A type of SNP which does not modify the amino acid sequence of a gene
T (Thymine)	One of the four nitrogenous bases in DNA that pairs with adenine

Term	Description
Transcription	A process where nucleotide sequences of a chromosome break apart and one strand produces a complimentary strand of RNA
U (Uracil)	A base which replaces a thymine in RNA

Chapter 1 Introduction

This chapter introduces genetic epidemiology and the biological concepts which underpin genetic epidemiological research. The basic statistical application to genetic association analysis will then be described, before introducing genome-wide association studies (GWAS) and the approaches used both pre and post analysis. Lung function and chronic obstructive pulmonary disease (COPD) will then be introduced, with a discussion of our current knowledge as to the role of genetics in COPD, poor lung function susceptibility, and smoking (the biggest risk factor for poor lung health). The UK Biobank, the core data resource for this thesis will also be described, before concluding the chapter by outlining the overall aims of this thesis, as well as the structure for the proceeding chapters.

1.1 Overview of genetic epidemiology

Genetic epidemiology is the study of the effect of familial inheritance and also heritance factors within populations, on disease aetiology (1,2). These concepts have resulted in discoveries of thousands of links between genetic variation and disease. Diseases are often segregated by genetic characteristics into two categories. Some diseases may either be largely or solely driven by a particular mutation or genetic variant, and these are referred to as monogenic or mendelian diseases (based on the work by Gregor Mendel in the 19th century) (3). Alternatively, diseases can be driven by multiple mutations or genes, and these are referred to as complex or polygenic diseases (4). The aim is to increase our knowledge of how an individual's genetic make-up can affect their risk of disease, initially from a statistical standpoint, but ultimately by producing biological insight to aid in risk prediction and the development of new treatment, whether preventative and/or curative. To do this requires knowledge and application of both the biological and statistical theory underpinning genetic epidemiology. These concepts are introduced in the proceeding sections.

1.1.1 Biological theory

The human genome (the complete genetic complement of a human) is made of DNA (deoxyribonucleic acid). DNA is characterised through the pairing of two nucleotide

sequences, with each nucleotide made up of a nitrogenous base, known as Thymines (T), Cytosines (C), Guanines (G) or Adenines (A), a phosphate group, and a sugar (deoxyribose). Strong covalent bonds are formed between the phosphate of one nucleotide and the sugar of the next to produce a nucleotide sequence (5). Two of these nucleotide sequences bind together via hydrogen bonds, with Adenines from one nucleotide sequence always pairing with Thymines from the other, and Guanines always pairing with Cytosines, to form base pairs. This in turn produces the double helix DNA structure (**Figure 1.1**). Each DNA strand's directionality is described by a 3' end and a 5' end, with each strand running in an opposite direction. There are an estimated 3.2 billion base pairs within the human genome (5), and these are partitioned into 46 chromosomes. There are 22 pairs of autosomes and 2 sex chromosomes (with a male carrying an XY sex chromosome pairing and a female carrying XX) (5,6).

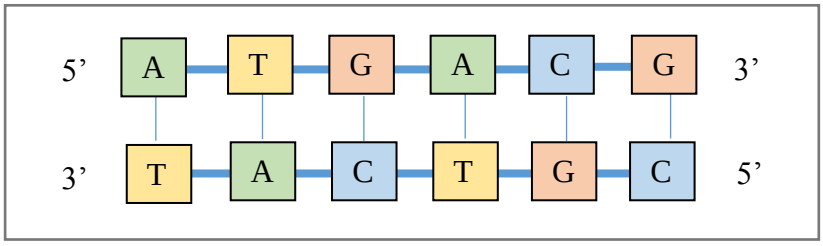


Figure 1.1 - Example of the construction of DNA

For each pair of chromosomes, one is contributed by the mother and one is contributed by the father as DNA is passed on from parents to offspring through reproductive cells. These reproductive cells, namely sperm from the male and ovum (or eggs) from the female, are referred to as gametes. The gametes, produced through a process of cell division called meiosis, contain one member of each pair of chromosomes i.e. one for each of the 22 autosomes, an X sex chromosome for an egg cell, and an X or Y sex chromosome for a sperm cell. As part of meiosis, chromosomal crossover takes place in which blocks of DNA from each chromosome member are exchanged, in a process called recombination (**Figure 1.2**). As a result, each gamete carries a copy of each merged chromosome pairing, rather than an identical replication of one of the chromosome's members. The chromosome copies from mother and father carried by the gametes then form the chromosomal pairing for the offspring (5,6).

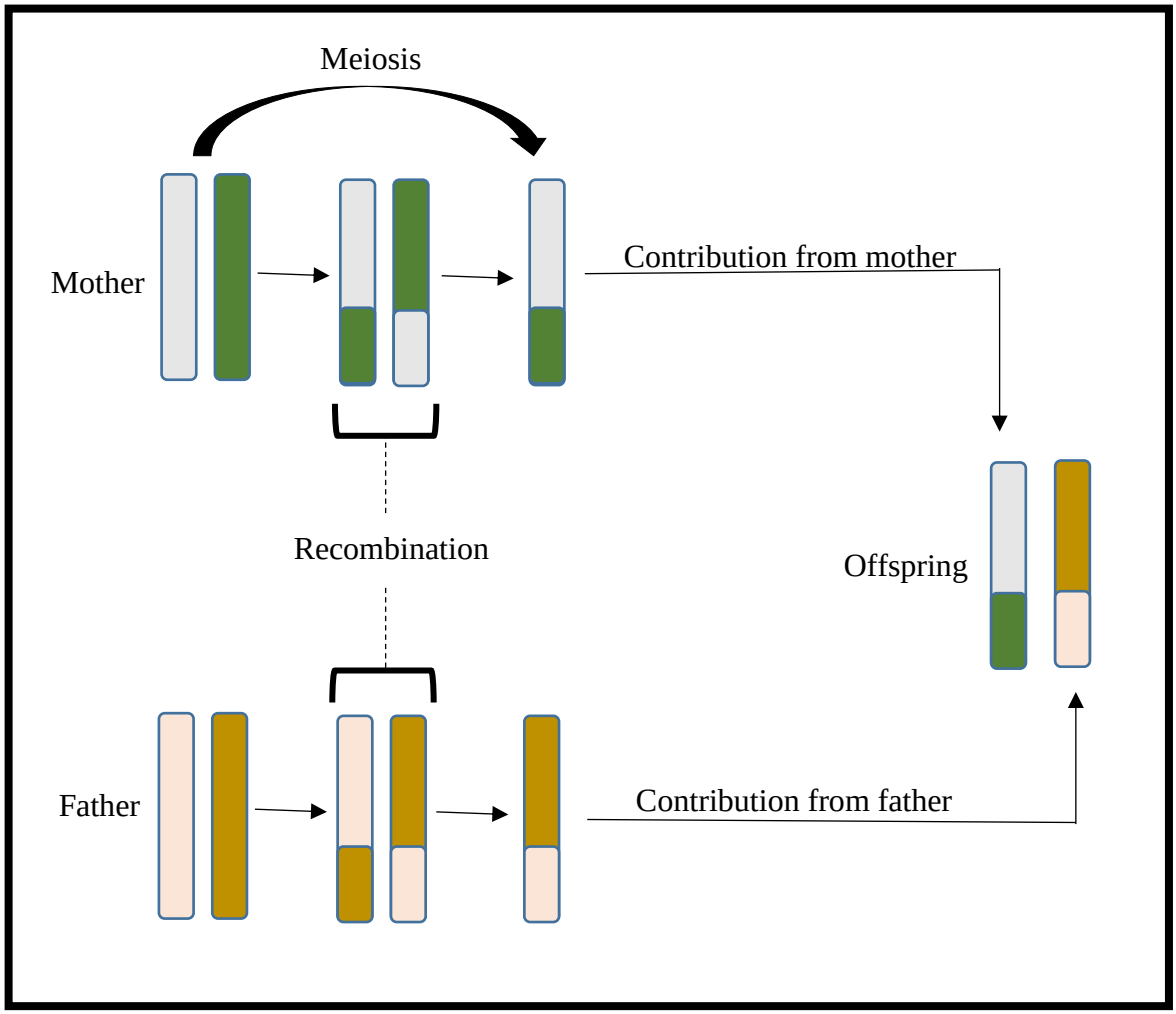


Figure 1.2 - An illustration of how DNA is passed from parent to offspring
The process of meiosis and recombination for a particular chromosome (the blocks represent chromosomes)

Different sections of the genome are responsible for the coding of different proteins, with these sections referred to as genes. Genes contain non-coding regions called introns and protein coding regions, called exons (**Figure 1.3A**). During a process known as transcription **(Figure 1.3B**), the two nucleotide strands of each chromosome carried by an individual break apart, and one strand is used to produce a complementary strand of RNA (ribonucleic acid). RNA is different from DNA in both molecular structure and also due to a replacement of base T for base U (uracil). After transcription, mature messenger RNA (mRNA) is produced by disposing of the introns and joining the exons together, in a process known as splicing (**Figure 1.3A**). Every three nucleotides of the mRNA molecule is referred to as a codon, with each codon responsible for an amino acid. The resulting amino acid chain is then responsible for

producing the protein. This process is called translation**,** and the resulting proteins are responsible for many human bodily functions (6).

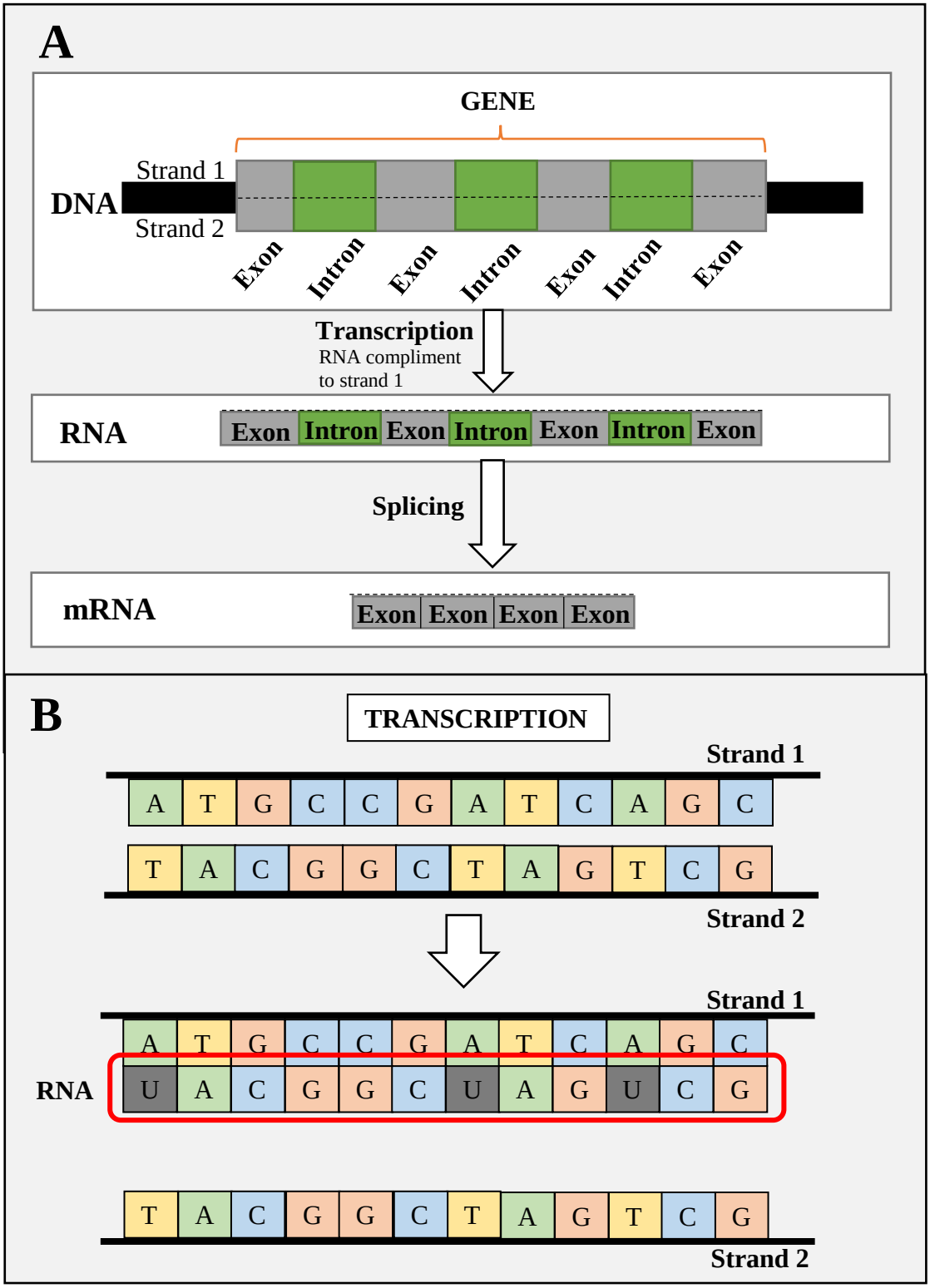


Figure 1.3 - Genes, splicing, and transcription.
Diagram which shows: A - the structure of a gene (RNA is produced complementary to strand 1) and the result of splicing and B - the transcription process in which RNA is produced (Thymine is replaced by Uracil).

Although the majority of human DNA is identical between individuals, there are instances where DNA can differ, and this can present itself in different ways. A single nucleotide polymorphism (SNP), which can also sometimes be referred to as a single nucleotide variant (SNV), is when a nucleotide at a particular locus in the human genome is substituted for another (**Figure 1.4**) (6,7). SNPs can be found all across the human genome, including both between genes (intergenic) and within genes. SNPs within genes can be located in an intron (intronic) or an exon (exonic) and can also be splice site SNPs i.e. located in a region where splicing occurs. Exonic SNPs can be either synonymous or non-synonymous. A non-synonymous SNP is a coding SNP which modifies the amino acid sequence of a gene and is therefore likely to severely alter the function and structure of the protein. This can happen by replacing one codon responsible for a certain amino acid with another codon responsible for a chemically different amino acid, in what is known as a missense mutation. Alternatively, a nonsense mutation replaces a codon for a stop codon, which prematurely ends translation of the protein. In contrast, a synonymous SNP will have a less severe or zero effect on the protein.

Other forms of genetic differences between individuals consist of copy number variations (CNV), insertion-deletions (indels) and inversions (**Figure 1.4**) (7), however the SNP will be the focus for this thesis.

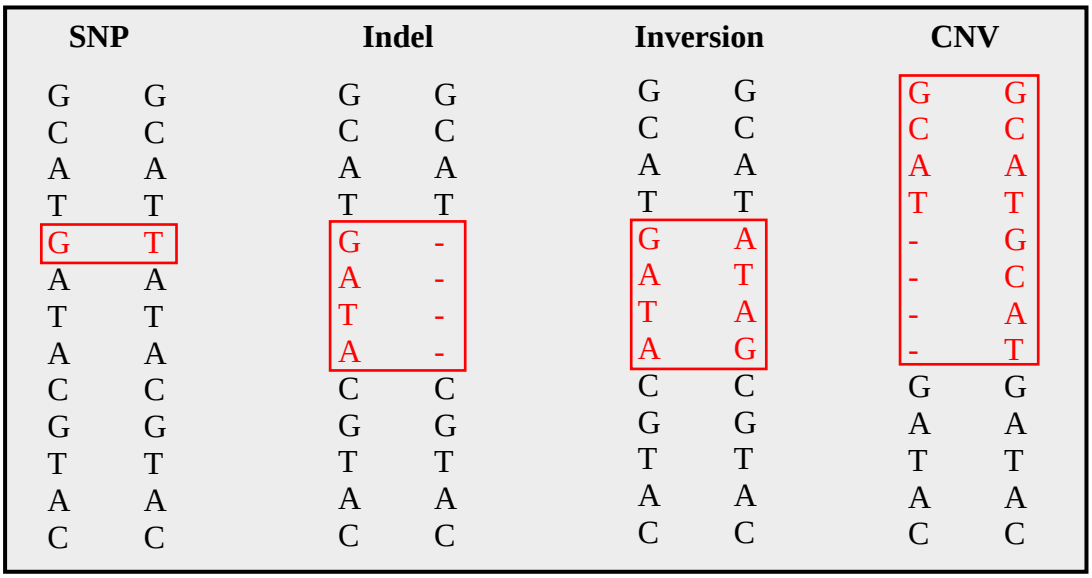


Figure 1.4 - Types of genetic differences between individuals.
Each nucleotide sequence is one strand of each comparable chromosome.

The two nucleotides present at a SNP locus are referred to as alleles. Taking the alleles across consistent strands of a chromosomal pair dictates an individual's genotype. Using the SNP example from **Figure 1.4** in which a G is replaced by a T, the three genotype groups are GG, GT or TT (**Figure 1.5**). Alternatively given the complementary characteristic of the two strands of DNA, one could also define the genotype groups using the other strand, giving the genotype groups CC, CA, AA. The choice of strand for categorising genotypes is arbitrary for certain SNPs, however confusion arises when considering those with C>G or A>T substitutions, for which genotype groups would be CC,CG,GG and AA,AT,TT respectively, regardless of the strand used for reference. Therefore, it is often the case that a consistent strand is used to genotype all SNPs, with reference strand information often given by the genotyping array manufacturer.

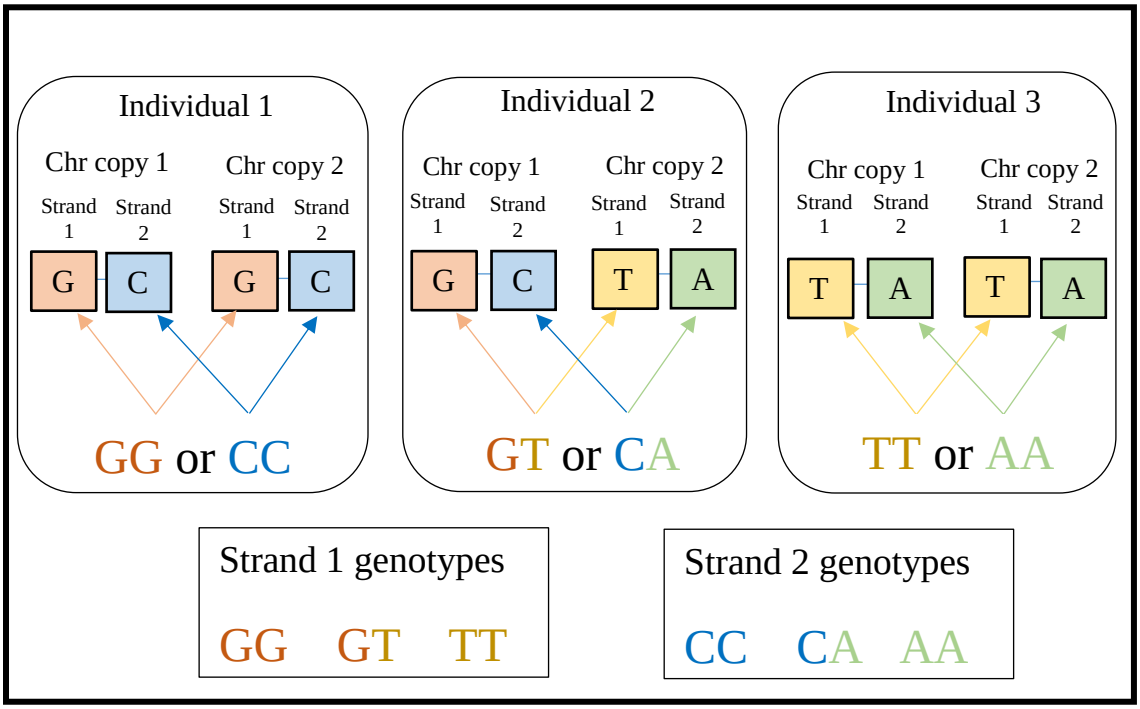


Figure 1.5 - Defining the possible genotype groups at a SNP
A Guanine is replaced by a Thymine for 3 individuals. Genotypes alter based on the stand used for reference.

The genotype in which both alleles are the same is called the homozygous genotype, with the genotype where alleles differ defined as the heterozygous genotype.

The frequency of a particular SNP, whether in the general population or in a sub sample, is defined by its minor allele frequency (MAF). That is, the frequency of the

less common or minor allele. For example, in a sample of 1000 individuals where 600 had genotype GG, 200 had genotype GT and 200 had genotype TT, the frequency of each allele would be calculated by taking the count of each allele, divided by the total number of alleles. Thus, the frequency of each allele would be:

$$G = \frac{(2\times600)+200}{2\times(600+200+200)} = 0.7 \qquad T = \frac{(2\times200)+200}{2\times(600+200+200)} = 0.3$$

Therefore, for this particular SNP, the minor allele is T and the MAF is 0.3 or 30%. Generally, a SNP with a MAF larger than 5% is considered common, a SNP with a MAF between 1% and 5% is considered low frequency, and a SNP with MAF below 1% is considered rare. MAF is an important consideration in study design, and the allele frequencies considered will have an effect on the method used to collect genotype data (rare variation is harder to capture), and also the statistical analysis used (statistical methods used for common variants are lower powered when applied to rare variants (8)).

During recombination, SNP alleles which are located closer together are often exchanged together in the same block, thus become correlated within populations. This genetic correlation is called linkage disequilibrium (LD), and genetic variants which occur on the same haplotype (section of DNA where several alleles are inherited together) are considered to be "in LD". This means that the presence of one allele can be inferred by the presence of another. Conversely, variants which are less common or further away from each other in the genome, are likely (but not necessarily) to be in low LD as are not inherited together.

The effect of genotype on disease susceptibility or a continuous trait can act in different ways, namely through a dominant, recessive, additive or genotypic model. In the SNP example from before, with alleles G and T, take G to be the coded allele and T to be the non-coded allele. In a dominant model considering a binary outcome (has disease or not), those that carry at least one copy of the coded allele (thus genotypes GG and GT) would carry the disease, as only one copy of the coded allele (G) is needed. In a recessive model, the individual would need to carry two copies of the coded allele to carry the disease (GG). If the model is additive or genotypic, the effect on the disease is

dependent on whether you carry one or two copies of the coded allele i.e. those with the GG genotype would carry increased risk of having the disease than those with the GT genotype. For an additive model the risk would increase multiplicatively for each effect allele increase, however in a genotypic model each genotype may carry independent risk, which is not multiplicative per effect allele increase.

1.1.2 Genotyping, sequencing and imputation

The technology available to accurately record variation across the human genome is becoming more readily available and is always improving in efficiency. Genotyping individuals requires the use of genotype arrays, which return allele probe intensities for each allele of the particular variant. Calling algorithms are used to convert these allele probe intensities into distinct genotype groups (9), as seen in **Figure 1.6**, one for the heterozygous pairing of alleles, and a group for each of the homozygous pairings of alleles (in the case of a SNP with two alleles).

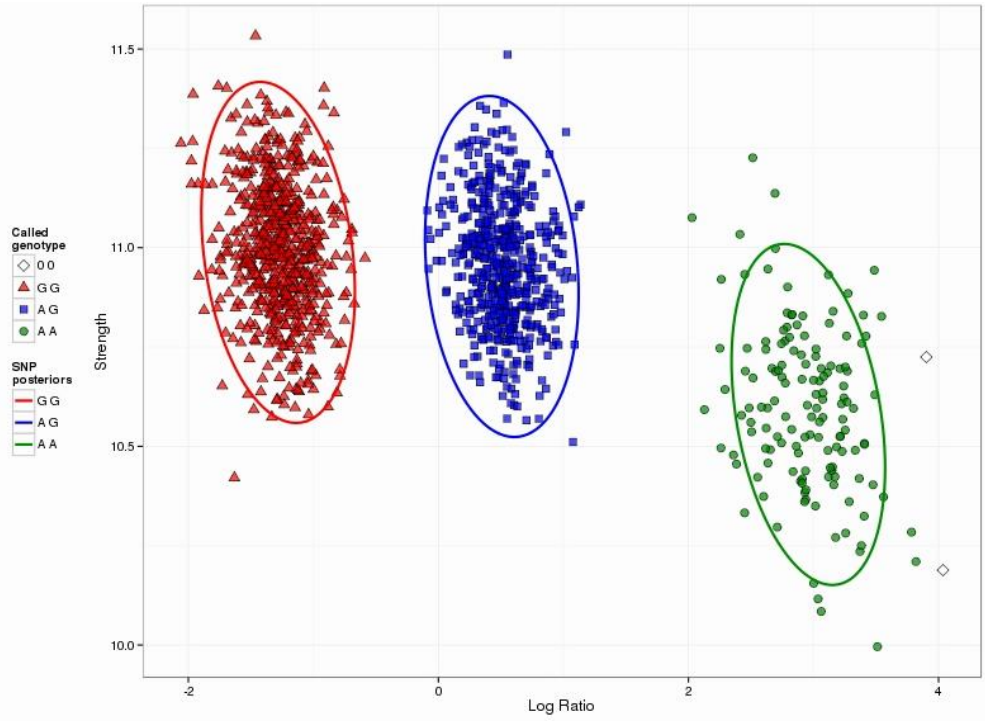


Figure 1.6 - An example cluster plot for a genotyped SNP

Each point on the graph is an individual. The x-axis is the Log of the ratio of the two allele probe intensities and the y-axis is the strength of intensities. The oval shapes are the posterior probability of where the groups should lie given the mean and variances of the observed intensities. The genotypes have been grouped such that the red triangles represent genotype group GG, the blue squares represent genotype group AG and the green circles represent genotype group AA. The white diamonds represent those individuals not assigned a genotype

Originally only providing the ability to capture a small number of SNPs, genotyping arrays can now accurately genotype millions of variants. Additionally, genotyping arrays can be customised specifically for the analysis being undertaken. For example, one may use a custom exome array if only wanting to analyse protein coding variants, or one may want to include additional SNPs (beyond the default array coverage) due to showing evidence of an association with the phenotype of interest in the past.

In contrast to using a genotype array in which a predefined number of variants are genotyped, sequencing allows for the ability to genotype the entire genome (can also be utilised to capture certain targeted regions of the genome, for example the exome) by essentially capturing and recording the full nucleotide sequence of an individual. This provides a denser genetic profile than that achieved by the use of an array. However, even though the cost of sequencing has comparatively reduced to previous years as technology and equipment has improved, it is still an expensive method, particularly when considering a large number of individuals (10).

The preferred strategy to avoid the expense of whole genome sequencing but improving the coverage of variation compared to genotyping arrays, is to genotype any missing variants not captured by an array using genetic imputation. Using particular software for imputation such as IMPUTE (11) or MaCH (12), one can use available reference panels, such as the 1000 genomes, UK10K project or the Haplotype Reference Consortium (13-15) to predict the missing genotypes. With reference panels using whole genome sequencing to genotype multiple individuals, they capture variants across the full MAF spectrum, including low frequency and some rare variation. Furthermore, such reference panels previously mentioned include data for multiple ancestries, such that samples of differing ancestries can be imputed with greater accuracy. Instead of producing a certain discrete score for each genotype (0,1,2) for the number of the chosen effect allele, imputed genotypes are presented in dosage format, taking a continuous value from 0 to 2, to reflect the uncertainty of estimating from a reference population. Also provided is an imputation score (a continuous value from 0 to 1), to reflect the software's confidence in the resulting imputed genotypes. The larger the value, the more accurately that SNP has been imputed.

1.1.3 Heritability, genetic association studies and the GWAS

Heritability is an estimate of the phenotypic variance that is attributable to genetic effect (16). Two statistics are used to define heritability, known as narrow-sense (h^2) and broad-sense (H^2). The former is deemed narrow due to only considering additive genetic effect, whilst the latter considers other genetic models and other types of genetic effect, such as interactions. The complexities involved in estimating heritability means that methods to do so are in abundance. Specifically, heritability can be highly dependent on phenotype, as some may have a stronger genetic component than others. Furthermore, heritability will be dependent on population, as estimated heritability may change from one ancestry to the next. However, these estimates are still vital in the field of genetic epidemiology. Using such a metric provides an understanding of current knowledge of the genetic architecture of particular traits, and by estimating the contribution of associations found to date, it also provides direction for future studies to determine the contributors of the remaining (missing) heritability.

The ability to understand the genetic contribution to disease has largely been dependent on the ability to accurately and efficiently measure variation in the human genome. Early research into the contribution of genetics on disease susceptibility made use of genetic linkage studies, which focussed specifically on related individuals (17). The idea behind the genetic linkage study, is that by observing related individuals with a particular disease, one could explore the unchanged regions of the genome passed down from parent to offspring (not altered by recombination) and thus any corresponding genetic markers, as a starting point to identifying potential contributing or causal genes (18). An early example of this is the identification of a gene involved in Huntingdon's disease (19). Linkage studies are now no longer widely used in genetic epidemiology and so will not be described further here.

Candidate gene studies have also been a popular approach for determining genetic association with disease, in which only those genes determined to be biologically plausible contributors to the phenotype of interest are chosen for analysis (17). Such studies remove the cost of expensive genotyping by focussing analyses on a finite number of genomic regions. Furthermore, such studies arguably benefit from the fact that given such regions already play a role in the susceptibility of the phenotype

biologically, any novel statistical findings will have more plausible meaning and greater context. These methods are however largely reliant on the strength of prior beliefs regarding the role of chosen candidate genes.

The genome-wide association study (GWAS) avoids this reliance on prior belief by simply analysing the whole genome to identify potential associated variants with the phenotype, thus requiring no prior hypothesis about contributing genes or linked genetic markers. Since the first published GWAS in 2002 (20), GWAS have provided an essential tool in uncovering genetic contributions for many diseases. These studies have largely been aided by resources such as the HapMap project (21), which has significantly increased our understanding of LD structure across the human genome in various populations. The following sections introduce the concepts regarding the preparation, analysis and follow-up of a typical GWAS.

1.1.3.1 Pre-analysis quality control

Prior to undertaking genetic association analysis, a number of quality control (QC) steps are implemented. This is often segregated into two main categories; sample QC (removal of individuals on the bases of phenotype and genotype data), and genotype QC (flagging or removing individual variants). This is an important stage in any GWAS pipeline prior to analysis, as inaccurate and error prone input data will inevitably produce unreliable and uninterpretable end results.

Firstly, phenotype QC ensures that the phenotypic data (corresponding to both outcome of interest and recorded covariates) contains no errors and is "clean" ready for any analysis. The presence of erroneous values recorded incorrectly due to human or computational error, are often common culprits for phenotype data failing QC. Such erroneous phenotype data could result in the attenuation of genetic signals through issues such as misclassification bias. These errors may initially be identified through the use of simple plots and summaries of the data available, particularly when extreme or suspicious values are easily identifiable. For variables such as age and height for example, a 200-year-old individual or one which is 5 metres tall, would be obvious errors and easy to identify. However, some errors may be harder to pinpoint and will require more complex methods to identify them beyond eye-balling the data, such as

using regression to identify outlying residuals or determining the reproducibility of lung function measures for example.

Sample removal due to genotype quality control

In terms of genotype quality control, samples may be removed for the following criteria:

- Genotype missingness
  - Individuals that have a large number of missing genotypes (SNPs not genotyped properly) which may be indicative of low DNA quality.

- Sex mismatch
  - Contradiction between the sex recorded and the sex inferred from the DNA e.g. phenotype and genetic data do not match due to mix-up during recruitment (such mismatches could be reduced by checking clinical records for the relevant individuals).

- Putative aneuploidy
  - Individuals that have a sex chromosome karyotype which was not XX (female) or XY (male), for example XXY.

- Heterozygosity
  - Individuals that have an extreme autosomal heterozygosity rate (the proportion of genotypes which are heterozygous) whether much larger than the mean (could indicate DNA contamination) or much smaller (could indicate inbreeding) (22).

- Duplicated samples
  - Identical samples which appear in the data twice (this may have been identified during phenotype QC).

- Relatedness
  - Related samples may or may not be removed depending on the analysis pipeline and the effect of removal on sample size. Some analysis software can account for relatedness in order to include all available samples regardless of relatedness and some cannot. Cryptic relatedness can be identified using metrics such as identity by state (IBS) and identity by descent (IBD). The former determines the number of shared alleles between individuals, which assists in determining the latter, whether these are shared through a common ancestor or by chance. The IBD metric can range from 0-1 and the higher the value the stronger the relatedness (23).

- Ancestry outlier
  - Individuals may need to be removed on the basis of ancestry, for example if the analysis was to be limited to European individuals only. Sample ancestry can be inferred from genotype data with the use of principal component analysis (PCA) (24). Principal component analysis reduces the complexity and multi-dimensionality of the genotype data into a number of linearly uncorrelated variables called principal components. A plot of principal components (PC1 vs PC2, **Figure 1.7**) will result in a number of visible clusters (some distinct and some overlapping) representing different ancestries. Reference populations (such as those available as part of the HapMap project (21)) can be used to aid in matching each cluster to its corresponding ancestry. Plotting the first two principal components for individuals in the study sample, over the top of reference clusters such as those in **Figure 1.7**, would enable analysts to accurately cluster the study sample by ancestry.

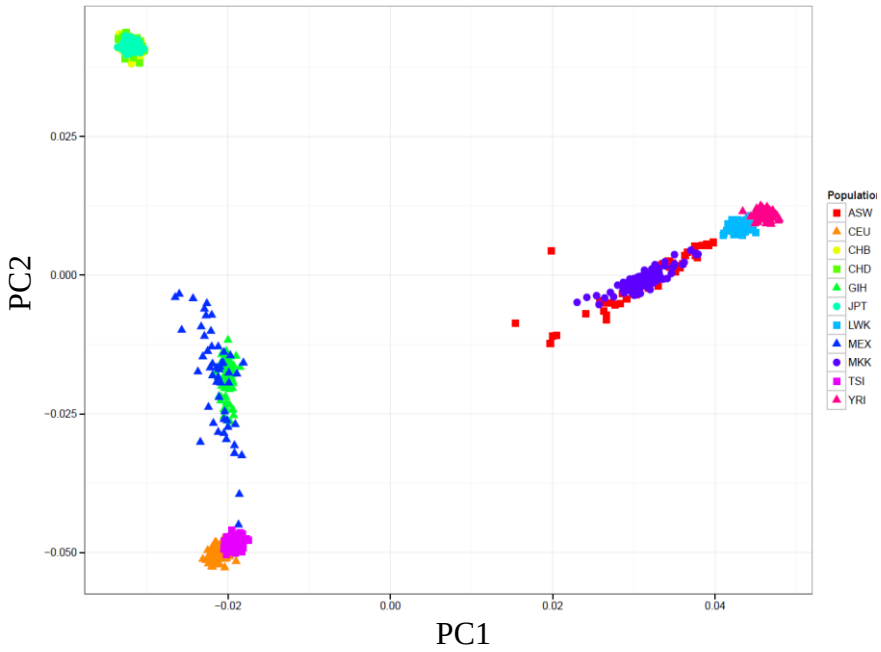


Figure 1.7- An example principal component plot with reference ancestry clusters. PC1 vs PC2 for the study sample can be plotted over the top of the reference points to identify ancestries in the study sample. The populations included in the plot are based on HapMap ancestry codes, ASW, LWK, MKK, and YRI are African populations, CHB, CHD and JPT are East Asian populations, GIH corresponds to South Asian ancestry, MEX is Admixed American ancestry and CEU and TSI are European ancestry (image from appendix of Wain et al. 2015 (11))

Variant removal due to genotype quality control

With regards to genotype quality control for individual variants, variants may be removed for the following criteria (criteria with an asterisk (*) may also be used to flag problematic SNPs to be explored further post-analysis, rather than for removal prior to analysis):

- Hardy-Weinberg equilibrium (HWE)
  - If mating is assumed random in a population and not affected by other influences, then SNPs should be in Hardy-Weinberg equilibrium, and the genotype probabilities between generations should be relatively stable. Specifically, given a SNP with two alleles A and a, with frequencies p and $1-p$ respectively, we would expect the frequency of individuals with genotypes AA, Aa and aa to be p^2, $2p(1-p)$ and $(1-p)^2$ respectively. SNPs not adhering to this may be removed.

- Minor allele frequency (MAF) and/or minor allele count (MAC)

- SNPs may be removed for having a very low minor allele frequency or minor allele count, for which genotyping can produce very low call rates and analysis can be underpowered.

❖ Genotype missingness

- SNPs with high levels of missingness (or low call rate) are removed due to poor genotyping.
- SNPs with differential missingness may also be removed (missing rate differing between cases and controls).

❖ Genotyping cluster plots

- Variants may be removed based on poor genotype clustering. As discussed in section 1.1.2, cluster plots are used to determine how well individuals are categorised into respective genotype groups. It can also be the case that the clusters are not as distinct as presented in **Figure 1.6**. An example of poor clustering is presented in **Figure 1.8.**

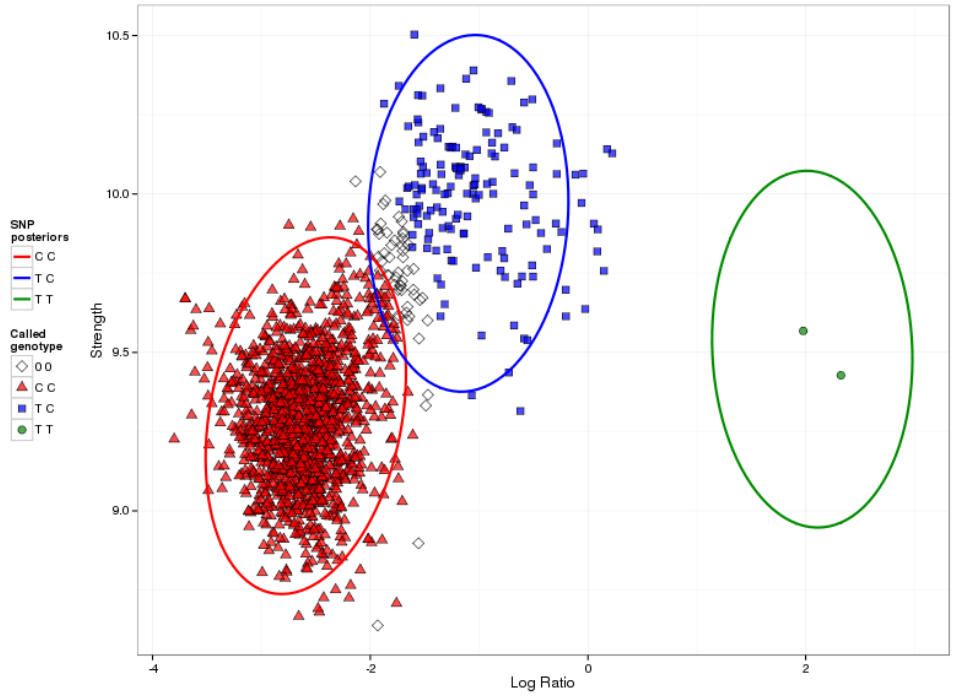


Figure 1.8 - An example of bad clustering for the genotyping of a SNP

Each point on the graph is an individual. The x-axis is the log of the ratio of the two allele probe intensities and the y-axis is the strength of intensities. The oval shapes are the posterior probability of where the groups should lie given the mean and variances of the observed intensities. The genotypes have been grouped such that the red triangles represent genotype group CC, the blue squares represent genotype group TC and the green circles represent genotype group TT. The white diamonds represent those individuals not assigned to a genotype

1.1.3.2 The basics of statistical analysis for GWAS

Given a particular trait (or phenotype) of interest and the recorded genetic information for each individual (categorising individuals by genotype using genotyping methods), analysis can then be undertaken to determine whether there is a link between genotype and disease e.g. comparing allele frequencies between disease statuses in the case of a binary outcome. It is often the case that determining the association between genetics and disease makes use of regression, which in addition to comparing the risk of disease across genotype groups, one can also adjust for potential confounders which may have an effect on the phenotype of interest. For example, principal components might be included in the regression model to adjust for population stratification (25). Other possible covariates of interest may also be included such as age and sex.

Given a continuous phenotype (where the outcome is recorded on a continuous scale e.g. height or weight), the following linear regression equation can be used:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 C1_i + \beta_3 C2_i + \cdots + \varepsilon_i \qquad (1.1)$$

Here, Y_i is the phenotype of interest for individual i, G_i is the genotype for individual i (coded depending on the genotype model used e.g. for additive coded as 0,1,2 for number of effect alleles present), $C1_i$, $C2_i$, … are the additional covariates adjusted for and ε_i is an error term with a normal distribution $\varepsilon_i \sim N(0, {\sigma_\varepsilon}^2)$. β_0 is the intercept term, $\beta_1$is the increase in risk on the disease per unit increase in effect allele, with β_2, β_3 … being the increase in risk on the disease per unit increase in the added covariate in the model.

Alternatively, if the phenotype is binary (you either have the disease or not) then the following logistic regression equation could be used:

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 G_i + \beta_2 C1_i + \beta_3 C2_i \ldots \qquad (1.2)$$

Where p_i is the probability of individual i having the disease, $logit(p_i)$ is the log odds of the probability p_i, and G_i, $C1_i$, $C2_i$…, and β_0 are defined as before. In this scenario β_1 is the increase in the log(odds) of having the disease per increase in number of effect

alleles, with e^{β_1} the odds ratio for the effect that SNP (for each increase in copies of the effect allele) has on the disease compared with the baseline group (G = 0). Similarly, for each of the additional covariates β_n would be the increase in log odds per unit increase in the covariate ($C1_i$, $C2_i$, …).

The test of $\beta_1 = 0$ in both equations would be used to determine whether the SNP is associated with the phenotype (using a statistical test such as the Wald test or the Likelihood Ratio Test (LRT)). In a GWAS, this would be used to test each variant individually across the genome.

Other analysis designs are also available for genetic analysis. For example, instead of analysing the disease or trait itself (whether binary or quantitative), it may be of interest to consider the survival rate for the particular disease as the outcome. This is however not the focus of this thesis and therefore will be not be explained in detail here.

1.1.3.3 Post GWAS analysis plots and metrics

Post analysis summary plots and metrics are used to explore the validity of any results and summarise the analysis. The quantile-quantile (QQ) plot is a plot of the observed $-log_{10}$ P-values against those expected under the null hypothesis (**Figure 1.9**). The ideal scenario is to observe a plot which follows the line $y = x$ for the majority of values, with an increase in the observed $-log_{10}$ P-values compared with the expected $-log_{10}$ P-values at the higher end of the distribution. This would be indicative of "true" phenotype associated SNPs. In the scenario where deviation from the line $y = x$ is a general theme throughout the plot, then this may suggest errors with the analysis, specifically the potential presence of unaccounted for population stratification. This plot is often used in conjunction with a metric called the genomic inflation factor (λ) (26,27) with equation given in 1.3.

$$\lambda = \frac{median(\chi_1^2 + \chi_2^2, \ldots, \chi_n^2)}{0.456} \qquad (1.3)$$

This is a ratio of the median of the observed Chi-squared test statistics for n tests (χ_1^2 ... χ_n^2) and the median of the expected test statistics using a Chi-squared distribution with one degree of freedom. Typically, a λ which represents a correctly controlled analysis (with regards to population stratification) would be expected to lie within the range of 0.9 and 1.1. Extreme values above 1.1 such as 1.5 for example, which would coincide with an overly inflated QQ plot, would suggest bias in the GWAS results caused by scenarios such as population stratification or cryptic relatedness, and this will need to be resolved. There could be cases where λ is above 1.1, but issues such as population stratification and cryptic relatedness are not present or have been identified and corrected. An example of natural inflation (not due to problematic data) would be the dependency of inflated test statistics on polygenic inheritance (for traits where multiple genes contribute) (28). In this case test statistics from the analysis can be scaled by the value of λ in a method called "genomic control", which will give unbiased "corrected" p-values.

Manhattan plots (**Figure 1.10**) can be used to visualise the results from a GWAS analysis. By plotting the $-log_{10}$ p-values for each chromosome, Manhattan plots can highlight those variants with more extreme p-values. Signals from significantly associated variants would be observed in the form of peaks (rather than individual points) due to the underlying LD structure.

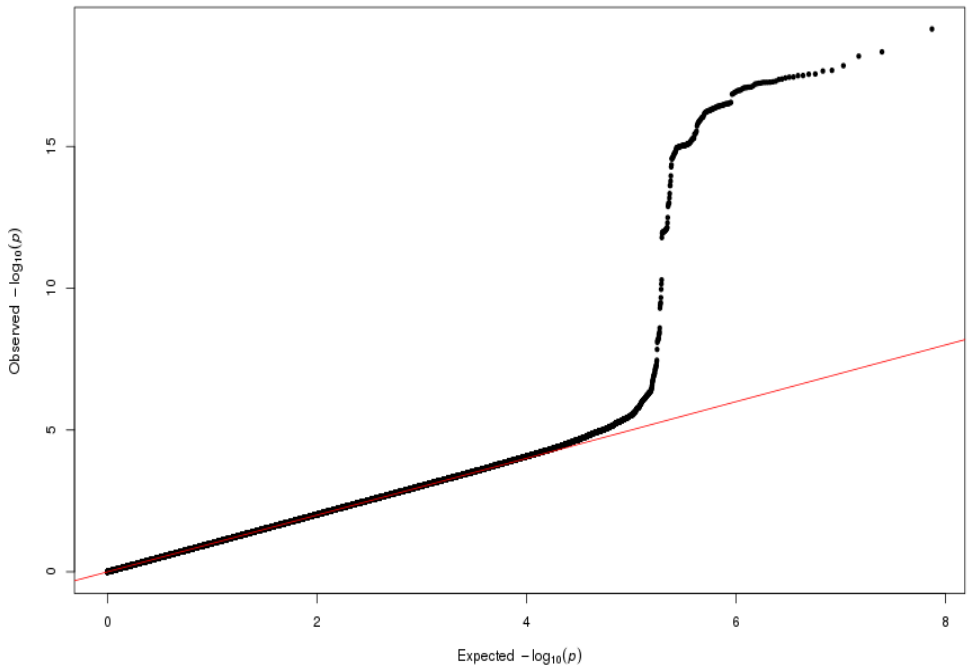


Figure 1.9 - An example QQ plot for a set of GWAS analysis results

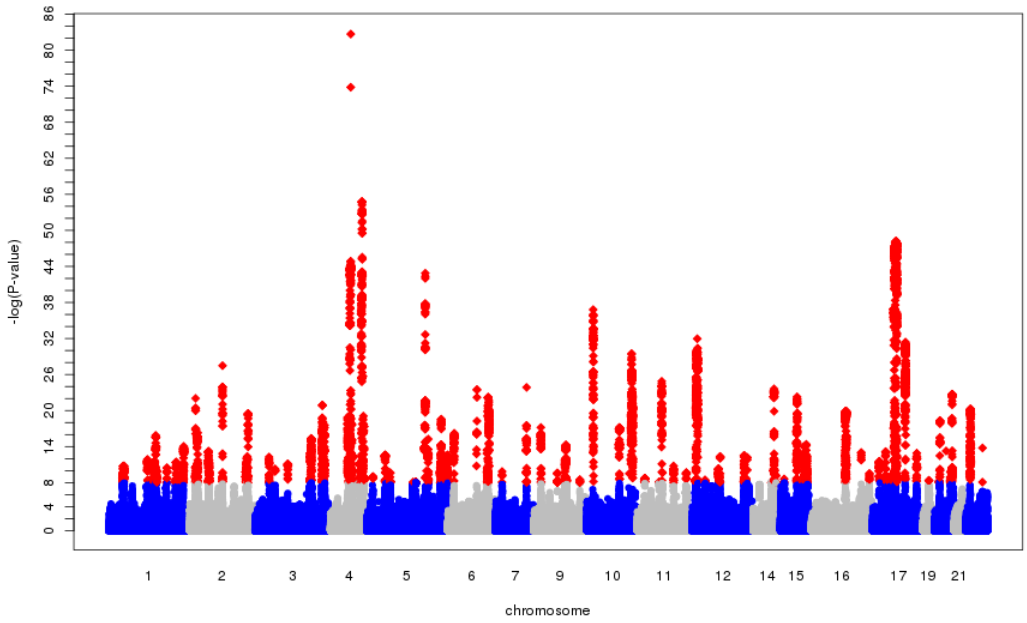


Figure 1.10 - An example Manhattan plot for GWAS

Each point represents an analysed variant, with variants for all chromosomes plotted. The x-axis is the chromosome number and the y-axis is the log of the p-value ($-log_{10}(P)$). Points in red are those which reach a predetermined significance level, in this case 5×10^{-8}.

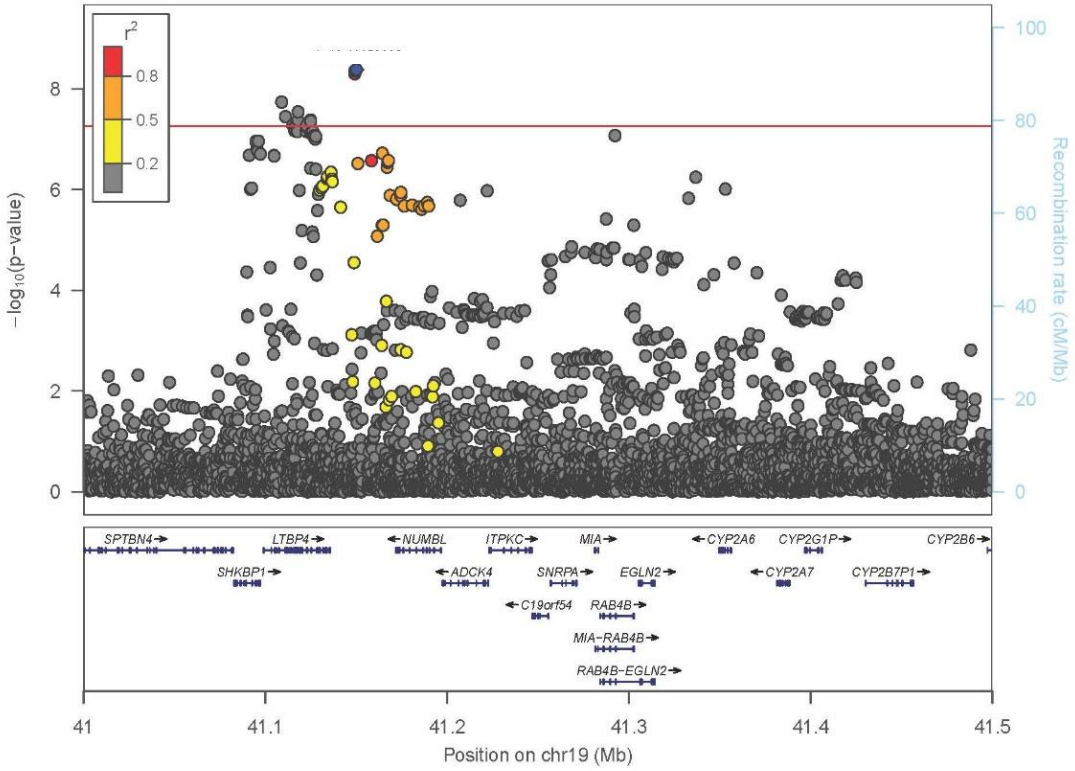


Figure 1.11 - An example magnified region plot.

Each point is a variant analysed and the blue point is the chosen reference SNP for the plot (also the most statistically significant) from which the LD for all other SNPs in the region are calculated. The colour of each point represents the magnitude of LD for each variant with the reference variant. Grey means the variant is in low LD with the reference variant (<0.2), whilst red infers high LD (> 0.8) between the two variants. The LD estimate for yellow and orange points falls between these two extremes of high and low LD. Below the plot are the gene names which span the genomic region.

Additionally, regions plots, such as those produced using locuszoom (29), provide a magnified image of certain areas of the genome to explore the distribution of p-values more closely (**Figure 1.11).** Region plots also provide further information, namely the observed LD structure, recombination rates and the genes which span the chosen region of interest.

1.1.3.4 Determining statistical significance and signal selection

When considering a single hypothesis test, one can simply choose the threshold at which to reject the null hypothesis, in favour of the alternative. For example, using a threshold or type I error rate of $\propto = 0.05$ in a single test, would mean that we would allow for a 5% chance of observing our pre-defined alternative hypothesis without rejecting the null hypothesis. A p-value less than this would produce a statistically significant result and we would reject the pre-defined null hypothesis in favour of the alternative hypothesis. As the number of tests increase, this type I error will increase such that if considering 2000 tests, 100 (5% of 200) could be statistically significant by chance. This multiple testing problem also applies in genetic association analysis when considering multiple variants across the genome with a particular phenotype. Therefore, we must make an adjustment to control the type I error. Although other methods are available to control this Type I error, Bonferroni correction is a simple adjustment, with the threshold simply divided by the number of (independent) tests undertaken. Due to the presence of Linkage Disequilibrium (LD), not all tests are independent between variants, thus adjusting for the total number of tests would be extremely stringent. To determine an appropriate threshold, one could therefore determine the number of independent SNPs in an analysis using software such as PLINK (30) to prune variants by LD, or alternatively 5×10^{-8} has widely been used for GWAS analysis, representing adjustment for 1,000,000 tests (31).

A signal is a group of SNPs which are associated with the given trait and are in strong LD with each other. After identifying SNPs with p-values meeting the pre-defined threshold outlined above, the aim is to then determine how many of the identified signals are independent or are dependent with regards to LD structure. For example, we may find two individual SNPs that meet the required threshold to conclude statistical significance, such as 5×10^{-8}, but if they are in "high LD" with each other, we could

not reject the possibility that these SNPs are part of the same signal. Identifying independent signals in GWAS firstly requires identifying the most statistically significant SNP (or the sentinel), then including those SNPs within an arbitrary distance either side as part of the same region e.g. ±1Mb (megabase). The next most statistically significant SNP outside of this region would then be allocated as a second signal within a new 2Mb region, and this would continue until all SNPs that have reached the pre-defined threshold have been accounted for. A second step implemented is to determine whether each region harbours only one signal or more than one. This can be done in various ways. Software can be used such as clumping in PLINK (30), which identifies independent signals using an LD cut off, by choosing a sentinel and allocating it along with all SNPs in high LD to one group, then choosing the next signal and so on. Alternatively, one can use conditional analysis, in which the first sentinel is added to the analysis model as a covariate, before re-analysing the data and observing whether there are any statistically significant p-values remaining in the region. If only one signal is present then the signal will "drop out" leaving no further statistically significant SNPs.

1.1.3.5 Replication of signals

After identification (or discovery) of novel independent genetic associations with the phenotype of interest, one would then look to replicate these associations. The aim of replication is to validate any discovered signals and ensure that they have not just arisen by chance. This is undertaken in an independent data set, and ideally with a sample size reminiscent of the discovery sample, so that there is no loss in power. To produce a set of replicated variants, one may select variants from the discovery stage using the stringent genome-wide threshold and conclude replication of SNPs using a more relaxed threshold in the independent data set. The threshold for statistical significance is often more relaxed at the replication stage due to following up a smaller number of SNPs than the full genome-wide set originally analysed, thus the applied Bonferroni correction is less conservative. Alternatively, one may look to select variants using a more relaxed threshold initially (than that used for genome-wide significance), combine the data with an independent dataset and determine whether the association reaches genome-wide statistical significance in a further analysis stage.

1.1.4 Missing heritability and the interaction

Although many associations between SNPs (predominantly common MAF) and diseases have been identified using GWAS, there is still a large amount of unexplained heritability using this common variant common disease approach. For example, known genetic associations for lung function traits only account for as little as ~10% (32,33) of the estimated heritability (estimated at 20-40%) (34). As a result, focus has turned to other genetic predictors of disease which previous exploration efforts have not been tailored to capture, in order to uncover this missing heritability (35). A potential contributor to the missing heritability are low frequency and rare variants with minor allele frequency (MAF) less than 5%. Furthermore, interactions (with regards to both gene by gene interactions (GxG) and gene by environment interactions (GxE)), are also attracting interest as contributors to disease beyond that of marginal effects, resulting in the proposal of numerous new methodological approaches to efficiently explore their effects. In this thesis, I will focus on gene by environment interactions.

The linear regression and logistic regression models in equations 1.1 and 1.2 can be expanded further to consider interactions, whether epistatic (GxG) or GxE. Interactions add a further dimension to the previous statistical models in that the SNP effect becomes dependent on another covariate, and its effect size is no longer interpretable on its own. For example, we may wish to know whether the SNP effect is dependent on categorical covariates, such as whether you are male or female, or which genotype you carry for an additional SNP. Alternatively, one could look at interactions with a continuous variable such as height or weight. The concept of interactions and the statistical theory and application in a genetic association setting is explored further in chapter 2, alongside an exploration of the current methods available for interaction analysis.

1.2 Lung function & COPD

1.2.1 Spirometry and measures of lung function

Spirometry is a method used to measure the flow and volume of air that can be exhaled by an individual after maximum inhalation. This can be observed by the individual forcibly exhaling into a spirometer (36). This method enables us to record several

pulmonary function metrics which can be used to diagnose specific lung diseases, in addition to providing an understanding of respiratory health in general (37). The key volumetric and flow measures recorded during spirometry for this purpose are:

- **FVC (forced vital capacity)** - the total volume of air forcibly expelled from the lungs
- **FEV_1 (forced expiratory volume in 1 second)** - the total volume of air expelled in the first second of observation
- **FEV_1/FVC** - the ratio of the two previously defined metrics which provide an indicator of airflow obstruction (low expiratory flow in relation to the total volume expelled)
- **PEF (peak expiratory flow)** - the maximum airflow achieved during the spirometry test

Respiratory health and function and the resulting measurements presented above are affected by a number of variables such as age, sex, height and ethnicity. There are studies which have provided reference equations using such predictor variables to predict the lung function of an individual given their characteristics. For example, one could use the estimated coefficients from the National Health and Nutrition Examination survey (NHANES III) (38). Such reference equations can be used to calculate an individual's percent-predicted FEV_1 (%pred FEV_1) which is defined as follows:

$$\%\text{pred } FEV_1 = \frac{\text{Actual } FEV_1}{\text{predicted } FEV_1} \times 100$$

This gives an understanding of whether an individual's airflow is as expected given their characteristics (or alternatively whether it is comparatively better or worse than expected).

1.2.2 COPD diagnosis

Chronic obstructive pulmonary disease (COPD) which encompasses diseases emphysema and chronic bronchitis, is characterised in an individual by a severe airflow obstruction. COPD is currently one of the top causes of worldwide mortality, thus

determining the factors of COPD susceptibility to aid discovery of effective treatment is a continuously active area of research (39).

Potential causes of COPD include environmental risk factors such as smoking, air pollution, exposure to dust during occupation and history of chest infection throughout childhood (36). A possible 1-2% of COPD cases are induced by a reduction in alpha1-antitrypsin (AAT) levels caused by mutations in the *SERPINA1* gene (40,41). A deficiency of AAT weakens the protection of pulmonary tissue to the effects of enzymatic degradation, thus increasing COPD susceptibility (40).

Diagnosing COPD can be difficult due to differing levels of severity and determining presence of the disease is aided by the use of the previously defined metrics collected during spirometry. A value of $\frac{FEV_1}{FVC} < 0.7$ is deemed indicative of limited airflow and COPD presence (32). In order to further categorise individuals by COPD severity, this ratio is often used in combination with the previously defined %pred FEV_1. The severity of the disease can be determined using criteria provided by the Global Initiative for Chronic Obstructive Lung Disease (GOLD). For example, the criteria which refer to FEV_1/FVC and %pred FEV_1 specifically are presented in **Table 1.1** (42).

Table 1.1 -The four severity stages of COPD defined by spirometry

Severity of COPD	**FEV_1/FVC**	**%predicted FEV_1**
Mild	< 0.7	≥ 80%
Moderate	< 0.7	≥ 50% & < 80%
Severe	< 0.7	≥ 30% & < 50%
Very severe	< 0.7	< 30%

1.2.3 Treatment for COPD

As of yet, COPD cannot be cured, with current treatment options offering symptomatic relief. For COPD patients who smoke, quitting is strongly advised and is effective in slowing down the progression of the disease and reducing symptoms, although may not reverse the damage already done (43-45). Quitting smoking can be facilitated with counselling supplemented by nicotine replacement therapy or varenicline. Bronchodilators are a form of medication which improve breathing by relaxing airway

muscles and typically administered via inhalation (46). These can be either short acting or long acting. The former is effective for four to six hours and administered to individuals who experience an episode of breathlessness. For individuals with regular breathlessness, a long-acting bronchodilator may be more appropriate, with effects lasting 12 – 24 hours after inhalation. Additionally, steroid inhalers may be issued, which reduce airway inflammation. As well as inhalers, oral theophylline may be prescribed for its bronchodilatory properties whilst mucolytic tablets can potentially thin and reduce the mucus that often accompanies COPD (47). Oral steroids have sometimes been used to manage exacerbations but carry substantial risks, especially with long-term use (48). For individuals in the advanced stages of the disease complaining of severe dyspnoea and pain, lung surgery may be considered (47), but often pallitative care is administered using a combination of oxygen therapy for hypoxia, non-invasive ventilation and opioids (49,50).

1.2.4 Known genetics of lung function and COPD

1.2.4.1 Genetic association studies undertaken prior to full UK Biobank release

GWAS exploring genetic associations in the absence of interactions (marginal effect) with the previously described lung function measures (FEV_1, FVC, FEV_1/FVC and PEF) has led to the identification of more than 250 signals of association. Wilk et al. (51) published the first evidence of genetic association with lung function from a GWAS in 2009. The GWAS undertaken, which utilised 7,691 individuals from the Framingham Heart Study, identified an association between variants in the 4q31 region and FEV_1/FVC, which mapped to a location nearby the hedgehog-interacting protein encoded by the *HHIP* gene. This locus was confirmed via meta-analyses conducted by SpiroMeta and CHARGE consortia which utilised substantial sample sizes (20,288 and 20,890 respectively) (52,53). Additionally, these studies identified a further 9 loci associated with FEV_1 or FEV_1/FVC. In 2012 and 2014, two large joint meta-analyses between the aforementioned consortia with sample sizes of 48,301 (with an additional 46,411 for follow up) and 52,253 (32,917 for follow-up) respectively, identified a further 22 lung function associated loci (54,55). The former reported 16 novel variants associated with FEV_1 and FEV_1/FVC, whilst the latter identified 6 novel variants associated with FVC. In 2015 the UK Biobank Lung Exome Variant Evaluation (UK

BiLEVE) study (56) sampled individuals at the extremes of the FEV_1 distribution from the substantially large UK Biobank (n≈500k), with comparison of individuals with high and low FEV_1 measures identifying 6 novel signals (2 of which mapped to previously identified gene regions). The SpiroMeta consortium utilised the 1000 genomes project imputation panel to provide increased and extensive coverage of low frequency variants for 38,199 individuals (with a further 54,550 used for follow-up of statistically significant signals) and reported a further 16 novel signals (57). In 2017, Wain et al. (58) increased the number of identified lung function associated signals from 54 to 97, using 48,943 individuals from the aforementioned UK BiLEVE study, following up results using an independent dataset of 95,375 individuals combined from UK Biobank, the SpiroMeta consortium and the UK Households Longitudinal Study (UKHLS). An exome chip analysis by Jackson et al. (59) found a further 7 novel loci, using data meta-analysed across the SpiroMeta and CHARGE consortia with 60,749 and 7,721 individuals of European and African ancestry respectively, before following up in an independent dataset of 111,556 individuals. A Multi-ethnic meta-analysis undertaken by Wyss et al. (60) identified a further 73 associated loci with lung function, using meta-analysed 1000 genomes imputed data for 90,715 individuals. As a result of the previously described studies the number of lung function signals identified was 177.

With COPD defined using lung function measurements, it is a sensible assumption that the two would share associated genetic loci. Indeed, this has proven to be the case, with 36 of the reported lung function signals so far also showing association with COPD (56,58,61-66). In addition to this, there are also 7 COPD signals independent of associations with lung function (61,62,67,68).

In addition to the signals discussed above, a specific region on chromosome 15 (the 15q25 region) which harbours the nicotine receptor genes *CHRNA3/5* have also been identified in studies of lung function and COPD (63,69), however this association is likely to be driven by this region's strong association with smoking behaviour (70-72).

In total, the number of signals identified due to an association with lung function, COPD or both (including the 15q25 signal) was 185.

1.2.4.2 GWAS of lung function using full UK Biobank sample

In 2018, the first GWAS of lung function which utilised the full release (~ 500k individuals) of UK Biobank (the largest GWAS of lung function undertaken to date) was undertaken (32,33).

Introduction to UK Biobank

The UK Biobank provides great potential to aid current and future research for numerous diseases and traits (73,74). This resource harbours both genetic and lifestyle information for over 500,000 UK based individuals aged 40-69, an age range which favours individuals at risk of developing disease over the years following their recruitment. The resource provides an extensive range of health related and phenotype data, from basic anthropometric measurements, to data on several serious diseases and conditions, such as cancer, stroke, heart disease, diabetes and depression. In addition, there is also an extensive observation of various lifestyle and risk factors, for example drinking and smoking behaviour. Of relevance to this research is both the detailed smoking information and the extensive spirometry data collection.

Recruitment and data collection

Using National Health Service (NHS) records (to source information such as patient date of birth), individuals which fit the age criteria of 40-69 years were contacted and invited to take part in the study, regardless of their current health status. A response rate of 5% was observed for the study, with 9,000,000 individuals contacted and approximately 500,000 individuals recruited (75). Those willing to take part in the study attended a local assessment centre. During the assessment (which lasted for approximately three hours), participants were requested to complete a 30 minute touch screen self-reported questionnaire, with more specific information such as medication and operation history collected using a ten minute computer-assisted interview with a member of staff. In addition to the questionnaire, individuals had baseline measurements taken. Examples of measurements collected were height, weight and grip strength as well as spirometry to measure each individual's lung function. Participants

also provided samples of blood and urine which were stored long term and used to capture genetic data.

Genetic data in UK Biobank

The first 50,000 individuals were genotyped using the Affymetrix Axiom UK BiLEVE array, whilst the remaining 450,000 individuals were genotyped using the Affymetrix Axiom UK Biobank array (both of which were custom-designed for UK Biobank). This provided coverage of 820,967 SNPs and indels (**Figure 1.12**) (74). This included approximately 100,000 SNPs chosen either for relevance to specific traits (including variants relevant to lung function phenotypes), due to being identified in previous GWAS, and for residing in areas of the genome that are of interest e.g. the HLA region. In addition, there was coverage for over 100,000 coding variants and approximately 600,000 variants accredited to producing substantial genome-wide coverage of various minor allele frequencies in individuals of European ancestry. The array has a large coverage of rare variation with approximately 300,000 variants genotyped having a minor allele frequency below 1%. The two arrays used to genotype the full 500,000 dataset shared 95% common content. To predict missing genotypes, the genetic data for UK Biobank was imputed using two reference panels, the Haplotype Reference Consortium (HRC) panel (76) and the combined UK10K + 1000 Genomes panel (77). This resulted in available genetic data for over 90,000,000 SNPs, indels and structural variants.

UK Biobank demographics and limitations

Demographic comparisons between UK Biobank and the general population it aims to represent suggest that the sample chosen is a much healthier one. Prevalence's of smoking, alcohol intake and obesity were all lower in the UK Biobank sample for example (78). Additionally, there were fewer health conditions self-reported. As expected, this resulted in lower rates of 5-year mortality than observed in the general population (75). Analysis undertaken in UK Biobank will therefore require careful consideration of these demographic differences when putting results into context and generalising any results for the UK population.

Analysis of lung function using the full UK Biobank release

Shrine et al. (32,33) used this resource and the SpiroMeta consortium to produce a sample size of 400,102 European individuals. From the 185 signals associated with lung function and COPD which were previously discussed, 140 passed signal QC (with regards ensuring independency between signals, no associations with smoking and a sex chromosome signal removed due to only undertaking autosomal analysis) and provided corroborative evidence of an association in the largest GWAS of lung function to date (which is expanded on further in chapter 5). In addition, the study identified a further 139 novel lung function associated signals. Thus, to date there are 279 lung function and COPD associated loci.

The work undertaken throughout this thesis references two milestones during the progress made in identifying genetic contributions to lung health. The first is the identification of the 97 lung function signals to date in 2017 (51-58) and the second is the additional replicated and novel signals (including COPD specific signals) contributing to the 279 signals identified in 2018 (32,33,59,60,62,67,68,79).

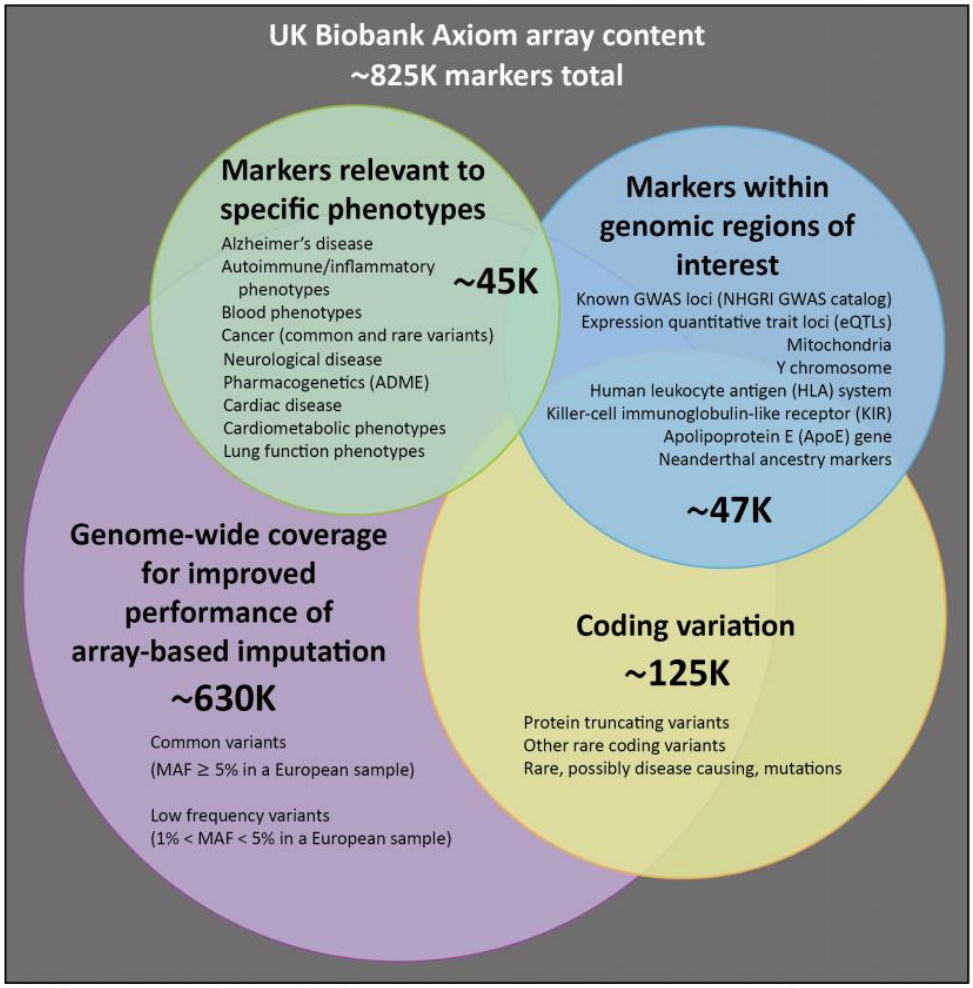


Figure 1.12 - UK Biobank genetic coverage

This is taken from the Bycroft et al. paper describing the genetic data for the full UK Biobank sample (74)

Reported SNPs and nearest gene (not necessarily the causal gene) are presented in **Table 1.2** for the 140 lung function and COPD signals showing corroborative evidence in the Shrine et al. paper. The additional 139 novel associations from the same study are presented in **Appendix A**.

Table 1.2 - Previously associated signals with lung function and COPD

These signals showed corroborative evidence for association in the largest GWAS of lung function to date (32,33). Chromosome (Chr) and position (Pos) are build 37.

Reported RSID	Chr	Pos	Trait	Gene	Reference	COPD reference
rs2284746	1	17306675	FEV_1/FVC	*MFAP2*	Soler Artigas et al. 2011 (54)	Wain et al. 2017 (58)
rs17513135	1	40035686	FEV_1/FVC	*LOC101929516*	Wain et al. 2017 (58)	-
rs1192404	1	92068967	FEV_1/FVC	*TGFBR3*	Wain et al. 2017 (58)	-
rs12140637	1	92374517	FEV_1/FVC	*TGFBR3*	Wain et al. 2017 (58)	-
rs200154334	1	118862070	FVC	*SPAG17*	Wain et al. 2017 (58)	-
rs6681426	1	150586971	FEV_1	*ENSA*	Soler Artigas et al. 2015 (57)	-
rs2821332	1	200085714	FVC	NR5A2	Wyss et al. 2017 (60)	-
rs12092943	1	204434927	FEV_1	PIK3C2B	Wyss et al. 2017 (60)	-
rs512597	1	215095003	FVC	CENPF/KCNK2	Wyss et al. 2017 (60)	-
rs4846480	1	218598469	COPD only	*TGFB2*	-	Cho et al. 2014 (61)
rs993925	1	218860068	FEV_1/FVC	*MIR548F3/TGFB2*	Soler Artigas et al. 2011 (54)	Cho et al. 2014 (61), Hobbs et al. 2017 (62)
rs4328080	1	219963088	FEV_1/FVC	*RNU5F-1*	Soler Artigas et al. 2015 (57)	-
rs6657854	1	221630555	FVC	C1orf140/DUSP10	Wyss et al. 2017 (60)	-
rs6688537	1	239850588	FEV_1/FVC	*CHRM3*	Wain et al. 2017 (58)	-
rs62126408	2	18309132	FEV_1/FVC	*KCNS3*	Soler Artigas et al. 2015 (57)	Wain et al. 2017 (58)
rs1430193	2	56120853	FVC	*EFEMP1*	Loth et al. 2014 (55)	-
rs2322659	2	136555659	FEV_1	*LCT*	Jackson et al. 2016 (59)	-
rs72904209	2	157046432	FEV_1/FVC	KCNJ3/NR4A2	Wyss et al. 2017 (60)	-
rs2571445	2	218683154	FEV_1	*TNS1*	Repapi et al. 2010 (52)	Soler Artigas et al 2011 (49) , Wain et al. 2017 (52,58)
rs10498230	2	229502503	FEV_1/FVC	*PID1*	Hancock et al. 2010 (53)	Wain et al. 2017 (58), Hobbs et al. 2017 (62)

Reported RSID	Chr	Pos	Trait	Gene	Reference	COPD reference
rs61332075	2	239316560	FEV_1/FVC	*TRAF3IP1*	Wain et al. 2017 (58)	-
rs12477314	2	239877148	FEV_1/FVC	*FLJ43879*	Soler Artigas et al. 2011 (54)	Wain et al. 2017 (58)
rs1529672	3	25520582	FEV_1/FVC	*RARB*	Soler Artigas et al. 2011 (54)	Wilk et al. 2012 (79), Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs17666332	3	29469675	FEV_1/FVC	RBMS3	Wyss et al. 2017 (60)	-
rs1458979	3	55150677	FEV_1/FVC	*CACNA2D3*	Wain et al. 2017 (58)	-
rs79294353	3	57494433	FEV_1	DNAH12	Wyss et al. 2017 (60)	-
rs1490265	3	67452043	FVC	*SUCLG2*	Wain et al. 2017 (58)	-
rs6778584	3	98815640	FEV_1	DCBLD2/MIR548G	Wyss et al. 2017 (60)	-
rs2811415	3	127991527	FEV_1/FVC	*EEFSEC*	Wain et al. 2017 (58,62)	Hobbs et al. 2017 (62)
rs1595029	3	158241767	FVC	*RSRC1/RP11-538P18.2*	Soler Artigas et al. 2015 (57)	-
esv2660202	3	168738454	FEV_1/FVC	*MECOM*	Wain et al. 2017 (58)	-
rs1344555	3	169300219	FEV_1	*MECOM*	Soler Artigas et al. 2011 (54)	-
rs28520091	4	7846240	FEV_1/FVC	AFAP1	Wyss et al. 2017 (60)	-
rs13110699	4	89815695	FEV_1/FVC	*FAM13A*	Wain et al. 2017 (58)	Wain et al. 2017 (58)
rs2045517	4	89870964	FEV_1/FVC	*FAM13A*	Hancock et al. 2010 (53)	Cho et al. 2010 (48) , Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs34480284	4	106064626	FEV_1	*TET2*	Wain et al. 2015 (56)	Wain et al. 2015 (56), Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs10516526	4	106688904	FEV_1	*GSTCD*	Repapi et al. 2010 (52), Hancock et al. 2010 (53)	Soler Artigas et al 2011 (49), Wain et al 2017 (58), Hobbs et al. 2017 (62)

Reported RSID	Chr	Pos	Trait	Gene	Reference	COPD reference
rs34712979	4	106819053	FEV_1	*NPNT*	Wain et al. 2015 (56)	Wain et al. 2015 (56), Wain et al. 2017 (58)
rs138641402	4	145445779	FEV_1	*HHIP-AS1*	Wilk et al. 2009 (51)	Pillai et al. 2009 (41), Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs111898810	4	146174040	FEV_1	OTUD4/SMAD1	Wyss et al. 2017 (60)	-
rs91731	5	33334312	FVC	*TARS*	Wain et al. 2017 (58)	-
rs1448044	5	44296986	FVC	*FGF10*	Jackson et al. 2016 (59)	-
rs1551943	5	52195033	FEV_1/FVC	*ITGA1*	Wain et al. 2017 (58)	-
rs2441026	5	53444498	FVC	*ARL15*	Wain et al. 2017 (58)	-
rs72776440	5	77440196	FVC	AP3B1	Wyss et al. 2017 (60)	-
rs153916	5	95036700	FEV_1/FVC	*SPATA9*	Soler Artigas et al. 2011 (54)	-
rs7713065	5	131788334	FEV_1/FVC	*C5orf56*	Wain et al. 2017 (58)	-
rs7715901	5	147856392	FEV_1	*HTR4*	Repapi et al. 2010 (52), Hancock et al. 2010 (53)	Soler Artigas et al. 2011 (49), Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs3839234	5	148596693	FEV_1	*ABLIM3*	Wain et al. 2017 (58)	-
rs10515750	5	156810072	FEV_1/FVC	*CYFIP2*	Wain et al. 2017 (58)	-
rs1990950	5	156920756	FEV_1/FVC	*ADAM19*	Hancock et al. 2010 (53)	Castaldi et al. 2011 (66), Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs1294421	6	6743149	FEV_1/FVC	*LY86*	Jackson et al. 2016 (59)	-
rs2076295	6	7562998	COPD only	*DSP*	-	Hobbs et al. 2017 (62)
rs6924424	6	7801611	FVC	*BMP6*	Loth et al. 2014 (55)	-
rs1928168	6	22017738	FEV_1/FVC	LINC00340	Wyss et al. 2017 (60)	-

Reported RSID	Chr	Pos	Trait	Gene	Reference	COPD reference
rs34864796	6	27459923	FEV_1	*ZNF184*	Soler Artigas et al. 2011 (54)	Wain et al. 2017 (58)
rs2070600	6	32151443	FEV_1/FVC	*AGER*	Repapi et al. 2010 (52), Hancock et al. 2010 (53)	Castaldi et al. 2011 (66), Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs114544105	6	32635629	FEV_1	*HLA-DQB1*	Wain et al. 2015 (56)	Wain et al. 2015 (56)
rs141651520	6	73670095	FEV_1/FVC	*KCNQ5*	Wain et al. 2017 (58)	-
rs2768551	6	109270656	FEV_1/FVC	*ARMC2*	Soler Artigas et al. 2011 (54)	Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs11759026	6	126792095	FVC	CENPW/RSPO3	Wyss et al. 2017 (60)	-
rs7753012	6	142745883	FEV_1/FVC	*GPR126/LOC153910*	Soler Artigas et al. 2015 (57)	Wilk et al. 2012 (79), Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs148274477	6	142838173	FEV_1/FVC	*GPR126/LOC153910*	Hancock et al. 2010 (53)	Wain et al. 2017 (58)
rs10246303	7	7286445	FEV_1/FVC	*C1GALT1*	Wain et al. 2017 (58)	-
rs55905169	7	15506529	FVC	AGMO	Wyss et al. 2017 (60)	-
rs72615157	7	99635967	FEV_1/FVC	*ZKSCAN1*	Wain et al. 2017 (58)	-
rs12698403	7	156127246	FEV_1	*LOC285889*	Wain et al. 2017 (58)	-
rs771924	9	1555835	FVC	DMRT2/SMARCA2	Wyss et al. 2017 (60)	-
rs7872188	9	4124377	FEV_1	*GLIS3*	Wain et al. 2017 (58)	-
rs10965947	9	23588583	FEV_1/FVC	FLJ35282/ELAVL2	Wyss et al. 2017 (60)	-
rs16909859	9	98204792	FEV_1/FVC	*PTCH1*	Hancock et al. 2010 (53)	-
rs2451951	9	109496630	FEV_1/FVC	TMEM38B/ZNF462	Wyss et al. 2017 (60)	-
rs803923	9	119401650	FEV_1/FVC	*ASTN2*	Soler Artigas et al. 2015 (57)	Wain et al. 2017 (58)
rs10858246	9	139102831	FVC	*QSOX2/LHX3*	Soler Artigas et al. 2015 (57)	-
rs10870202	9	139257411	FVC	*DNLZ*	Wain et al. 2017 (58)	-
rs7090277	10	12278021	FEV_1/FVC	*CDC123*	Soler Artigas et al. 2011 (54)	Wain et al. 2017 (58)

Reported RSID	Chr	Pos	Trait	Gene	Reference	COPD reference
rs3847402	10	30267810	FEV_1/FVC	*KIAA1462*	Wain et al. 2017 (58)	-
rs7899503	10	65087468	FEV_1	JMJD1C	Wyss et al. 2017 (60)	-
rs7095607	10	69957350	FVC	*MYPN*	Wain et al. 2017 (58)	-
rs3849969	10	75525999	FEV_1	*SEC24C*	Jackson et al. 2016 (59)	-
10:77002679:TC:T	10	77002679	FVC	COMTD1/ZNF503-AS1	Wyss et al. 2017 (60)	-
rs2637254	10	78312002	FEV_1	*C10orf11*	Soler Artigas et al. 2011 (54)	Wilk et al. 2012 (79), Wain et al. 2017 (58)
rs721917	10	79946567	COPD only	*SFTPD*	-	Hobbs et al. 2017 (62)
rs2293871	10	124273671	FEV_1/FVC	HTRA1	Wyss et al. 2017 (60)	-
rs4237643	11	43648368	FVC	*MIR129-2*	Loth et al. 2014 (55)	-
rs2863171	11	45250732	FVC	*PRDM11*	Loth et al. 2014 (55)	-
rs2509961	11	62310909	FEV_1	*AHNAK*	Wain et al. 2017 (58)	-
11:73280955:GA:G	11	73280955	FEV_1/FVC	FAM168A	Wyss et al. 2017 (60)	-
rs11234757	11	86443072	FEV_1	*PRSS23*	Wain et al. 2017 (58)	-
rs567508	11	126008910	FEV_1	*RPUSD4*	Wain et al. 2017 (58)	-
rs2348418	12	28689514	FVC	*CCDC91*	Soler Artigas et al. 2015 (57)	-
rs772920	12	56390364	FEV_1	RAB5B	Wyss et al. 2017 (60)	-
rs11172113	12	57527283	FEV_1/FVC	*LRP1*	Soler Artigas et al. 2011 (54)	-
rs1494502	12	65824670	FEV_1	*MSRB3*	Wain et al. 2017 (58)	-
rs7971039	12	85724305	FVC	ALX1/RASSF9	Wyss et al. 2017 (60)	-
rs11107184	12	94184082	FVC	CRADD	Wyss et al. 2017 (60)	-
rs113745635	12	95554771	FEV_1/FVC	*FGD6*	Wain et al. 2017 (58)	-
rs12820313	12	96255704	FEV_1/FVC	*SNRPF/CCDC38*	Soler Artigas et al. 2011 (54)	Wain et al. 2017 (58)
rs10850377	12	115201436	FEV_1	*TBX3*	Soler Artigas et al. 2015 (57)	-

Reported RSID	Chr	Pos	Trait	Gene	Reference	COPD reference
rs35506	12	115500691	FVC	*TBX3*	Wain et al. 2017 (58)	-
rs4444235	14	54410919	FEV_1/FVC	DDHD1/MIR5580	Wyss et al. 2017 (60)	-
rs1698268	14	84309664	FEV_1/FVC	*LINC00911*	Wain et al. 2017 (58)	-
rs7155279	14	92485881	FEV_1	*TRIP11*	Soler Artigas et al. 2015 (57)	Wain et al. 2017 (58)
rs117068593	14	93118229	FEV_1	*RIN3*	Soler Artigas et al. 2015 (57)	Cho et al. 2014 (61), Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs1200345	15	41819716	FEV_1/FVC	*RPAP1*	Jackson et al. 2016 (59)	-
rs72724130	15	41977690	FEV_1/FVC	*MGA*	Wain et al. 2017 (58)	-
rs8025774	15	67483276	FVC	SMAD3	Wyss et al. 2017 (60)	-
rs10851839	15	71628370	FEV_1/FVC	*THSD4*	Repapi et al. 2010 (52), Hancock et al. 2010 (53)	Wilk et al. 2012 (79), Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs12591467	15	71788387	FEV_1/FVC	*THSD4*	Wain et al. 2017 (58)	-
rs66650179	15	84261689	FEV_1/FVC	*SH3GL3*	Wain et al. 2017 (58)	-
rs12149828	16	10706328	FEV_1/FVC	*TEKT5*	Soler Artigas et al. 2015 (57)	-
rs181206	16	28513403	COPD only	*IL27*	-	Hobbs et al. 2016 (67), Hobbs et al. 2017 (62)
rs12447804	16	58075282	FEV_1/FVC	*MMP15*	Soler Artigas et al. 2011 (54)	-
rs3973397	16	70040398	FVC	PDXDC2P	Wyss et al. 2017 (60)	-
rs3743609	16	75467021	FEV_1/FVC	*CFDP1*	Soler Artigas et al. 2011 (54)	Wain et al. 2017 (58), Hobbs et al. 2017 (62)
rs1079572	16	78187138	FVC	*WWOX*	Loth et al. 2014 (55)	-
rs59835752	17	28265330	FEV_1/FVC	*EFCAB5*	Wain et al. 2017 (58)	-
rs62070631	17	29087285	FEV_1	SUZ12P1	Wyss et al. 2017 (60)	-
rs11658500	17	36886828	FEV_1/FVC	*CISD3*	Wain et al. 2017 (58)	-

Reported RSID	Chr	Pos	Trait	Gene	Reference	COPD reference
rs8067511	17	37611352	FVC	MED1/CDK12	Wyss et al. 2017 (60)	-
rs35524223	17	44192590	FEV_1	*KANSL1*	Wain et al. 2015 (56)	Wain et al. 2015 (56)
rs6501431	17	68976415	FVC	*CASC17*	Loth et al. 2014 (55)	-
rs1859962	17	69108753	FEV_1	*CASC17*	Jackson et al. 2016 (59)	-
rs7218675	17	73513185	FEV_1	*TSEN54*	Wain et al. 2015 (56)	Wain et al. 2015 (56)
rs647097	18	8808465	COPD only	*MTCL1*	-	Hobbs et al. 2017 (62)
rs7243351	18	20148531	FEV_1	CTAGE1/RBBP8	Wyss et al. 2017 (60)	-
rs7238093	18	20728158	FVC	CABLES1	Wyss et al. 2017 (60)	-
rs8089865	18	50957922	FVC	DCC	Wyss et al. 2017 (60)	-
rs9636166	19	31829613	FEV_1/FVC	TSHZ3	Wyss et al. 2017 (60)	-
rs113473882	19	41124155	FEV_1/FVC	*LTBP4*	Soler Artigas et al. 2015 (57)	-
rs6140050	20	6632901	FVC	*BMP2*	Wain et al. 2017 (58)	-
rs6138639	20	25669052	FEV_1	ZNF337	Wyss et al. 2017 (60)	-
rs1737889	20	31042176	FEV_1	C20orf112	Wyss et al. 2017 (60)	-
rs6088813	20	33975181	FVC	*UQCC1*	Jackson et al. 2016 (59)	-
rs2236519	20	45529571	FVC	EYA2	Wyss et al. 2017 (60)	-
rs72448466	20	62363640	FEV_1	*ZGPAT*	Wain et al. 2017 (58)	-
rs2834440	21	35690499	FEV_1/FVC	*KCNE2*	Soler Artigas et al. 2011 (54)	Wain et al. 2017 (58)
rs11704827	22	18450287	FEV_1	*MICAL3*	Wain et al. 2017 (58)	-
rs4820216	22	20854161	FEV_1/FVC	KLHL22/MED15	Wyss et al. 2017 (60)	-
rs2283847	22	28181399	FEV_1	*MN1*	Wain et al. 2017 (58)	-

1.2.4.3 Interactions associated with lung function and other diseases

Interactions associated with lung function

To date, research into the role of gene-smoking interactions and their effect on the observed variation of pulmonary function has produced limited results. Exploring these effects is often hindered by the need for larger sample sizes. It has been suggested that a four-fold increase in sample size is needed for studies of interactions comparative to marginal effect GWAS sample sizes (80,81). With smoking identified as the biggest risk factor for poor lung function and COPD, the interest around gene-smoking interactions is driven by the fact that not all smokers develop restrictive lung problems. This could be suggestive that interactions are at play and genetic effects are dependent on smoking behaviour.

Early work accredited to the exploration of gene by smoking interactions predominantly focused on candidate genes from previous studies of lung function (including related phenotypes such as Asthma) for interaction analyses. He et al. (82) found an interaction effect between smoking (pack-years) and Glutathione S-Transferase (*GST*) variants contributing to poor lung function susceptibility, whilst Sadeghnejad et al. (83) discovered an interaction between the same smoking variable and the *IL13* gene. In addition, Hunninghake et al. (84) found *MMP12* only to effect lung function for particular exposure groups, such as those who smoke or asthmatics. However, these findings were all identified using small study sample sizes and replication has not been achieved.

In contrast to the candidate gene approach, Hancock et al. (85) studied the effect of gene-environment interactions on a much larger scale. To do this, a genome-wide joint meta-analysis, which jointly tests for both main genetic effect and interaction effect was undertaken, using an overall sample size of 50,047 contributed by 19 studies. Although results did not suggest evidence of any strong interactions as such, it provided useful direction for the methodological approach of large-scale gene-environment interaction analysis, highlighting the potential drawback of analyses ignoring interactions, where genuine associations could be missed. For example, a genetic effect which is modest in a particular environmental exposure group and absent in another, may be flagged by an

interaction analysis, but not by a marginal effect analysis, as the effect may be nullified by combining exposed and unexposed individuals. Another large-scale lung function gene-environment interaction study has since been undertaken in 2013, utilising 4,785 individuals from the Framingham Heart Study (86). Although considering occupational exposure rather than smoking for interaction analysis, results suggested that variant rs9931086 in *SLC38A8* modified the effect of occupational exposure on poor lung function, however replication analysis was not undertaken. Finally, a genome-wide interaction study (GWIS) by Park et al. (87) assessed the role of gene-smoking interactions in poor lung function susceptibility (using FEV_1 as phenotype) using individuals of Korean ancestry. The study used a 3 degrees of freedom joint test of main genetic effect and two interaction effects (SNP by pack-years and SNP by smoking status) to filter variants, testing implicitly for interaction effect in a second "testing" step. The most statistically significant signal (approaching but didn't meet a Bonferroni corrected significance threshold of $p < 1.61 \times 10^{-7}$ for 310,515 SNPs) was identified near the *SOX9* gene on chromosome 17. The second interaction testing step showed evidence of an interaction with smoking behaviour (never vs former and never vs current smokers) at the 5% significance threshold. Although there appeared evidence of replication in independent data, it was not consistent with regards smoking phenotype (for example identified by smoking status for interaction but replicated for pack-years interaction). Furthermore, with being undertaken in a sample of Korean ancestry, it's unclear how applicable results will be for other ancestries.

These studies clearly provide evidence that lung function variability can be better understood by considering interactions between gene and the environment, in a quest to uncover the missing heritability. However, our understanding is still very limited, and it is likely that analyses in large scale data will be needed to really understand the role of interactions in poor lung function and COPD susceptibility with any certainty.

Interactions associated with other diseases or traits

The low success rate (with regards replicating identified interaction effects) in detecting gene-environment interactions is not unique to studies of lung function and genetic interactions with smoking behaviour, but is also observed when considering other diseases and conditions. As previously discussed, this is likely due to small sample sizes

and issues with power, but also because of other problems with interaction analysis, such as how to measure environmental exposures, the measurement error than can result, and the requirement of large sample sizes which in turn requires large replication sample sizes to confirm any interaction discoveries. As a result, although a number of gene-environment interactions have been detected for other traits and environmental exposures, few have been replicated in independent datasets and populations. An interaction between the *FTO* gene and physical activity for BMI (88), an interaction between smoking behaviour and *NAT2* for bladder cancer (89) and an interaction between alcohol and *ALDH2* for esophageal cancer (90-94) are some notable examples, which have been replicated in independent populations. These replicated findings are discussed in a thorough review by Ritz et al. (95) with consideration of how future interaction studies can benefit from previous interaction analysis designs and conclusions.

1.2.4.4 Known Genetics of Smoking

When exploring the effect of gene-smoking interactions, it is important to consider the possibility that any statistical interaction could be explained by an association between the signal and smoking, rather than by a true dependency of genetic effect on smoking behaviour. Smoking has an evident genetic component with research undertaken to date resulting in 13 strong genetic signals of association with smoking phenotypes (56,70-72), with the strongest accredited to the previously mentioned 15q25 locus, which includes the nicotine receptor genes *CHRNA3/5*. Of the 13 associations, 6 were associated with the number of cigarettes smoked per day (CPD), 6 were associated with smoking initiation (SI, ever vs never smoker) and 1 was associated with smoking cessation (SC, current vs former smoker). In addition, Hancock et al. 2015 (96) found an association between a variant in the nicotine receptor gene *CHRNA4* and the Fagerström Test for Nicotine Dependence (FTND), a set of 6 questions which determines an individual's dependence on nicotine (**Table 1.3**) (97).

Table 1.3 - The Fagerström Test for Nicotine Dependence (FTND)

Questions	Answers	Points
How soon after you wake up do you smoke your first cigarette?	Within 5 minutes	3
	6 – 30 minutes	2
	31 – 60 minutes	1
	After 60 minutes	0
Do you find it difficult to refrain from smoking in places where it is forbidden e.g. in church, at the library, in cinema, etc.?	Yes	1
	No	0
Which cigarette would you hate most to give up?	The first one in the morning	1
	All others	0
How many cigarettes/day do you smoke?	10 or less	0
	11 – 20	1
	21 – 30	2
	31 or more	3
Do you smoke more frequently during the first hours after waking than during the rest of the day?	Yes	1
	No	0
Do you smoke if you are so ill that you are in bed most of the day?	Yes	1
	No	0

1.3 Thesis aims and objectives

The overall aim of this thesis is to explore and potentially identify variants whose effect on lung function is dependent on the largest risk factor for poor lung health, smoking. This research will aim to improve on our current knowledge of the genetic architecture of poor lung function and COPD, which has largely been dominated by marginal effects from common variants and provide some insight into questions such as:

- Why don't all smokers develop COPD?

- Do gene-smoking interaction effects account for a proportion of the missing heritability for lung function and COPD?

In chapter 2, "Statistical methods for gene-environment interaction analysis", the concept of an interaction effect will be discussed in more detail, in particular both the statistical and biological interpretation of an interaction, and how they can be modelled in a statistical analysis. Beyond the basic statistical application to interactions, the chapter will then proceed to explore the plethora of methods available for interaction analyses, segregated into 6 broad groups based on application and design.

Chapter 3, "Power available for gene-environment interaction analysis utilising UK Biobank", will use simulation to determine the power available when using the unprecedented sample size in UK Biobank to observe interaction effects. This will consider two methods (applicable to quantitative traits representative of lung function), namely a joint test of main and interaction effect, and a test of interaction effect only. To determine power, the simulation produces a comparison with previously attainable sample sizes, as well as determining the effect of differing interaction effect magnitudes and minor allele frequencies (an important consideration given the dense coverage of low frequency and rare variants in UK Biobank).

Chapter 4, titled "Prioritising regions previously associated with lung function for gene-smoking interaction analysis" focuses gene-smoking interaction analysis on candidate regions chosen to represent the 97 signals identified by 2017 (the number identified at time of writing) due to producing signals of marginal genetic effect (section 1.2.3.1). The rationale behind this is that these regions may be more likely to harbour interaction effects given that they are associated with lung function already. The chapter presents a gene-smoking interaction analysis (using ever/never smoking as the binary exposure variable) of over 300,000 European individuals in UK Biobank (post quality control) of lung function phenotypes FEV_1 and FEV_1/FVC. Each previous association is positioned centrally in a 3Mb "known region", using HRC imputed variants and no MAF filter. A two-stage approach was used for analysis. In the first stage, variants were chosen using a joint test of marginal and interaction effect and this was applied for all SNPs (individually) in each known region. In a second stage, presence of interaction effect was determined using a test of interaction effect only. To evaluate whether the

interaction was explained by a true dependency of genetic effect on smoking behaviour rather than by an association between the signal and smoking, each signal was explored for association with smoking traits. Replication was sought with the use of an independent meta-analysed dataset of up to 71,000 individuals.

Chapter 5, titled "Gene-smoking interaction analysis for lung function associated SNPs in UK Biobank", focusses on the 139 novel and 140 previously reported marginal genetic effect associations which were associated with lung function (FEV_1, FVC, FEV_1/FVC, PEF), as reported in a recent study by Shrine et al. (32,33). The aim was to determine whether variants showing strong marginal genetic effect for lung function and COPD, interact with smoking behaviour. In addition, the interaction effect between all 279 variants combined and smoking behaviour was explored using a weighted genetic risk score, to determine whether the combined effect of all 279 variants (rather than the individual effect) was modified by smoking behaviour.

Chapter 6, titled "Genome-wide gene-smoking interaction analysis in UK Biobank", presents the largest genome-wide gene-smoking (ever smoking) interaction analysis of lung function undertaken to date, utilising the HRC imputed data from UK Biobank and FEV_1, FVC, FEV_1/FVC and PEF as phenotypes. The driving mechanism behind any identified effect was determined through exploration of association with smoking behaviour and replication was attempted for any remaining associations using an independent data set with approximately 71,000 individuals.

Finally, chapter 7 discusses the work undertaken as part of this thesis, the conclusions found and the potential for future work moving forward.

Chapter 2 Statistical methods for gene-environment interaction analysis

With continuing interest into the contribution of gene-environment interactions in disease susceptibility, there is now a plethora of methods aimed at capturing such effects efficiently. This chapter aims to provide a comprehensive insight into methods which have been proposed for gene-environment interaction analysis to date. It begins by introducing the analytical concept using methodology originally proposed for standard single variant GWAS, and proceeding into a discussion of the more advanced gene-environment interaction methods, such as screening approaches, data mining and gene-based analysis. Although not an exhaustive list, this chapter aims to provide an overview of each methodological approach.

2.1 The concept of interaction

In terms of genetic association analysis, an interaction can either be between a gene and the environment (GxE) or between genes/SNPs (GxG), which are given the name epistatic effects. As discussed in section 1.1.4, the interaction adds a new complexity to understanding and observing the role of genetics in disease susceptibility, beyond considering marginal genetic effect only. No longer interpretable by itself, the genetic effect now has dependency on other factors. For example, take lung function to be the phenotype of interest, a SNP which is known to be associated under an additive genetic model, and two exposure groups, such as smokers and non-smokers. The SNP is coded such that the effect allele has a deleterious effect on lung function, as shown in **Figure 2.1**, so as the number of effect alleles increase, the observed lung function decreases.

The former example may correspond to a situation where smokers and non-smokers are pooled together, and thus one genetic effect applies to both groups. The concept can be extended further to consider smokers and non-smokers separately. In **Figure 2.2A**, we can allow for the effect of smoking group on the outcome. Here, lung function is higher for non-smokers compared with smokers, suggesting a smoking effect on lung function, however this difference is consistent as the effect allele increases, as characterised by the parallel lines for smoker and non-smoker groups. There is no interaction. However, in **Figure 2.2B**, although non-smokers generally have better lung function as before, the

genetic effect on lung function differs for each smoking group. The deleterious effect of the SNP on lung function is larger as the number of effect alleles increase for smokers

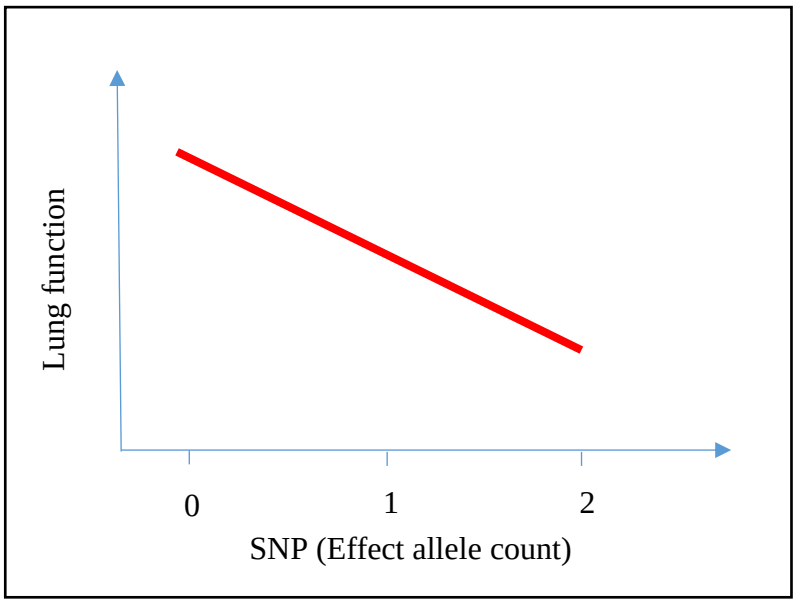


Figure 2.1 - An example of the deleterious effect of a SNP on lung function using an additive genetic model

compared with non-smokers, such that the difference in lung function between groups is not consistent (characterised by the differing line gradients or non-parallel lines observed when comparing smoker and non-smoker groups). This scenario would insinuate that there is an interaction between genetic effect and smoking exposure, as the genetic effect on lung function is dependent on smoking behaviour. Additionally, interactions may not necessarily be characterised by effects that are consistent in direction for the two groups, as shown in **Figure 2.2C,** where the effect is deleterious in smokers but protective in non-smokers.

From a biological and epidemiological standpoint, the identification and replication of interactions will benefit the development of disease treatment, both curative and preventative. Knowing how exposures and genes interact, or how genes interact with each other can aid in the development of personalised medicine, where treatment can be tailored towards individuals exposed to a certain environmental exposure, or carrying certain gene combinations. However, the evident complexity in understanding the biological mechanism behind any observed statistical association in marginal effect studies is well documented (80,81), and such issues will still apply, if not increase, when exploring the contribution of interactions (82). Prior to analysis, in addition to

defining the variants or genes to analyse, careful consideration will be needed when choosing interacting factors to study, such as additional genes/variants or environmental exposures. Exploring environmental exposures that are unlikely to feature in a known biological pathway for the targeted disease phenotype, will make any statistical associations difficult to put into context. Beyond this, even when considering an exposure that is of relevance, interacting SNPs or variants may still reside in regions where determining their role biologically is complex, particularly when undertaking analysis genome-wide. Nonetheless, as with marginal effect studies, identification of interaction signals is the first step on the complex pathway to treatment development, and the methodology to do so is explored in the remainder of this chapter.

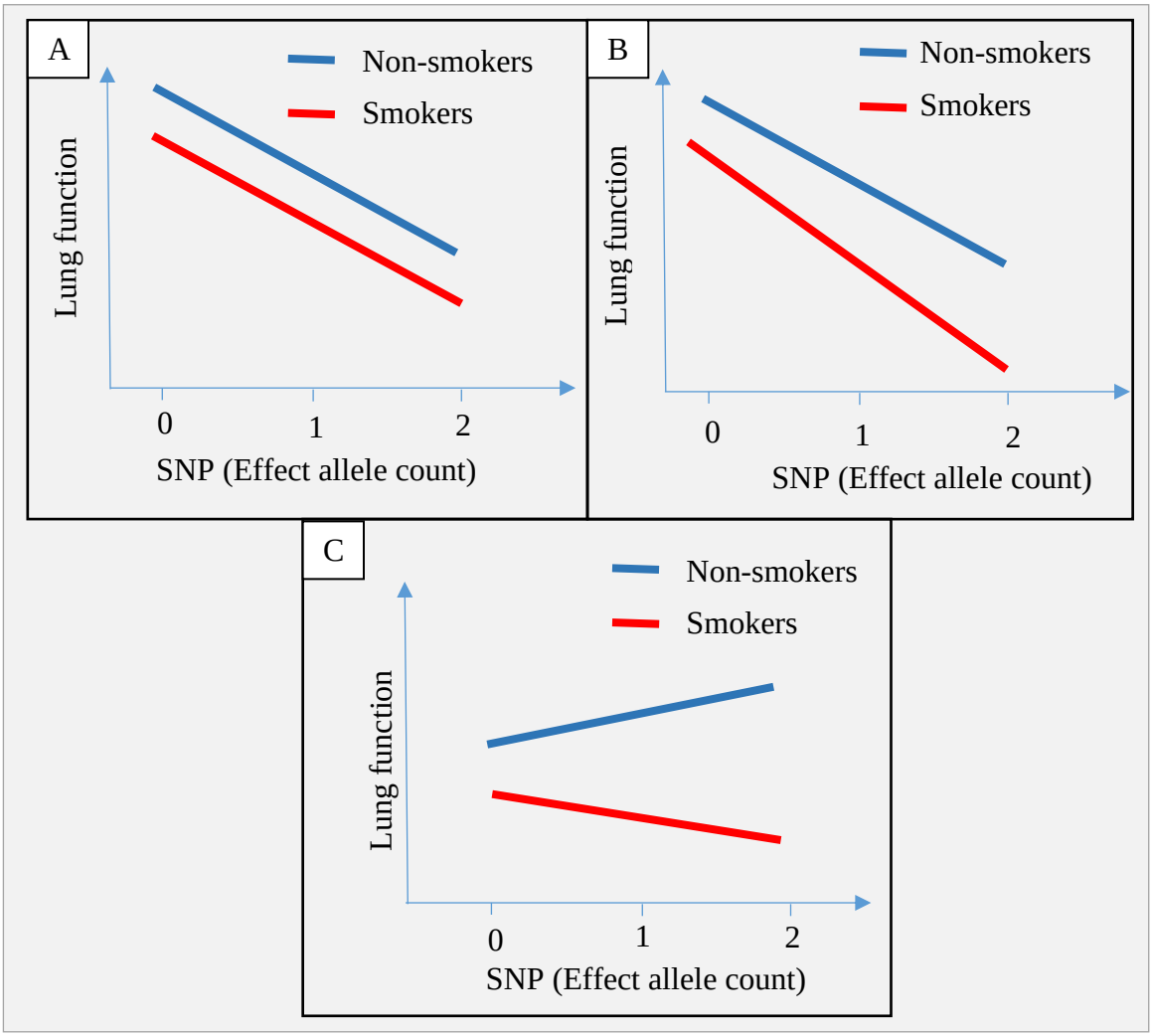


Figure 2.2 - An example of a deleterious effect of a SNP on lung function
(A) A smoking effect (B) An interaction effect between smoking and SNP such that effects are consistent in direction across exposure groups (C) An interaction effect between smoking and SNP such that effects are opposite in direction across exposure groups

2.2 Methods for interaction analysis

The incorporation of gene-environment interaction for genetic analysis in a statistical framework can be conveyed by extending the common GWAS approach used to test each variant individually (8). To recap, using logistic regression, the following model could be constructed to test association between genotype and a binary trait for each SNP:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 G_i \qquad (2.1)$$

Where p is the probability of the disease or trait, β_1 is the log(odds) effect size for the SNP tested and G_i corresponds to the categorisation of genotypes for the i^{th} SNP. Thus, e^{β_1} gives an odds ratio for the effect that SNP has on the disease, with statistical significance determined using a p-value generated from various statistical hypothesis tests. As previously discussed, this is a simplified representation and in practice other covariates may be added, for example principal components to address population stratification, but the underlying concept of marginal effect genetic association analysis is the same. Using an additive genetic model where the genotype, G, for each SNP, i, is categorised using $G = 0,1,2$ (to represent number of effect alleles present), the odds ratio for those with one copy of the effect allele (compared with those with the homozygous reference allele genotype) would be e^{β_1}, whilst for those with the homozygous effect allele genotype it would be $e^{\beta_1+\beta_1}$ (thus an additive log(odds) effect or conversely a multiplicative odds ratio effect).

2.2.1 Extending the methodology to consider interactions

Given a binary environmental exposure, one can extend the methodology to assess presence of interaction in the simplest sense by applying equation (2.1) to test marginal genetic effect on the binary phenotype, across environmental exposure groups (**Table 2.1**). This allows each group to have an individual independent genetic effect, as is the case in **Figure 2.2** in the smokers/non-smokers lung function example.

Table 2.1- Logistic regression equations applied across exposure groups to demonstrate a simple representation to determine presence of interaction

Logistic regression equation	Environmental Exposure
$logit(p) = \beta_0 + \beta_1 G_i$	Exposed individuals
$logit(p) = \beta_2 + \beta_3 G_i$	Unexposed individuals

Here, β_1 corresponds to the marginal genetic effect on the binary trait for individuals with the exposure, whilst β_3 is the equivalent effect for those without. In theory, in the absence of a gene-environment interaction β_1 and β_3 would be the same, and conversely significantly different values for β_1 and β_3 would suggest a differing genetic effect depending on exposure status, thus an interaction between SNP and exposure. The consistent difference observed in the example scenario in **Figure 2.2A** would be characterised by differing values of the intercept terms β_0 and β_2 but the same genetic effect, suggesting an exposure effect on the outcome which is independent of genetic effect. In contrast **Figure 2.2B** and **Figure 2.2C** show scenarios where both β_0 and β_2 differ, aswell as β_1 and β_3, suggesting an interaction is present. Determining presence of interaction could alternatively be undertaken by testing trait and environmental exposure association across genotype groups (which would also be the approach taken when considering a quantitative exposure variable), and comparing resulting exposure effect sizes. Although useful for explanation, this simplified approach would be cumbersome in application when addressing multiple SNPs (e.g. genome-wide) and thus can be reduced to a single regression equation with an interaction effect term as follows:

$$logit(p) = \beta_0 + \beta_1 G_i + \beta_2 E + \beta_3 G_i * E \qquad (2.2)$$

where $\beta_1$is the genetic main effect, β_2 is the environmental exposure (E) main effect (for some binary or quantitative environmental exposure – thus allowing the intercept to be different across groups as in **Figure 2.2A**) and β_3 is the interaction effect, where statistical significance can be determined as for marginal genetic effect in (2.1). Here, the interaction effect size, e^{β_3}, is the ratio of two odds ratios (i.e. the ratio of the odds ratio for the genetic effect for exposed individuals and the odds ratio for the genetic

effect in unexposed individuals) (42). This is equivalent to $\frac{e^{\beta_1}}{e^{\beta_3}}$ in **Table 2.1** and therefore concluding no interaction in **Table 2.1** when $\beta_1 = \beta_3$ is the same as $\beta_3 = 0$ ($\frac{e^{\beta_1}}{e^{\beta_3}} = 1$) in equation (2.2) (i.e. no dependence of genetic effect on exposure group).

A further point to note here is that in the scenario where individual level data is not available for the analyst and an equation such as (2.2) cannot readily be applied, there are tests available which use summary statistics from a marginal effect subgroup analysis (such as the genetic effects, β_1 and β_3, and corresponding standard errors from **Table 2.1**) to determine whether genetic effect significantly differs between the groups. An example of this is the Welch test or unequal variances t-test.

The above could be considered the simplest way to test for interaction effect for multiple variants across the genome. However, to model interactions more efficiently by addressing key issues such as sample size and power to detect interaction effects, number of variants analysed, and the minor allele frequency of analysed variants, various methodological approaches have been suggested. Interaction methods can be broadly grouped into 6 methodological categories based on their primary characteristics, although some methods may have characteristics attributable to more than one category. These are discussed in further detail for the remainder of this chapter.

2.2.2 Joint tests

Methods which allow for, rather than implicitly test for, an interaction effect (as was the case above), were proposed in the form of joint tests, and were introduced in a bid to improve the power of interaction analysis which implicitly tests for an interaction effect. These methods test for genetic main effect whilst allowing for interactions (i.e. jointly testing both effects). Kraft et al. (83) proposed such an approach in the form of the joint 2 degrees of freedom test. This test differs to the standard 1 degree of freedom interaction test applied in a logistic or linear regression setting, in that genetic main effect and interaction effect are tested simultaneously rather than the interaction effect only (**Figure 2.3**).

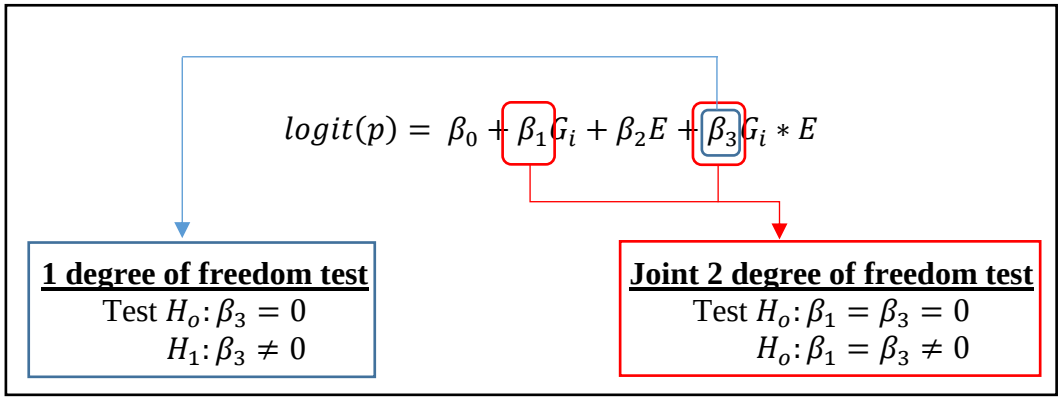


Figure 2.3 - Contrast between the standard interaction test hypothesis and the 2 degrees of freedom joint test analysis

The major difference between the 1 degree of freedom interaction term test and the 2 degree of freedom joint test, is that the latter does not give any information regarding direction or magnitude of effect (for the interaction or the marginal effect) but simply a statistical p-value. This means that the method determines whether the variant is associated using information on both main and interaction effect, but does not identify which is the driving mechanism behind the association. However, Kraft et al. showed how the joint test benefits from improved power over implicitly testing for GxE interaction only, when both main genetic effect and interaction effect are strong (have large magnitude) and also in situations where the genetic effect is only present in exposed individuals. Even when this is not the case, the joint test maintains similar power to that of the standard case-control analysis in the previous section, thus providing argument that this could be an efficient approach when the underlying assumption of genetic and/or interaction effect are unknown.

Dai et al (84) extended the single model joint test approach of Kraft et al. using two regression models. In contrast to testing two parameters in the same model as Kraft et al. did using genetic **main** effect and interaction effect, here the approach jointly tests genetic **marginal** effect and interaction effect (**Figure 2.4**).

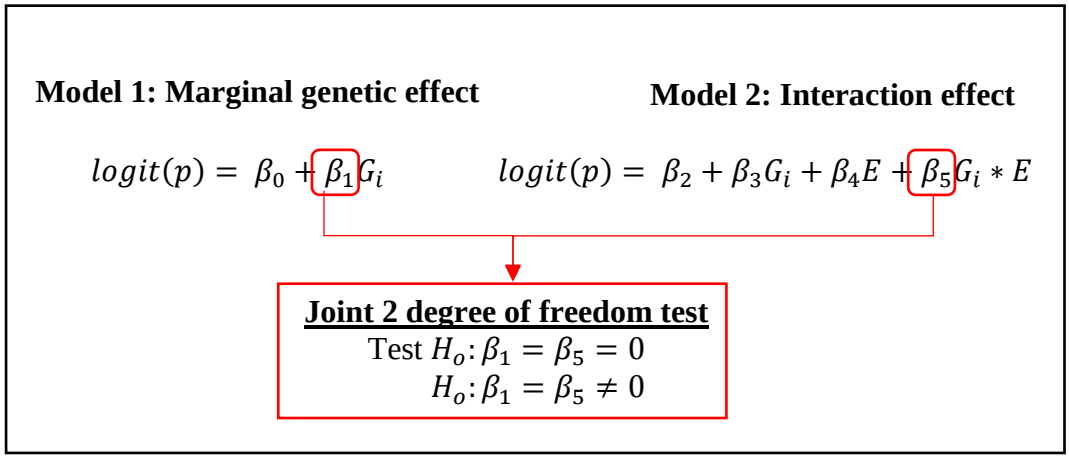


Figure 2.4 - The two model joint test approach

This approach has been shown to outperform the single model approach, specifically due to its flexible characteristics which allow for the use of testing approaches beyond the standard case-control analysis approach. Methods such as case-only analysis and the empirical Bayes estimator can be incorporated into the analysis, which are more powerful methods when considering a case-control design. These are discussed in the next section.

2.2.3 Case-only analysis

For binary traits, a particular discussion point regarding the superior methodology for GxE analysis has been the case-only study design, introduced by Piegorsch WW, Weinberg CR and Taylor JA in 1994 (85). Rather than comparing cases and controls for a particular trait to implicitly test or allow for an interaction effect, one would simply exclude (or avoid genotyping) controls. Presence of an interaction is then determined via association testing between environmental exposure and genotype. For example, one could use the equation:

$$logit(p) = \beta_0 + \beta_1 E_i \qquad (2.3)$$

Where p is the probability of having the mutation and E is the continuous or dichotomous environment variable. A test of $\beta_1 = 0$ would determine whether there is a

statistical association between the SNP and environmental exposure variable. Alternatively, one could use the exposure E as the outcome variable and test for differences in exposure across genotype groups to infer interaction. Should the null hypothesis of equal exposure prevalence across genotype groups be rejected, then this would be indicative of an interaction effect. Although showing increased power over case-control analysis (86), this design does however firmly rely on an assumption of gene-environment independence in the general population, with detrimental increases in type I error rates observed if the assumption is violated (87). For example, if considering variants which interact with smoking to affect COPD susceptibility, type I errors would increase for variants within the 15q25 region due to the association of 15q25 with smoking behaviour (thus violating the gene-environment independence assumption) in the general population. Typical to the characteristics of any assumption, having 100% confidence that it holds is difficult, and prior knowledge of the particular disease in the population may prove to be vital when relying on such an approach to deliver reliable conclusions. In response to the concerns regarding the case-only analysis, Albert et al (86) provided a comprehensive review of this method using simulation. Their work suggested that the case-only approach should only be considered in scenarios where knowledge for a particular gene–environment association is dense, either to confirm the gene-environment independence assumption or alternatively so that the results can be corrected if the assumption is false (using the reciprocal of the gene-environment association results, to scale the case-only results and thus account for the gene-environment association). Furthermore, they suggested the method should only be used if controls are difficult to obtain, or as a preliminary exploration stage to provide insight, prior to the more robust case-control analysis.

If the assumption of gene-environment independence severely affects the case-only approach, then one could perhaps implement a two-stage analysis. In the first stage, one could undertake a crude test of gene-environment independence in the population, by determining whether there is an association between the gene or variants in a set of controls. The second stage would then be to apply the case-only analysis should the assumption hold, otherwise considering another statistical approach if not. Although this would not remove the gene-environment independence limitation, it would reduce the risk of inflated type I errors by only analysing those genes or variants which have empirical evidence of independence from the environmental variable.

Further methods were created in order to benefit from the increased power of the case-only analysis over standard case-control methods, whilst minimising the negative effect of its sensitivity to gene-environment association in the population. Mukherjee & Chaterrjee (88) and Li & Conti (89) proposed methods that used both approaches in combination. Mukherjee & Chaterrjee suggested that although applying a two-stage design (gene-environment independence test as previously discussed and then case-only analysis) is one way of potentially weighting results (using a discrete weight of 0 or 1 based on whether or not the assumption was violated), it is not the least biased. They suggested use of an Empirical Bayes (EB) estimator, which used a continuous weighting method (thus not taking discreet values 0 and 1), weighting results from the two designs based on the belief of an underlying association between gene and environment. In contrast, Li & Conti suggested using Bayes model averaging, which consisted of averaging over estimators from each of the two methods. These methods are more robust to the gene-environment independence assumption, resulting in better controlled type I error rates, whilst also ensuring that variants or genes are not completely disregarded by an arbitrary cut off for statistical significance of gene-environment association. They are however methods which are tailored towards a candidate gene analysis, and their efficiency and practicality when conducting interaction analysis on a genome-wide scale needs further research.

2.2.4 Screening approaches

An immediate problem with simply analysing all possible SNPs and/or interactions, is the loss in power that inevitably comes from testing multiple loci individually due to the need for multiple testing adjustments. This issue is especially prominent in a genome-wide setting, where variants across the whole genome are included in the analysis. For example, use of a conservative multiple testing method such as Bonferroni correction, where the significance level is reduced according to the number of tests, could result in excluding potential genuine interaction associations. To avoid this penalisation, methods to reduce the exploratory search space for gene by environment interaction analysis have been proposed, commonly referred to as “screening” or “two step” methods (71). Firstly, SNPs are categorised or excluded via an initial requirement, for example, marginal trait association, with those associated taken forward for interaction analysis in step 2 (**Figure 2.5**). All such methods rely on independence between screening and analysis stages, thus ensuring multiple testing penalties need only be

considered when testing the refined number of variants passing the screening stage (90). This ensures that the penalisation is much smaller than if having to correct for the number of tests undertaken when analysing the "full search space".

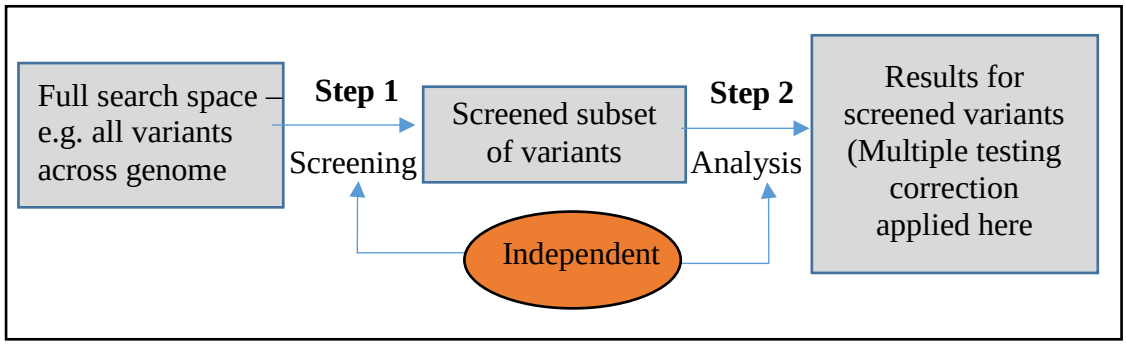


Figure 2.5 - The screening approach for GxE interaction analysis

Kooperberg & Leblanc (91) and Murcray et al. (92) were responsible for publishing early screening methods, which in time have formed the basis of since published and extended screening methodology. Kooperberg & Leblanc suggested choosing variants which showed statistical evidence of genetic marginal effect on the disease of interest, using the logistic regression equation in (2.1), or alternatively linear regression if considering a quantitative phenotype (presented in equation 1.1 in chapter 1 section 1.1.3.2). This therefore used an assumption that variants contributing interaction effects are more likely to lie within a subset of variants which produce a marginal genetic effect, with statistical evidence inferred by a pre-determined test threshold. In contrast, Murcray et al. suggested variants be screened on the basis of a correlation between SNP and environmental exposure in combined cases and controls, which they demonstrated as a characteristic indicative of GxE interaction. For example, should an interaction between environmental exposure and SNP be present in observation of effect on a binary trait, such that disease risk is only affected for exposed individuals with the SNP, then we would expect to see more cases with the environmental exposure and the SNP. This therefore would lead to correlation between SNP and environmental exposure in the cases. Furthermore, this method still maintains the requirement of independence between screening and testing stages in order for multiple testing to only be applied once. Although screening based on marginal effect does succeed in reducing the search space and increasing power over consideration of all variants by relaxing the multiple testing penalisation, it suffers due to the assumption that interaction effect must be

accompanied by marginal effect, which is not necessarily the case (35). Although generally a more powerful approach, the correlation method suffers larger losses in power when a particular disease is common in the population (prevalence close to 0.5) (93). This is because the method benefits from obtaining a sample which has a case prevalence larger than the population prevalence, which induces associations in the sample between genes and the environmental exposure. Producing this "extreme" sample would be more difficult when considering diseases which are common in the population. Therefore, Murcray et al. (93) suggested a hybrid screening method combining the strengths of the two, which aimed to provide a robust approach (to the often unknown underlying genetic architecture of disease) for GxE analysis. In this approach, a proportion of the genome-wide significance level is then allocated to each method, via an arbitrary probability p (set at 0.5 during simulation). Specifically, define the marginal effect screening method as M-screen and the correlation screening method as C-screen with respective variant selection thresholds of $\alpha_{\text{M-Screen}}$ and $\alpha_{\text{C-screen}}$, then should any variant meet such thresholds it will be formally tested in stage 2 for interaction effect. If the number of variants screened by M-screen and C-screen are N_m and N_c respectively, then the thresholds for determining statistical significance for interaction effect will be $\rho\alpha / N_m$ and $(1-\rho)\alpha / N_c$ for marginal and correlation screened variants respectively. Those variants achieving statistical significance for both screening methods are assessed for GxE statistical significance using the maximum of $\rho\alpha / N_m$ and $(1-\rho)\alpha / N_c$.

Hsu et al. (94) expanded on the screening methods above with the **cocktail** screening and GxE analysis approach. The aim was to produce a flexible 'pick and choose' three step screening approach, after noticing the potential rigidity of the three previous methods, which only considered a case-control design to test screened variants for interaction effect. In a bid to improve the power of screening approaches, the three steps consisted of a screening step, a multiple testing correction step and an analysis step (**Figure 2.6**).

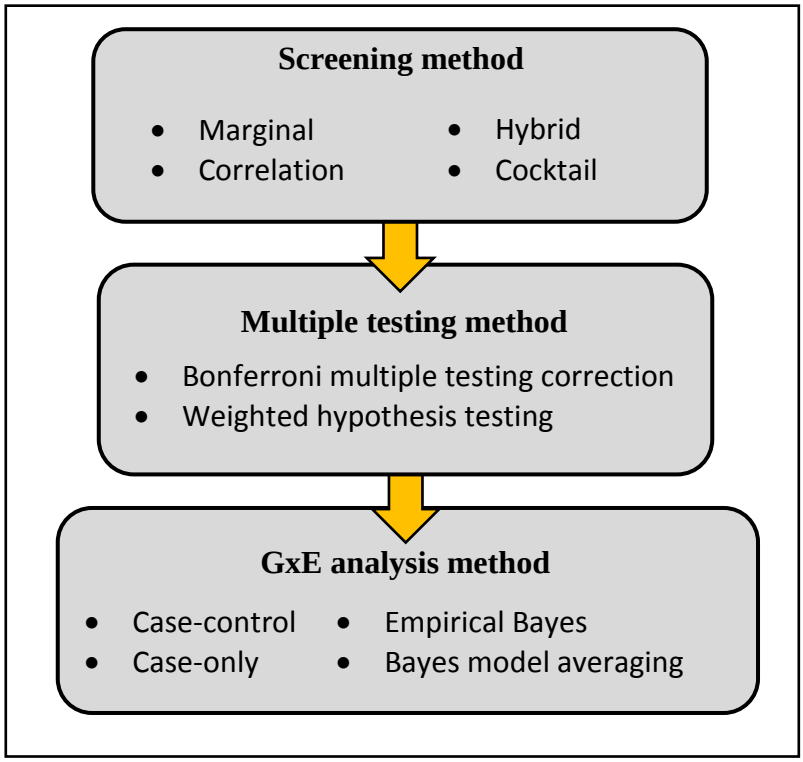


Figure 2.6 - The three module approach to screening for GxE

The testing approaches were expanded to consider the case-only method as well as the case-only/case-control hybrid methods i.e. the empirical Bayes and Bayes model averaging. In addition, weighted hypothesis testing was suggested in addition to the widely used Bonferroni correction for multiple testing correction. The cocktail method of screening in **Figure 2.6** refers to the two cocktail methods suggested (**Figure 2.7**), similar in design to the hybrid screening method of Murcray et al (93), in that both marginal and correlation methods were used in combination. Cocktail method one consisted of choosing the correlation or marginal p-value dependent on whether the marginal p-value was above or below a predefined threshold. The second cocktail method took the minimum of the two p-values to screen variants. Weighted hypothesis testing was used to correct for multiple testing, and the Empirical Bayes or case-control method was used to test variants depending on which method was used to screen the variant. This was to ensure that screening and testing steps remained independent.

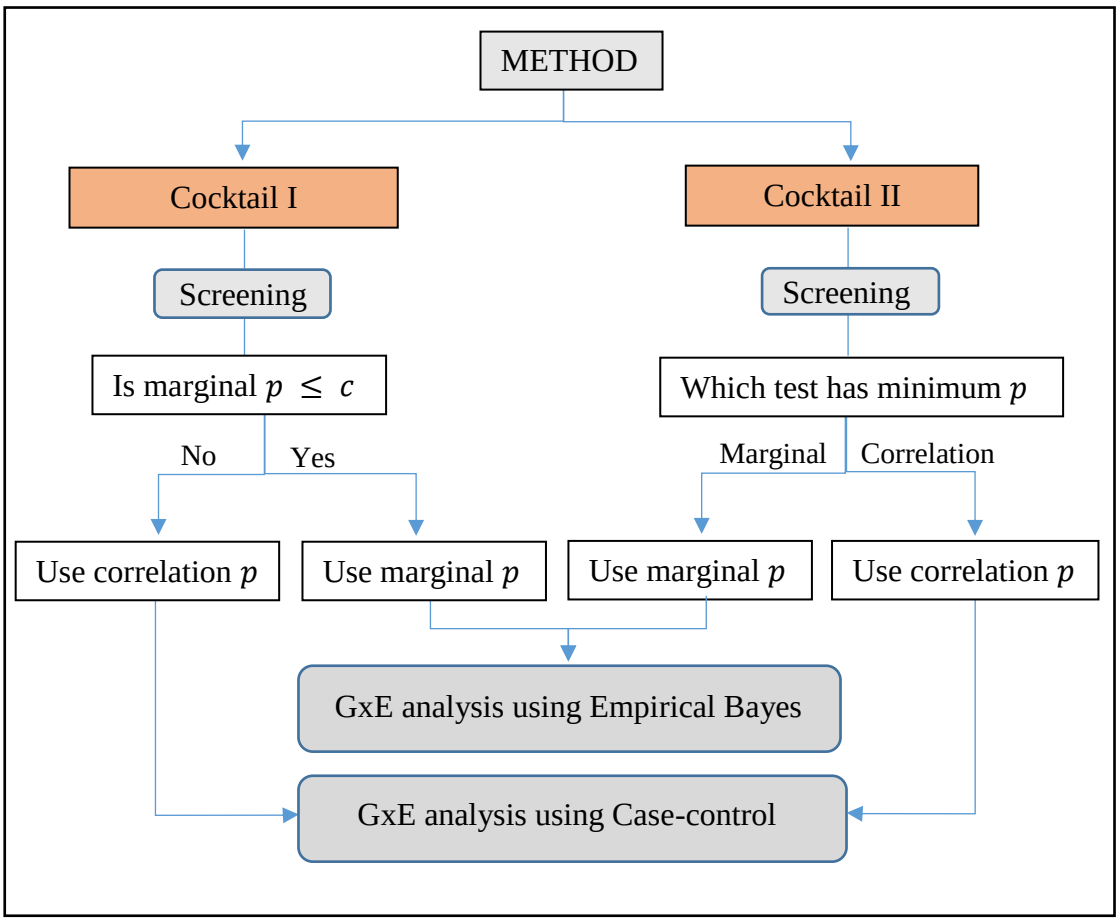


Figure 2.7 - The two cocktail approaches for screening variants prior to GxE analysis

The cocktail methods proposed provided a flexible and robust approach to suit many interaction scenarios, matching powers of the previous screening approaches discussed, whilst also allowing for use of the powerful case-only testing approaches such as the empirical Bayes estimator where possible. Absence of screening was also considered, with severe power losses witnessed in comparison to screening SNPs in the majority of scenarios.

Since the work undertaken by Hsu et al., Gauderman et al. (95) questioned whether, rather than ultimately choosing SNPs for analysis via marginal screening *or* correlation screening as suggested by Murcray et al. and Hsu et al., power could be improved by choosing SNPs using a combined test statistic. They suggested summing the test statistics of the marginal and correlation screening analyses to minimise loss of information, thus providing a more efficient and reliable method of screening. Simulation suggested this method was more powerful when genetic marginal effects of

those SNPs screened are modest. The authors argued that this approach benefits from utilising all available information to screen variants, in contrast to previously, where inclusion of SNPs showing evidence of marginal and correlation effects was dependent on one or the other (for example the larger test statistic of the two).

2.2.4.1 Quantitative trait specific screening

With the methods previously discussed largely focussed on binary traits, a method specifically aimed towards quantitative trait analysis was suggested by Pare et al. (96). They proposed that when analysing a biallelic SNP (a SNP in which there are only two alleles present at the locus), presence of interaction could be flagged via variation differences in the quantitative phenotype between the three genotype groups. Thus, in a "variation prioritisation" screening step, those SNPs showing evidence of differences in phenotypic variance between genotype groups were chosen and analysed using standard linear regression. Although an observed variation is not a confirmation of interaction presence, this method removes the reliance on screening by marginal effect which, as previously discussed, could fail to detect genuine interactions which do not produce a marginal association with the disease. In addition, this method does not require the environmental variable to be explicitly defined, and thus can be used for exploratory analysis when the contributing environmental exposure is not known.

2.2.4.2 Screening based on statistical and biological plausibility

Finally, in contrast to the potentially complex approach of incorporating both screening and analysis stages into the same GxE exploration, and to avoid the debate of which statistical approach to use, Chirag J. Patel (97) proposed reducing the search space manually using prior information from GWAS hits. Specifically, this refers to reported and since replicated genetic associations with the disease of interest. Evidently, this approach is less computationally intensive, utilising previous GWAS efforts as a substitution for computationally screening variants. Also, it provides analysts with a set of replicated variants for which functional and biological relevance will have already been explored, allowing for greater context should an interaction be detected. In addition to this, one could also consider variants in the literature associated with related traits, for example if lung function is of interest, one may look at associated variants

with asthma. There are however two main drawbacks with such an approach. Firstly, although this method may successfully reduce the search space when testing for interactions, the extent to which it is reduced will depend on what is currently known about the phenotype of interest. For example, for lung function, we may have a search space of over 200 associated signals for which we focus our interaction analysis efforts (see chapter 1). However, for a rare condition such as idiopathic pulmonary fibrosis (IPF) which has a significantly fewer number of identified associations (98-100), the search space may be reduced so significantly that the possibility of identifying interactions will be slim. Secondly, with the majority of identified associations previously reported due to exhibiting a marginal genetic effect on disease, the same assumption applies to that of applying a marginal screening method i.e. a SNP with a marginal effect has larger potential of producing an interaction effect. There is therefore a risk of excluding potentially strong interaction effects by ignoring SNPs with marginal effects which were not statistically or genome-wide significant in previous studies.

In addition to screening variants due to their statistical significance for a given trait from previous GWAS efforts, one could instead reduce the full search space by functional relevance. For example, analysis could be restricted only to SNPs which code for proteins for example or only intronic SNPs. Software such as PolyPhen-2 (101), Sorting Intolerant From Tolerant (SIFT) (102), Functional Analysis Through Hidden Markov Models (FATHMM) (103) and DeepSEA (104) can be used to prioritise protein-coding SNPs on the basis of whether the resulting amino acid changes severely affect protein structure.

2.2.5 Data mining

A methodological category which shares similar philosophies to the screening approach, is data mining. The non-parametric approach (making no assumption about the distribution of the data) consists of reducing the multidimensional workspace of genetic architecture, and looking for patterns in higher dimensional data to assist with further analysis. These methods are more accredited to generating hypotheses to explore further rather than testing them, as they allow for interactions rather than implicitly assess their effect (similar to the concept behind the joint tests in section 2.2.2). Therefore, much like screening methods, they can provide useful information and

direction, before proceeding with a secondary interaction analysis stage. Random Forests (RF) and Multifactor Dimensionality Reduction (MDR) proposed by Breiman (105) and Ritchie et al. (106) respectively, are two common approaches that adhere to the data mining philosophy.

Random forest analysis is an exploratory technique used in machine learning. The concept behind machine learning is to make use of previous available data in order to build automated models for prediction. For random forests specifically, a regression tree or decision tree is created to assess the role of certain predictors, with the random element introduced by allowing for **multiple** regression trees which may harbour differing prior beliefs (**Figure 2.8**). In its simplest form, a decision tree is a flow chart constructed of a number of questions or classifications (labelled criteria levels in **Figure 2.8**). It allows us to use previous knowledge and data to form a final prediction or answer a specific question, with each question conditional on the previous answer. For example, determining the temperature tomorrow will be dependent on the season, the weather today, the average weather over previous years for this season etc. Whilst one decision tree may use one source for previous weather data, another tree may use another source. This will lead to variable answers between trees and all trees can be collated to provide a final prediction, such as the average of prediction temperatures computed, or the most common answer, as examples.

Random forest theory can be applied in a genetic framework by considering SNPs/genes and environmental exposures as the predictors (or questions/classifications in the flow chart analogy) to form final predictions of how they affect disease susceptibility. Alternatively, prediction may not be the aim and instead determining the importance of particular predictors would be more useful to provide direction, often defined by variable importance measures (35,107). Given the conditional nature of regression trees, such an approach could be useful in the exploration of simple or higher order gene-gene or gene-environment interactions, assessing whether the effect of genes or SNPs are still present given the effect of previously considered SNPs/genes or environmental exposures (108). Although not producing a formal test of interaction, random forests could be used to capture and highlight variants with potential for interaction effects. For example, Maenner et al. (109) used random forests to explore gene-environment interaction effects on coronary heart disease. Alternatively, random forests could be used to screen potentially interacting variants for a further testing stage

to explore, such as the approach implemented in a study by De Lobel et al. (110), who used random forests to screen for gene-gene interactions.

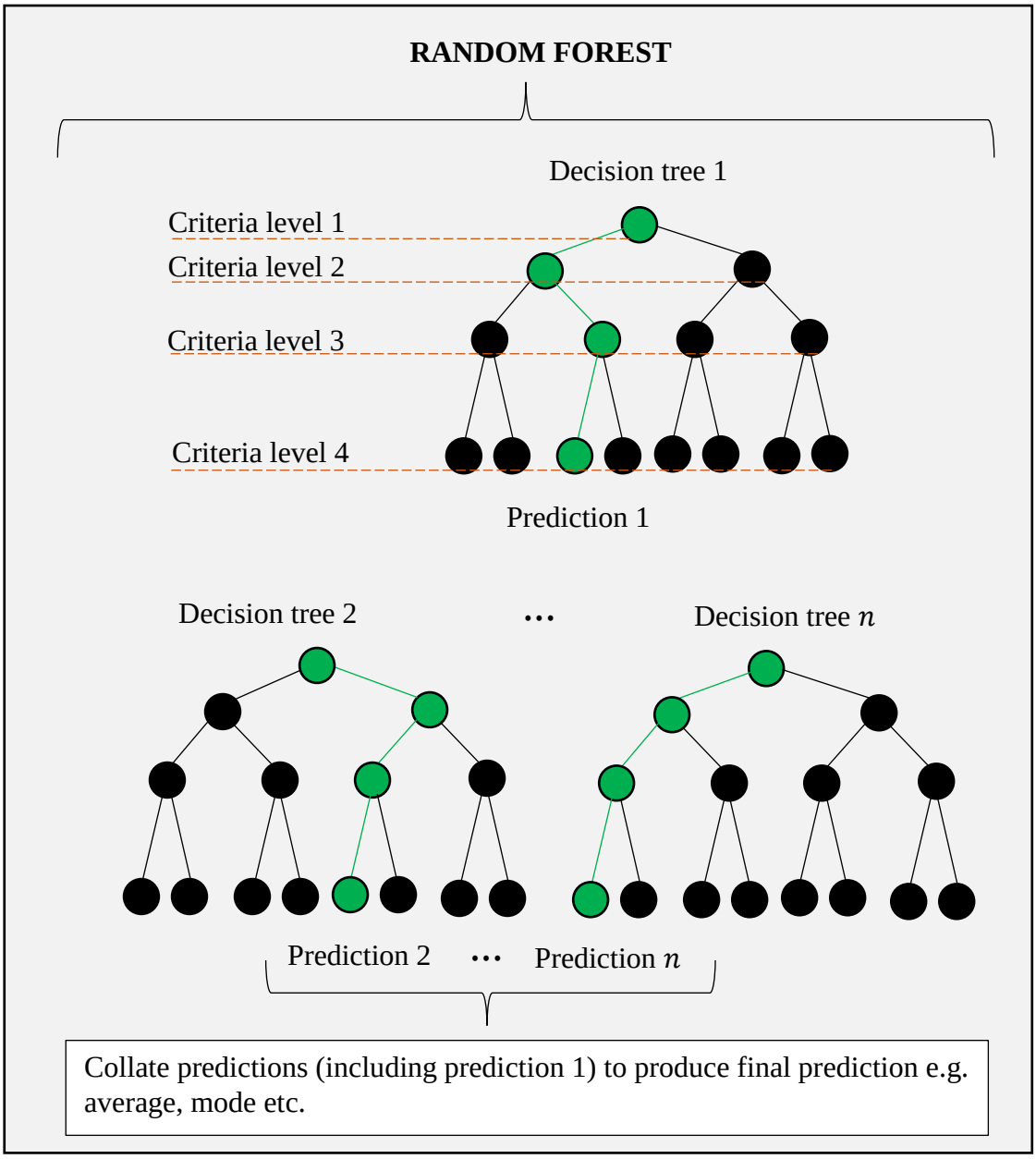


Figure 2.8 - The concept of decision trees and random forests
The green circles represent the path taken to reach the prediction for each individual decision tree based on the criteria used

The multifactor dimensionality reduction method was originally proposed to explore epistatic effects. It works by collapsing information on multiple markers into two discrete groups of high risk and low risk genotypes, by computing case control ratios (**Figure 2.9**). As a result it reduces the multidimensionality of genotype groups into two dimensions (high risk vs low risk) and then explores the effect of the collapsed variable on the disease of interest.

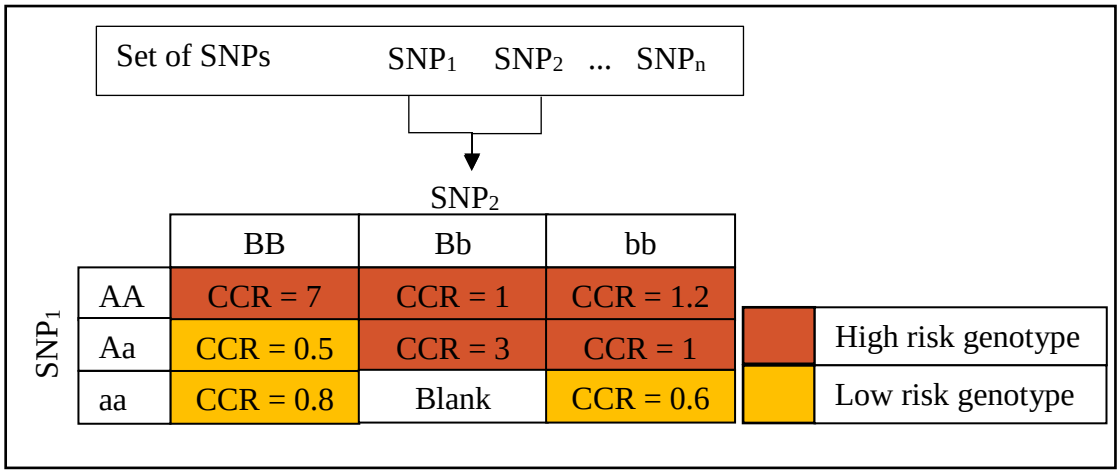


Figure 2.9 - A simple example of multifactor dimensionality reduction
CCR = case-control ratio

To do this, the accuracy of the MDR model chosen is assessed by predicting the disease status of a subset of the sample multiple times, with a prediction error defined as the proportion of individuals incorrectly categorised. The MDR model with the lowest prediction error is then used, with the consistency of this chosen model (how many times it arises) compared with the consistency under a null hypothesis of no disease association, using permutation to generate a p-value.

An extension to MDR, namely model based multifactor dimensionality reduction (MBMDR) (111), was also introduced. The aim of this method is to provide a model based approach with flexibility for addressing quantitative traits, whilst also allowing for the inclusion of covariates, which the standard MDR method could not incorporate. In contrast to the standard MDR approach which uses case-control ratios to define multi-locus genotypes into high risk and low risk categories, the MBMDR method assesses the effect of each genotype cell on the disease using regression and a pre-defined threshold (by determining the effect of that genotype on disease compared to all other individuals), before categorising genotype cells into three categories; high risk, low risk and no evidence. The effect of the collapsed groups on the disease of interest are then determined with the use of regression.

Although these methods could prove to be an effective tool for directing future GxE analyses, their suitability for this purpose needs further research. Alterations need to be made to be able to efficiently incorporate these methods in any exploratory or screening GxE stage with confidence. This is in part due to their original focus on GxG and for the random forest approach in particular, incorporation of environmental exposures into

the model brings added complexities. Specifically, it is unclear whether the best approach is to use an outcome variable which is pre-adjusted for the environmental exposure, or whether the environmental exposure should be incorporated within the decision tree using an unadjusted outcome. As a result, application of these methods for the purpose of GxE analysis is at the moment unclear (35,107).

2.2.6 Gene based methods

Another methodological area for the analysis of GxE, is the gene-based approach. Gene-based approaches have been explored in depth in the context of marginal genetic effect analysis and have often been used for the analysis of rare variants (particularly for sequencing data which gives the ability to directly genotype rare variants rather than estimating them using imputation), where the single marker approach may lack power (112). Additionally, gene-based approaches offer the incentive of improved biological context when following up statistical associations over individual SNP analysis, for which determining the biological context and assigning the variant to a particular gene or function can be difficult. Although not the focus of this thesis, gene-based approaches are described in the proceeding paragraphs for completeness.

The primary characteristic of the gene-based approach is to pool or combine variants within a specified gene or region and test them together, rather than testing each one individually. For example, a pooled association or burden test collapses information over several genetic variants into a single score. Such methods can benefit from increased power when regions harbour more than one causal variant and direction of effects on the analysed trait are in the same direction (112). However, with such methods unable to capture individual variant effects within the gene or tested region, power is lost when variants have effects in opposite directions with differing magnitudes or when there is a large contribution of non-causal variants within the gene or tested region (8). The burden test gene-based approach can be introduced as an extension of the previously discussed regression framework used to explore the marginal effect of variants on the outcome of interest in GWAS. For example, for the sum test, which is a method originally proposed for common variants referred to in studies by Pan (113) and Chapman and Whittaker (114), the single score is the sum of the allele counts for all variants within the region. Given k variants under an additive

genetic model such that $X_{ij} = 0,1,2$ is the allele count for variant j for individual i, the sum test can be characterised using logistic regression as:

$$logit(p) = \beta_{0,C} + \sum_{j=1}^{k} \beta_C X_{ij} \qquad (2.4)$$

where $\beta_{0,C}$ is the underlying log-odds of having the disease and β_C is the common association strength (log-odds ratio) for the summed allele counts. A test of $H_o: \beta_C = 0$ can then be used to determine whether there is a statistically significant effect of the combined genotype score on the outcome. This regression concept was further extended with introduction of the Cohort Allelic Sums Test (CAST), the Combined Multivariate and Collapsing method (CMC) and the weighted sum test (115-117), all of which had different caveats to their application. The CAST method used an indicator variable (0 or 1) for whether a rare variant was present, rather than using a collapsed sum of all the variants. The CMC method segregated common and rare variants for analysis. The weighted sum test suggested weighting variants according to their minor allele frequency, rather than equally (as was the case with the sum test), giving larger weight to rare variants. This is simply an introduction to the vast plethora of rare variant methods, with such pooled association or burden tests one of the original methodological groups available. Other methodological groups, some of which also use a regression framework for analysis such as data adaptive, variance-component, and combination methods are also available (118-127). The benefit of variance-component methods over burden methods is the removal of any implicit assumptions regarding all variants within the region being causal or that the magnitude and direction of effects are consistent across variants. An example of such a method is the Sequence Kernel Association Test (SKAT) (127), which uses a variance-component score test to test the coefficients for individual variants in a regression framework. This provides a flexible method allowing variants to be causal/non-causal and to have differing directions and magnitudes of effect. However in the scenario where the majority of variants in a region are causal with consistent direction, the burden test is still more powerful (112). Further information regarding the methods suggested for efficiently and accurately analysing gene-based marginal effects can be found in thorough reviews such as those produced by Asimit & Zeggini (8) and Lee et al. (112).

For interaction analysis, difficulties with rare variants still apply, and the appealing prospect of generating more biologically plausible results mean that gene-based approaches tailored towards interaction analysis are in demand. Although such methods proposed to date are more directed towards epistatic effects (interactions between genes rather than gene-environment), they do have the flexibility to be extended for GxE interaction analysis (35).

Naturally, a starting point would be to extend the gene-based marginal effect approach in equation (2.4), to incorporate gene-based interaction analysis by summing over both the SNP effect and the interaction effects using the following equation:

$$logit(p) = \beta_{0,C} + \beta_C \mathrm{CV} + \beta_E \mathrm{E} + \beta_{CE} CV * \mathrm{E} \qquad (2.5)$$

Where $CV = \sum_{j=1}^{k} X_{ij}$ is the collapsed variable for a set of k SNPs, β_E is the effect on the outcome for the environmental exposure, and β_{CE} is the interaction effect term for the interaction between the collapsed variable and the environmental exposure. However, it is unclear as to how such an approach would behave, given that the method was originally tailored towards rare variants with assumptions that may not necessarily adhere to the analysis of GxE effects (128). For example, as previously discussed, the power of burden gene-based tests for marginal effect rely on an assumption of consistent direction for variants pooled into the single score statistic, an assumption which may not apply to GxE effects. To combat this, Jiao et al. (128) suggested the Set Based gene EnviRonment InterAction test (SBERIA) in which the characteristics of the combined interaction effects could be inferred with the use of correlation screening as described in section 2.2.4. The correlation screening step enables variants to be weighted -1, 1, or 0 depending on their deleterious, protective or non-causal effect on the disease. As before, the screening and analysis steps are independent, thus type I error need only be addressed in the testing stage.

A further method devised with the aim to incorporate the complexity of multiple variants in an interaction testing stage, and also applicable in a regression framework, is the Tukey's 1 degree of freedom test proposed by Chatterjee et al. (129). This method

utilises the test of non-additivity proposed by John Tukey (130) to determine the presence of interaction between 2 factors. In summary, for two genes, pair-wise interaction effects between multiple SNPs are incorporated into a 1 d.f. test by first regressing each gene on an underlying phenotype. The two phenotypes are assumed to be causal for the disease of interest. The resulting phenotypes are then incorporated within a logistic regression framework to determine the association with the disease of interest, for which the interaction effect between the two can be observed. With the attractive characteristic of providing a test with a low number of degrees of freedom, the method aims to be a powerful approach to test the genetic effects of multiple variants within a gene, whilst allowing for GxE. However, its restrictive assumption of proportionality between interaction and marginal genetic effects, and its computational complexity potentially limiting it to candidate gene analysis, led Wang et al. (131) to extend its methodology. In this method, the authors proposed using a Least Squares Kernel Machine (LSKM) as an efficient method to simultaneously test the effect of multiple SNPs within a region. Kernel machines are a machine learning technique, used to transform data into a higher dimensional space, to explore underlying patterns and aid in the fitting of statistical models. LSKM, a non-parametric regression approach, benefits from a more computationally simple approach than that offered by the Tukey's 1 degree of freedom. With use of the score test, the method is less computationally demanding allowing for genome-wide application, and various interaction effects could be modelled via simply altering the kernel function using different distributions (quadratic, Gaussian etc.).

Further gene-based approaches have been suggested which are more primarily tailored towards incorporating the contribution of rare variants. Methods such as gene-trait similarity regression and an extended sequence kernel association test (SKAT), proposed by Tzeng et al. (70) and Chen et al. (132) respectively aimed to provide powerful approaches for both common and rare variants. In the first of these methods, Tzeng et al. proposed collapsing information via genetic similarity (in which rarer variants carry more weight) between pairs of individuals, rather than across multiple genotypes within a region. An interaction effect can then be assessed between this collapsed genetic similarity variable and an environmental exposure (or alternatively another collapsed genetic similarity variable using a separate set of SNPs), by including an interaction term in the model, for which the coefficient can be tested. Although

originally proposed for quantitative traits, this method has since been implemented in software with a binary trait option. The second method uses the statistical framework of the revered SKAT method for gene-based marginal genetic effect analysis devised by Wu et al. (127), to produce two methods of testing for interactions (one allowing for fixed genetic effects and one allowing for random genetic effects), alongside a joint test of main and interaction effect. The method benefits from modelling main and interaction effects individually rather than via a single kernel function, as was proposed in previous gene-environment interaction joint test methods incorporating the SKAT approach (133). Chen et al. also highlighted the adaptability of the gene-environment set association test (GESAT), a common variant gene-based method devised by Lin et al. (134). This method can be adapted through the incorporation of weights for rare variant interaction analysis. Using a variance-component score test within a regression framework allows for random interaction effects between the SNPs in a set and the environmental exposure variable. The method incorporates ridge regression to account for high LD between variants in a set, for which the alternative least squares estimates can produce large estimates of parameter variances.

With data mining techniques previously discussed showing potential to aid the understanding of the complex genetic mechanisms driving disease susceptibility, the methodology has also been incorporated into novel gene-based interaction methods. Oh et al. (135) put forward the gene-based multi factor dimensionality reduction approach. The MDR methodology accredited to Ritchie et al. (106) discussed in section 2.2.5 is incorporated and extended for gene-based analysis in 2 steps. Firstly, within-gene MDR applied to all SNPs within a gene produces a summarised effect for each gene. It essentially does this by determining the SNP subset which best fits the observed data. Methodologically this follows the example in **Figure 2.9**, however considers multiple sets of SNPs rather than a single set. MDR is then applied at the gene-level in the same fashion in a second step, to explore interaction effects. Gauderman et al. (136) suggested genetic characteristics could better be captured through the use of principal components, a statistical technique often used to assess the effect of population stratification in large GWAS. One could reduce the multidimensionality of SNPs within multiple genes or loci into a number of principal components, which can then be included in a regression model as covariates, including interaction terms (i.e. principal component by environment interaction analysis). Similarly, the canonical correlation

based approach accredited to Peng et al. (137), sought to reduce the complex dimensionality of genotype data by producing estimates of correlation between genes, utilising linear transformation for SNPs contained within their region along the genome. Although these methods show great potential for interaction exploration, they are very much hypothesis generating approaches rather than hypothesis testing. As a result of using dimension reduction techniques, interpreting and putting any generated results into context can be difficult and thus these methods become strong contenders for use in a preliminary stage before analysis, to aid direction. Furthermore, only the method proposed by Gauderman et al. appears to be easily applied to GxE analysis, whilst application of the others for this purpose is currently unclear.

2.2.7 Pathway-level analysis

A further subgroup of methods with potential application to GxE analysis are pathway-level analysis methods. These methods heavily rely upon previous biological knowledge of a disease or trait, and the consideration of more than one gene. Such gene-set analysis (GSA) approaches were suggested to address the potential lack of functional relevance resulting from identifying small numbers of *single* SNPs/genes associated with a disease or trait, rather than exploring how numerous variants or genes working in unison affect disease susceptibility. This means that successfully uncovering the driving mechanisms for disease can be difficult (138). Gene Set Enrichment Analysis (GSEA) is a method which relies upon this approach, and was originally proposed for the efficient analysis of genome-wide gene-expression studies (139). The approach involved determining whether pre-defined gene sets were randomly or systematically distributed amongst all genes studied, when ranked by expression difference between cases and controls. This was suggested as an alternative to simply focusing in on those genes which showed the largest expression differences, with the selection of such genes then producing difficulties when seeking a biological explanation for why they are influential. This approach has since been altered for application in association studies (140-142) and Zhang et al. (138) created the improved GSEA for GWAS (i-GSEA4GWAS) web server, which is a free to access online database, providing potential for identifying disease correlated gene-sets. Currently there is limited published work of such an approach being used in a gene-environment interaction setting. Most notable was the study undertaken by Wei et al. (143), where

the approach was used to explore pathway-level interactions between asbestos exposure and gene sets in lung cancer susceptibility, which found a potential interaction between asbestos exposure and the Fas signalling pathway. There is potential for pathway-level methods to be incorporated into other methodological categories previously discussed. For example, one may use pathway-level approaches as a screening tool to prioritise biologically relevant SNPs for interaction analysis in single marker gene-environment interaction tests. Alternatively, these methods could also be incorporated to screen for functionally relevant genes to test using a gene-based approach, such as those discussed in section 2.2.6.

2.3 Summary of methods

The methods discussed in this chapter are summarised in **Table 2.2** alongside a brief description. The methods are grouped into the broad categories discussed and advantages and disadvantages for each methodological category are presented.

Table 2.2 - Summary of methods for gene by environment interaction analysis

Category and description	Method (Reference)	Method description	Advantages	Disadvantages
Interaction test	-	Implicitly test interaction effect size in a regression framework	Computationally simple and produces a measurable interaction effect	Underpowered for rare variants
Summary statistics test e.g. Welch test	-	Tests for genetic effect difference between two samples or subgroups	Can be used to determine presence of interaction effect when individual level data is not available	Not as informative as methods that can be applied on individual level data and analyses will need to be consistent across subgroups (same covariates used etc.)
Joint tests Joint test of main/marginal effect and interaction effect	One model 2 d.f. joint test (83)	Jointly testing main genetic effect and interaction effect within the same regression model	Has better power than testing for interaction effect only in certain scenarios e.g. when both main and interaction effects are strong	Does not produce an interaction effect or test for interaction individually (unclear whether the marginal genetic effect or interaction effect is the driving mechanism behind the association)
	Two model 2 d.f. joint test (84)	Jointly testing marginal genetic effect and interaction effect from separate regression models		
Case-only analysis Analysis in cases only	Case-only regression analysis (85)	Analyse association between gene/SNP and environmental exposure in cases to determine presence of interaction	Increased power over case-control analysis for interactions	Large type I error rates if gene and environment independence assumption in population is violated

Category and description	Method (Reference)	Method description	Advantages	Disadvantages
Case-only and case-control Hybrid of case-only and case-control study designs	Empirical Bayes (EB) estimator (88)	Uses both case-control and case-only analysis, weighting results based on belief of underlying dependence between gene and environment in the population	More robust than case-only analysis to gene and environmental exposure independence assumption	Aimed at candidate genes and not GWAS
	Bayes model averaging (89)	Averaging over the case-only and case-control estimators		
Screening Reducing the search space of variants by implementing a screening step prior to analysis	Marginal screening (91)	SNPs screened for showing a marginal effect with the disease	Reduces the penalisation of multiple testing correction due to analysing a subset of SNPs	Can suffer from strong assumptions e.g. only marginal effect SNPs will produce interaction effects
	Correlation based screening (92)	SNPs screened for showing a correlation with the environmental exposure in combined cases and controls		
	Hybrid method (93)	Uses both correlation and marginal screening with case-control analysis		
	Cocktail methods (94)	Use of both correlation and marginal screening, with combination of analysis methods (EB and case-only)		

Category and description	Method (Reference)	Method description	Advantages	Disadvantages
Screening Reducing the search space of variants by implementing a screening step prior to analysis	Gauderman et al. 2 step screening method - EDGxE method (95)	Use sum of test statistics from marginal and correlation screening to screen SNPs	Reduces the penalisation of multiple testing correction due to analysing a subset of SNPs	Can suffer from strong assumptions e.g. only marginal effect SNPs will produce interaction effects
	Variance prioritisation method (96)	Screen SNPs on the basis of an observed phenotype variance difference between genotype groups		
	Previously associated SNPs (97)	Screen SNPs based on their reported associations in the literature		

Category and description	Method (Reference)	Method description	Advantages	Disadvantages
Data mining methods Hypothesis generating methods consisting of exploring higher dimensional data to provide direction for analysis	Random Forests (RF) (105)	Looking for patterns in high dimensional data to determine the predictors of the disease or outcome using regression/decision trees	Complexity of high dimensional data can be simplified to identify possible predictors for the trait of interest	Hypothesis generating not hypothesis testing so would need an additional analysis stage
	Multifactor Dimensionality Reduction (MDR) (106)	Reducing combinations of genotypes at multiple loci into a single low risk/ high risk variable, and exploring association between the binary phenotype and collapsed variable		
	Model Based Multifactor Dimensionality Reduction (MBMDR) (111)	Extended MDR in regression framework to consider covariate adjustment and quantitative traits		

Category and description	Method (Reference)	Method description	Advantages	Disadvantages
Gene-based methods Methods in which instead of considering a single SNP, several SNPs are considered together e.g. SNPs within a gene	Set Based gene EnviRonment InteAction test (SBERIA) (128)	Uses correlation screening to assign weights to SNPs within a gene (based on their direction of effect), to aid interaction analysis between genes and an environmental exposure variable	More powerful in the presence of rare variation	Often unclear how to implement genome-wide and some methods aimed at gene-gene interactions rather than gene-environment interactions Data-mining gene-based methods more hypothesis generating than testing and results may be difficult to put into context (e.g. principal component interaction)
	Tukey's 1 degree of freedom model of interaction (129)	Incorporates pairwise interactions of multiple SNPs in a region into a 1 degree of freedom test for GxG/GxE interaction		
	Extension to Tukey's 1 degree of freedom model of interaction (131)	Exploits both marginal and interaction effects using a partial least squares algorithm applied to data from multiple SNPs in a gene		
	GESAT (134)	Uses a variance component test to analyse snp-set by environment interactions and estimates genetic main effects using ridge regression		

Category and description	Method (Reference)	Method description	Advantages	Disadvantages
Gene-based methods Methods in which instead of considering a single SNP, several SNPs are considered together e.g. SNPs within a gene	PCA interaction analysis (136)	Derive principal components for each gene of interest and use a logistic regression model including PC-PC interaction terms	More powerful in the presence of rare variation	Often unclear how to implement genome-wide and some methods aimed at gene-gene interactions rather than gene-environment interactions Data-mining gene-based methods more hypothesis generating than testing and results may be difficult to put into context (e.g. principal component interaction)
	Canonical correlation based method (137)	Obtains systematic correlations between genes using a linear transformation of SNPs within them		
	Gene-based MDR method (135)	Two step procedure of within and between-gene MDR 1. Use MDR to jointly model effects of multiple SNPs within a gene 2. Perform MDR analysis on these summarised gene-level effects		
	Gene-trait similarity regression (70)	Evaluate similarity between pairs of individuals to assess gene-level GxE interaction. Regress trait similarity on genetic similarity using an adaptive weighting method to decipher common and rare variants		
	Extended SKAT method (132,144)	Extension of the SKAT method for rare variants to incorporate interactions		

Category and description	Method (Reference)	Method description	Advantages	Disadvantages
Pathway level analysis Exploring the effect of multiple pre-defined functionally relevant gene-sets	Gene Set enrichment analysis (GSEA) (139)	Determine whether pre-defined gene sets are randomly or systematically distributed amongst all genes studied, when ranked by expression difference between cases and controls - can be extended to study pathway level interactions	Potential for more functionally relevant results which have biological plausibility	Application to gene-environment interactions unclear and needs further research

2.4 Discussion

It is well understood that interactions (whether gene-environment or gene-gene) play a role in disease susceptibility, allowing genetic effects to be dependent on the presence of another gene, or exposure to an environmental variable. What is not known, is the best statistical approach to use to identify such effects, particularly for genome-wide interaction analysis. This has led to an extensive number of novel methods being proposed, with the aim of producing the optimum approach to efficiently capture their effects. Although every available method is not documented as an exhaustive list here, the aim of the chapter was to present methods from each available methodological category, in order to represent the scope of the toolset available to undertake interaction analysis.

The method chosen for any interaction analyses will largely depend on the aims and research questions/hypotheses for any study undertaken. For example, one should first determine whether research is to focus on the involvement of individual SNPs and their interaction effects (whether between the SNPs themselves or with environmental exposures), or alternatively regions or genes. Additionally, one would need to consider whether analysis will be applied across all SNP/genes genome-wide or just a subset, using screening to minimise the search space, either through functional relevance or trait association for example. Further considerations are how the trait is characterised (whether binary or quantitative), the number of individuals to be studied, and whether to implicitly test for an interaction effect or simply allow for it, to provide a more powerful genetic association test to detect main genetic effects. Upon defining the research question and using this discussed criteria, an informed decision can be made as to which method is most suitable.

For the purpose of this thesis, the aim is to extend the marginal effect analysis undertaken in GWAS to consider gene-environment interactions. This research will therefore focus on the interaction effect of individual SNPs rather than snp-sets grouped, for example, by genes. The eventual aim is to apply a genome-wide interaction analysis in UK Biobank, including the entire (imputed) genome, thus any method chosen will need to be computationally efficient. Additionally, from a statistical standpoint, the method must be able to incorporate quantitative traits such as lung

function (the primary aim of this thesis). Therefore, methods revolved around case-only approaches and logistic regression will not be applicable. In terms of undertaking an implicit test of interaction or simply allowing for interaction, both methodologies will be considered. Upon pruning the available methods using this aforementioned criteria, the standard interaction test and the joint test of Kraft et al. (83) will be considered in the proceeding chapters.

2.5 Conclusion

In conclusion, there are a number of methods available for the analysis of interaction effects. To determine the method suitable in application, one must consider the research question, the study design, and the computational complexity of the approach, as primary selection criteria. On this basis, the standard interaction test and the joint test of Kraft et al. (83) were chosen for the work in the proceeding chapters of this thesis.

The next chapter considers the use of these two methods, using simulation to determine the power available for interaction analysis in UK Biobank.

Chapter 3 Power available for gene-environment interaction analysis utilising UK biobank

3.1 Introduction

With large scale data resources now becoming more readily accessible, such as UK Biobank (see chapter 1 section 1.2.3 for information), previously utilised sample sizes for genetic association analysis are now being eclipsed. Boosting sample sizes and therefore power for association analysis in recent years has largely relied on the use of meta-analysis efforts, which consist of combining multiple genetic datasets. This approach has been essential for improving our understanding of the genetic architecture of common complex diseases, and is still a prominent method for ensuring powerful genetic association analyses of combined cohort data, where individually studies would struggle to attain the power needed to detect modest genetic effects. Such efforts are however limited by the circulation of summarised rather than individual study statistics, so although rewarding, it can often be a time-consuming process from analysis plan to final results stage. It also relies on regular and reliable communication with individual study analysts. As a result, larger data sources such as UK Biobank, which provide analysts with immediate access to individual-level data hold large appeal, which is further increased by the availability of sample sizes that exceed many meta-analysis efforts. These resources are therefore likely to be an instrumental tool in the future of genetic research to improve our understanding of the role of both marginal genetic effects and interaction effects in disease susceptibility, and thus aid in treatment development.

The sample size available in UK Biobank will have a direct effect on power for association analysis to detect associations between disease and genes. Not only will such resources open the door to the potential detection of novel modest marginal effect signals, but they provide an opportunity to explore the effect of interactions, where detection requires substantial sample sizes (1-4). More specifically as part of this thesis, this resource could help to understand the effect of gene-smoking interactions on lung function, for which our knowledge and the literature is sparse, as discussed in section 1.2.3.2. UK Biobank provides analysts with genetic information for variants across a wide MAF spectrum (see information regarding array content in chapter 1 section 1.2.3.1), including coverage of those variants with rarer minor allele frequency (MAF <

1%), for which our understanding of their effect on common complex diseases continues to be a focus of genetic research, as we strive to understand their contribution to disease susceptibility. Additionally, the array was designed to provide good coverage for imputation in the lower frequency MAF range of 1 – 5%.

It follows then that a pertinent area for exploration is the power that resources such as UK Biobank can provide for gene by environment interaction analysis, and a useful method to explore this is simulation. Furthermore, with a large emphasis on rare variant genotyping as part of the UK Biobank project, it is important to understand the detectable effects for such variants of lower minor allele frequency.

3.2 Recap of the single variant analysis approach

As discussed in section 1.1.3.2, the single variant analysis approach consists of testing each SNP individually. As in a GWAS, this often involves testing the marginal effect of each variant on the disease of interest. This approach can be extended, and in contrast to exploring marginal effects, interactions between each SNP and a variable of interest (i.e. exposure variable) can be explored. When association testing multiple variants individually, we must consider chance findings, and significance levels which determine statistical significance are controlled with the use of adjustments such as the Bonferroni correction, where the significance level is often split between the number of (independent) variants considered for the analysis.

Detection of rare variant effects requires larger sample sizes than needed for common variants or relies on the assumption that rare variants contribute larger effect sizes. Due to the novelty of the sample size contributed by UK Biobank, it would be of interest to determine the power available for interactions and furthermore interactions with rare variants when using a single variant analysis approach.

3.3 Previous gene-environment interaction simulations

The novelty of UK Biobank's substantially large sample size produces new questions in regard to the available power for genetic association analysis, which previous

simulations have not been designed to answer. With previously attainable sample sizes much smaller in comparison, this has been reflected in simulation studies considering interactions (**Table 3.1**). In addition, previous simulations focussed on testing each variant individually, have predominantly focussed on variants of common MAF (MAF > 0.05) due to a lack of power and confidence for rare variant effect detection in smaller sample sizes, often opting for gene-based approaches instead (for example the simulation undertaken by Chen et al. (132) for the extended SKAT approach – **Table 3.1** - discussed in chapter 2). However, given how large UK Biobank is, the available power to detect gene-environment interactions of low frequency MAF (0.01 < MAF < 0.05) and rare MAF (MAF < 0.01) needs further exploration, and perhaps one should not dismiss individually testing rare variants completely when using a sample size of this magnitude. In addition, many of the previous gene-environment interaction simulations undertaken have been published with the purpose to present a new method, thus in terms of method comparison, conclusions could contain an element of bias. Following these reasons, presented here is a simulation which aims to provide further clarity regarding the available power for gene-environment interaction analysis. The following section succinctly outlines the simulation's core objective and aims.

Table 3.1 – Genetic interaction simulations in the literature (not an exhaustive list)

Author	Maximum considered sample size	MAF range considered	Ref.
Kraft et al.	40,000	0.1 - 0.25	(83)
Pare et al.	15,000	0.1 - 0.4	(96)
Dai et al.	4,000	0.2 - 0.5	(84)
Mukherjee et al.	25,000	0.1 – 0.3	(145)
Hsu et al.	2,000	0.5	(94)
Wason and Dudbridge	5,000	0.05 - 0.5	(146)
Ege and Strachan	10,000	0.1 - 0.9	(147)
Aschard et al.	2,000	0.3	(148)
Gauderman et al.	7,000	0.134-0.4	(95)
Marigorta and Greg Gibson	40,000	0.05 - 0.95	(149)
Chen et al.	2,000	0.005 - 0.05	(132)
Boonstra et al.	40,000	0.1 - 0.3	(150)

3.4 Simulation objectives and aims

3.4.1 Primary aim

The primary aim for this simulation was to determine the power available for GxE single variant interaction analysis when considering the sample size of UK Biobank (post phenotype and genotype quality control exclusions).

3.4.2 Secondary aims

To address the simulation's primary aim, further aims were incorporated to determine the power dependence on a number of factors. The first aim was to determine the power dependency on which single variant interaction approach was used to test for genetic effects. With the majority of interaction specific individual variant testing methods addressed in the previous chapter only suitable for case-control data, the choice of test for quantitative data is limited. Considered here was the standard interaction term test, which consisted of testing the interaction effect implicitly in a regression framework, and the joint test of Kraft et al. (5), in which the main genetic effect and the interaction effect are tested simultaneously. As well as being able to incorporate a quantitative trait, these two methods are also computationally and methodologically simple, which are characteristics which adhere to future plans to undertake genome-wide interaction analysis. Further secondary aims were to determine power dependency on the sample size utilised for analysis, the minor allele frequency of the variant being analysed and the effect size of the variants being analysed.

3.5 Methods

Consider a linear regression model for testing the interaction effect of each individual SNP with an exposure variable (e.g. ever/never smoking) on a continuous trait (e.g. lung function) as follows:

$$P_i = \beta_0 + \beta_1 G_i + \beta_2 S_i + \beta_3 G_i S_i + \varepsilon_i \qquad (3.1)$$

where P_i, G_i and S_i are the respective measurements of phenotype, genotype (using an additive model i.e. G = 0,1,2) and smoking exposure, i.e. ever smoker (1) and never

smoker (0), for individual i. The term G_iS_i denotes the interaction between genotype and smoking status, defined as the difference in effect of genotype on phenotype for ever smokers compared with never smokers. ε_i is a normally distributed error term with mean 0 and variance σ^2. This is methodologically the same as the mathematical theory introduced for the analysis of interactions in chapter 2 section 2.2.1 however assigning ever-smoking as the binary environmental exposure variable.

A brief recap of the methodology behind each of the two methods is as follows:

3.5.1 Standard interaction term test

The interaction term test corresponds to implicitly testing for an interaction effect i.e. testing the null hypothesis $\beta_3 = 0$ in equation (3.1) for statistical significance. If the null hypothesis is rejected, the alternative hypothesis of $\beta_3 \neq 0$ is accepted, and we could conclude that there is statistical evidence of an interaction effect. This is undertaken with the use of the student t-test to determine statistical significance for individual linear regression effect sizes in R.

3.5.2 Joint test (2 degrees of freedom)

In contrast to implicitly testing the interaction effect for statistical significance, one can instead explore the joint effect contributed by both the genetic main effect, β_1, and the interaction effect, β_3, using the null hypothesis $\beta_1 = \beta_3 = 0$. To determine whether to reject or accept the null hypothesis, one could fit two models to the data as follows:

Model 1: $P = \beta_0 + \beta_1 S + \varepsilon$
Model 2: $P = \gamma_0 + \gamma_1 G + \gamma_2 S + \gamma_3 GS + \varepsilon$

The joint effect could then be determined with the use of analysis of variance (ANOVA), by comparing the two models for goodness of fit to the data, using a likelihood ratio test, F test or chi-squared test. Therefore, testing the variance explained between model 2 and the nested model, model 1, the effect of the addition of both γ_1 and γ_3, which are the genetic (G) main effect and the interaction (GS) effect respectively, can be observed. Here, the default testing method in R is used, which is the F test.

3.5.3 Simulation Criteria

Phenotype data was be created for N individuals. A single SNP was generated from a binomial distribution using N observations taking the value 0,1 or 2 (to mimic the number of effect alleles for each genotype for an additive model approach) and probability p, with the value of p used to determine the MAF for the simulated SNP. It therefore follows from this that the variant is in Hardy-Weinberg equilibrium. The ever/never smoking variable was generated from a binomial distribution as a 1 (ever) / 0 (never) indicator variable. This value was kept constant and was determined via exploration of the prevalence of ever/never smokers from the provided UK Biobank phenotype data. Approximately, 44% of individuals in UK Biobank were defined as ever smokers, thus $p = 0.44$ was chosen to simulate the smoking exposure variable. An interaction variable was created using the product of allele count and the ever/never smoking variable. Linear regression was used to simulate phenotype data using the model in equation (3.1).

3.5.3.1 Primary scenarios

Two primary scenarios were considered:

Scenario 1 - Main genetic effect and interaction effect present
This means that the SNP considered produced both a main effect on the outcome, as well as producing an interaction effect with the smoking exposure variable. Thus, from equation (3.1), both β_1 and β_3 are non-zero.

Scenario 2 - Interaction effect present only i.e. only exposed individuals at genetic risk
This means that the SNP considered produced no main effect on the outcome but produced an interaction effect with the smoking exposure variable. Thus, from equation (3.1), $\beta_1 = 0$ and $\beta_3 \neq 0$.

3.5.3.2 Secondary scenarios

Within each of the above scenarios, the following variables were altered to address the secondary aims in section 3.4.2:

- **Sample size**
    - N = 300,000 was used to represent a conservative estimate for the UK Biobank sample size post quality control and N = 100,000 was used to represent an approximate previously attainable sample size for lung function association analysis (or follow-up) prior to UK Biobank release (50,51).
- **Minor allele frequency**
    - A SNP from each of the three MAF categorisations was simulated. A common SNP was simulated with MAF = 0.4, a low frequency SNP with MAF = 0.02, and a rare SNP with MAF = 0.005.
- **SNP effect size**
    - 15 different interaction effects were simulated to mimic effects through the full spectrum, from modest to large. To do this, interaction effects were calculated as multiples of the main genetic effect using 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 1.5, 2, 3, 4, 5, 6, 7, 8 and are presented in the results alongside the inferred ratio of genetic effect for exposed individuals compared with unexposed individuals. For example, given a multiple of 1 and using equation (3.1), β_3 would be equivalent to $(1 \times \beta_1)$ and thus an exposed and an unexposed individual would be subject to the following genetic effects (using the genetic main and interaction effect terms):

      Exposed: $(S = 1) \rightarrow \beta_1 G_i + (1 * \beta_1) G_i * 1 \rightarrow \beta_1 + \beta_1 = 2\beta_1$

      Unexposed: $(S = 0) \rightarrow \beta_1 G_i + (1 * \beta_1) G_i * 0 = \beta_1$

Thus, an interaction effect calculated as 1 times the main effect would mean a 2 times larger effect in exposed individuals compared with unexposed individuals ($\frac{2\beta_1}{\beta_1} = 2$).

Using the same theory, a multiple of 0.05 implies an interaction effect with a magnitude which is 5% of the main effect, leading to a 1.05 times larger effect in exposed individuals compared with unexposed individuals. A multiple of 0.1 infers 1.1 times larger and so on until the maximum multiple of 8, which infers an effect in exposed individuals 9 times larger than for unexposed individuals.

When interaction effect was present only, interaction effect sizes were still calculated as percentage increases from the predicted genetic main effect based on the MAF using equation (1.2) for continuity, regardless of the main effect being absent.

3.5.4 Phenotype distribution and effect sizes for the linear regression model

The phenotype was created with mean 0 and variance 1, with variance for the error term (ε) chosen to explain the remaining variance unexplained by covariate effects within the model. The remainder of this section explains how effect sizes within the regression model were simulated.

3.5.4.1 Genetic effect (main effect)

Genetic effect sizes were simulated using an additive model (such that G from equation (3.1) takes the values 0, 1 and 2 to represent the number of effect alleles present) in two ways to represent the relationship between MAF and effect size:

Approach 1 - Larger effect size for low MAF variants

Effect sizes contributed by variants at the lower extremity of the MAF spectrum is still a debated topic (151-153), although it is often suggested that effect size and minor allele frequency are inversely proportional, such that the rarer a variant is, the larger its effect size. This was explored and suggested as part of a study by Park et al. (154) using population genetic models for a number of complex traits.

In the first approach then, the documented relationship between the genetic variance contributed by a variant, the MAF, and the effect size (154) was used, and is shown in equation (3.2).

$$gv = 2\beta^2 MAF * (1 - MAF) \quad (3.2)$$

Here, gv is the fraction of the genetic variance of a simulated trait (where genetic variance refers to a trait's estimated heritability) contributed by the individual SNP. β is the corresponding effect size for the given SNP and *MAF* is its minor allele frequency. From the literature at the time of undertaking the simulations (50), with approximately 9.6% of lung function phenotypic variance (FEV_1) accounted for by 97 common signals (and under the assumption that each SNP accounted for an equal share of this variance), a single SNP was given a phenotypic variance contribution of 0.099%. With upper limits of the observed range of heritability estimates for lung function deemed to be 54% (155,156), the fraction of the total genetic variance accounted for by a single SNP (*gv*) was calculated as 0.099/54 = 0.002. Using this figure for *gv*, beta values were altered dependent on *MAF* (with MAF values presented in section 3.5.3.2).

Approach 2 - Constant effect size across MAFs

In contrast to the first approach, it may be the case that rare variants do in fact contribute modest rather than large effect sizes, however limitations in power due to using smaller sample sizes (than offered by UK Biobank) have meant that detection has been difficult. Therefore, in a second approach the relationship in equation (3.2) was used to estimate the common variant effect size, with the calculated effect size maintained as the MAF decreased.

3.5.4.2 Smoking effect

Ever/never smoking effect size (β_2) was determined through exploration of ever/never smoking effect on FEV_1 in UK Biobank. With the discovery of an ever-smoking effect of 0.077 litres and an observed FEV_1 standard deviation of 0.759, the figure was chosen to represent an effect size approximate to 0.077/0.759 = 0.101 standard deviations, rounded and chosen to be 0.1.

Methods were assessed for controlled type I error at the 5% significance level, and power was recorded for each method in each of the scenarios and for each interaction effect size. Upon determining whether tests were controlled at the 5% significance level, power was calculated using a threshold resembling a genome-wide analysis (5×10^{-8}).

3.5.5 Number of simulations

Power for each scenario was calculated using 1000 simulations, providing a compromise between error in power estimation and run time for the simulation (**Table 3.2**). For example, with 1000 simulations, the error for a power estimate of 80% was 1.3%, thus, providing a 95% confidence interval of (77.5%, 82.5%).

Table 3.2 - Simulation run time

Number of individuals	Approximate run time
300,000	11 - 15 hours
100,000	2 - 5 hours

The R code for this simulation is presented in **Appendix B**.

3.6 Results

3.6.1 Type I Error

Both the joint test and the interaction test had controlled type I errors at the 5% level based on 1000 simulations for all MAFs (**Table 3.3**), with no statistically significant deviation from 0.05 using a test of proportions in R.

Table 3.3 - Type I errors for the interaction test and joint test at the 5% significance level (*proportion test H_0:type I error=0.05, n=1000)

Method	MAF	Type I Error	P value*
Interaction test	0.4	0.046	0.7537
	0.02	0.045	0.6741
	0.005	0.053	0.8396
Joint test	0.4	0.047	0.8351
	0.02	0.046	0.7537
	0.005	0.053	0.8396

3.6.2 Scenario 1 – Main genetic effect and interaction effect present

3.6.2.1 Larger effect sizes for low MAF variants in 300,000 individuals

For the scenario where a SNP had both a main and interaction effect for a sample size of 300,000 individuals (**Table 3.4 and Figure 3.1A**) and a common variant of MAF 0.4, power for the interaction test was low for interaction effects at the lower extremity of the considered range. For example, there was effectively zero power to detect interaction effects corresponding to a 1.05, 1.1 and 1.2 times larger genetic effect in exposed individuals compared to unexposed individuals. Power increased as the interaction effect increased, and to achieve a power of approximately 80% in the same scenario, suggested here was a required genetic effect in exposed individuals of 1.5 times larger than in unexposed individuals (a standard deviation effect size of 0.032). In contrast, the joint test produced significantly higher powers than the interaction test for most interaction effect scenarios with power consistently 1, even for the smallest interaction effect (1.05 times genetic effect in exposed individuals compared with unexposed – an effect of 0.003 standard deviations). For the interaction test to have achieved the same power in this scenario, there would have needed to be a 2 times larger genetic effect for exposed individuals compared to unexposed individuals (equivalent to an effect of 0.065 standard deviations).

Under the assumption that rare variants contributed larger effect sizes (**Table 3.4 and Figure 3.1A**), the performance for both methods, as well as the comparison between methods, had the same conclusions as the variant's MAF decreased. The joint test continued to be consistently high powered across most interaction effect scenarios as before, and the interaction test suffered large losses in power as the interaction effect magnitude reduced. Individual method powers for each scenario of interaction effect magnitude barely differed between consideration of a common variant and a rare variant, as effect sizes were adjusted (increased) accordingly as suggested by the relationship between MAF and effect size, due to genetic variance contribution. However, this conclusion is reliant on much larger effect sizes for rare variants. For example, for the interaction test to maintain 80% power to detect a 1.5 times larger effect in exposed individuals compared with unexposed individuals, the interaction

effect required was 0.224 for a rare variant (MAF = 0.005), 7 times larger than the 0.032 required by a common variant (MAF = 0.4).

3.6.2.2 Constant effect size across MAFs in 300,000 individuals

In the scenario where rare and low frequency variants contributed effect sizes consistent with common variants (**Table 3.5 and Figure 3.1B**) and thus produced a main effect size accredited to a common variant of 0.065 regardless of MAF, there was a loss of power for both methods across all interaction effect magnitudes. For the interaction test, power was negligible when the genetic effect was 1.05 to 2 times higher for exposed individuals (effects between 0.003 and 0.065 standard deviations) and 1.05 – 3 times higher for exposed individuals (effects between 0.003 and 0.129 standard deviations) for a low frequency (MAF = 0.02) and rare (MAF = 0.005) variant respectively. This was in comparison to achieving 80% power for a 1.5 times larger effect in exposed individuals for all MAFs previously, when rarer variants produced larger effects. To summarise, the interaction test needed the same standard deviation effects to produce the same power as the previous scenario (varying effect size for varying MAF), however when considering a constant effect size this corresponded to an increased ratio of genetic effects between exposed and unexposed individuals. For the joint test, the reduction in power was small for the low frequency variant (MAF = 0.02), still producing 92.6% power for a 1.05 times larger genetic effect for exposed individuals, even though the standard deviation effect reduced from 0.011 to 0.003. However, power loss was noticeable when considering a rare variant, and the same scenario only achieved a power of 0.022. To achieve approximately 80% power for a rare variant, a genetic effect 2.5 times larger in exposed individuals would be required (an effect of 0.097 standard deviations). As for the interaction test, where comparable standard deviation effects (even though this resulted in larger genetic effect ratios between exposed/unexposed individuals) produced comparable power, this was not the case for the joint test, largely due to its dependence on the main effect as well as the interaction effect.

3.6.2.3 Varying and constant effect size across MAFs in 100,000 individuals

When considering a smaller sample size of 100,000 individuals (**Tables 3.6, 3.7 and Figure 3.2)**, the conclusions for method comparison were consistent with the scenario of 300,000 individuals. The joint test was more powerful across both scenarios of increasing and consistent effect sizes for lower MAFs, and over the majority of interaction effect magnitudes. Individual method performance was however affected in some cases as expected, such that larger interaction effect sizes were needed to achieve the same power, when considering 300,000 individuals. For example, for the interaction test and a genetic effect 1.5 times higher in exposed individuals compared with unexposed individuals, when the effect size increased as MAF decreased, the power achieved was approximately 2-3%, compared with approximately 75% when analysing 300,000 individuals (interaction effect of 0.032, 0.113 and 0.224 standard deviations for a common, rare and low frequency variant respectively). The joint test behaved similarly for the increasing effect size for decreasing MAF scenario over the two sample size scenarios. When considering a constant effect size for reducing MAF however, power for both methods appeared much more sensitive to sample size. For example, when considering a large ratio of genetic effect between exposed and unexposed individuals of 3 for a low frequency variant (MAF = 0.02), which for 300,000 individuals the interaction test produced power of 0.943, for 100,000 individuals the corresponding power was 0.092 (standard deviation effect of 0.129). Furthermore, for a rare variant (MAF=0.005), the interaction test required an interaction effect of 0.452 for 100,000 individuals to achieve the same power as an effect of 0.258 for 300,000 individuals (ratio of 8 and 5 respectively to achieve power of ~0.94). Similarly, for the joint test, equivalent interaction effect produced lower powers between the two sample size scenarios. For example, for a low frequency and rare variant respectively and a genetic effect 2 times larger in exposed individuals compared to unexposed individuals (interaction effect of 0.065 standard deviations), powers for 300,000 individuals were 1 and 0.388 compared with 0.666 and 0.004 for 100,000 individuals.

Table 3.4 - Power for each method when considering both main and interaction effect when rare variants produce larger effect sizes for 300,000 individuals.
Ever-smoking prevalence of 0.44. IT - Interaction test and JT- Joint test. Repeated rows where power is 1 for both tests are not presented.

Interaction Effect (Multiple of main effect, interaction effect ratio between exposure groups and corresponding interaction effect sizes for each MAF in standard deviations)					**MAF = 0.4** Main effect = 0.065		**MAF = 0.02** Main effect = 0.226		**MAF = 0.005** Main effect = 0.448	
Multiple	**Ratio**	**Common**	**Low**	**Rare**	**IT**	**JT**	**IT**	**JT**	**IT**	**JT**
0.05	1.05	0.003	0.011	0.022	0	1	0	1	0	1
0.1	1.1	0.006	0.023	0.045	0	1	0	1	0	1
0.2	1.2	0.013	0.045	0.090	0.003	1	0.001	1	0.003	1
0.3	1.3	0.019	0.068	0.135	0.037	1	0.039	1	0.030	1
0.4	1.4	0.026	0.090	0.179	0.304	1	0.271	1	0.284	1
0.5	1.5	0.032	0.113	0.224	0.750	1	0.747	1	0.764	1
1	2	0.065	0.226	0.448	1	1	1	1	1	1

Table 3.5 - Power for each method when considering both a main and interaction effect with effect size constant as MAF decreases for 300,000 individuals.
Ever-smoking prevalence of 0.44. IT - Interaction test and JT - Joint test. Repeated rows where power is 1 for both tests are not presented. Note: common variant with MAF = 0.4 is not presented as this is equivalent to the common variant results in Table 3.4.

Interaction Effect (Multiple of main effect, interaction effect ratio between exposure groups and corresponding interaction effect sizes for each MAF in standard deviations)			**MAF = 0.02** Main effect = 0.065		**MAF = 0.005** Main effect = 0.065	
Multiple	**Ratio**	**Effect**	**IT**	**JT**	**IT**	**JT**
0.05	1.05	0.003	0	0.926	0	0.022
0.1	1.1	0.006	0	0.953	0	0.021
0.2	1.2	0.013	0	0.981	0	0.041
0.3	1.3	0.019	0	0.979	0	0.054
0.4	1.4	0.026	0	0.997	0	0.066
0.5	1.5	0.032	0.001	0.999	0	0.107
1	2	0.065	0.035	1	0	0.388
1.5	2.5	0.097	0.410	1	0.001	0.767
2	3	0.129	0.943	1	0.023	0.966
3	4	0.194	1	1	0.419	1
4	5	0.258	1	1	0.937	1
5	6	0.323	1	1	1	1

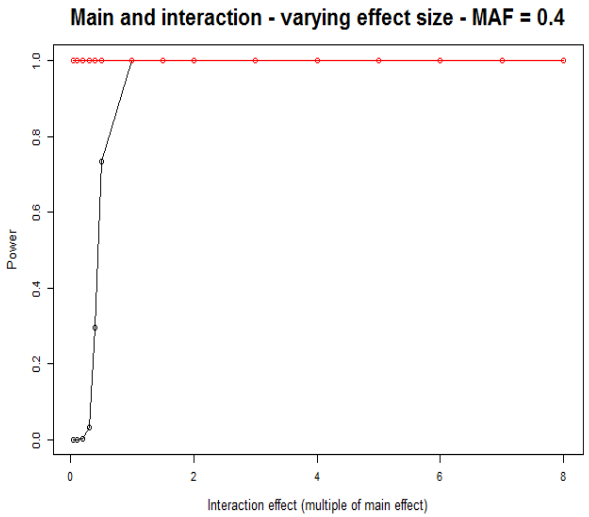
Table 3.6 - Power for each method when considering both main and interaction effect and larger effect sizes for rare variants for 100,000 individuals.
Ever-smoking prevalence of 0.44. IT - Interaction test and JT- Joint test. Repeated rows where power is 1 for both tests are not presented.

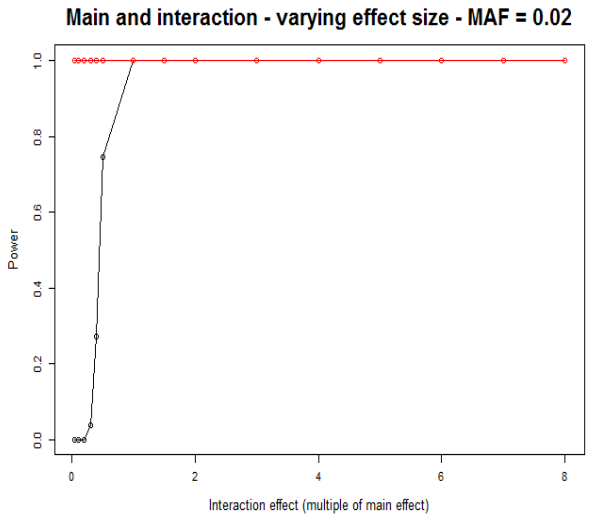
Interaction Effect (Multiple of main effect, interaction effect ratio between exposure groups and corresponding interaction effect sizes for each MAF in standard deviations)					**MAF = 0.4** Main effect = 0.065		**MAF = 0.02** Main effect = 0.226		**MAF = 0.005** Main effect = 0.448	
Multiple	**Ratio**	**Common**	**Low**	**Rare**	**IT**	**JT**	**IT**	**JT**	**IT**	**JT**
0.05	1.05	0.003	0.011	0.022	0	1	0	1	0	1
0.1	1.1	0.006	0.023	0.045	0	1	0	1	0	1
0.2	1.2	0.013	0.045	0.090	0	1	0	1	0	1
0.3	1.3	0.019	0.068	0.135	0.001	1	0.001	1	0	1
0.4	1.4	0.026	0.090	0.179	0.003	1	0.004	1	0.004	1
0.5	1.5	0.032	0.113	0.224	0.029	1	0.034	1	0.021	1
1	2	0.065	0.226	0.448	0.940	1	0.949	1	0.953	1
1.5	2.5	0.097	0.339	0.673	1	1	1	1	1	1

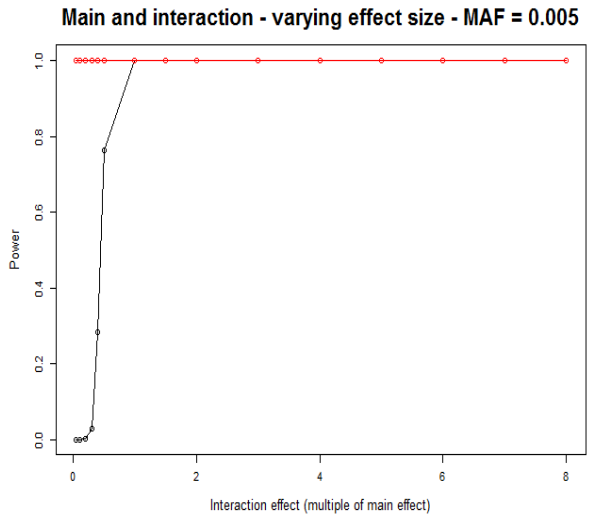
Table 3.7 - Power for each method when considering both a main and interaction effect with effect size constant as MAF decreases for 100,000 individuals
Ever-smoking prevalence of 0.44. IT –Interaction test and JT- Joint test). Note: common variant with MAF = 0.4 is not presented as this is equivalent to the common variant results in Table 3.6.

Interaction Effect (Multiple of main effect, interaction effect ratio between exposure groups and corresponding interaction effect sizes for each MAF in standard deviations)			**MAF = 0.02** Main effect = 0.065		**MAF = 0.005** Main effect = 0.065	
Multiple	**Ratio**	**Effect**	**IT**	**JT**	**IT**	**JT**
0.05	1.05	0.003	0	0.059	0	0
0.1	1.1	0.006	0	0.067	0	0
0.2	1.2	0.013	0	0.092	0	0
0.3	1.3	0.019	0	0.143	0	0
0.4	1.4	0.026	0	0.176	0	0.001
0.5	1.5	0.032	0	0.273	0	0
1	2	0.065	0	0.666	0	0.004
1.5	2.5	0.097	0.013	0.955	0	0.019
2	3	0.129	0.092	0.995	0.001	0.087
3	4	0.194	0.727	1	0.008	0.470
4	5	0.258	0.995	1	0.078	0.881
5	6	0.323	1	1	0.311	0.996
6	7	0.387	1	1	0.734	0.999
7	8	0.452	1	1	0.938	1
8	9	0.516	1	1	0.996	1

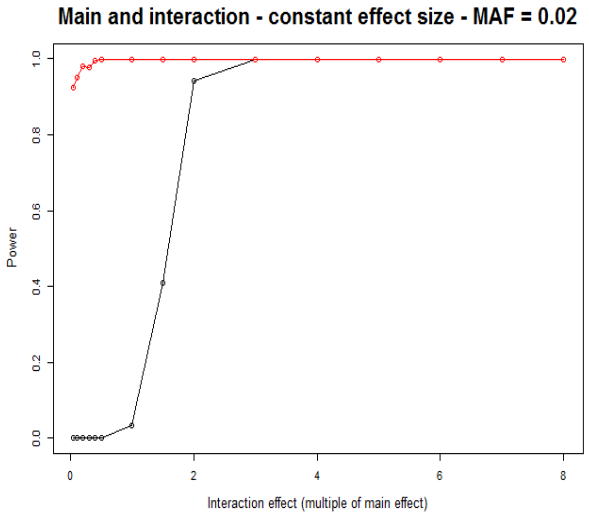
A

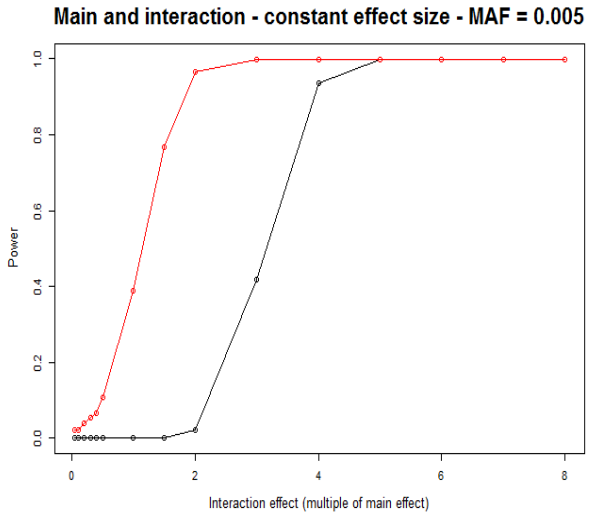






B





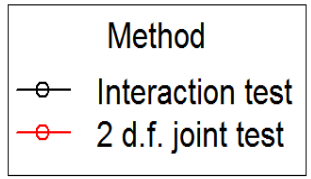


Figure 3.1- Power curves for each method when considering both main and interaction effect for 300,000 individuals
For (A) Larger effect size for rare variants and (B) constant effect size. Ever-smoking prevalence of 0.44

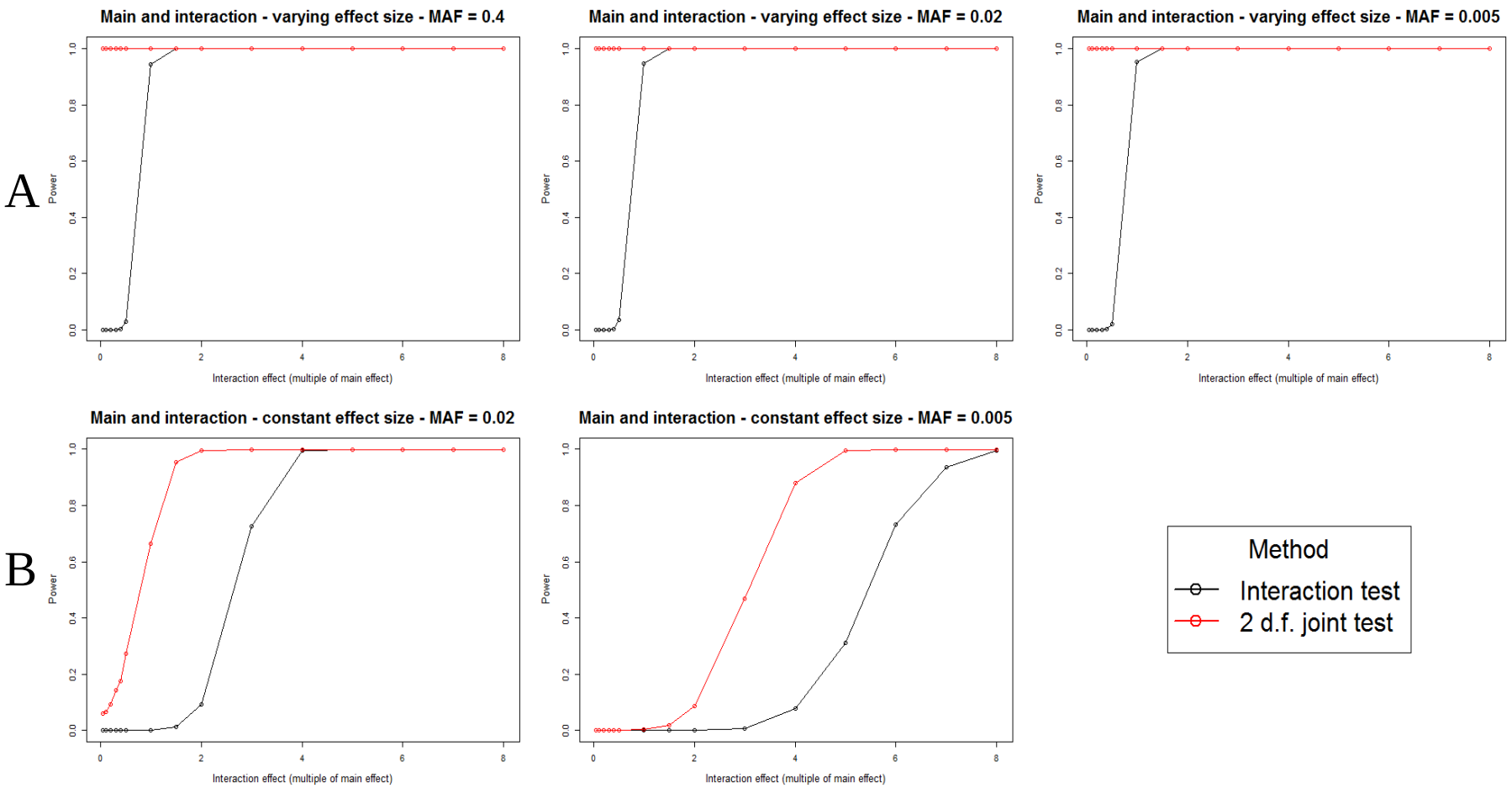


Figure 3.2 - Power curves for each method when considering both main and interaction effect for 100,000 individuals

For (A) Larger effect size for rare variants and (B) constant effect size. Ever-smoking prevalence of 0.44

3.6.3 Scenario 2 – Interaction effect present only (no genetic main effect)

3.6.3.1 Larger effect sizes for low MAF variants in 300,000 individuals

For the scenario where only the interaction effect was present and the main effect was absent (**Table 3.8, Table 3.9 and Figure 3.3**), as expected, the interaction test, which tests solely the interaction effect, performed similarly to the scenario where the main effect was present. This was the case for both varying and constant effect sizes for varying MAFs. A noticeable difference however was the reduction in performance for the joint test method. For increasing effects for decreasing MAFs **(Table 3.8 and Figure 3.3A)**, it no longer produced a consistently large power of 1. For example, to achieve the power consistently displayed across all interaction effect scenarios (including the smallest effect of 0.003 standard deviations – ratio of 1.05) when the main effect was present as before, here the effect required was 0.032 (or a ratio of genetic effect between exposed and unexposed individuals of 1.5). Furthermore, there was now zero power for the joint test to detect the aforementioned smallest interaction effect size considered. This comparison was the same across all MAFs when effect sizes increased as MAF decreased. However, regardless of the severe power loss for the joint test, this method still outperformed the interaction test in many interaction effect magnitude scenarios (particularly for small interaction effects with ratios between 0.2 – 0.5 and corresponding effects between 0.013 – 0.032).

3.6.3.2 Constant effect size across MAFs in 300,000 individuals

As before, when rare variants were assumed to have consistent effect sizes with common variants **(Table 3.9 and Figure 3.3B)**, the same effect size in standard deviations was needed for the interaction test to produce comparable power for lower MAF variants, but as a result the ratio for the genetic effect in exposed compared with unexposed individuals was larger. Powers for the interaction test are consistent with when the main effect is present, as expected. With no main effect present, the joint test's performance is solely dependent on the interaction effect, thus here, like for the interaction test, the joint test required the same standard deviation interaction effect, which increases the genetic effect ratio between exposed and unexposed individuals.

For example, for an interaction effect of ~0.2 for a rare (MAF = 0.005) variant, the joint tests produced ~90-99% power, however the ratio was ~1.5 compared with ~5 for varying and constant effects respectively. Furthermore, as expected, the joint test suffers losses in power for the constant effect size scenario across main effect and no main effect scenarios. As before, this will be due to its dependence on the main effect. Finally, the joint test remained the more powerful method, in particular for the smaller standard deviation effects. For example, when considering a standard deviation effect of 0.194 for a rare (MAF = 0.005) variant (ratio of 4) powers produced for the interaction and joint test were 0.443 and 0.919 respectively.

3.6.3.3 Varying and constant effect size across MAFs in 100,000 individuals

When considering a reduced sample size of 100,000 individuals **(Table 3.10, Table 3.11 and Figure 3.4),** the conclusions drawn for the comparison in performance for the interaction test between the sample size scenarios in section 3.6.2 were the same here. For the joint test, when the effect size varied with MAF **(Table 3.10 and Figure 3.4A)**, larger effects were needed for the 100,000-individual scenario for comparable power. For example, for effects of 0.026, 0.090 and 0.179 for a common, low frequency and rare variant respectively (genetic effect ratio of 1.4 between exposed and unexposed groups), the power produced was ~0.75 for 300,000 individuals compared with ~0.03 for 100,000 individuals. The same conclusion of requiring larger effects for the same MAF was drawn when considering constant effect for varying MAF, across the two sample size scenarios **(Table 3.11 and Figure 3.4B)**.

Table 3.8 - Power for each method when considering interaction effect only and larger effect sizes for rare variants for 300,000 individuals

Ever-smoking prevalence of 0.44. IT - Interaction test and JT- Joint test. Repeated rows where power is 1 for both tests are not presented.

Interaction Effect (Multiple of main effect, interaction effect ratio between exposure groups and corresponding interaction effect sizes for each MAF in standard deviations)					**MAF = 0.4** Main effect = 0.065		**MAF = 0.02** Main effect = 0.226		**MAF = 0.005** Main effect = 0.448	
Multiple	**Ratio**	**Common**	**Low**	**Rare**	**IT**	**JT**	**IT**	**JT**	**IT**	**JT**
0.05	1.05	0.003	0.011	0.022	0	0	0	0	0	0
0.1	1.1	0.006	0.023	0.045	0	0	0	0	0	0
0.2	1.2	0.013	0.045	0.090	0.001	0.013	0	0.008	0.002	0.011
0.3	1.3	0.019	0.068	0.135	0.030	0.196	0.036	0.218	0.032	0.209
0.4	1.4	0.026	0.090	0.179	0.274	0.764	0.252	0.770	0.276	0.768
0.5	1.5	0.032	0.113	0.224	0.723	0.993	0.755	0.996	0.720	0.990
1	2	0.065	0.226	0.448	1	1	1	1	1	1

Table 3.9 - Power for each method when considering interaction effect only with effect size constant as MAF decreases for 300,000 individuals and an

Ever-smoking prevalence of 0.44. IT - Interaction test and JT- Joint test. Repeated rows where power is 1 for both tests are not presented. Note: common variant with MAF = 0.4 is not presented as this is equivalent to the common variant results in Table 3.8.

Interaction Effect (Multiple of main effect, interaction effect ratio between exposure groups and corresponding interaction effect sizes for each MAF in standard deviations)			**MAF = 0.02** Main effect = 0.065		**MAF = 0.005** Main effect = 0.065	
Multiple	**Ratio**	**Effect**	**IT**	**JT**	**IT**	**JT**
0.05	1.05	0.003	0	0	0	0
0.1	1.1	0.006	0	0	0	0
0.2	1.2	0.013	0	0	0	0
0.3	1.3	0.019	0	0	0	0
0.4	1.4	0.026	0	0	0	0
0.5	1.5	0.032	0	0.001	0	0
1	2	0.065	0.022	0.143	0	0.001
1.5	2.5	0.097	0.413	0.909	0.002	0.016
2	3	0.129	0.914	1	0.029	0.152
3	4	0.194	1	1	0.443	0.919
4	5	0.258	1	1	0.926	0.999
5	6	0.323	1	1	1	1

Table 3.10 - Power for each method when considering interaction effect only and larger effect sizes for rare variants for 100,000 individuals

Ever-smoking prevalence of 0.44. IT - Interaction test and JT- Joint test. Repeated rows where power is 1 for both tests are not presented.

Interaction Effect (Multiple of main effect, interaction effect ratio between exposure groups and corresponding interaction effect sizes for each MAF in standard deviations)					**MAF = 0.4** Main effect = 0.065		**MAF = 0.02** Main effect = 0.226		**MAF = 0.005** Main effect = 0.448	
Multiple	**Ratio**	**Common**	**Low**	**Rare**	**IT**	**JT**	**IT**	**JT**	**IT**	**JT**
0.05	1.05	0.003	0.011	0.022	0	0	0	0	0	0
0.1	1.1	0.006	0.023	0.045	0	0	0	0	0	0
0.2	1.2	0.013	0.045	0.090	0	0	0	0	0	0
0.3	1.3	0.019	0.068	0.135	0	0.003	0	0.003	0.002	0.002
0.4	1.4	0.026	0.090	0.179	0.002	0.024	0.008	0.024	0.004	0.030
0.5	1.5	0.032	0.113	0.224	0.023	0.152	0.028	0.152	0.031	0.168
1	2	0.065	0.226	0.448	0.951	1	0.945	0.999	0.932	1
1.5	2.5	0.097	0.339	0.673	1	1	1	1	1	1

Table 3.11 - Power for each method when considering interaction effect only with effect size constant as MAF decreases for 100,000 individuals

Ever-smoking prevalence of 0.44. IT - Interaction test and JT- Joint test. Note: common variant with MAF = 0.4 is not presented as this is equivalent to the common variant results in Table 3.10.

Interaction Effect (Multiple of main effect, interaction effect ratio between exposure groups and corresponding interaction effect sizes for each MAF in standard deviations)			**MAF = 0.02** Main effect = 0.065		**MAF = 0.005** Main effect = 0.065	
Multiple	**Ratio**	**Effect**	**IT**	**JT**	**IT**	**JT**
0.05	1.05	0.003	0	0	0	0
0.1	1.1	0.006	0	0	0	0
0.2	1.2	0.013	0	0	0	0
0.3	1.3	0.019	0	0	0	0
0.4	1.4	0.026	0	0	0	0
0.5	1.5	0.032	0	0	0	0
1	2	0.065	0	0	0	0
1.5	2.5	0.097	0.008	0.047	0	0.001
2	3	0.129	0.068	0.341	0	0.002
3	4	0.194	0.717	0.995	0.010	0.053
4	5	0.258	0.994	1	0.066	0.385
5	6	0.323	1	1	0.332	0.853
6	7	0.387	1	1	0.706	0.989
7	8	0.452	1	1	0.947	1
8	9	0.516	1	1	0.998	1

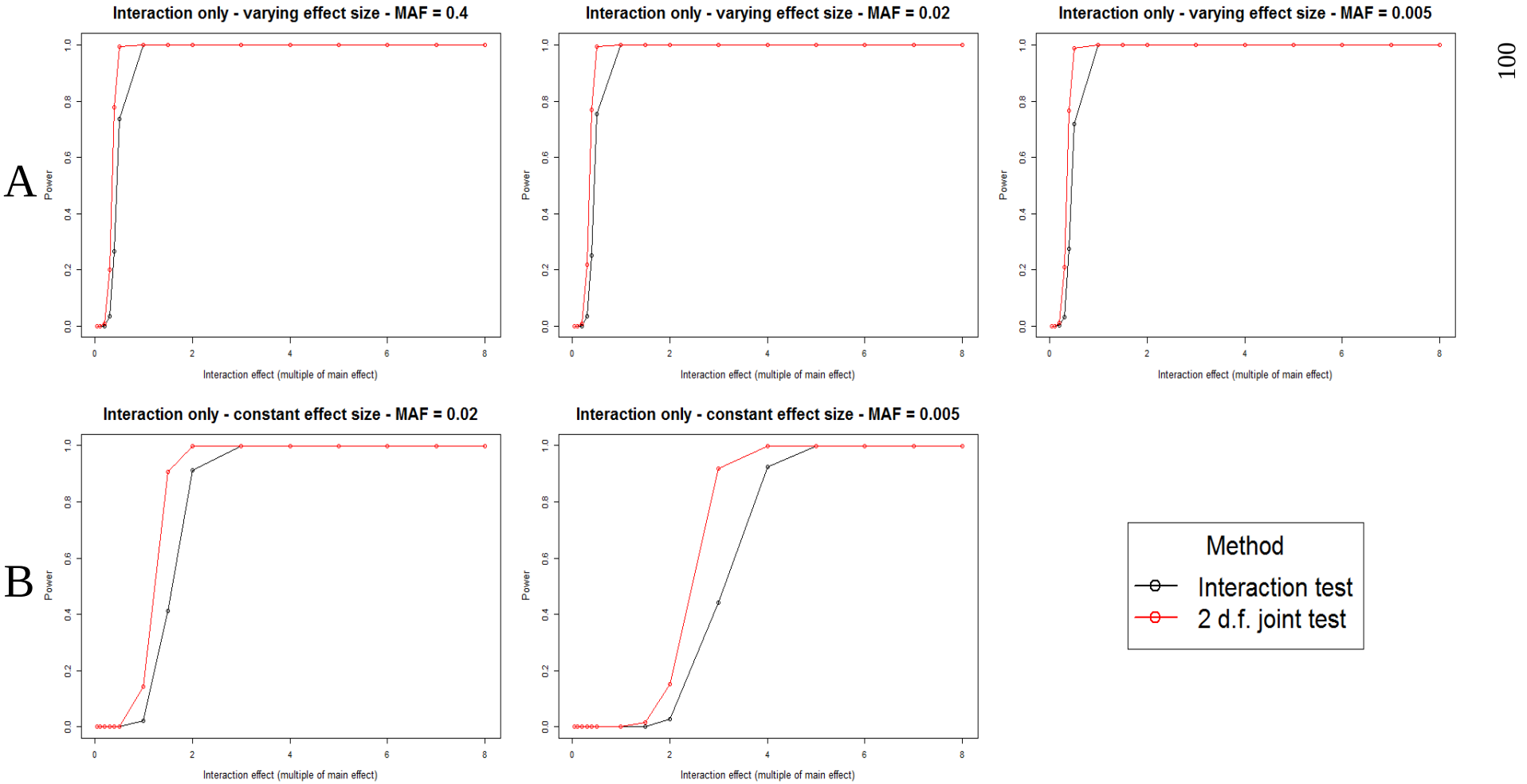
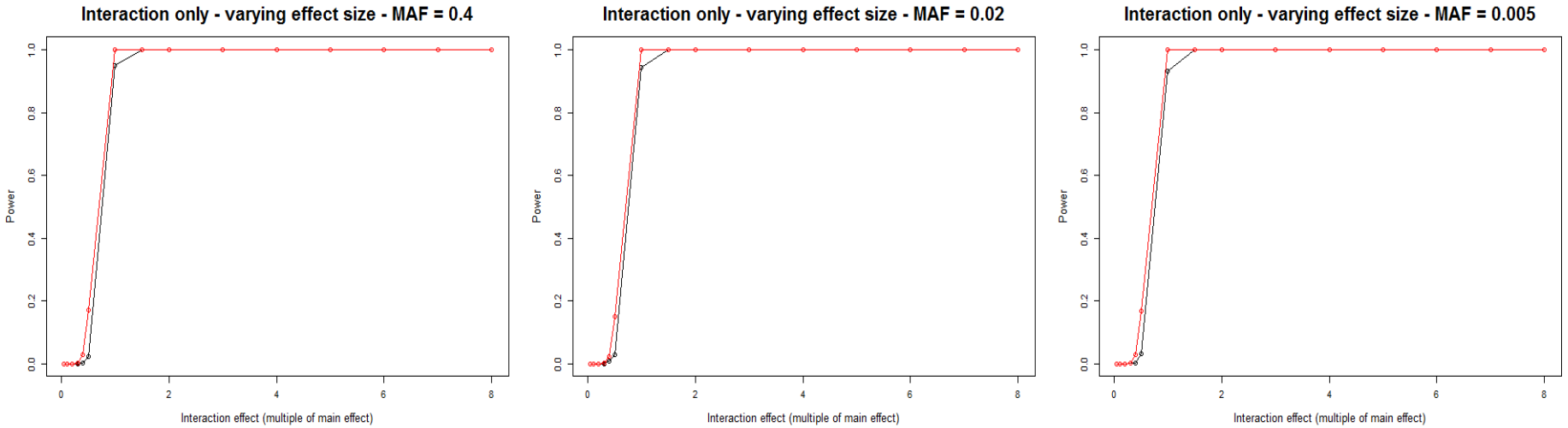


Figure 3.3 - Power curves for each method when considering interaction effect only for 300,000 individuals

For (A) Larger effect size for rare variants and (B) constant effect size. Ever-smoking prevalence of 0.44

A



B

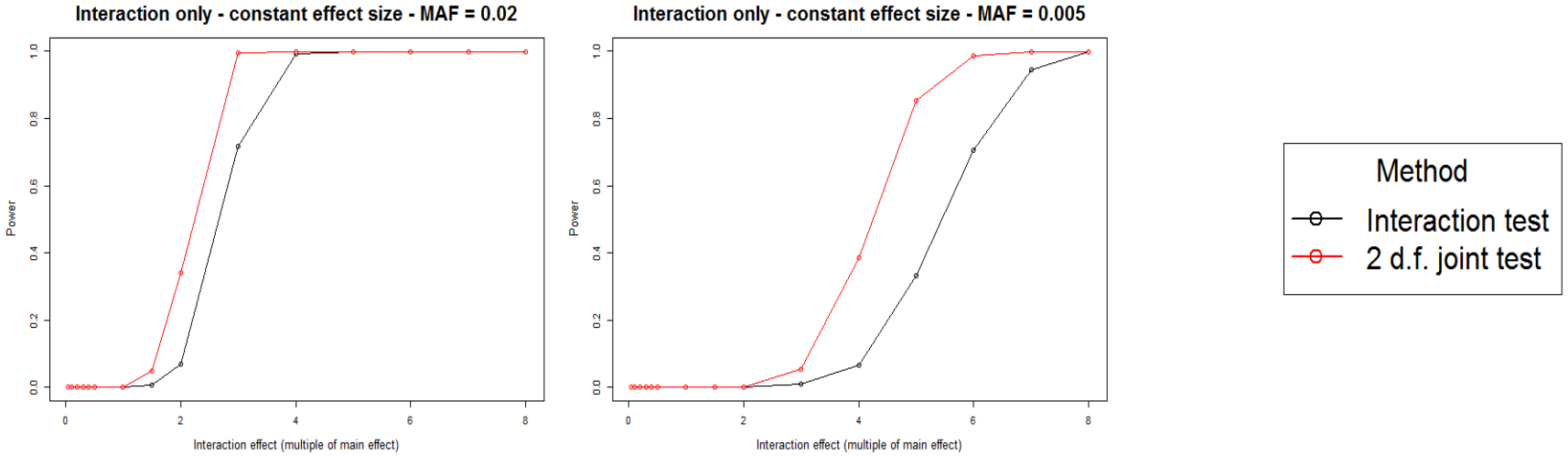


Figure 3.4 - Power curves for each method when considering interaction effect only for 100,000 individuals
For (A) Larger effect size for rare variants and (B) constant effect size. Ever-smoking prevalence of 0.44

3.7 Discussion

The aim of this simulation was to explore the power available for gene-smoking interaction analysis for lung function phenotypes. This is in anticipation of UK Biobank releasing phenotypic and genotypic data for over 500,000 individuals, a sample size which eclipses previously attainable sample sizes for lung function genetic research. Additionally, the genetic array provides coverage of both low frequency and rare variation, as well as being designed for effective imputation for variants at the lower extremity of the MAF spectrum. Thus, in order to determine the power available to detect gene-smoking interaction effects, presented here is a simulation which contributes two novel aspects. Firstly, it considers a substantially large sample size (300,000 individuals) which previous literature has not (due to previously smaller attainable sample sizes likely to be used for analysis). Secondly, with previous gene-environment interaction simulations dominated by the exploration of common variation (the primary focus of genome-wide association analysis in previous years), explored here is the effect of low frequency and rare variation. This was to allow for the exploration of interaction analysis method dependency on effect sizes contributed by variants at the lower end of the MAF spectrum.

For method performance when a main and interaction effect acted together on the phenotype of interest, an effect 1.5 times larger in exposed individuals (compared with unexposed individuals) was required for ~80% power for the interaction test (when rare variants produce larger effect sizes). For constant effect sizes (applying common variant effect size for all MAFs), genetic effects would need be in the range of 2.5 – 5 times larger for exposed individuals, depending on whether a low frequency or rare variant is the target. In the same scenario, the joint test is highly powered (power of 1) for even the smallest interaction effects (1.05 times the effect in exposed individuals), and this is consistent even when considering a low frequency variant with common variant effect. For a rare variant, suggested here is a ratio of approximately 2.5 for 80% power. For the scenario where the interaction effect is present but the main effect is absent, the interaction test as expected performs similarly, with the joint test losing power due to its dependency on the main effect. For 80% power the joint test requires ratios between 1.4 – 1.5, when considering larger effects for rare variants and between 2 – 4 for low frequency or rare variants when assuming constant effect size for lower MAF variants.

With regards to method comparison, the joint test is consistently the higher-powered method (as concluded by Kraft et al. (83)), over both scenarios of main effect/no main effect and also varying/constant effect sizes. Although expected in presence of a main effect, the joint test was also the most powerful method when only the interaction effect was present, suggesting the joint test to be a robust method when lacking knowledge of the underlying effect of a variant. Furthermore, when considering comparison to a smaller sample size (100,000 individuals) to mimic previously attainable sample sizes, the effects detectable were reduced. This suggests that the large sample sizes provided by resources such as UK Biobank could produce the opportunity to detect previously undetectable interaction effects.

It is important to note here however that although the two methods' ability to detect an interaction effect in the presence or absence of a main effect has been explored, these two methods do answer different statistical and biological questions. As discussed before, the joint test will determine whether there is a genetic effect present, allowing for the contribution of an interaction during analysis, but not revealing which of the two is the driving mechanism behind the association. In contrast, the interaction test is specifically testing for an interaction effect and thus should it return a statistically significant result, it suggests that the association is driven by an interaction. To ensure a direct comparison, a possible extension would be to compare the power of the interaction test to the power of a 2 stage process of the joint test and then the interaction test. Determining the power of this 2 stage approach would however bring added complexities. Firstly it must be noted that the joint test and interaction test are dependent on one another. Secondly, power will be dependent on the thresholds applied for the joint test and interaction test in each stage to determine statistical significance, as in application one may apply a more lenient threshold for the stage 1 test (the joint test) and a more conservative threshold for the stage 2 test (interaction test). Thirdly, calculating power requires a clear definition of the criteria required for concluding that the variant is statistically significant for both joint test and interaction test. Methods for combining p-values from two or more dependent tests have been discussed in the literature (157-159), however is still a topic for debate. Here the primary aim was to determine the power available for interaction analysis in UK Biobank, with method

comparison a secondary aim. Therefore, incorporating a method which utilises a combination of p-values warrants further research for application in this setting.

For the simulation undertaken in this chapter, the interaction effect sizes considered were calculated using multiples of the main genetic effect, with a range chosen to determine power estimates. Changing the magnitude of the interaction effect will in turn determine the proportion of variation explained in the regression model by the interaction. This association can therefore be used in reverse order as an alternative method for calculating interaction effect magnitudes, by determining its contribution to the model in terms of variance, and then calculating a corresponding effect size. This concept has been discussed in the literature, with methodology such as the Pratt index or Pratt's index (160,161) suggested to determine effect sizes of correlated predictors in a model, based on variance contribution.

For this simulation, a strict significance level was used to calculate power, with $\propto = 5 \times 10^{-8}$ chosen to represent p-value adjustment to determine statistical significance when undertaking a genome-wide association analysis. The adjustment is made based on a Bonferroni correction of 0.05 shared between the number of independent tests (or independent SNPs being analysed). Thus, if considering a reduced number of SNPs such as screened SNPs prior to analysis or SNPs within previously known regions of association with the phenotype, this restriction could be relaxed. Therefore, power here could be underestimated, but provides a good basis for knowledge of power to detect interaction effects genome-wide in a substantial sample size of 300,000 individuals.

A further point to consider here is the observed dependency by Kraft et al. (83) of the method comparison on exposure prevalence, with the difference in power between these two methods minimised for exposures with smaller prevalence. With this simulation utilising an estimate of ever-smoker prevalence in a UK population, it is therefore important to take into account the exposure prevalence for the analysed sample when generalising results across other ancestries.

3.8 Conclusion

In summary, the sample size available does provide the opportunity to detect modest interaction effects with good power and with a sample size 3 times larger than that available previously, smaller effects can be detected. Although power available is dependent on the assumed effect size for rare variation, it is clear that the joint test appears the more robust option regardless, providing significantly higher power than the interaction test when there is a main effect. The joint test also provides slightly greater power without a main effect present. The main drawback for this method is the lack of a quantifiable interaction effect size; however, this could be explored post detection of a variant jointly tested for both main and interaction effect. Furthermore, conservative multiple testing corrections will influence power, however their effect can be reduced with the use of screening methods, or by analysing associated regions previously detected from independent studies.

The next chapter will use both the joint test and the interaction test to search for novel variants which interact with ever smoking to affect lung function in UK Biobank. The analysis will focus on regions which we already know to be associated with lung function traits.

Chapter 4 Prioritising regions previously associated with lung function for gene-smoking interaction analysis

4.1 Introduction

The gene-smoking interaction analysis undertaken in this chapter focusses on the 97 regions known to be associated with lung function (to date at the time of analysis), identified through previous GWAS efforts (43-50). These were introduced in chapter 1 section 1.2.3 and are presented in **Table 4.1.**

Table 4.1 - The 97 lung function associated signals identified at the time of analysis in this chapter (to date in 2017)

SNP (Chr:Pos)	Trait	Gene	Reference
rs2284746 (1:17306675)	FEV_1/FVC	*MFAP2*	Soler Artigas et al. 2011 (46)
rs17513135 (1:40035686)	FEV_1/FVC	*LOC101929516*	Wain et al. 2017 (50)
rs1192404 (1:92068967)	FEV_1/FVC	*TGFBR3*	Wain et al. 2017 (50)
rs12140637 (1:92374517)	FEV_1/FVC	*TGFBR3*	Wain et al. 2017 (50)
rs200154334 (1:118862070)	FVC	*SPAG17*	Wain et al. 2017 (50)
rs6681426 (1:150586971)	FEV_1	*ENSA*	Soler Artigas et al. 2015 (49)
rs993925 (1:218860068)	FEV_1/FVC	*MIR548F3/TGFB2*	Soler Artigas et al. 2011 (46)
rs4328080 (1:219963088)	FEV_1/FVC	*RNU5F-1*	Soler Artigas et al. 2015 (49)
rs6688537 (1:239850588)	FEV_1/FVC	*CHRM3*	Wain et al. 2017 (50)
rs62126408 (2:18309132)	FEV_1/FVC	*KCNS3*	Soler Artigas et al. 2015 (49)
rs1430193 (2:56120853)	FVC	*EFEMP1*	Loth et al. 2014 (47)
rs2571445 (2:218683154)	FEV_1	*TNS1*	Repapi et al. 2010 (44)
rs10498230 (2:229502503)	FEV_1/FVC	*PID1*	Hancock et al. 2010 (45)
rs61332075 (2:239316560)	FEV_1/FVC	*TRAF3IP1*	Wain et al. 2017 (50)
rs12477314 (2:239877148)	FEV_1/FVC	*FLJ43879*	Soler Artigas et al. 2011 (46)
rs1529672 (3:25520582)	FEV_1/FVC	*RARB*	Soler Artigas et al. 2011 (46)
rs1458979 (3:55150677)	FEV_1/FVC	*CACNA2D3*	Wain et al. 2017 (50)
rs1490265 (3:67452043)	FVC	*SUCLG2*	Wain et al. 2017 (50)
rs2811415 (3:127991527)	FEV_1/FVC	*EEFSEC*	Wain et al. 2017 (50)
rs1595029 (3:158241767)	FVC	*RSRC1/RP11-538P18.2*	Soler Artigas et al. 2015 (49)
esv2660202 (3:168738454)	FEV_1/FVC	*MECOM*	Wain et al. 2017 (50)
rs1344555 (3:169300219)	FEV_1	*MECOM*	Soler Artigas et al. 2011 (46)
rs13110699 (4:89815695)	FEV_1/FVC	*FAM13A*	Wain et al. 2017 (50)

SNP (Chr:Pos)	Trait	Gene	Reference
rs2045517 (4:89870964)	FEV_1/FVC	*FAM13A*	Hancock et al. 2010 (45)
rs34480284 (4:106064626)	FEV_1	*TET2*	Wain et al. 2015 (48)
rs10516526 (4:106688904)	FEV_1	*GSTCD*	Repapi et al. 2010, Hancock et al. 2010 (44,45)
rs34712979 (4:106819053)	FEV_1	*NPNT*	Wain et al. 2015 (48)
rs138641402 (4:145445779)	FEV_1	*HHIP-AS1*	Wilk et al. 2009 (43)
rs91731 (5:33334312)	FVC	*TARS*	Wain et al. 2017 (50)
rs1551943 (5:52195033)	FEV_1/FVC	*ITGA1*	Wain et al. 2017 (50)
rs2441026 (5:53444498)	FVC	*ARL15*	Wain et al. 2017 (50)
rs153916 (5:95036700)	FEV_1/FVC	*SPATA9*	Soler Artigas et al. 2011 (46)
rs7713065 (5:131788334)	FEV_1/FVC	*C5orf56*	Wain et al. 2017 (50)
rs7715901 (5:147856392)	FEV_1	*HTR4*	Repapi et al. 2010, Hancock et al. 2010 (44,45)
rs3839234 (5:148596693)	FEV_1	*ABLIM3*	Wain et al. 2017 (50)
rs10515750 (5:156810072)	FEV_1/FVC	*CYFIP2*	Wain et al. 2017 (50)
rs1990950 (5:156920756)	FEV_1/FVC	*ADAM19*	Hancock et al. 2010 (45)
rs6924424 (6:7801611)	FVC	*BMP6*	Loth et al. 2014 (47)
rs34864796 (6:27459923)	FEV_1	*ZNF184*	Soler Artigas et al. 2011 (46)
rs28986170 (6:31556155)	FEV_1/FVC	*LST1*	Wain et al. 2017 (50)
rs2857595 (6:31568469)	FEV_1/FVC	*NCR3*	Soler Artigas et al. 2011 (46)
rs2070600 (6:32151443)	FEV_1/FVC	*AGER*	Repapi et al. 2010, Hancock et al. 2010 (44,45)
rs114544105 (6:32635629)	FEV_1	*HLA-DQB1*	Wain et al. 2015 (48)
rs114229351 (6:32648418)	FEV_1	*HLA-DQB1*	Wain et al. 2017 (50)
rs141651520 (6:73670095)	FEV_1/FVC	*KCNQ5*	Wain et al. 2017 (50)
rs2768551 (6:109270656)	FEV_1/FVC	*ARMC2*	Soler Artigas et al. 2011 (46)
rs7753012 (6:142745883)	FEV_1/FVC	*GPR126/LOC153910*	Soler Artigas et al. 2015 (49)
rs148274477 (6:142838173)	FEV_1/FVC	*GPR126/LOC153910*	Hancock et al. 2010 (45)
rs10246303 (7:7286445)	FEV_1/FVC	*C1GALT1*	Wain et al. 2017 (50)
rs72615157 (7:99635967)	FEV_1/FVC	*ZKSCAN1*	Wain et al. 2017 (50)
rs12698403 (7:156127246)	FEV_1	*LOC285889*	Wain et al. 2017 (50)
rs7872188 (9:4124377)	FEV_1	*GLIS3*	Wain et al. 2017 (50)
rs16909859 (9:98204792)	FEV_1/FVC	*PTCH1*	Hancock et al. 2010 (45)
rs803923 (9:119401650)	FEV_1/FVC	*ASTN2*	Soler Artigas et al. 2015 (48)
rs10858246 (9:139102831)	FVC	*QSOX2/LHX3*	Soler Artigas et al. 2015 (48)

SNP (Chr:Pos)	Trait	Gene	Reference
rs10870202 (9:139257411)	FVC	*DNLZ*	Wain et al. 2017 (50)
rs7090277 (10:12278021)	FEV_1/FVC	*CDC123*	Soler Artigas et al. 2011 (46)
rs3847402 (10:30267810)	FEV_1/FVC	*KIAA1462*	Wain et al. 2017 (50)
rs7095607 (10:69957350)	FVC	*MYPN*	Wain et al. 2017 (50)
rs2637254 (10:78312002)	FEV_1	*C10orf11*	Soler Artigas et al. 2011 (46)
rs4237643 (11:43648368)	FVC	*MIR129-2*	Loth et al. 2014 (47)
rs2863171 (11:45250732)	FVC	*PRDM11*	Loth et al. 2014 (47)
rs2509961 (11:62310909)	FEV_1	*AHNAK*	Wain et al. 2017 (50)
rs11234757 (11:86443072)	FEV_1	*PRSS23*	Wain et al. 2017 (50)
rs567508 (11:126008910)	FEV_1	*RPUSD4*	Wain et al. 2017 (50)
rs2348418 (12:28689514)	FVC	*CCDC91*	Soler Artigas et al. 2015 (48)
rs11172113 (12:57527283)	FEV_1/FVC	*LRP1*	Soler Artigas et al. 2011 (46)
rs1494502 (12:65824670)	FEV_1	*MSRB3*	Wain et al. 2017 (50)
rs113745635 (12:95554771)	FEV_1/FVC	*FGD6*	Wain et al. 2017 (50)
rs12820313 (12:96255704)	FEV_1/FVC	*SNRPF/CCDC38*	Soler Artigas et al. 2011 (46)
chr12:114743533 (12:114743533)	FEV_1	*TBX5*	Wain et al. 2015 (48)
rs10850377 (12:115201436)	FEV_1	*TBX3*	Soler Artigas et al. 2015 (48)
rs35506 (12:115500691)	FVC	*TBX3*	Wain et al. 2017 (50)
rs1698268 (14:84309664)	FEV_1/FVC	*LINC00911*	Wain et al. 2017 (50)
rs7155279 (14:92485881)	FEV_1	*TRIP11*	Soler Artigas et al. 2015 (48)
rs117068593 (14:93118229)	FEV_1	*RIN3*	Soler Artigas et al. 2015 (48)
rs72724130 (15:41977690)	FEV_1/FVC	*MGA*	Wain et al. 2017 (50)
rs10851839 (15:71628370)	FEV_1/FVC	*THSD4*	Repapi et al. 2010, Hancock et al. 2010 (44,45)
rs12591467 (15:71788387)	FEV_1/FVC	*THSD4*	Wain et al. 2017 (50)
rs66650179 (15:84261689)	FEV_1/FVC	*SH3GL3*	Wain et al. 2017 (50)
rs12149828 (16:10706328)	FEV_1/FVC	*TEKT5*	Soler Artigas et al. 2015 (48)
rs12447804 (16:58075282)	FEV_1/FVC	*MMP15*	Soler Artigas et al. 2011 (46)
rs3743609 (16:75467021)	FEV_1/FVC	*CFDP1*	Soler Artigas et al. 2011 (46)
rs1079572 (16:78187138)	FVC	*WWOX*	Loth et al. 2014 (47)
rs59835752 (17:28265330)	FEV_1/FVC	*EFCAB5*	Wain et al. 2017 (50)
rs11658500 (17:36886828)	FEV_1/FVC	*CISD3*	Wain et al. 2017 (50)
rs35524223 (17:44192590)	FEV_1	*KANSL1*	Wain et al. 2015 (48)
rs6501431 (17:68976415)	FVC	*CASC17*	Loth et al. 2014 (47)
rs7218675 (17:73513185)	FEV_1	*TSEN54*	Wain et al. 2015 (48)
rs113473882 (19:41124155)	FEV_1/FVC	*LTBP4*	Soler Artigas et al. 2015 (48)

SNP (Chr:Pos)	Trait	Gene	Reference
rs6140050 (20:6632901)	FVC	*BMP2*	Wain et al. 2017 (50)
rs72448466 (20:62363640)	FEV_1	*ZGPAT*	Wain et al. 2017 (50)
rs2834440 (21:35690499)	FEV_1/FVC	*KCNE2*	Soler Artigas et al. 2011 (46)
rs11704827 (22:18450287)	FEV_1	*MICAL3*	Wain et al. 2017 (50)
rs134041 (22:28056338)	FEV_1	*MN1*	Soler Artigas et al. 2015 (48)
rs2283847 (22:28181399)	FEV_1	*MN1*	Wain et al. 2017 (50)
rs7050036 (23:15964845)	FEV_1/FVC	-	Soler Artigas et al. 2015 (48)

The primary reason to focus on these regions is because they have already produced replicated marginal effect signals of association with lung function, implicating genes as having a role in lung health and disease. As a result of this we could expect such areas of the genome which contain at least one marginal effect signal, more likely to produce additional signals that represent interaction effects. Therefore, these regions provide plausible candidate areas of the genome to focus initial efforts for gene-environment interaction analysis. Furthermore, with the consideration of a smaller finite set of regions than considered genome-wide, one should be able to benefit from a more relaxed statistical significance threshold for determining presence of interaction effect, than the 5×10^{-8} deemed appropriate for genome-wide genetic association analysis. The aim is to identify novel gene-smoking interactions for lung function traits FEV_1 and FEV_1/FVC, which as previously described are the primary lung function phenotypes used to define COPD. Such research aims to increase our knowledge of the contribution of genetic effect towards poor lung function with only 9.6% and 14.3% of heritability explained for FEV_1 and FEV_1/FVC respectively at the time of analysis (with predicted heritability ≈ 40%) (34).

The analysis will use a two-step procedure for identifying SNPs producing interaction effect, for which the summarised pipeline is presented in **Figure 4.1** (more detailed methods are described in the methods section, section 4.6). Firstly, the joint test of both main and interaction effect will be used to screen the full SNP search space (within those regions containing at least one marginal genetic effect signal for lung function) and produce a subset of SNPs for interaction analysis. In the second step, the SNPs screened by the joint test will be tested for an interaction effect only, to determine the driving mechanism behind the association.

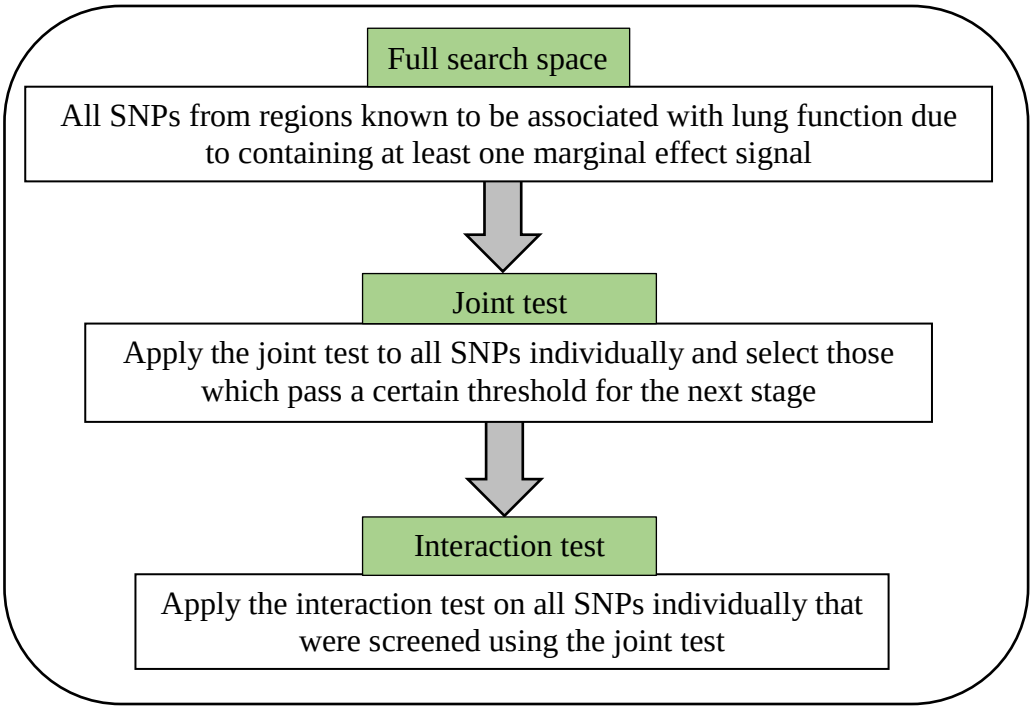


Figure 4.1 - Summary of analysis strategy for analysing the regions associated with lung function for interaction effects

4.2 Phenotype quality control for UK Biobank sample

4.2.1 Introduction

Spirometry for the UK Biobank data resource (introduced in chapter 1 section 1.2) was undertaken using a Vitalograph Pneumotrac 6800. Each individual produced a minimum of two blows (unless unable to due to extenuating circumstances) with a third blow performed if the first two were not consistent or reproducible, as determined by the spirometry equipment (reproducibility expanded on in section 4.2.2.1). As well as producing measures of lung function traits as default output, such as FEV_1 and FVC, the spirometry software also recorded the volume-time data for each individual at regular time intervals (every 10 milliseconds), in order to visualise the cumulative volume expelled during the test. An example volume-time curve is presented in **Figure 4.2**.

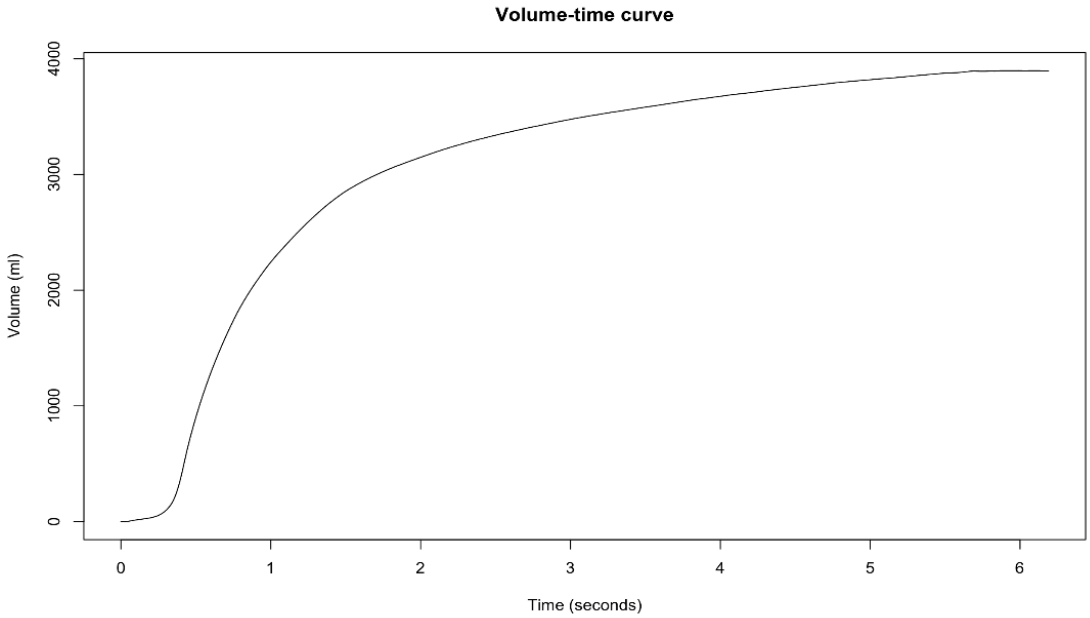


Figure 4.2 - An example volume-time curve for an individual who has undertaken spirometry

4.2.2 Spirometry QC methods

Using the volume-time data (introduced in the previous section) for each individual, we can further evaluate the quality of spirometry data beyond the default software output, by deriving variables informative for identifying unacceptable blows (**Table 4.2** and **Figure 4.3**). The additional QC metrics that were derived from the blow curves firstly required the calculation of the peak expiratory flow (PEF) for each individual. PEF is determined by calculating the flow (gradient of the volume-time curve) at each 80-millisecond interval and taking the maximum value (blue dashed line in **Figure 4.3**). By calculating PEF and in turn the time it was achieved (PEF time) and the volume exhaled at this point (PEF volume), we can then back extrapolate to determine the individual's true start time for each blow assuming maximum effort. This is defined as the "new time zero" (the new start time if constant flow was observed until PEF time). The volume exhaled at new time zero (the back extrapolated volume) was then calculated and used as an additional QC metric. Additionally, after determining new time zero, FEV_1 and FVC could also be derived, by recording the volume 1 second from this time point, and the maximum volume exhaled respectively. These measures could then be used to assess consistency with those output by the spirometry software.

The full UK Biobank data set has phenotypic data for 502,682 individuals and this was the sample used as input for spirometry quality control. All quality control was undertaken in R.

Table 4.2 - Variables derived from blow curves for each individual and each blow (in order of derivation)

Variable	Definition
PEF (Peak expiratory flow)	The maximum flow in ml/s calculated via averaging over each 80ms interval
PEF time	The time at which the maximum flow was attained (seconds)
PEF volume	The volume exhaled at the time PEF is reached (ml)
New time zero	The new time zero calculated by extrapolating back from max flow gradient to zero volume (seconds)
Back extrapolated volume	The volume exhaled by new time zero (ml)
FEV_1	Volume expired 1 second after new time zero (L)
FVC	Maximum volume expired during blow (L)

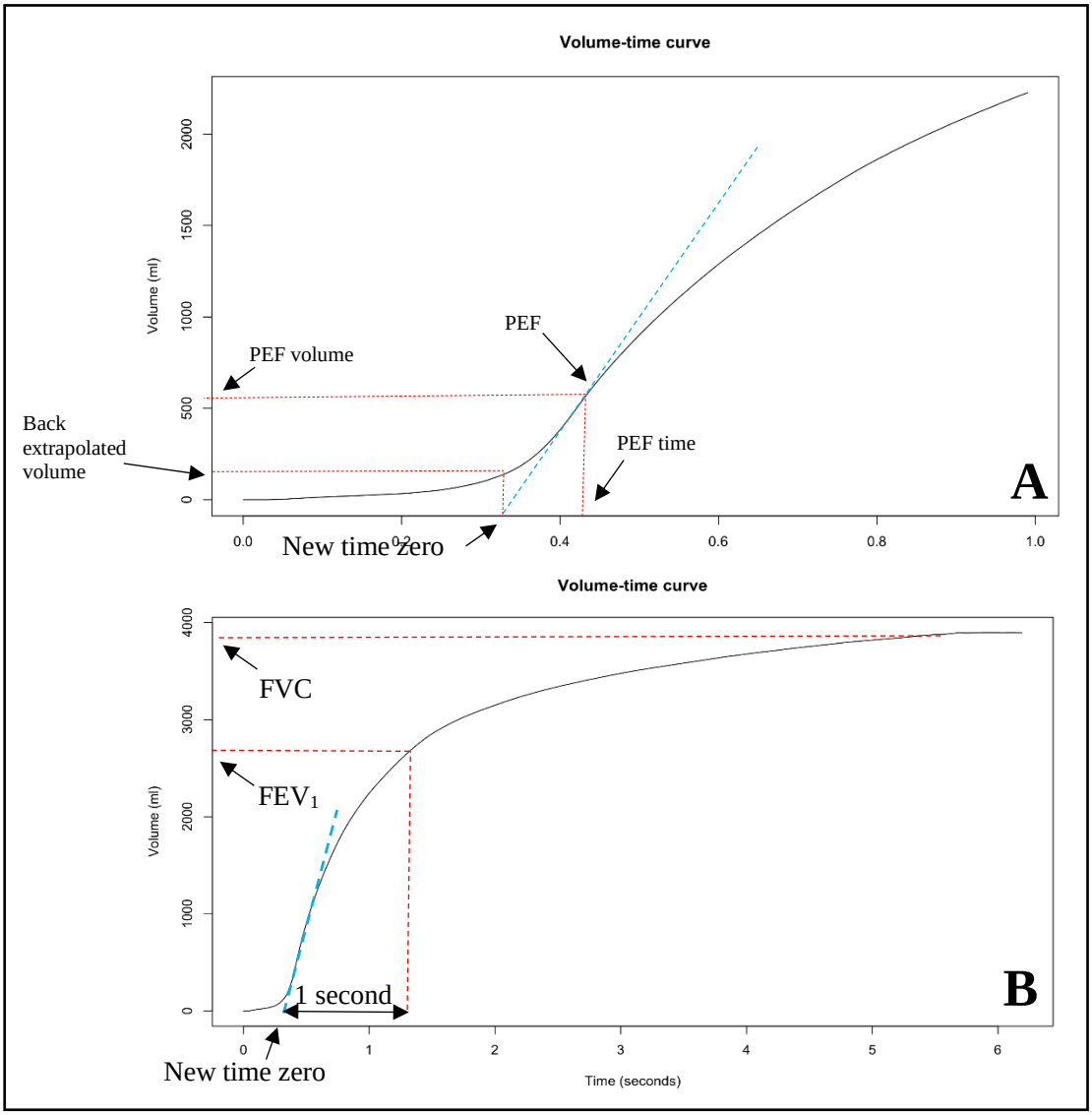


Figure 4.3 - An example volume time curve for the UK Biobank data used to illustrate the derivation of variables in Table 4.2

Fig. A presents a magnified plot of the first second of an individual's blow whilst Fig. B presents the full blow ($\approx 6\ seconds$)

4.2.2.1 Criteria for blow and individual removal

The following criteria were used to remove problematic blows and individuals, to produce a sample with clean lung function data:

Minimum two blows and relevant covariates

Individuals were first removed if they did not contribute at least two blows (two values for FEV_1 and FVC) and have data for spirometry method used, age, sex, standing height, and smoking variables (discussed in more detail in section 4.5).

Blow acceptability

Spirometry blows were then assessed for acceptability using the hexadecimal acceptability field (converted to decimal) returned by the spirometry equipment (UK Biobank field ID 3061). This indicated whether or not blows were problematic or deemed acceptable. For example an entry of "COUGHING" indicates that a cough was observed during the spirometry test for that blow. Blows were defined as acceptable if the field contained any of the following (with their decimal identifiers):

- **Blank - 0 -** This indicates no entry for the acceptability field
- **"ACCEPT" - 32 -** This indicates that the investigator accepts the blow
- **"BELOW6SEC ACCEPT" - 48 -** This indicates that the blow was less than 6 seconds and accepted by the investigator
- **"BELOW6SEC" - 16 -** This indicates that the blow was less than 6 seconds

Excessive back extrapolated volume

Excessive back extrapolated volume (defined in **Table 4.2**) suggested that the blow was not carried out with maximal effort. Back extrapolated volume was deemed excessive if larger than 5% of the observed FVC or 150ml, whichever of the two was greater, and thus excluded.

Consistency between UK Biobank output and derived lung function phenotypes

Values of FEV_1 and FVC derived from the volume-time data were checked for consistency with the FEV_1 and FVC values given by UK Biobank. Those which differed between the two sources were considered inconsistent. Inconsistency was determined by a difference between UK Biobank and derived values for FEV_1 and FVC of more than 5%. Blows which had an inconsistent value of FEV_1, FVC or both were removed.

Selecting FEV_1 and FVC for each individual

All blows that passed previous quality control criteria were deemed acceptable. From these remaining blows, each individual's FEV_1 and FVC measurement was chosen as the maximum value produced (not necessarily from the same blow).

Reproducibility of blows

To determine whether blows were reproducible (i.e. values of FEV_1 and FVC could be replicated and were not a chance set of measurements), the FEV_1 and FVC measurements chosen for each individual were compared with the corresponding values from other blows (did not necessarily have to be acceptable blows from previous criteria). To conclude reproducibility the chosen maximum values of FEV_1 and FVC had to be within 250ml of any other blow.

At each blow removal step, individuals were removed if they no longer contributed at least one blow passing the quality control step.

4.2.3 Spirometry QC results

56,928 individuals either did not have data available for at least two blows or had missing data for spirometry method used, age, sex, standing height or smoking variables and thus were removed, leaving 445,754 individuals. 777,676 blows from 387,430 individuals remained after removing unacceptable blows using the UK Biobank acceptability criteria. After assessing blows for excessive back extrapolated volume, 776,923 blows from 387,278 individuals remained. The UK Biobank provided FEV_1 and FVC values were checked for consistency with the values derived from the blow-curve in **Figure 4.3** and are presented in **Figure 4.4.** Values from the two sources were predominantly consistent, however 614 blows were identified for inconsistency. This left 776,309 blows from 387,051 individuals.

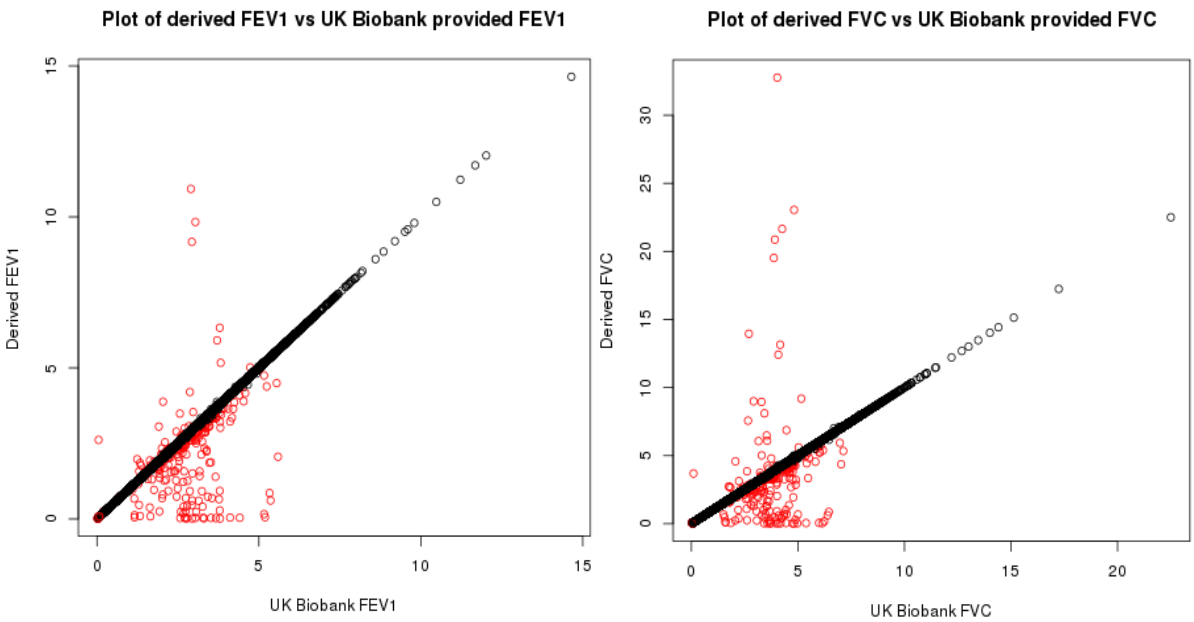


Figure 4.4 - Plots of derived FEV_1 and FVC from the volume-time curve for each blow against the provided values of FEV_1 and FVC from UK Biobank

This was used to identify inconsistent measurements. Those which differ by more than 5% are highlighted in red.

Finally, assessing the reproducibility of the blows identified 348,937 of the 387,051 individuals that had an acceptable and reproducible measurement for both FEV_1 and FVC.

Figure 4.5 illustrates the number of individuals removed at each phenotype QC stage.

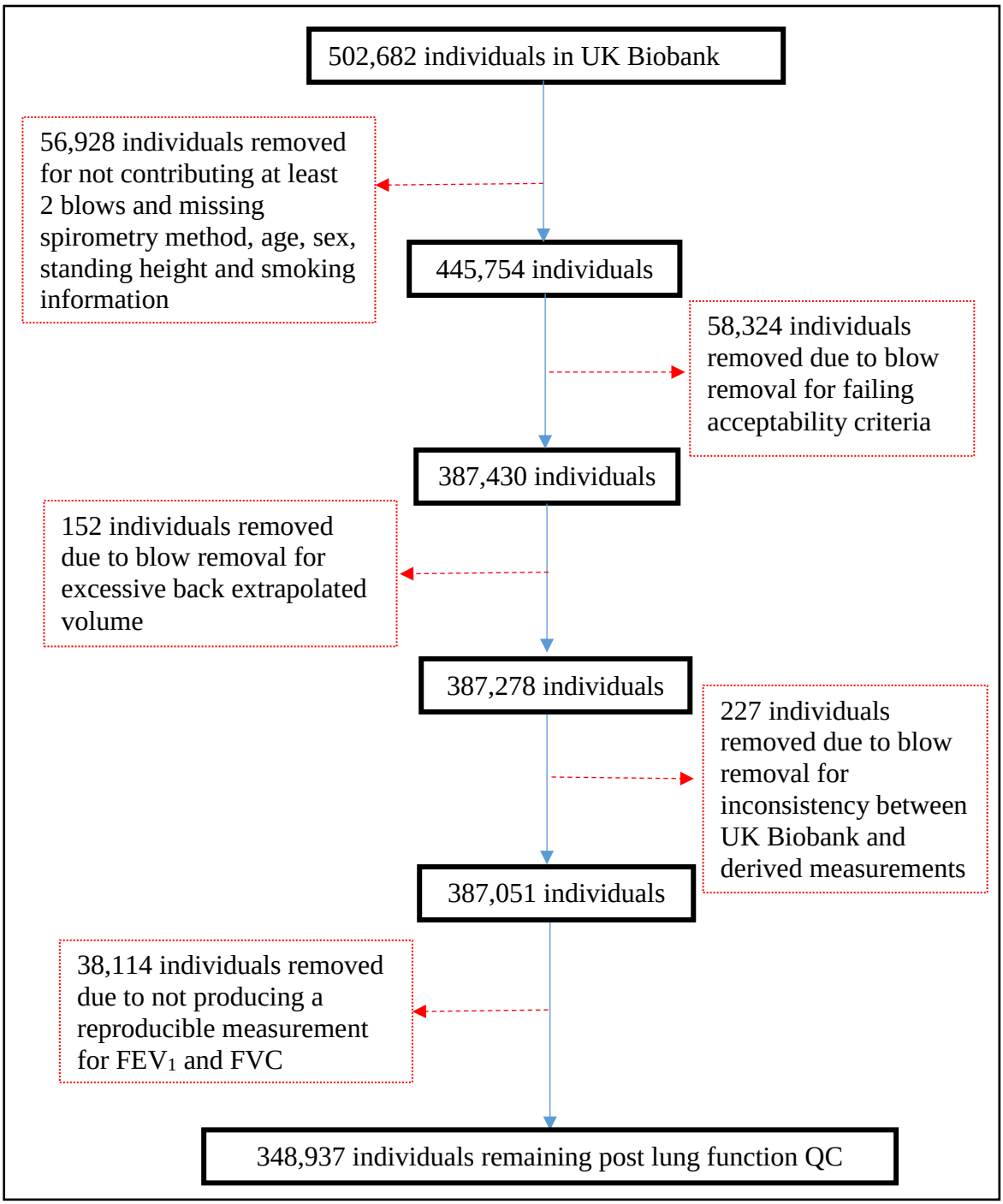


Figure 4.5 - Phenotype (lung function) QC sample removal

4.3 Genotype quality control and further sample exclusion for UK Biobank sample

4.3.1 Genotype QC methods

For the individuals passing phenotype QC, exclusions were then made on the basis of genotype QC. This was undertaken using both the provided sample QC metrics from

UK Biobank, and also by clustering individuals into ancestry groups using plots of principal components (also provided by UK Biobank).

Individuals were removed for the following criteria (introduced in chapter 1 section 1.1.3.1) to remove the risk of erroneous analysis results due to problematic samples:

- **There was no available genotype data for the individual**

- **Sex mismatch**
    - Sex submitted by participant and sex inferred by Affymetrix array sex chromosome marker intensity contradicts

- **Putative aneuploidy**
    - These individuals were not sex mismatches but had a sex chromosome karyotype which was not XX (female) or XY (male).

- **Heterozygosity or missingness outliers**
    - Individuals removed for extreme heterozygosity (proportion of non-missing genotypes that were heterogeneous)

- **Extreme relatedness**
    - These individuals were removed for showing relatedness with an excessive number of other individuals (> 10)

- **Ancestry clustering** (inferred from genotype with use of principal components supplied by UK Biobank - UK Biobank field ID 22009)
    - A further genotype QC measure used was principal component plots and K-means clustering to determine a European subset, rather than relying solely on self-reported ancestry.

4.3.2 Genotype QC results

Firstly 6,565 individuals were removed due to not having available genotype data (thus not included in the sample QC file). This left 342,372 individuals.

For the remaining 342,372 individuals, the QC metrics provided by UK Biobank identified further individuals for removal. **Table 4.3** gives the number of samples flagged for each of the QC metrics provided by UK Biobank presented in the previous section, as well as the intersecting number of samples (across all metrics) that were removed.

These exclusion metrics used to remove problematic samples resulted in the removal of a further 1,269 individuals, leaving 341,103 individuals.

Table 4.3 - Sample QC metrics in UK Biobank and number excluded for each
*Intersection of all individuals flagged not the sum, as some individuals flagged for more than one QC metric.

Reason for sample exclusion	Description of exclusion criteria	Number of individuals
Sex mismatch	Sex submitted by participant and sex inferred by Affymetrix array sex chromosome marker intensity contradict	228
Individuals with putative aneuploidy	These individuals were not sex mismatches but had a sex chromosome karyotype which was not XX (female) or XY (male)	368
Heterozygosity or missingness outliers	Individuals removed for extreme heterozygosity (proportion of non-missing genotypes that were heterogeneous) defined as having a heterozygosity above the average value (0.1903) and missing rate > 0.05.	628
Extreme relatedness	These individuals were removed for showing relatedness with an excessive number of other individuals (> 10), with 6 individuals related to > 200 others.	147
Total number of individuals*		**1,269**

Based on *self-reported* ancestry, approximately 95% of the 341,103 remaining individuals were of white ethnicity (**Table 4.4**). These self-reported ancestries are presented in a plot of principal components (PC1 vs PC2) in **Figure 4.6** and with emphasis on the "white British" group for more clarity in **Figure 4.7**. See section 1.1.3.1 for the methodology behind the use of principal components for exploring population structure.

Table 4.4 - Breakdown of self-reported ethnicity for 341,103 individuals passing previous phenotype and genotype QC

Ethnicity	Number of individuals
White	324,262
Mixed	1,988
Asian	5,837
Black	4,376
Chinese	969
Unknown (participant did not know)	2,590
Missing (participant did not answer)	1,081

In contrast to clustering individuals by self-reported ancestry only, a subset defined by UK Biobank using both self-reported ancestry and genotype information (provided as a pre-computed sample QC metric (162)) was used to more accurately define white-British individuals (**Figure 4.8**). To determine an accurate final subset of European individuals, this was used in combination with K-means clustering, with the chosen number of clusters adhering to the following criteria:

- All UK Biobank defined white-British individuals contained within one cluster
- The number of other individuals in the cluster are minimised

K-means clustering using 6 clusters identified 321,057 individuals as European (**Figure 4.9**). The number of individuals allocated to each cluster, alongside the number of UK Biobank defined white British, self-reported white British, and self-reported other ancestries are presented in **Table 4.5**.

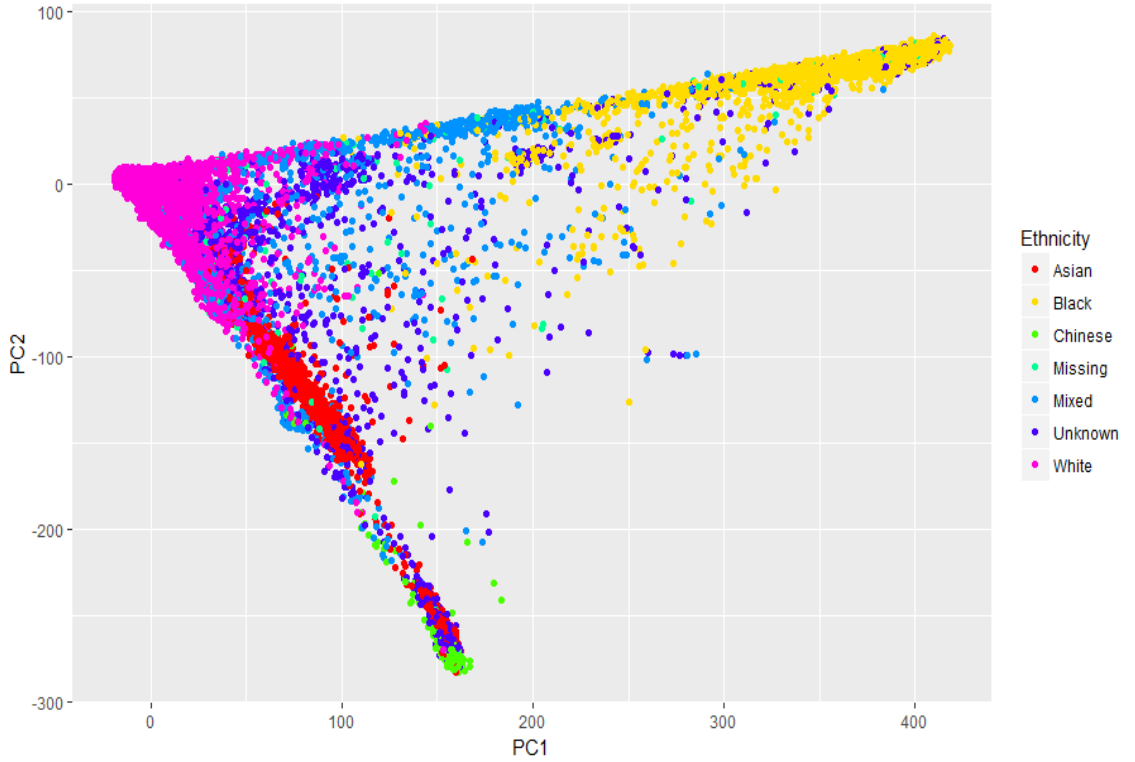


Figure 4.6 - Plot of PC1 vs PC2 with individuals categorised by self-reported ancestry

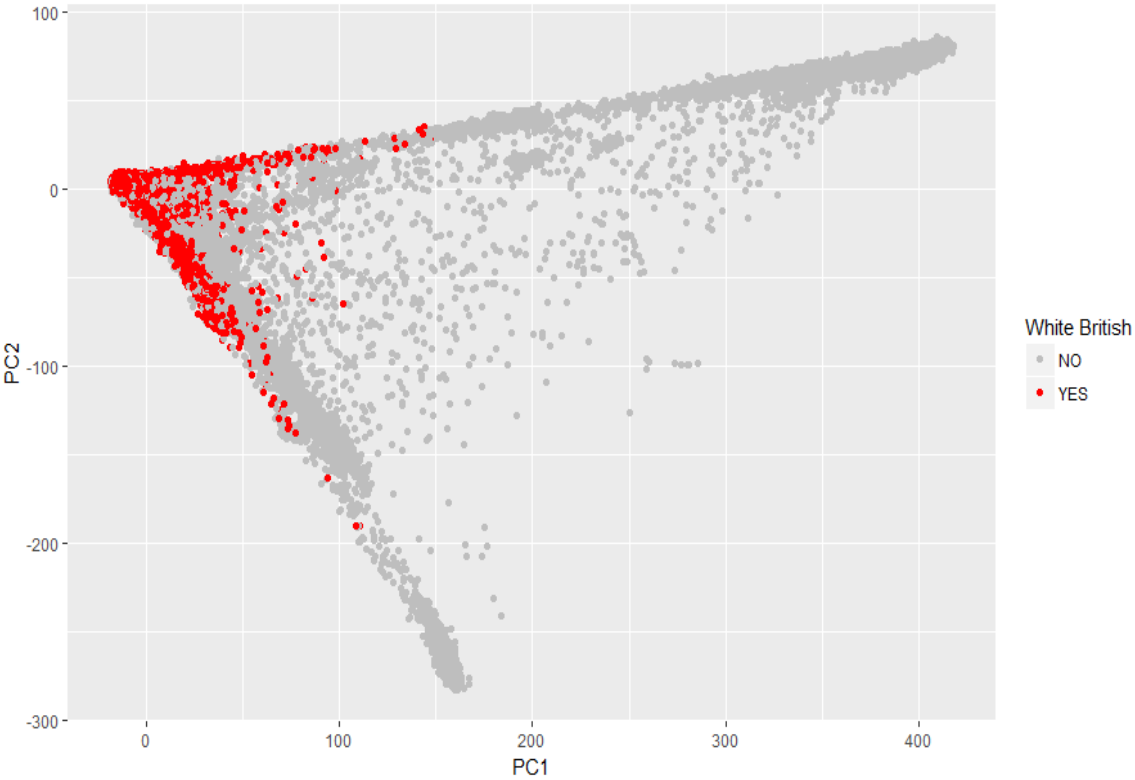


Figure 4.7 - Plot of PC1 vs PC2 for white British and non-white British groups based on self-reported ancestry

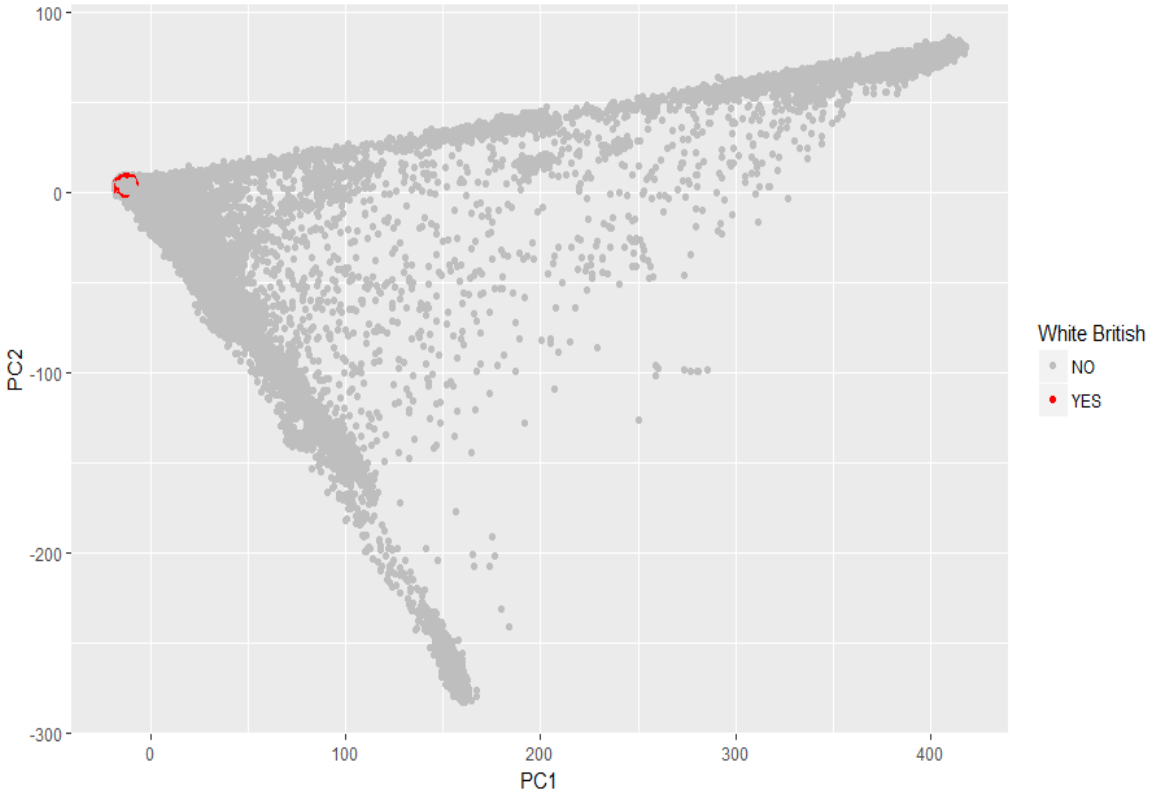


Figure 4.8 - Plot of PC1 vs PC2 for white British and non-white British groups based on criteria defined by UK Biobank

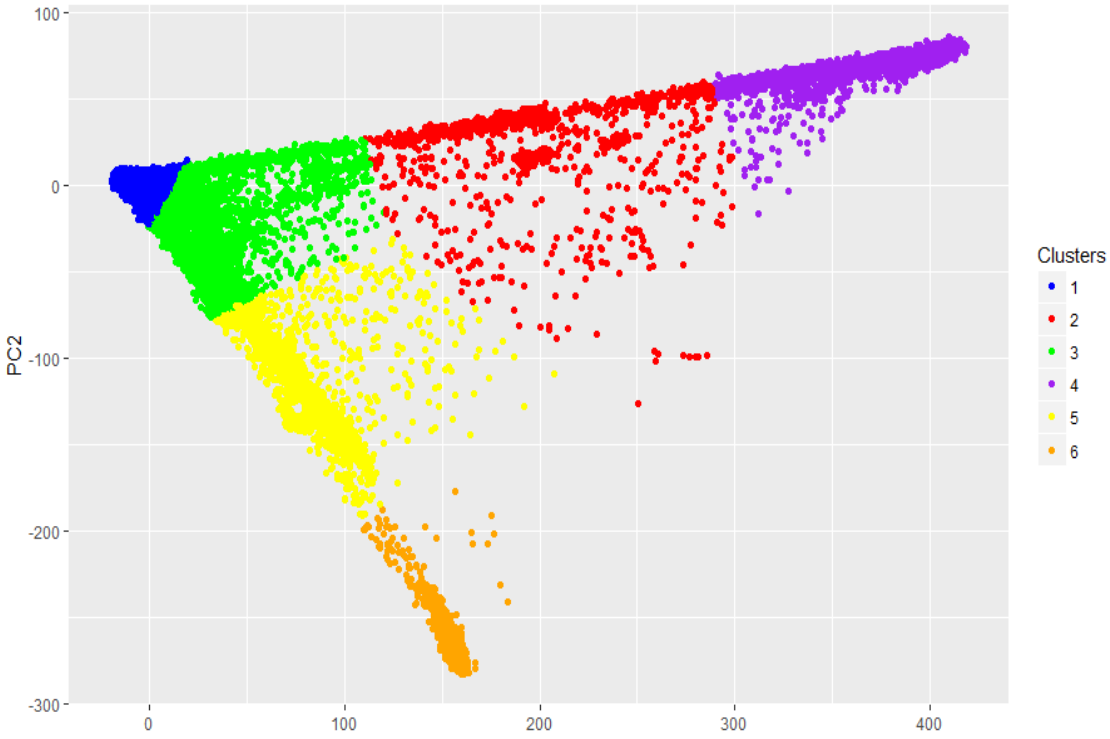


Figure 4.9 - K-means clustering with 6 clusters used to define the European subset (cluster 1)

Table 4.5 - The number of individuals in each K-means cluster
This presents the number defined as white British by UK Biobank, the number self-reported white British, and the number that self-reported an ethnicity other than white British.

Cluster	Number of individuals	UK Biobank defined white British	Self-reported white British	Self-reported other
1 (European cluster)	321,057	289,646	319,816	1,241
2	1,254	0	12	1,242
3	4,350	0	1	4,349
4	6,578	0	134	6,444
5	1,626	0	1	1,625
6	6,238	0	4,298	1,940

4.3.3 Additional exclusions to determine final European subset

A further ten samples were removed from the 321,057 individuals passing genotype QC. These individuals were flagged due to outlying residuals when regressing five of the available lung function phenotypes, FEV_1, FVC, FEV_1/FVC, PEF and FEF(25-75) (Forced Expiratory Flow at 25-75% of the volume exhaled, additionally derived for another project by another member of the university of Leicester genetic epidemiology group) on sex, age, age^2, height, smoking status and genotype array. Thus, the final sample of European individuals post phenotype and genotype QC was 321,047.

4.4 Relatedness removal for UK Biobank sample

4.4.1 Methods for identifying and removing related individuals

To produce a final sample of unrelated European individuals (removed due to the limitations of joint test analysis software, discussed further in section 4.6.3), relatedness was assessed within the 321,047 individuals passing previous QC requirements. Amongst the descriptive sample measures provided with the UK Biobank full data release, were the kinship estimates for individuals related up to the third degree (where first degree would be a parent-offspring relationship and examples of second and third degree relationships would be grandparent-grandchild or full siblings and aunt/uncle-nephew/niece relationships respectively).

Individuals related to 2^{nd} degree and above were removed, however 3^{rd} degree relatives were retained, to maximise sample size without causing negative effect on results due to

relatedness which is unaccounted for. Such an approach has been suggested for GWAS analysis (163). By retaining 3rd degree related individuals the sample size and thus power for analysis is maximised, without inflating type I error from using an unrelated analysis method on a sample of related individuals (164). The KING software (165) used to calculate kinship estimates proposed the kinship estimate range for a second-degree related pair is (0.0884, 0.177), thus taking the lower boundary, any related pair with a kinship estimate of 0.0884 or above was considered 2nd degree related or closer.

The following iterative process was used to produce an unrelated sample whilst ensuring the minimum number of exclusions:

Step 1

Prioritise removing individuals which share minimum 2nd degree relatedness with **more than one** other individual to retain as many individuals as possible using the below iterative process:

a. Pull out all kinship estimates for pairs which contain a sample ID that has more than one 2nd degree or above relationship (i.e. the IDs appear more than once in the relatedness file). This will leave "independent pairs" which are addressed in step 2.
b. For this subset of pairs, calculate the number of times each sample ID arises
c. Remove the individual (and resulting pairs which include these individuals) with the highest number of 2nd degree or above relationships.
d. Repeat process from b. until the largest number of times any ID arises is 2.
e. For the sample IDs which appear twice, remove the one (and the corresponding pairs) with the lowest call rate (calculated using the directly genotyped data and using PLINK (30)).
f. Continue step e. until no individuals remain which have more than one 2nd degree or above relationship, and therefore only independent pairs remain. These pairs will be collated with the other independent pairs and addressed in step 2.

Step 2

With independent pairs left from stage 1 (those initially identified and the ones remaining from iteratively removing individuals with more than one 2nd degree or above relationship), remove the individual with the lowest call rate based on directly genotyped data, with call rate calculated using PLINK (30).

4.4.2 Results for identifying and removing related individuals

Of the 107,162 related pairs of third degree or closer in UK Biobank, 48,712 remained when removing those related pairs for which one or both individuals were not present in the 321,047 individuals passing previous QC steps. When defining relatedness groups using the kinship ranges proposed by KING (165), the majority of the remaining pairs were related by third degree (29,456, **Figure 4.10**), with only 94 pairs with a kinship estimate suggesting that these two individuals were twins (or possibly duplicates).

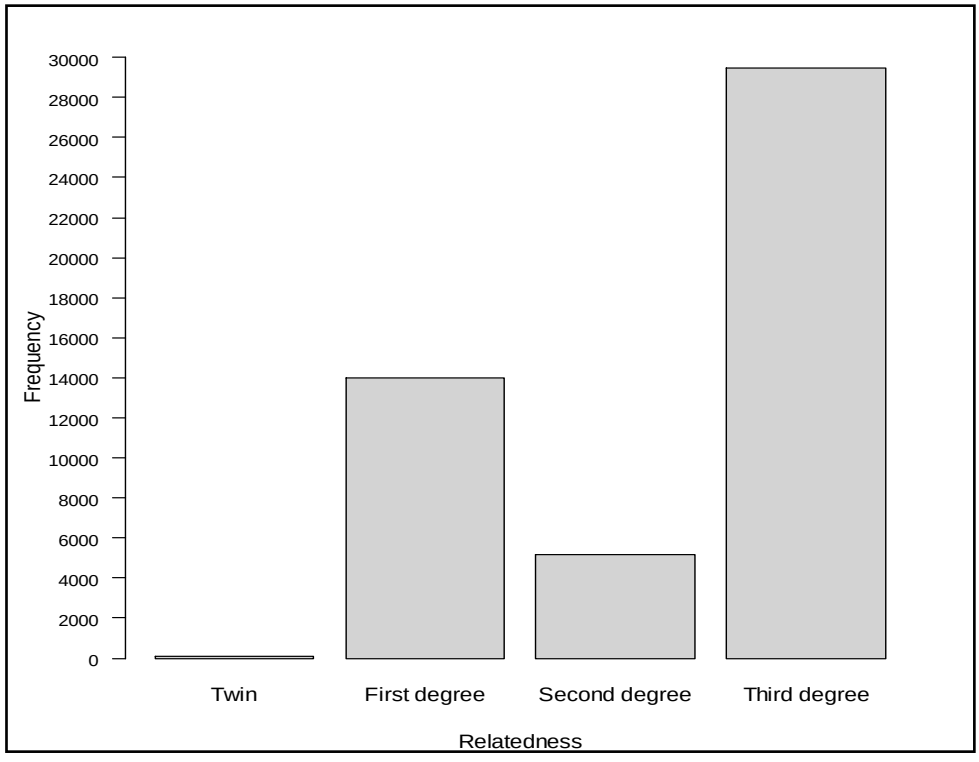


Figure 4.10 - Bar plot presenting the number of twins (or duplicates), first degree, second degree and third degree related pairs from the 321,047 individuals passing previous QC steps
Note: These are not independent pairs and individuals may contribute to more than one group. Groups were defined using a Kinship value (KV) range of KV > 0.354, 0.177 ≤ KV ≤ 0.354, 0.0884 ≤ KV < 0.177 and 0.0442 ≤ KV < 0.0884 for twins, first degree, second degree and third degree related pairs respectively.

The corresponding number of samples and related pairs removed at each stage is presented in **Figure 4.11**.

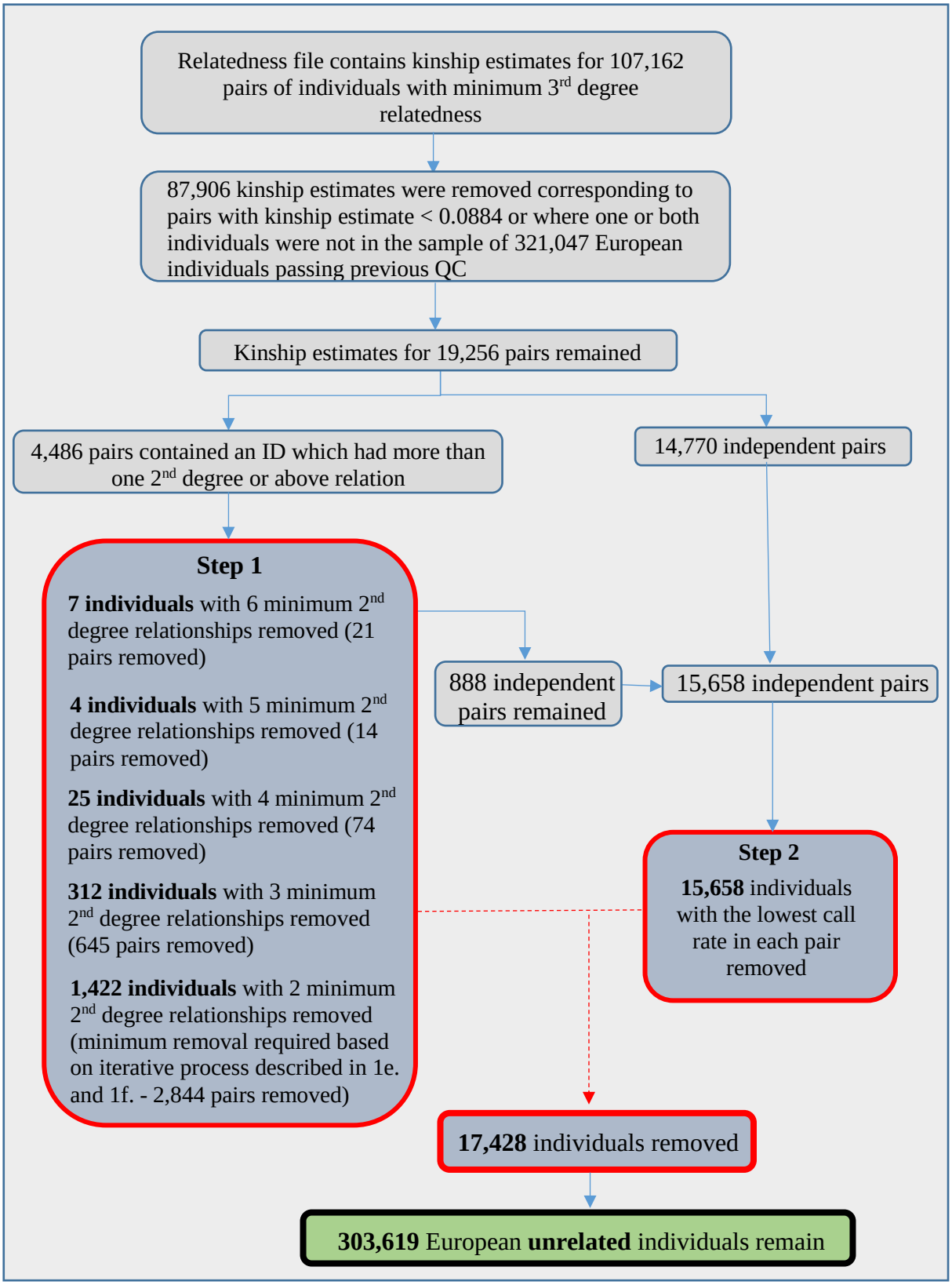


Figure 4.11 - Removal of related samples in UK Biobank (post phenotype and genotype QC) to produce an unrelated European subset

After removal of related individuals up to the second degree, 303,619 individuals remained. Of these 303,619 individuals, there were 45,339 individuals remaining that were related by 3rd degree with at least one other from the same sample of 45,339 individuals. The majority of related individuals (88%) were related to only one other individual in the analysed sample (**Table 4.6).**

Table 4.6 - A summary of the number of individuals that were related to at least one other individual in the sample

Number of 3rd degree relatives	**Number of individuals**
1	39,968
2	4,803
3	509
4	56
5	3
Total	45,339

4.5 Defining ever and never smokers

Three UK Biobank fields were used to define ever and never smokers for the 303,619 unrelated European UK Biobank samples. These fields were "current tobacco smoking" (UK Biobank field ID – 1239), "previous tobacco smoking" (UK Biobank field ID – 1249), and "at least 100 smokes in lifetime" (UK Biobank field ID – 2644). Ever and never smokers were defined using the combinations of answers from these three fields as presented in **Table 4.7**. Amongst the 303,619 unrelated samples there were 139,288 ever smokers and 164,331 never smokers.

Table 4.7 - Defining ever and never smokers for the unrelated European sample of 303,619 individuals

Current tobacco smoking (FIELD ID – 1239)	**Previous tobacco smoking** (FIELD ID – 1249)	**At least 100 smokes in lifetime** (FIELD ID – 2644)	**No. Samples**	**Categorisation**
Do not currently smoke	Never smoked in the past	-	118,925	Never smoker (n = 164,331)
	Smoked occasionally or once/twice in the past	Less than 100 smokes in lifetime	45,406	
Do not currently smoke	Smoked on most or all days in the past	-	74,788	Ever smoker (n = 139,288)
	Smoked occasionally or once/twice in the past	More than 100 smokes in lifetime	34,040	
Currently smoke on most days or occasionally	-	-	30,416	
Prefer not to answer about current smoking	Smoked most days or occasionally in the past	More than 100 smokes in lifetime	44	

4.6 Methods: Association analysis

4.6.1 Defining the regions containing at least one marginal effect signal associated with lung function and SNP QC

Selection of the regions of the genome involved in lung function (due to containing at least one associated marginal effect signal) required extraction of the 97 associated loci with traits FEV_1, FVC and the ratio FEV_1/FVC published to date (43-50). Using the reported sentinel SNP for each signal, the regions were then defined by taking an equidistant length of ± 1.5Mb around each one, to create 97 regions of length 3Mb. However, in the case of closely located independent signals and sentinel SNPs, where allocated regions would overlap, region sizes were increased accordingly to contain all corresponding signals for the analysis and maintaining the 1.5Mb distance either side (by taking the lowest base position SNP minus 1.5Mb and the highest base position SNP plus 1.5Mb). Of the 97 SNPs documented, seven were removed from further analysis, six due to being within the MHC region on chromosome 6 (an area of the

genome known for its difficult interpretation due to large variability and high LD) and one non-autosomal signal for which genotype data in UK Biobank was not available.

For all regions, SNPs were removed from the raw HRC-imputed UK Biobank genotype data if not biallelic (have more than two alleles at their locus), had a minor allele count (MAC) ≤ 3, and if imputation quality (the metric used to determine how well the SNP was imputed) was < 0.5 regardless of MAF and < 0.8 for SNPs with MAF < 5%.

4.6.2 Defining a significance threshold

Concluding statistical significance requires consideration of an appropriate significance threshold. Adjustment is needed to account for multiple testing to reduce the risk of false-positive associations (observed associations which are in fact false). When considering a finite number of regions along the genome, we want to apply a threshold which is more relaxed than that applied in genome-wide analysis of imputed data. Instead of selecting a relaxed threshold arbitrarily, the aim was to determine the number of independent SNPs across all defined regions, and thus adjust the 5% significance level accordingly using a Bonferroni correction.

Independent SNPs in each region were determined with the use of LD pruning in PLINK (30) (using LD < 0.2, the suggested default cut off, to determine independence).

4.6.3 Software for analysis

To determine the software appropriate for the joint test stage of the two-step analysis required in this chapter, multiple software packages were considered. The software needed to meet the following three criteria:

- **Criteria 1** - Can run the joint test of both main and interaction effect
- **Criteria 2** - Can analyse non-integer imputed genotype dosage data and not just directly genotyped (integer) data
- **Criteria 3** - Can analyse related individuals (account for the effects of relatedness within the sample and adjust results accordingly)

Various software packages were considered for suitability based on the three requirements above (**Table 4.8**).

Table 4.8 - Software considered for the joint test interaction analysis

Software	Ref.	Criteria 1	Criteria 2	Criteria 3	Does it meet all requirements?
Plink	(30)	YES	YES	NO	NO
SNPTEST	(11)	NO	YES	YES	NO
BOLT-LMM	(166)	NO	YES	YES	NO
CASSI	(167)	NO	NO	NO	NO
CGEN R	(168)	YES	YES	NO	NO
snpStats	(169)	NO	YES	NO	NO
ProbABEL	(170)	YES	YES	YES	YES
QUICKTEST	(171)	NO	YES	NO	NO

Only the ProbABEL package (170) of the several considered appeared to meet all three criteria. However, relatedness could not be accounted for due to computer memory restrictions, due to the dependency on R for computing kinship matrices for the 321,047 individuals passing phenotype and genotype QC. Therefore, related individuals were excluded and the joint test was applied in ProbABEL using unrelated individuals only.

4.6.4 Adjustment of phenotypes for analysis

Phenotypes (FEV_1 and the ratio FEV_1/FVC) were adjusted for sex, age, age^2, height and ten principal components using a linear regression model. Traits were then inverse normalised (the residuals from the regression model were ranked and converted into normal z-scores) and used as the phenotype for the joint test SNP by smoking interaction analysis, with the following model:

$$Phenotype_i = \beta_0 + \beta_1 G_i + \beta_2 S_i + \beta_3 G_i S_i + \varepsilon_i \qquad (4.1)$$

where G_i is the genotype for individual i, and S_i is the ever/never smoking status. The p-value returned corresponds to the joint test of both β_1 and β_3, the main effect of each SNP (G) and the interaction effect (GS) between SNP and ever/never smoking (S)

respectively. Genotypes were input in dosage format (taking a continuous value from 0 to 2).

4.6.5 Signal selection from joint test analysis

For both traits FEV_1 and FEV_1/FVC, regions which contained at least one SNP reaching the defined significance threshold were identified and explored further to determine the number of independent signals of association. A signal is defined here by a group of SNPs which show association with the phenotype of interest, but are not however independent of one another due to LD structure. For each signal, we can allocate a sentinel SNP i.e. the most statistically significant SNP in that LD dependent block of SNPs. Due to the possibility of each region, which has at least one signal, having further independent signals, the following process was applied:

1. Initially, each region harbouring statistically significant SNPs was allocated a sentinel SNP
2. This SNP and all SNPs in LD (with threshold LD > 0.2) were considered one signal
3. For any SNPs reaching the significance threshold and yet to be accounted for, a new sentinel was allocated (the most statistically significant unaccounted for SNP)
4. Stage 2 and 3 was repeated until no SNPs remained.

This LD clumping procedure was applied using PLINK software (30). Region plots were also produced using LocusZoom (29), using the sentinels returned by the PLINK clumping algorithm as reference SNPs. These plots were produced to examine the plausibility of the defined independent signals in each region assigned to each sentinel.

4.6.6 Interaction test for joint test signals

Independent signals from the joint testing stage were taken forward to determine whether the signal has been identified due to contributing a main genetic effect, an interaction effect with smoking, or potentially both. This was concluded using a test of interaction effect only (testing β_3 only from linear regression equation (4.1) in section 4.6.4) using the mean genotype (expected allele dosage) method in QUICKTEST

(171,172). The mean genotype method is closely related to the widely used Score test, however utilises the exact variance of the score test statistic rather than an estimation, maintaining power but reducing the false positive rate (171). The 5% significance threshold was Bonferroni corrected for the number of SNPs reaching this stage. For any signals which showed statistical evidence of an interaction effect, the marginal effect in each smoking group (ever and never smokers) was also calculated, and this was also undertaken using QUICKTEST.

4.6.7 Replication and smoking association for interaction signals

For any SNPs which produce a statistically significant interaction effect, two follow-up stages were implemented. Firstly, lung function phenotype and genome-wide imputed genotype data collected from multiple cohorts as part of the SpiroMeta consortium, was used to seek replication of the SNPs identified for producing an interaction effect. Secondly, a look up was undertaken for each of the interaction signals to determine whether any were associated with smoking behaviour. This was used to conclude whether the signal was driven by a smoking association, rather than a true differing genetic effect between smoking groups.

4.6.7.1 SpiroMeta consortium meta-analysis look-up

The SpiroMeta consortium consists of 22 studies (with a total sample size of 71,067 individuals) with both lung function (including FEV_1 and FEV_1/FVC) and smoking (ever/never smoking) data (**Table 4.9**). Genotypes for each study were imputed to either 1000G (14) or HRC (15) reference panels. The genome-wide results of a marginal genetic effect analysis undertaken for each study (using the same adjustment and phenotype transformation described in section 4.6.4) was available for the total sample, and also for ever and never smoking subgroups separately. For replication of results from this study, the ever and never smoking subgroup analysis was meta-analysed across all studies by myself, and genetic effects were pulled out for each subgroup for each SNP which showed a statistically significant interaction effect outlined in section 4.6.6. Welch's t-test was used to produce a test statistic for testing interaction effect (testing equality of marginal genetic effect sizes in ever and never smoking subgroups) with test statistic (t) and degrees of freedom ($d.f.$) as follows (using the t-distribution to calculate a p-value):

$$t = \frac{\beta_1 - \beta_2}{\sqrt{{se_1}^2 + {se_2}^2}} \qquad d.f. = \frac{({se_1}^2 + {se_2}^2)^2}{\frac{{se_1}^2}{n_1 - 1} + \frac{{se_2}^2}{n_2 - 1}}$$

Table 4.9 - SpiroMeta consortium studies and sample sizes

Study	Sample size	Ever Smokers	Never Smokers	Imputation panel
B58C	5934	4225	1709	1000G
BHS1&2	4355	2054	2301	1000G
CROATIA-Korcula	826	423	403	1000G
CROATIA-SPLIT	493	254	239	1000G
CROATIA-Vis	925	537	388	1000G
EPIC	20771	11239	9532	HRC
Generation Scotland	8093	3774	4319	1000G
Health 2000	820	572	248	1000G
KORA F4	1474	918	556	1000G
KORA S3	1147	627	520	1000G
LBC1936	991	554	437	1000G
NFBC1966	5078	2600	2478	HRC
NFBC1986	3210	734	2476	HRC
NSPHS	871	121	750	1000G
ORCADES	1804	782	1022	1000G
Pivus+ULSAM	806	413	393	HRC
SAPALDIA	1373	747	626	1000G
SHIP	1759	941	818	HRC
SHIP-TREND	804	462	342	HRC
UKHLS	7442	4504	2938	HRC
VIKING	1672	744	928	HRC
YFS	419	186	233	1000G
TOTAL	**71067**	**37411**	**33656**	

Statistical significance was determined with the use of a Bonferroni corrected threshold for the number of SNPs tested.

4.6.7.2 Association of interaction SNPs with smoking

The purpose here was to determine whether any of the observed SNPs associated with interaction effects, were in fact driven by an association with smoking behaviour (with smoking an intermediate variable on the causal pathway). This could amplify the interaction effect greatly, increasing the comparative genetic effect between ever and never smokers. This was achieved with the use of a genome-wide association study of

smoking behaviour undertaken in UK Biobank (personal communication with Dr Chiara Batini at the University of Leicester).

Smoking phenotypes available from the UK Biobank GWAS were smoking initiation (SI) with N = 275,596 (123,890 ever smokers and 151,706 never smokers), smoking cessation (SC) with N = 123,851 (25,905 current smokers and 97,946 former smokers) and number of cigarettes smoked per day (CPD) with N = 80,015. The sample sizes for analysis included European individuals only. Sample sizes were smaller than the lung function phenotype sample size (N = 303,619) for two reasons. Firstly, a different QC pipeline was used to remove individuals, and the large difference in the SI phenotype was specifically accredited to much more stringent removal of related individuals. Secondly, for SC and CPD phenotypes, there was a larger proportion of missing data, with individuals either not knowing or not willing to provide information. All traits were adjusted for age, age^2, sex and the first 15 PC's and inverse normalised prior to analysis. Analysis was undertaking using a linear-mixed model in BOLT-LMM version 2.3 (166). Statistical significance was determined with the use of a Bonferroni corrected threshold for the number of SNPs tested.

A flow chart summarising the two-step analysis and follow-up methods used to identify SNPs that contribute interaction effects is given in **Figure 4.12**.

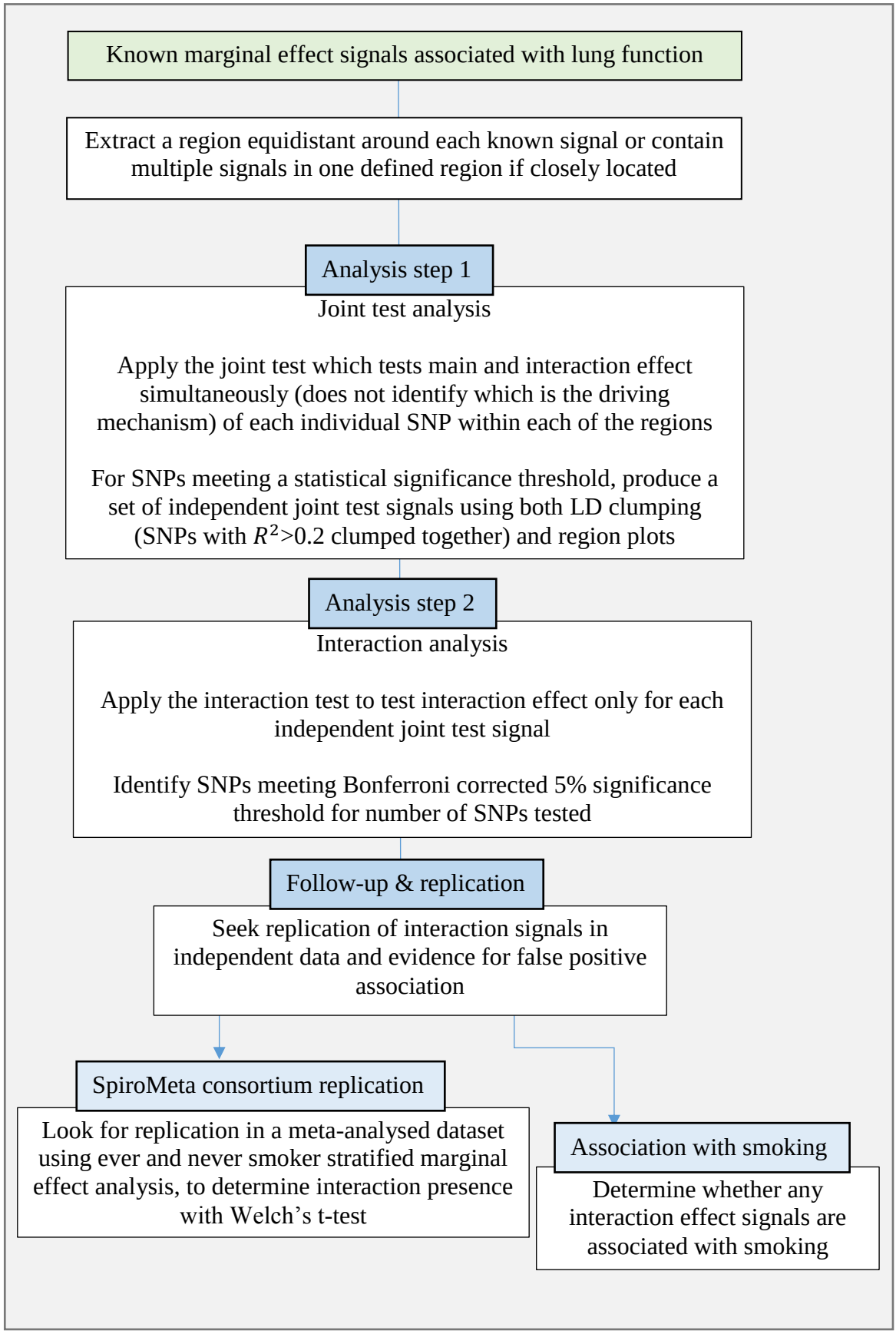


Figure 4.12 - Flow chart to summarise the methods for interaction analysis and follow-up

4.7 Results: Association analysis

4.7.1 Summary of lung function phenotypes and covariates

Table 4.10 presents the phenotype summary for the 303,619 individuals included in the gene-smoking interaction analysis, both in total and also stratified by ever and never smoking subgroups. There were less males (~45%) in the analysed sample and this was more extreme in the never smoker subgroup (~39% male). The average age was 56.5. Average values for lung function phenotypes were marginally lower for the ever smoker subgroup, when compared with the never smoker subgroup (statistically significant difference for both FEV_1 and the ratio of FEV_1/FVC with $p = 0$ when using a two sample t-test in R).

Table 4.10 - Phenotype summary for the 303,619 unrelated individuals included in the gene-smoking interaction analysis

Variable	Total (n = 303,619)	Ever smokers (n = 139,288)	Never smokers (n = 164,331)
Sex - n(%) male	135,478 (44.6)	71,525 (51.4)	63,953 (38.9)
Age - mean(sd)	56.5 (8.0)	57.3 (7.8)	55.8 (8.0)
Standing height – cm, mean(sd)	168.61 (9.13)	169.41 (9.05)	167.92 (9.14)
FEV_1 – Litres, mean(sd)	2.84 (0.76)	2.83 (0.77)	2.86 (0.75)
FEV_1/FVC – mean(sd)	0.76 (0.06)	0.75 (0.07)	0.77 (0.06)

4.7.2 Region definition, SNP QC and threshold

For the 90 signals previously reported for lung function, this equated to 70 3Mb+ regions of the genome. Of the 70 defined regions, 52 were 3Mb in length and contain one known lung function signal, with 18 regions larger than 3Mb and harbouring two or more closely located signals. The largest region was ≈5.7Mb in length. **Figure 4.13** presents a flow chart to illustrate the number of defined regions for analysis, and the number of SNPs removed at each QC stage as described in section 4.6.1. SNP QC resulted in a total of 1,831,014 SNPs taken forward for analysis. A large proportion of the 1,831,014 SNPs analysed were of rare MAF (MAF < 1%) (**Figure 4.14**). Of the 1,831,014 SNPs, 445,003 (~25%) were common with MAF ≥ 0.05 with 1,386,011 of low MAF (MAF < 0.05). Of the 1,386,011 low frequency SNPs 1,196,486 (~ 85%) were rare with 1,098,166 (~80%) having MAF < 0.005.

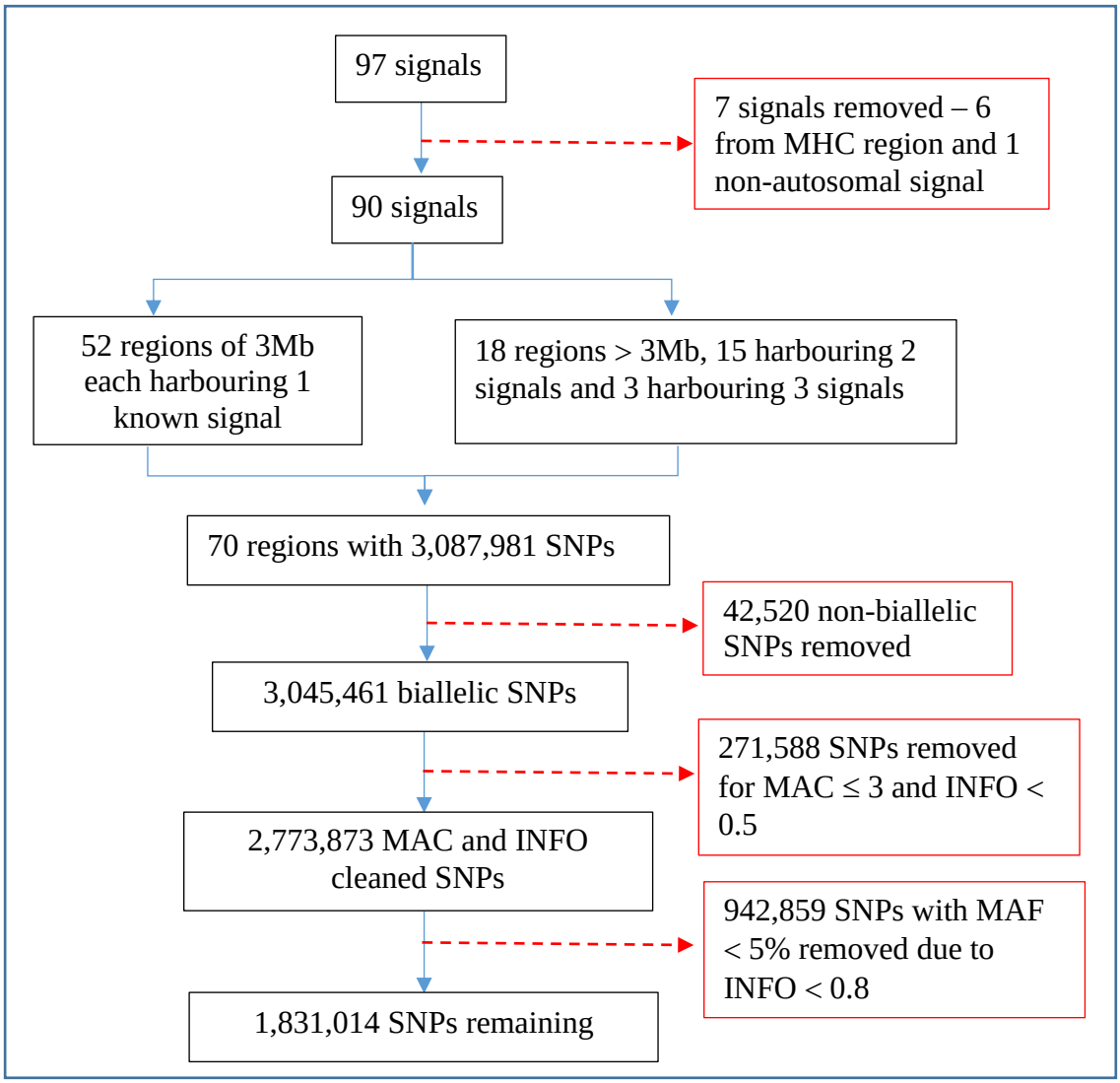


Figure 4.13 - Region selection and SNP QC process

MAC – minor allele count and INFO – imputation score.

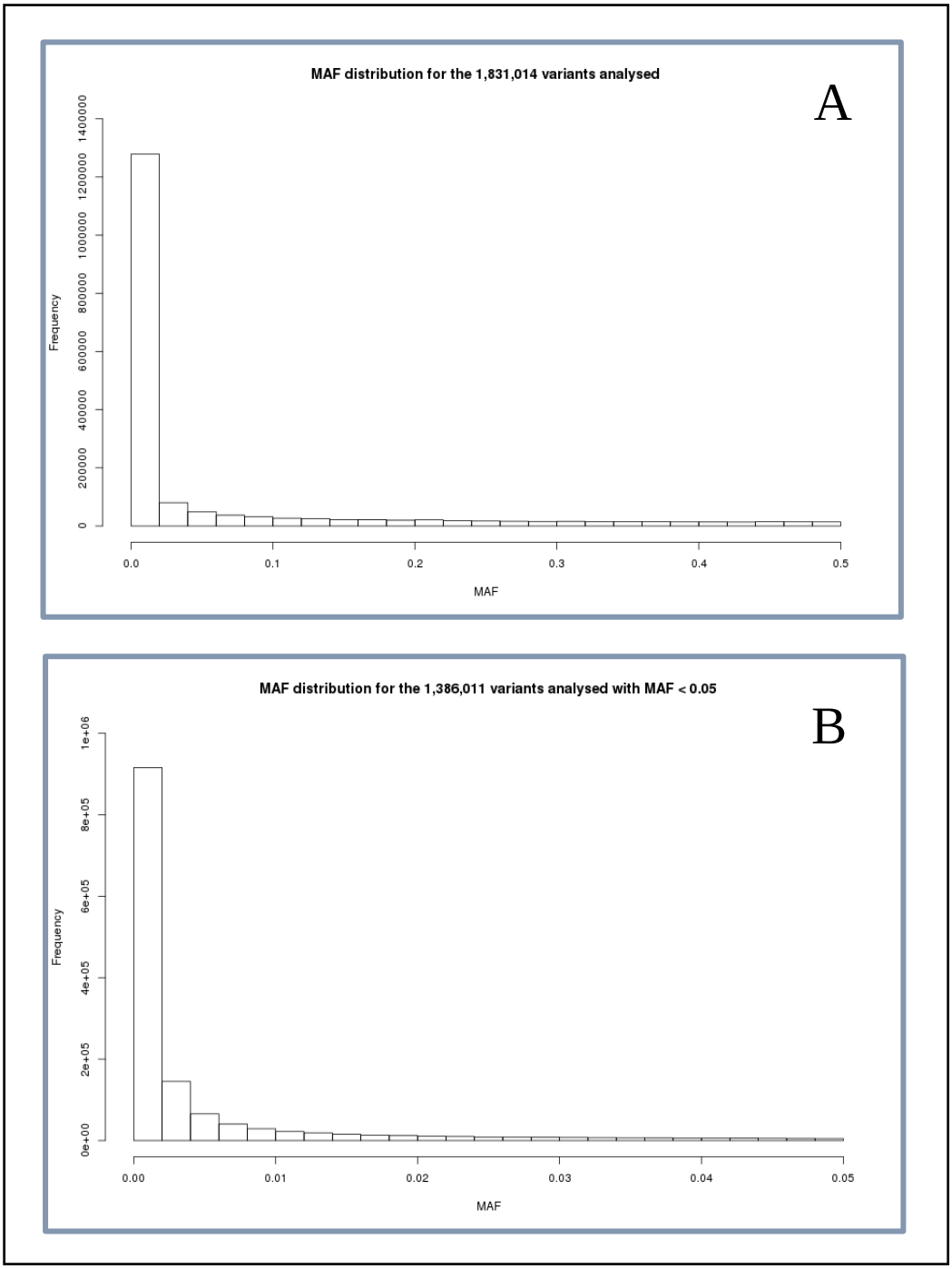


Figure 4.14 - MAF distribution for the 1,831,014 SNPs passing QC and taken forward for gene-environment interaction analysis

For (A) All analysed MAFs and (B) Low frequency MAFs i.e. MAF < 0.05

Across all regions 916,999 SNPs were identified as independent (LD < 0.2), with 761,972 and 155,027 with MAF ≤ 1% and MAF > 1% respectively (**Table 4.11**). For the analysis of all SNPs (all MAFs), adjustment was made for 916,999 independent tests, thus producing a significance threshold of $p < 5.5 \times 10^{-8}$.

Table 4.11 - Characteristics of the defined regions which contain one or more marginal effect signals used in the gene-environment interaction analysis

Region No.	Chr	Region size (Mb)	Reported SNP/s	Ref.	SNPs passing QC	Independent SNPs		
						All	MAF ≤1%	MAF >1%
1	1	3	rs2284746	(46)	21606	9526	7662	1864
2	1	3	rs17513135	(50)	21696	10741	8702	2039
3	1	3.31	rs1192404, rs12140637	(50)	24963	13243	11202	2041
4	1	3	rs200154334	(50)	24723	12397	10370	2027
5	1	3	rs6681426	(49)	14598	7015	5824	1191
6	1	4.1	rs993925, rs4328080	(46,49)	34337	16732	13951	2781
7	1	3	rs6688537	(50)	26218	11426	9178	2248
8	2	3	rs62126408	(49)	25699	12607	10391	2216
9	2	3	rs1430193	(47)	30813	15828	13201	2627
10	2	3	rs2571445	(44)	22474	10139	8377	1762
11	2	3	rs10498230	(45)	24488	11411	9263	2148
12	2	3.56	rs61332075, rs12477314	(46,50)	32756	14234	11431	2803
13	3	3	rs1529672	(46)	28948	13877	11608	2269
14	3	3	rs1458979	(50)	25612	12604	10360	2244
15	3	3	rs1490265	(50)	27134	12768	10609	2159
16	3	3	rs2811415	(50)	23775	11872	9864	2008
17	3	3	rs1595029	(49)	24050	11373	9418	1955
18	3	3.56	esv2660202, rs1344555	(46,50)	27807	13378	11074	2304
19	4	3.06	rs13110699, rs2045517	(45,50)	24590	11671	9772	1899
20	4	3.75	rs34480284, rs10516526, rs34712979	(44,45,48)	28558	14329	11880	2449
21	4	3	rs138641402	(43)	22520	10654	8845	1809
22	5	3	rs91731	(50)	23392	9743	7854	1889
23	5	4.25	rs1551943, rs2441026	(50)	39159	19538	16294	3244
24	5	3	rs153916	(46)	22740	11224	9331	1893
25	5	3	rs7713065	(50)	24405	12108	10403	1705
26	5	3.74	rs7715901, rs3839234	(44,45,48)	30302	13871	11434	2437
27	5	3.11	rs10515750, rs1990950	(45,50)	28604	14445	12105	2340
28	6	3	rs6924424	(47)	26371	11654	9395	2259
29	6	3	rs141651520 ;rs67163390	(50)	24215	11389	9422	1967
30	6	3	rs2768551	(46)	22132	9988	8187	1801

Region No.	Chr	Region size (Mb)	Reported SNP/s	Ref.	SNPs passing QC	Independent SNPs		
						All	MAF ≤1%	MAF >1%
31	6	3.09	rs113096699 ;rs7753012, rs148274477	(45,49)	26108	13421	11414	2007
32	7	3	rs10246303	(50)	29455	12870	10335	2535
33	7	3	rs72615157	(50)	21967	10840	8980	1860
34	7	3	rs12698403	(50)	29032	12930	10446	2484
35	9	3	rs7872188	(50)	31667	14814	12293	2521
36	9	3	rs16909859	(45)	24029	11921	9942	1979
37	9	3	rs803923	(49)	25230	12697	10534	2163
38	9	3.15	rs10858246, rs10870202	(49,50)	29658	14419	11642	2777
39	10	3	rs7090277	(46)	28443	13110	10655	2455
40	10	3	rs3847402	(50)	26400	11266	9171	2095
41	10	3	rs7095607	(50)	27410	13334	11298	2036
42	10	3	rs2637254	(46)	23004	11530	9534	1996
43	11	4.60	rs4237643, rs2863171	(47)	39457	20018	16699	3319
44	11	3	rs2509961	(50)	22498	10937	9171	1766
45	11	3	rs11234757	(50)	26859	13096	10857	2239
46	11	3	rs567508	(50)	25153	12250	9994	2256
47	12	3	rs2348418	(49)	24945	11579	9355	2224
48	12	3	rs11172113	(46)	19107	9433	7780	1653
49	12	3	rs1494502	(50)	21566	10925	9123	1802
50	12	3.70	rs113745635 , rs12820313	(46,50)	30990	14845	12085	2760
51	12	3.76	chr12:114743533, rs10850377, rs35506	(48-50)	31177	14709	12019	2690
52	14	3	rs1698268	(50)	25807	23186	20548	2638
53	14	3.63	rs7155279, rs117068593	(49)	30573	29303	26095	3208
54	15	3	rs72724130	(50)	23444	11992	10273	1719
55	15	3.16	rs10851839, rs12591467	(44,45,50)	23115	11764	9824	1940
56	15	3	rs66650179;rs79370947	(50)	18205	18247	16243	2004
57	16	3	rs12149828	(49)	32556	14698	12084	2614
58	16	3	rs12447804	(46)	25732	11897	9646	2251
59	16	5.72	rs3743609, rs1079572	(46,47)	68955	32943	27087	5856
60	17	3	rs59835752;rs80145911	(50)	18662	18445	16680	1765
61	17	3	rs11658500	(50)	19919	9969	8319	1650
62	17	3	rs35524223	(48)	17343	7581	6243	1338

Region No.	Chr	Region size (Mb)	Reported SNP/s	Ref.	SNPs passing QC	Independent SNPs		
						All	MAF ≤1%	MAF >1%
63	17	3	rs6501431	(47)	23388	11122	9151	1971
64	17	3	rs7218675	(48)	25936	11882	9757	2125
65	19	3	rs113473882	(49)	23666	11236	9300	1936
66	20	3	rs6140050	(50)	22275	8577	6562	2015
67	20	3	rs72448466	(50)	19037	8915	7193	1722
68	21	3	rs2834440	(46)	24146	11847	9744	2103
69	22	3	rs11704827	(50)	20375	8182	6372	1810
70	22	3.13	rs134041, rs2283847	(49,50)	25040	12484	10117	2367
Total					**1831014**	**916999**	**761972**	**155027**

4.7.3 Joint test results

After phenotype QC, relatedness removal, and SNP exclusions, the final joint test analysis consisted of 303,619 individuals and a total of 1,831,014 SNPs, across regions chosen for containing at least one of the 90 marginal effect signals for lung function.

Figure 4.15 and **Figure 4.16** present Manhattan plots for the joint test results across all of the aforementioned regions of the genome associated with lung function, for the traits FEV_1 and FEV_1/FVC respectively. The p-value for each of the 90 marginal effect sentinels are represented by a red diamond on each of the Manhattan plots. For the sentinels appearing non-statistically significant, this is due to the sentinel being reported for association with a different lung function trait than the one presented here (please see the following section on results for marginal effect signals).

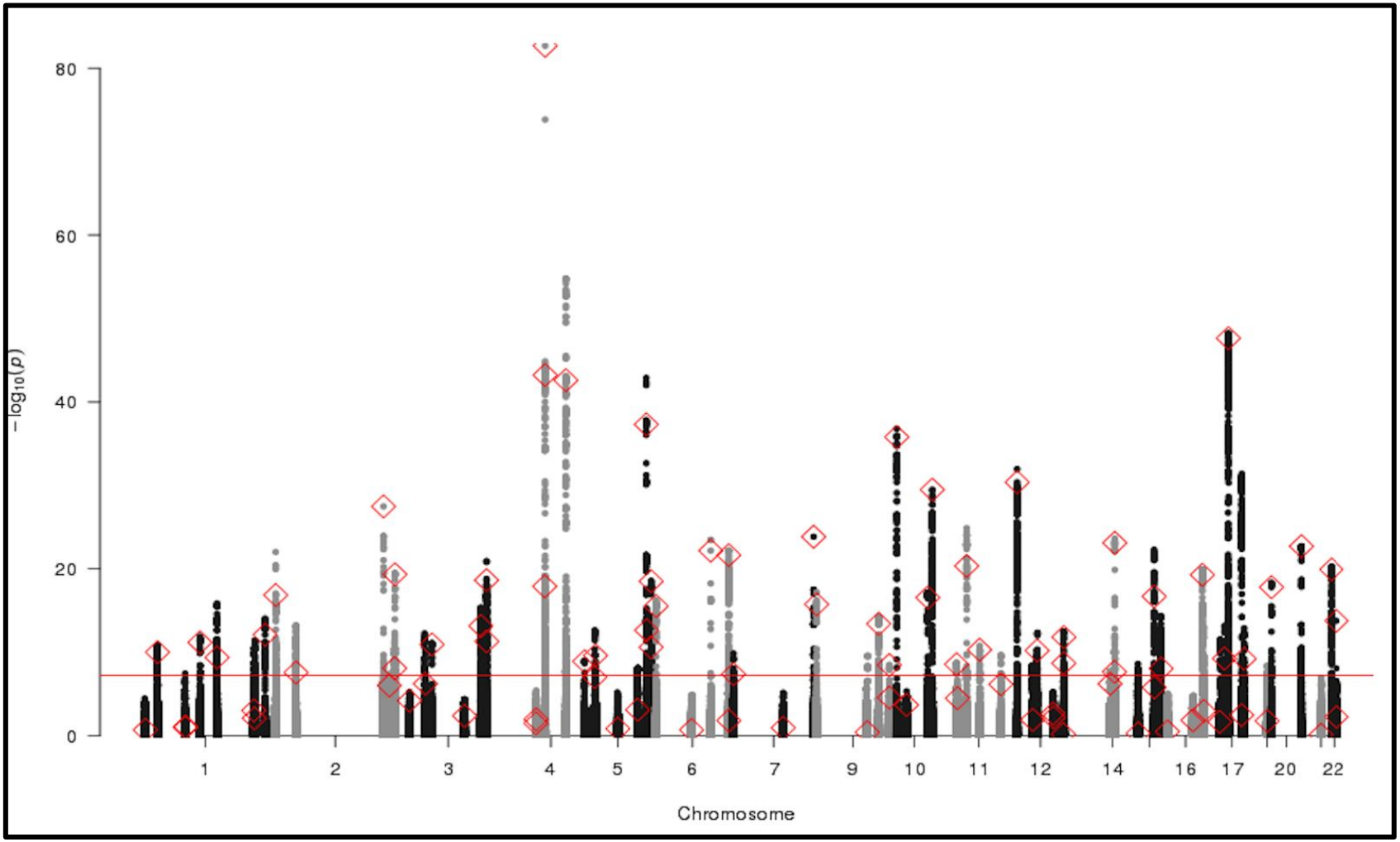


Figure 4.15 - Manhattan plot for the FEV_1 joint test results

Red horizontal line represents the threshold for statistical significance of 5.5×10^{-8} and the red diamonds represent the 90 marginal effect signals for each region.

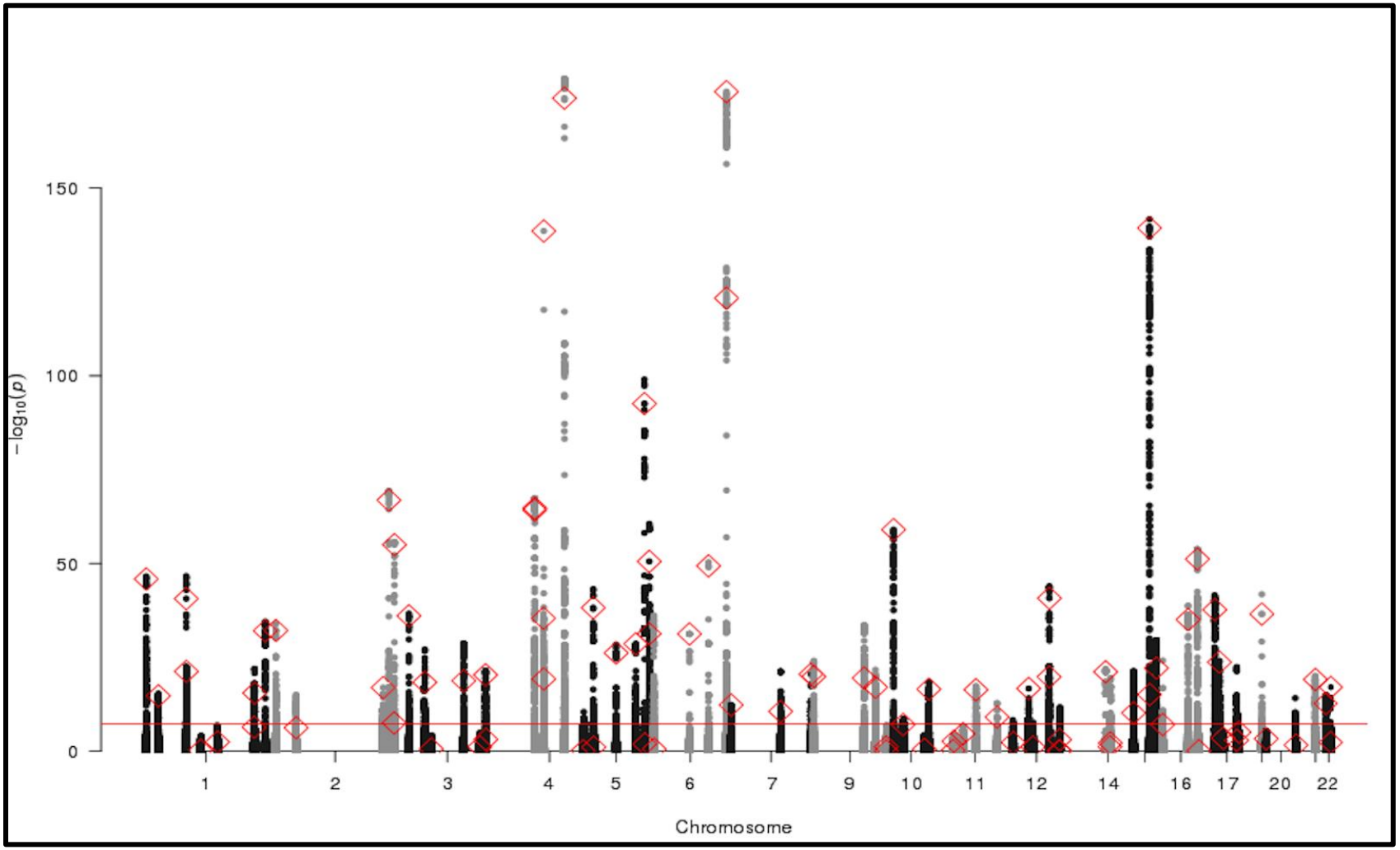


Figure 4.16 - Manhattan plot for the FEV_1/FVC joint test results

Red horizontal line represents the threshold for statistical significance of 5.5×10^{-8} and the red diamonds represent the 90 marginal effect signals for each region.

4.7.3.1 Joint test results for previously reported SNPs

Joint test results for the 90 previously reported marginal effect sentinels for lung function (which were used to define the regions for joint test interaction analysis in section 4.6.1) are presented in **Appendix C**, alongside the traits they were identified for. 73 of the 90 signals were previously identified for the phenotypes used in the joint test analysis, FEV_1 and FEV_1/FVC. For the 24 of these which were known FEV_1 signals, 21 reached a significance threshold of $p < 5 \times 10^{-8}$ for FEV_1 in the joint test analysis. For the remaining three signals, one signal reached $p < 5 \times 10^{-6}$ (but $p < 5 \times 10^{-8}$ for FEV_1/FVC), one signal reached $p < 5 \times 10^{-2}$, whilst one SNP, chr12_114743533, produced joint test $p > 0.5$ (for both traits). This is a very rare SNP (MAF ~ 0.008%) reported as part of the UK BiLEVE study specifically in ever-smokers only (48). For the 49 FEV_1/FVC signals, 46 reached a threshold of $p < 5 \times 10^{-8}$ for the FEV_1/FVC joint test analysis, with the remaining three signals reaching $p < 5 \times 10^{-7}$. Of the remaining 17 signals that were previously identified for FVC, 13 reached $p < 5 \times 10^{-8}$, two reached $p < 5 \times 10^{-5}$ and two reached $p < 5 \times 10^{-3}$ for at least one of the two phenotypes tested. 46 of the 90 signals were originally identified using UK Biobank data (14 for FEV_1, 24 for FEV_1/FVC and 8 for FVC, highlighted in **Appendix C**).

4.7.3.2 Selected signals from joint test analysis

Using the process outlined in 4.6.5 for signal selection (which included the known signals used to define the regions), and a significance threshold of 5.5×10^{-8} (Bonferroni corrected threshold for the number of tests equating to the number of independent SNPs across all defined regions), a preliminary total of 117 signals were found for FEV_1 (58 signals from 58 regions showing association) and FEV_1/FVC (59 signals from 59 regions showing association). Upon applying both the PLINK LD clumping procedure within each region containing statistically significant SNPs, and observing region plots for the identified clumps (for which **Figure 4.17** illustrates the process for the 4th region on chromosome 1), a further 153 and 362 signals were found for FEV_1 and FEV_1/FVC respectively. Thus, as a result of the joint test analysis undertaken in step 1, 632 independent joint test signals reached statistical significance at a threshold of $p < 5.5 \times 10^{-8}$, 211 for FEV_1 and 421 for FEV_1/FVC respectively **(Figure 4.18)**. The full results for the SNPs identified in the joint test analysis are in

Appendix D. Of the 632 SNPs identified, 20 were rare (MAF < 1%), 66 were low frequency (1% < MAF < 5%) and 546 were common (MAF > 5%). Imputation scores for the 632 SNPs ranged from 0.78 to 1.

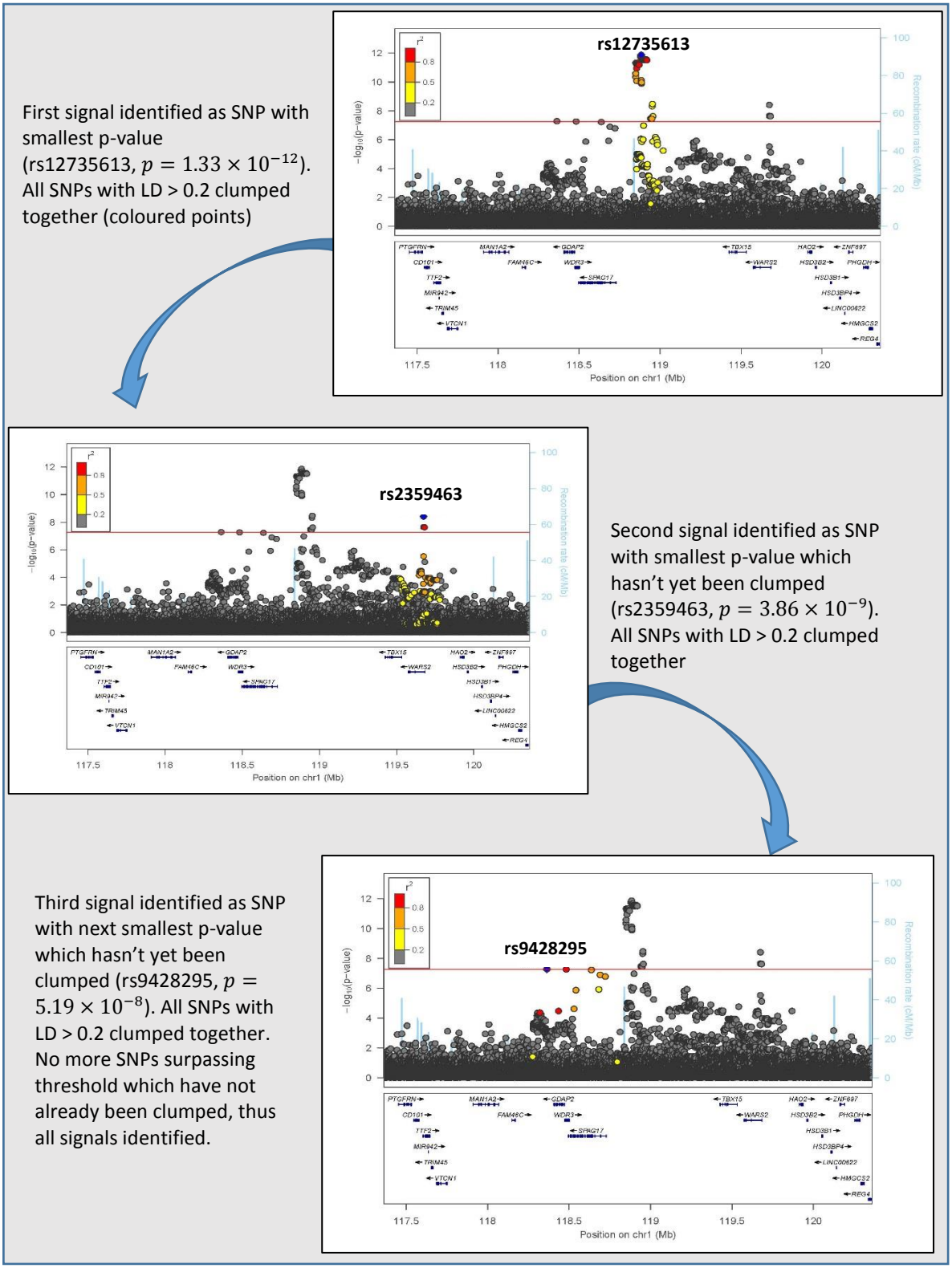


Figure 4.17 - Example of PLINK LD clumping process (for region 4 from Table 4.11) used to identify signals in each region

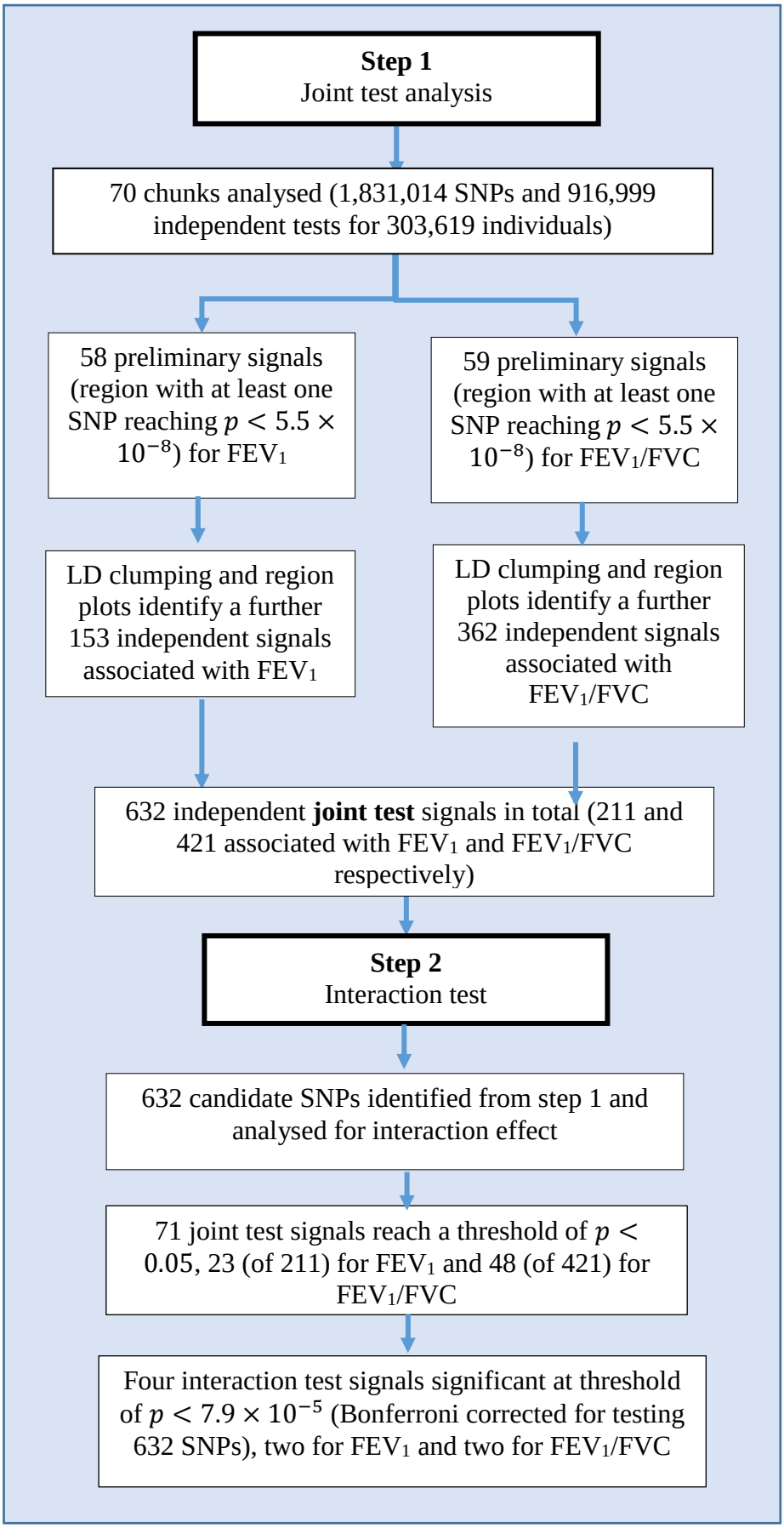


Figure 4.18 - Joint test analysis (step 1) and interaction test analysis (step 2) process for the 90 regions with previously known signals

4.7.4 Testing joint test candidate SNPs for interaction effect

The 632 screened candidate SNPs from step 1, chosen through use of the joint test of main and interaction effect, were taken forward for the interaction testing stage in step 2, to determine whether the signal was identified due to main effect or interaction effect. Of the 632 SNPs tested, 23 and 48 reached a suggestive statistical significance of $p < 0.05$ for FEV_1 and FEV_1/FVC respectively (**Figure 4.18**). The 71 SNPs which reached a threshold of $p < 0.05$ for the interaction test are presented in **Table 4.12** and **Table 4.13** for traits FEV_1 and FEV_1/FVC respectively, alongside their joint test p-values. The interaction test results for all 632 joint test screened SNPs are presented in the same table as the aforementioned joint test results in **Appendix D.** Of the 71 SNPs reaching this threshold, 63 were common (MAF > 5%), 5 were low frequency (1% < MAF < 5%) and 3 were rare (MAF < 1%).

Of the SNPs reaching $p < 0.05$, 4 SNPs additionally reached a Bonferroni corrected threshold (for 632 SNPs) of $p < 7.9 \times 10^{-5}$ for the test of interaction, two for FEV_1 and two for FEV_1/FVC. These are presented in **Table 4.14** and **Table 4.15**, alongside the marginal effect in ever and never smokers subgroups individually. The most significant SNP identified for FEV_1 was rs2561562 (MAF = 49%, Interaction beta = 0.0211, P = 3.98×10^{-5}), an intergenic SNP downstream of gene *LTBP4*. This SNP was identified from region 65 in **Table 4.11** with previously reported gene *LTBP4*. The second SNP identified for FEV_1 was rs3865527 (MAF = 43%, Interaction beta = 0.0206, P = 7.52×10^{-5}). This is an intronic SNP in *KCNQ2* identified from region 67 in **Table 4.11** with previously reported gene *ZGPAT*. For FEV_1/FVC the two SNPs showing statistical evidence of an interaction effect were rs9303283 (MAF = 42%, Interaction beta = 0.0213, P = 3.7×10^{-5}) and rs9618700 (MAF = 22%, Interaction beta = 0.0258, P = 3.7×10^{-5}) from regions 61 and 69 in **Table 4.11** respectively, with respective previously reported genes *CISD3* and *MICAL3*. Both are intronic SNPs, the first in *MED24* and the second in *GNB1L*. All four SNPs across both traits appeared to be identifiable using an interaction test due to a strong association in ever smokers ($p < 5 \times 10^{-8}$), and no association in never smokers ($p > 0.05$). All SNPs were not in LD with the respective known signals within their defined region ($r^2 < 0.01$).

The joint test region plots (to show the LD clumped independent signals) with the four SNPs highlighted as sentinels are given in **Figure 4.19** and **Figure 4.20** for the FEV_1 and FEV_1/FVC interaction signals respectively.

Table 4.12 - SNPs from the interaction analysis for FEV_1 for which the interaction test reached a threshold of $p < 0.05$.
CAF = Coded allele frequency, INFO = Imputation score, and MAF = Minor allele frequency. Betas (β) and standard errors (SE) to 4 d.p and are on the inverse normalised scale. Chromosome (Chr) and position (Pos) are build 37.

SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Nearest gene	Joint test P	Interaction test P	Interaction test β	Interaction test SE
rs66773125 (1:150174366)	A / G	0.134	0.134	0.99	*ANP32E*	2.749×10^{-12}	4.240×10^{-02}	-0.0153	0.0075
rs6680689 (1:221462038)	A / G	0.429	0.429	1.00	*DUSP10*	1.997×10^{-09}	3.972×10^{-02}	0.0107	0.0052
rs34811804 (2:18886374)	T / C	0.306	0.306	1.00	*NT5C1B*	1.813×10^{-08}	1.587×10^{-03}	0.0174	0.0055
rs1430190 (2:56031305)	G / C	0.270	0.270	0.99	*PNPT1*	5.558×10^{-14}	1.934×10^{-02}	-0.0138	0.0059
rs11125611 (2:56173907)	T / A	0.368	0.368	0.99	*RN7SKP208*	4.623×10^{-12}	4.348×10^{-02}	-0.0107	0.0053
rs9883125 (3:168748911)	A / T	0.372	0.372	1.00	*MECOM*	1.253×10^{-21}	2.539×10^{-02}	-0.0119	0.0053
rs11945032 (4:106897596)	A / G	0.357	0.357	0.99	*NPNT*	2.873×10^{-11}	2.287×10^{-02}	0.0123	0.0054
rs72660503 (4:107162069)	A / G	0.031	0.031	0.92	*TBCK*	3.676×10^{-10}	2.750×10^{-02}	0.0336	0.0153
rs4434191 (4:145024452)	T / G	0.271	0.271	0.94	*GYPB*	2.696×10^{-10}	1.954×10^{-02}	0.0138	0.0059
rs35466090 (5:131622836)	T / C	0.091	0.091	0.99	*P4HA2*	7.914×10^{-09}	4.328×10^{-04}	0.0317	0.0090
rs7733410 (5:147856522)	A / G	0.442	0.442	1.00	*HTR4*	1.328×10^{-43}	5.691×10^{-03}	0.0143	0.0052
rs9376680 (6:142553740)	T / C	0.261	0.261	1.00	*GPR126*	6.315×10^{-23}	2.639×10^{-02}	0.0131	0.0059
rs11594905 (10:77659733)	A / G	0.132	0.132	1.00	*C10orf11*	3.034×10^{-10}	1.665×10^{-03}	-0.0235	0.0075
rs61921171 (12:58662712)	T / C	0.212	0.212	1.00	*CTDSP2*	3.587×10^{-08}	7.740×10^{-04}	0.0211	0.0063
rs1511318 (14:84298899)	C / T	0.158	0.158	1.00	*OTX2-AS1*	1.889×10^{-08}	3.402×10^{-02}	-0.0152	0.0072
rs12910520 (15:41481420)	T / A	*0.619*	0.381	0.99	*EXD1*	2.400×10^{-09}	3.891×10^{-02}	-0.0110	0.0053
rs12932007 (16:75428556)	T / C	0.131	0.131	0.78	*CFDP1*	1.280×10^{-11}	1.352×10^{-02}	0.0213	0.0086
rs34724124 (17:44902516)	A / G	0.473	0.473	0.99	*WNT3*	5.824×10^{-10}	1.944×10^{-02}	-0.0121	0.0052
rs10410606 (19:41108975)	C / A	0.446	0.446	1.00	*LTBP4*	1.833×10^{-08}	4.398×10^{-03}	0.0147	0.0052
rs2561562 (19:41150086)	A / C	*0.509*	0.491	1.00	*NUMBL*	4.159×10^{-09}	3.980×10^{-05}	0.0211	0.0051
rs4809548 (20:62001148)	G / A	*0.066*	0.066	0.99	*CHRNA4*	1.519×10^{-10}	1.161×10^{-03}	-0.0336	0.0103
rs71325435 (20:62011379)	T / C	*0.271*	0.271	0.98	*CHRNA4*	4.195×10^{-09}	1.588×10^{-03}	0.0184	0.0058
rs3865527 (20:62047925)	C / T	*0.567*	0.433	0.98	*KCNQ2*	3.638×10^{-08}	7.523×10^{-05}	0.0206	0.0052

Table 4.13 - SNPs from the interaction analysis for FEV_1 /FVC for which the interaction test reached a threshold of $p < 0.05$.
CAF = Coded allele frequency, INFO = Imputation score, and MAF = Minor allele frequency. Betas (β) and standard errors (SE) to 4 d.p and are on the inverse normalised scale. Chromosome (Chr) and position (Pos) are build 37.

SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Nearest gene	Joint test P	Interaction test P	Interaction test β	Interaction test SE
rs12082710 (1:92155337)	C / T	0.377	0.377	0.99	*TGFBR3*	5.603×10^{-10}	1.471×10^{-02}	-0.0129	0.0053
rs2799097 (1:218524632)	G / A	0.849	0.151	0.99	*TGFB2*	1.307×10^{-17}	2.570×10^{-02}	-0.0160	0.0072
rs62817 (1:218746863)	C / G	0.435	0.435	0.99	*SLC30A10*	6.264×10^{-13}	1.983×10^{-02}	-0.0120	0.0052
rs79274749 (1:219500754)	T / C	0.075	0.075	0.99	*SLC30A10*	8.754×10^{-19}	3.053×10^{-02}	-0.0209	0.0097
rs8179795 (2:217613285)	T / C	0.859	0.151	0.99	*AC007563.5*	1.019×10^{-11}	4.898×10^{-02}	0.0142	0.0072
rs1286772 (3:25580776)	G / C	0.600	0.400	0.97	*RARB*	1.117×10^{-10}	9.568×10^{-03}	0.0137	0.0053
rs358079 (3:55110307)	G / T	0.314	0.314	0.98	*CACNA2D3*	1.560×10^{-08}	4.759×10^{-02}	0.0110	0.0055
rs9865871 (3:55120948)	T / C	0.554	0.446	0.99	*CACNA2D3*	3.087×10^{-13}	2.916×10^{-02}	-0.0113	0.0052
rs710834 (4:89618837)	C / T	0.379	0.379	1.00	*HERC3*	9.662×10^{-11}	4.039×10^{-02}	0.0109	0.0053
rs75106620 (4:89807613)	T / C	0.046	0.046	0.98	*FAM13A*	1.136×10^{-22}	2.531×10^{-03}	0.0376	0.0125
rs2609279 (4:89855495)	C / T	0.781	0.219	1.00	*FAM13A*	3.984×10^{-68}	1.677×10^{-04}	-0.0236	0.0063
rs185574798 (4:89863626)	A / G	0.005	0.005	0.82	*FAM13A*	2.026×10^{-09}	3.957×10^{-03}	0.1102	0.0383
rs2869966 (4:89869078)	T / C	0.407	0.407	1.00	*FAM13A*	4.831×10^{-65}	2.181×10^{-02}	-0.0119	0.0052
rs2904262 (4:89917060)	T / C	0.475	0.475	1.00	*FAM13A*	1.619×10^{-24}	2.337×10^{-02}	0.0116	0.0051
rs9884482 (4:106081636)	C / T	0.368	0.368	1.00	*TET2*	2.412×10^{-15}	1.628×10^{-02}	0.0127	0.0053
rs7669987 (4:145161554)	T / G	0.086	0.086	0.98	*ANAPC10*	7.339×10^{-17}	2.172×10^{-02}	-0.0208	0.0091
rs189268532 (4:145375696)	T / C	0.009	0.009	0.96	*HHIP*	2.058×10^{-08}	1.934×10^{-02}	-0.0646	0.0276
rs115004137 (4:145434756)	A / C	0.037	0.037	1.00	*RP11-361D14.2*	3.595×10^{-14}	4.854×10^{-02}	0.0264	0.0134
rs11724319 (4:145511040)	G / A	0.228	0.228	0.97	*HHIP*	2.834×10^{-74}	2.814×10^{-03}	-0.0182	0.0061
rs2353397 (4:145517578)	T / C	0.554	0.445	1.00	*HHIP*	1.205×10^{-27}	2.996×10^{-03}	0.0152	0.0051
rs12189242 (5:34627666)	A / C	0.242	0.242	1.00	*C1QTNF3*	3.643×10^{-11}	2.682×10^{-02}	-0.0131	0.0059
rs986494 (5:147449067)	G / C	0.675	0.325	0.98	*SPINK5*	7.314×10^{-11}	2.742×10^{-02}	0.0121	0.0055
rs7733410 (5:147856522)	A / G	0.442	0.442	1.00	*HTR4*	9.650×10^{-100}	1.700×10^{-02}	0.0123	0.0051

SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Nearest gene	Joint test P	Interaction test P	Interaction test β	Interaction test SE
rs2421471 (5:156631551)	G / A	0.409	0.409	0.99	*ITK*	4.065×10^{-09}	1.797×10^{-03}	0.0163	0.0052
rs10076465 (5:156801180)	C / T	0.419	0.419	0.99	*CYFIP2*	9.517×10^{-16}	1.257×10^{-02}	0.0130	0.0052
rs13361953 (5:156926442)	C / T	0.344	0.344	1.00	*ADAM19*	2.761×10^{-61}	2.051×10^{-02}	0.0125	0.0054
rs943613 (6:6832811)	T / C	0.445	0.445	0.99	*LY86-AS1*	3.467×10^{-09}	3.797×10^{-02}	-0.0107	0.0052
rs17672837 (6:7257023)	C / G	0.038	0.038	1.00	*RREB1*	3.809×10^{-08}	2.590×10^{-02}	0.0293	0.0132
rs35305377 (7:99938955)	A / G	0.549	0.451	0.99	*STAG3L5P*	2.021×10^{-16}	1.916×10^{-02}	0.0121	0.0052
rs6602570 (10:12320571)	C / A	0.812	0.188	0.99	*CDC123*	1.100×10^{-26}	4.416×10^{-02}	-0.0132	0.0066
rs4749652 (10:31270278)	T / C	0.523	0.477	1.00	*ZNF438*	2.781×10^{-09}	1.678×10^{-02}	-0.0122	0.0051
rs11594905 (10:77659733)	A / G	0.132	0.132	1.00	*C10orf11*	3.018×10^{-09}	1.112×10^{-02}	-0.0189	0.0074
rs2579762 (10:78318879)	C / A	0.464	0.464	1.00	*C10orf11*	5.773×10^{-19}	2.278×10^{-02}	0.0117	0.0051
rs186767801 (12:57727682)	A / G	0.003	0.003	0.89	*R3HDM2*	2.953×10^{-11}	2.414×10^{-02}	0.1034	0.0459
rs4900195 (14:94245652)	T / C	0.488	0.488	1.00	*PRIMA1*	2.619×10^{-10}	2.587×10^{-02}	-0.0114	0.0051
rs80019083 (15:71572350)	T / C	0.045	0.045	1.00	*THSD4*	3.606×10^{-12}	1.548×10^{-02}	-0.0292	0.0121
rs4238437 (15:71741779)	C / T	0.764	0.236	0.95	*THSD4*	1.457×10^{-12}	4.032×10^{-02}	0.0127	0.0062
rs28650139 (15:71812163)	C / T	0.599	0.401	1.00	*THSD4*	4.207×10^{-08}	4.604×10^{-02}	0.0104	0.0052
rs34660045 (15:71816660)	T / C	0.130	0.130	0.99	*THSD4*	2.847×10^{-10}	2.465×10^{-02}	0.0172	0.0076
rs11648508 (16:58063513)	T / G	0.681	0.319	0.99	*MMP15*	1.750×10^{-39}	4.786×10^{-02}	0.0109	0.0055
rs74470468 (16:58132535)	A / G	0.059	0.059	1.00	*C16orf80*	1.713×10^{-11}	3.738×10^{-02}	-0.0226	0.0109
rs8072345 (17:28604289)	T / C	0.445	0.445	1.00	*BLMH*	6.528×10^{-15}	4.995×10^{-02}	-0.0101	0.0052
rs216450 (17:28878817)	G / A	0.581	0.419	1.00	*RP11-271K11.1*	4.537×10^{-09}	1.582×10^{-02}	0.0125	0.0052
rs9303283 (17:38192633)	T / C	0.575	0.425	0.99	*MED24*	1.181×10^{-09}	3.705×10^{-05}	0.0213	0.0052
rs8067763 (17:70012939)	A / G	0.601	0.399	0.99	*AK094963*	2.360×10^{-08}	2.142×10^{-02}	-0.0121	0.0052
rs9913936 (17:70074148)	C / T	0.670	0.330	0.99	*SOX9-AS1*	8.344×10^{-11}	1.098×10^{-02}	-0.0139	0.0055
rs2834440 (21:35690499)	A / G	0.613	0.387	1.00	*AP000320.7*	8.052×10^{-20}	1.124×10^{-02}	-0.0133	0.0053
rs9618700 (22:19803382)	A / C	0.224	0.224	0.98	*GNB1L*	1.716×10^{-12}	3.703×10^{-05}	0.0258	0.0062

Table 4.14 - SNPs which were statistically significant ($p < 7.9 \times 10^{-5}$) for the interaction test in step 2 for FEV_1.
CAF = Coded allele frequency, INFO = Imputation score, and MAF = Minor allele frequency. Betas (β) and standard errors (SE) to 4 d.p and are on the inverse normalised scale. Chromosome (Chr) and position (Pos) are build 37.

										Ever Smokers			Never Smokers		
SNP (CHR:POS)	**Coded / Non coded allele**	**CAF**	**MAF**	**INFO**	**Nearby gene**	**Joint test P**	**Interaction test P**	**Interaction test β**	**Interaction test SE**	**β**	**SE**	**P**	**β**	**SE**	**P**
rs2561562 (19:41150086)	A / C	0.51	0.49	1.00	*LTBP4*	4.16×10^{-09}	3.98×10^{-05}	0.0211	0.0051	0.0226	0.0038	2.367×10^{-09}	0.0019	0.0035	0.593
rs3865527 (20:62047925)	C / T	0.57	0.43	0.98	*KCNQ2*	3.64×10^{-08}	7.52×10^{-05}	0.0206	0.0052	0.0213	0.0039	3.075×10^{-08}	0.0023	0.0035	0.521

Table 4.15 - SNPs which were statistically significant ($p < 7.9 \times 10^{-5}$) for the interaction test in step 2 for FEV_1/FVC.
CAF = Coded allele frequency, INFO = Imputation score, and MAF = Minor allele frequency. Betas (β) and standard errors (SE) to 4 d.p and are on the inverse normalised scale. Chromosome (Chr) and position (Pos) are build 37.

										Ever Smokers			Never Smokers		
SNP (CHR:POS)	**Coded / Non coded allele**	**CAF**	**MAF**	**INFO**	**Nearby gene**	**Joint test P**	**Interaction test P**	**Interaction test β**	**Interaction test SE**	**β**	**SE**	**P**	**β**	**SE**	**P**
rs9303283 (17:38192633)	C / T	0.58	0.42	0.99	*MED24*	1.18×10^{-08}	3.70×10^{-05}	0.0213	0.0052	0.0232	0.0038	1.447×10^{-09}	0.0041	0.0035	0.244
rs9618700 (22:19803382)	C / A	0.22	0.22	0.98	*GNB1L*	1.72×10^{-12}	3.70×10^{-05}	0.0258	0.0062	0.0314	0.0046	1.035×10^{-11}	0.0085	0.0043	0.046

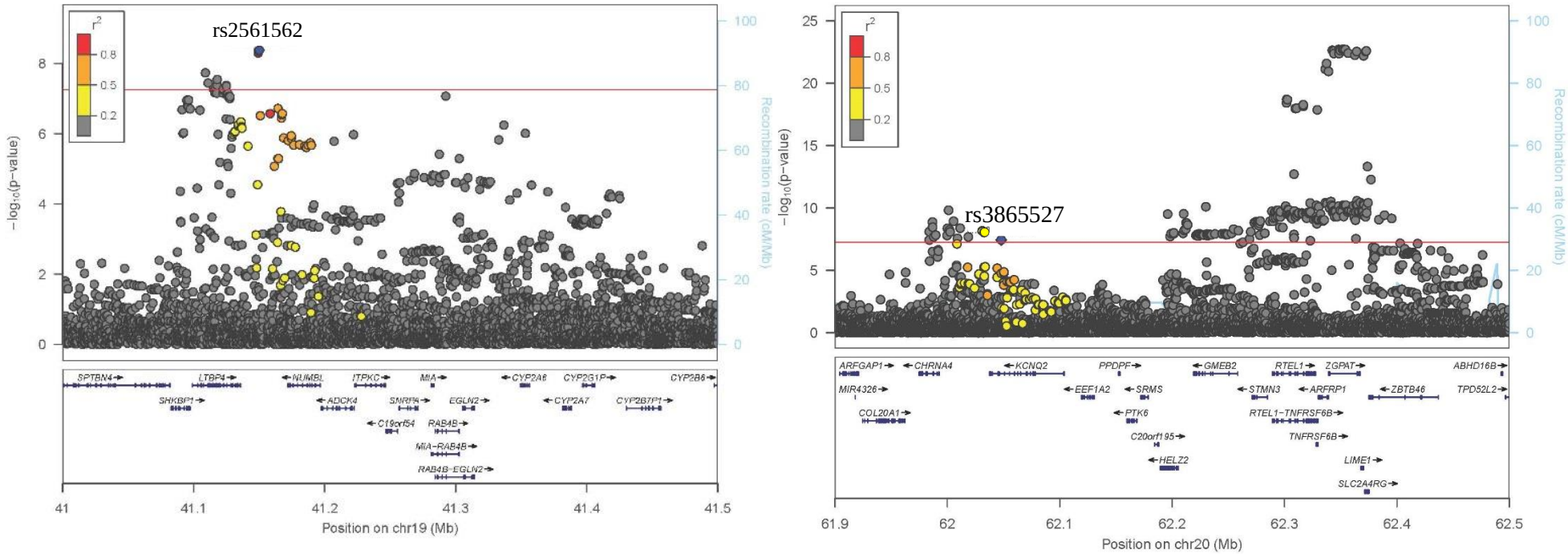


Figure 4.19 - Region plots for joint test results for FEV_1 signals rs2561562 and rs3865527

Note: rs3865527 is not the most significant SNP in the LD block due to differences in LD calculation by PLINK clumping and the LD calculation for the region plot. More specifically, the SNP in yellow that has a smaller p-value than the sentinel was actually assigned to a different clump in the clumping process, but given a borderline LD of ~0.2 with the sentinel it appears in LD in the region plot.

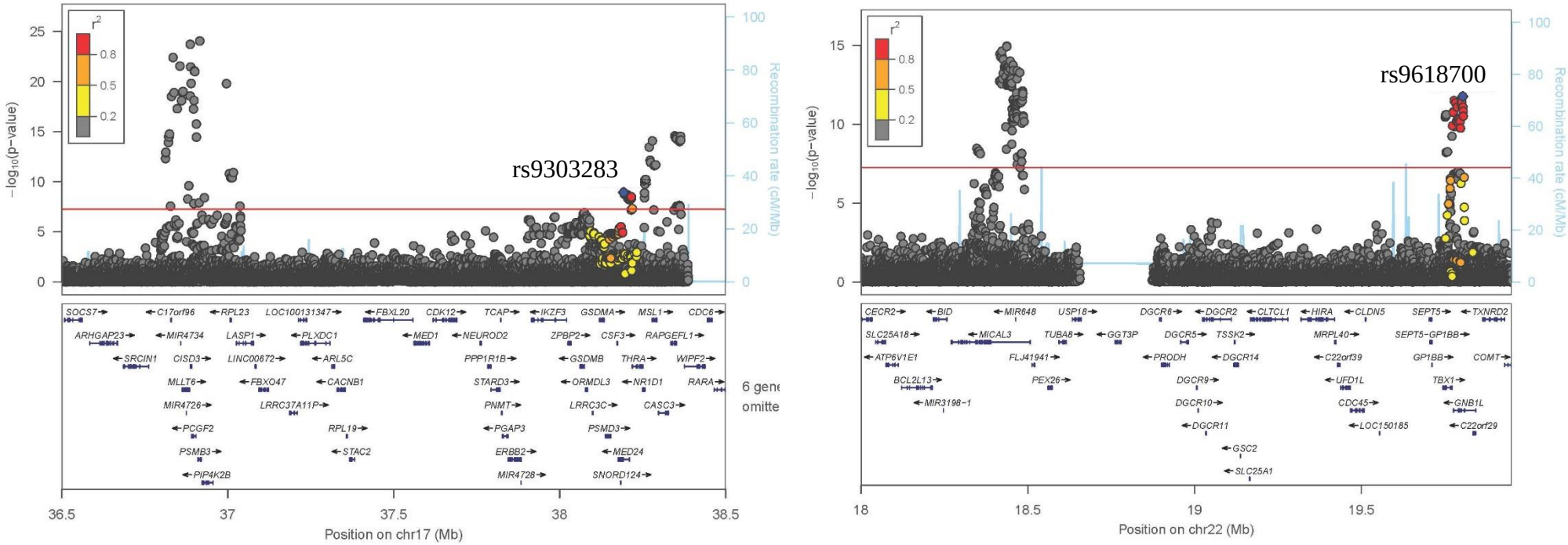


Figure 4.20 - Region plots for joint test results for FEV_1/FVC signals rs9303283 and rs9618700

4.8 Follow up of interaction signals

4.8.1 SpiroMeta consortium follow-up

For the four SNPs with a statistically significant interaction in section 4.7.4, there was no evidence of interaction effect in the SpiroMeta consortium meta-analysis (**Table 4.16**), using a Bonferroni corrected threshold of 0.0125 (corrected for 4 tests). SNPs rs3865527 ($p = 0.058$) and rs9303283 ($p = 0.057$) had p-values close to a nominal significance of 0.05, but were not statistically significant. Direction of effects were consistent for all four SNPs between UK Biobank and SpiroMeta for ever smokers (protective), but were only consistent for never smokers for SNPs rs2561562 and rs9618700, with never smoker effects for the remaining two SNPs in opposite direction between UK Biobank and SpiroMeta.

4.8.2 Exploring interaction signals for association with smoking

The four interaction signals were explored for association with smoking phenotypes smoking initiation (SI), smoking cessation (SC) and cigarettes per day (CPD). None of the signals showed a statistically significant genetic effect for any of the smoking traits (**Table 4.17**), using a Bonferroni corrected threshold for four SNPs to conclude statistical significance i.e. $p < 0.0125$. This suggests that the ever smoker effect for the four interaction signals was not driven by smoking behaviour (even though the marginal effect was extremely statistically significant in ever smokers). For CPD both rs2561562 ($p = 0.023$) and rs3865527 ($p = 0.028$) reached a threshold of $p < 0.05$, with SNP rs9618700 meeting the same threshold for both SI and SC (approaching Bonferroni corrected threshold for SI with $p = 0.013$).

Table 4.16 - Replication in SpiroMeta consortium meta-analysis for the four SNPs which showed statistical evidence of an interaction effect
CAF = coded allele frequency, MAF = minor allele frequency, TS = test statistic, DF = degrees of freedom and N_{eff} = effective sample size (considers imputation quality). Betas (β) and standard errors (SE) to 4 d.p and are on the inverse normalised scale.

			Ever Smoker						Never Smoker						Welch's t-test		
SNP	**Trait**	**Coded / Non coded allele**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**TS**	**DF**	**P**
rs2561562	FEV_1	A / C	0.51	0.49	36284	0.0043	0.0076	0.568	0.53	0.47	32748	0.0030	0.0079	0.704	0.12	4.26	0.910
rs3865527	FEV_1	C / T	0.57	0.43	30225	0.0217	0.0083	0.009	0.57	0.43	28023	-0.0094	0.0086	0.276	2.60	4.50	0.058
rs9303283	FEV_1/ FVC	T / C	0.57	0.43	36940	0.0127	0.0075	0.090	0.57	0.43	33263	-0.0158	0.0080	0.046	2.61	4.20	0.057
rs9618700	FEV_1/ FVC	A / C	0.22	0.22	37023	0.0196	0.0090	0.029	0.22	0.22	33344	0.0177	0.0094	0.060	0.15	5.93	0.886

Table 4.17 - UK Biobank smoking marginal effect analysis look up for the four SNPs which showed statistical evidence of an interaction effect
CAF = coded allele frequency, MAF = minor allele frequency, INFO = imputation score. Betas (β) and standard errors (SE) for all traits are to 4 d.p. and are on the inverse normalised scale

SNP	**Coded / Non coded allele**	**INFO**	**Smoking trait**	**CAF**	**MAF**	**β**	**SE**	**P**
rs2561562	A / C	1	CPD	0.504	0.496	-0.0114	0.0050	2.30×10^{-2}
			SI	0.503	0.497	-0.0019	0.0027	4.80×10^{-1}
			SC	0.503	0.497	0.0024	0.0040	5.40×10^{-1}
rs3865527	C / T	0.98	CPD	0.549	0.451	-0.0114	0.0050	2.80×10^{-2}
			SI	0.552	0.448	-0.0001	0.0027	9.60×10^{-1}
			SC	0.552	0.448	-0.0014	0.0040	7.30×10^{-1}
rs9303283	T / C	0.99	CPD	0.566	0.434	-0.0114	0.0050	7.70×10^{-1}
			SI	0.568	0.432	-0.0028	0.0027	3.10×10^{-1}
			SC	0.567	0.433	-0.0003	0.0040	9.30×10^{-1}
rs9618700	A / C	0.98	CPD	0.219	0.219	-0.0114	0.0050	6.90×10^{-2}
			SI	0.219	0.219	0.0081	0.0033	1.30×10^{-2}
			SC	0.220	0.220	-0.0108	0.0048	2.50×10^{-2}

4.9 Sensitivity analysis for identified interaction signals

When regressing phenotypes on potential confounding covariates for analysis, as outlined in section 4.6.4, adjustment for array was not included (~50k individuals genotyped using the UK BiLEVE array and the remaining individuals using the UK Biobank array). Furthermore, there has been suggestion that adjustment of principal components prior to inverse normalising residuals could potentially re-introduce confounding during the analysis stage (173). To account for this a sensitivity analysis was undertaken (for the second stage interaction test) for those interaction signals found in the previous section, using adjustment for array, and adjustment for principal components post inverse normalisation of phenotype residuals (during the analysis stage rather than during the inverse normalisation of phenotypes). This sensitivity analysis made little difference for two of the four interaction effect signals (**Table 4.18**), however two signals (rs3865527 for FEV_1 and rs9303283 for FEV_1/FVC) no longer reached the threshold for statistical significance ($p < 7.9 \times 10^{-5}$). The effect was primarily dependent on adjustment for array, with all four interaction signals still reaching statistical significance when the only revision to the analysis was to adjust for principal components *after* rather than *before* inverse normalising traits.

Table 4.18 - Effect of sensitivity analysis on p-values for the four interaction effect signals

SNP (Chr:Pos)	Trait	Original P Value	Adjustment for PCs post trait normalisation P value	Adjusting for Array in addition to PCs P value
rs2561562 (19:41150086)	FEV_1	3.98×10^{-5}	4.06×10^{-5}	5.65×10^{-5}
rs3865527 (20:62047925)	FEV_1	7.52×10^{-5}	7.59×10^{-5}	2.06×10^{-4}
rs9303283 (17:38192633)	FEV_1/FVC	3.7×10^{-5}	3.52×10^{-5}	2.28×10^{-4}
rs9618700 (22:19803382)	FEV_1/FVC	3.7×10^{-5}	4.10×10^{-5}	1.37×10^{-5}

4.10 Discussion

This chapter presents a candidate region gene-smoking interaction analysis. The aim was to identify SNPs whose effect on lung function was dependent on ever/never smoking status (thus producing a gene-smoking interaction effect). The primary rationale behind the analysis process was that we may be more likely to observe interaction effects in regions which already harbour lung function associated signals, thus these areas of the genome were targeted first in a "candidate region" approach. In addition, it also allowed for application of a potentially less stringent threshold to conclude association than one applied genome-wide, although this benefit was not as evident here due to the inclusion of very rare SNPs, which may have been excluded for the analyses of smaller sample sizes and marginal genetic effects, but included here due to the sample size available. Due to the large contribution of such SNPs, had they been removed, the threshold would have been much less stringent (for example $\frac{0.05}{155027} = 3.2 \times 10^{-7}$ when only considering SNPs > 1% MAF rather than $\frac{0.05}{916999} = 5.5 \times 10^{-8}$ as applied here for all SNPs). Furthermore, with the computational complexity which accompanies the application of the joint test (which was used to screen SNPs) to the UK Biobank data (for which software is very limited), it was more computationally and time efficient to prioritise a discrete number of regions rather than undertake a genome-wide joint test analysis.

Across the 70 defined regions which encompassed the 90 previously reported signals for lung function traits, 632 independent signals were statistically significant when using the joint test, 211 for FEV_1 and 421 for FEV_1/FVC. Of these 632 independent joint test signals, four produced a statistically significant interaction (SNPxsmoking) effect (all of which were common with MAF > 5%), with an interaction effect detected due to a clear marginal genetic effect in ever smokers only. For FEV_1, an interaction signal was identified downstream of *LTBP4*, whilst a second interaction signal was identified in *KCNQ2,* a gene involved in the functioning of potassium channels in the brain with a well-documented link to epilepsy (174). For FEV_1/FVC, one interaction signal was in *MED24 (*Mediator Of RNA polymerase II transcription subunit 24) which plays a role in the mediator complex, believed to be required in regulating expression for most genes (175), whilst the second signal was in *GNB1L* (Guanine nucleotide-

binding protein subunit beta-like protein 1) which has been shown to have a role in disorders such as autism and schizophrenia (176). All four signals were independent of the known lung function signals within their respected defined region ($r^2 < 0.01$). There was however no statistical evidence (based on a Bonferroni threshold of $p < 0.0125$) of replication in the SpiroMeta consortium data, although the signals in *KCNQ2* and *MED24* did approach a nominal statistical significance level of 0.05.

To address the possible situation that interaction effect could be driven by an association between SNP and smoking behaviour, each of the four SNPs were explored for association with smoking initiation, smoking cessation and cigarettes per day, using the results from a UK Biobank smoking GWAS. There was no statistical evidence for any of the four signals (using a Bonferroni corrected threshold for 4 SNPs of $p < 0.0125$) that smoking was an intermediate variable on the causal pathway leading to their identification, even in light of a strong effect in ever smokers only for all SNPs. However, for SNP rs9618700, the p-value for association with smoking initiation did approach Bonferroni corrected threshold, with $p \approx 0.05$ for smoking cessation and cigarettes per day and similarly for cigarettes per day with rs2561562 and rs3865527. This therefore might suggest the need for further exploration of the association of these signals with smoking in other independent datasets.

A sensitivity analysis was undertaken for two reasons. Firstly, there is a documented dependency of regression results for inverse normalised traits on whether principal components are adjusted for before, or after, the inverse normalising process (173). Secondly, with two different genotyping arrays used to genotype individuals, the UK BiLEVE array (n=44,460) and the UK Biobank array (n=259,159) it was important to account for an array effect. The sensitivity analysis which adjusted for both genotyping array and ten principal components post inverse normalisation of traits, resulted in two of the four interaction signals dropping below the threshold for replication ($p < 7.52 \times 10^{-5}$), namely rs3865527 in *KCNQ2* for FEV_1 (from $p = 7.52 \times 10^{-5}$ to $p = 2.06 \times 10^{-4}$) and rs9303283 in *MED24 (*from $p = 3.7 \times 10^{-5}$ to $p = 2.28 \times 10^{-4}$). The remaining two signals remained statistically significant. With the effect of adjusting for covariates before or after normalising the results deemed much less extreme for minimally skewed continuous phenotypic data (173) and with FEV_1 and FEV_1/FVC traits having little

skew (absolute values of 0.33 and 0.38 respectively), the effect was in fact only dependent on adjustment for array (all interaction signals remained statistically significant when PCs were adjusted for after inverse normalisation of traits, and array was not adjusted for at all). Time permitting, re-analysis of all stages of the interaction analysis would be beneficial and such steps are thus implemented in chapter 6 when analysing the data on a genome-wide scale.

It is of interest to note here that for the 71 signals that had suggestive statistical evidence of an interaction effect ($p < 0.05$), approximately 90% of them were of common MAF, with only eight signals of low frequency or rare MAF. This may suggest one of two things. Firstly, it may be the case that interaction effects are more likely to be contributed by SNPs with large MAFs, rather than SNPs that are low frequency or rare. Alternatively, a second hypothesis might be that even with sample sizes as large as that offered by UK Biobank, power is still too low to pick up interaction effects for very low frequency SNPs.

A limitation of this analysis is the applied significance threshold used to determine statistically significant joint test signals, which is arguably very conservative. The threshold was determined by finding the number of independent SNPs tested and adjusting accordingly. As a result of considering all MAFs for analysis, the defined significance threshold for this analysis (5.5×10^{-8}) was heavily penalised by the number of rare SNPs considered (with 83% of the 919,999 independent SNPs having MAF < 1%). Therefore, it is possible that by over adjusting to account for rare SNPs, detecting joint test effects for higher frequency SNPs (MAF > 1%) will have been penalised. The alternative strategy could have been to relax the threshold further to increase the possibility of identifying true interaction effects, based on an a priori belief that regions containing known marginal effect signals are more likely to produce interaction effects. This could have been chosen arbitrarily, or with the use of other methods to determine the true number of independent tests undertaken in a more intuitive manor, than using PLINK LD pruning with LD < 0.2 chosen to determine independent SNPs. M_{eff}-based methods are one such example, where M_{eff} (the effective number of independent tests) is estimated with the use of correlation matrix eigenvalues (177). Ultimately, the perfect multiple testing correction approach would be one which provides maximum power to detect effects for rarer SNPs (which was of interest here

due to the substantial sample size for analysis), without identifying false-positive associations, whilst additionally not heavily penalising the identification of true higher frequency SNP interaction effects. Often in GWA studies this problem of rare SNP penalisation would be avoided with the use of a MAF filter to remove low frequency variation (163). Another approach that could be considered is the application of individual thresholds for common and rare SNPs, which would also address the issue of penalising common SNPs.

Additionally, a further limitation of the analysis undertaken is the dependency between both step one and step two of the analysis design. It is required that screening and analysis steps be independent if one wishes to only apply multiple testing adjustment to the testing step (35,71). With the joint test and the interaction test not independent, a multiple testing correction for both screening and analysis steps was required, which will have consequently effected power.

With respect to replication of identified interaction signals using data from the SpiroMeta consortium, it is important to be aware of the contrasts between discovery and replication sample sizes. It may be that the relatively small sample size contributed by the SpiroMeta consortium (approximately 20% of that in UK Biobank) which in turn will result in significantly less power to detect genetic effect, may have been the reason that the interaction signals did not replicate. Direction of effects were only consistent for two of the four SNPs in both smoking groups across UK Biobank and the SpiroMeta consortium. Thus replication should be sought from independent data with larger sample sizes in the future, if attainable.

With regards to the selection of signals post joint test analysis, it must be pointed out that the use of LD clumping has dependency on the arbitrary choice of LD cut off. Here, a value of 0.2 was chosen to clump SNPs in LD and thus identify independent signals. However, using a more stringent or relaxed cut off would have resulted in a decrease or increase of identified independent signals respectively. It may therefore be the case that the number of true independent signals were over-estimated, and signals that should have been consolidated together were not. An alternative signal selection process could have been applied, in which after choosing the first sentinel SNP in a region (the most statistically significant SNP), the corresponding SNP is then added to the model as a

covariate and the region is re-analysed for statistically significant results. This iterative process, referred to as conditional analysis (re-analysing SNPs conditional on the sentinel) is applied until all statistically significant SNPs within the region have been accounted for. This would remove the issue of selecting an arbitrary LD cut off when using a clumping procedure, but can be more computationally difficult and time-consuming when considering a large number of regions, with multiple potentially independent signals. This method of signal selection is applied in the proceeding chapter.

4.11 Conclusion

In conclusion, the analysis here uses the rationale that we may be more likely to observe interaction effects when prioritising analysis to regions which contain at least one marginal genetic effect association with lung function, in a "candidate region" approach. Using an appropriately defined threshold for statistical significance, 2 SNPs were identified for interaction with ever/never smoking (one for FEV_1 and one for FEV_1/FVC), that were not associated with smoking behaviour and had consistent direction of effect in an independent dataset (the SpiroMeta consortium).

The next chapter presents an interaction analysis focusing on an updated set of lung function associated SNPs, which reflects the progress made in understanding the role of marginal effects in lung function, since the work undertaken in this chapter.

Chapter 5 Gene-smoking interaction analysis for lung function associated SNPs in UK Biobank

5.1 Introduction

Since the work undertaken in chapter 4, the number of signals associated with lung function and COPD has more than doubled (see chapter 1 for the genetic determinants of lung function identified through GWAS). In the most recent GWAS of lung function utilising both UK Biobank and the SpiroMeta consortium, undertaken at the University of Leicester (32,33) (introduced in chapter 1 section 1.2.3.1), 279 signals were reported for association with lung function and COPD. Of these 279 signals, 139 were novel and 140 were previously reported signals which showed evidence of association in the study. This is presented in more detail in section 5.2. None of these 279 signals showed any association with smoking, the largest risk factor for poor lung function and COPD.

This chapter explores the interaction effect between SNP and ever/never smoking for the 279 identified signals associated with lung function and COPD to date, through means of a marginal genetic effect. The aim is to determine whether any of these identified marginal effect SNPs interact with smoking behaviour. This approach is reminiscent of that suggested by Chirag J. Patel (97) and discussed in Chapter 2, in that SNPs previously identified in the literature for a particular phenotype could produce a subset of SNPs to focus on for interaction analysis.

In addition, this chapter explores whether there is evidence that SNPs simultaneously interact with smoking behaviour to affect lung function, rather than producing individual interaction effects. Specifically questioning whether the combined effect of these 279 signals identified for marginal genetic effect interact with smoking behaviour, with the total contribution of all signals represented for each individual by a genetic risk score (GRS). The concept of genetic risk score by smoking interaction (GRSxS) and its effect on lung function has been explored before in a study by Aschard et al. (178). The authors considered 26 SNPs identified for marginal effect association with FEV_1 and FEV_1/FVC, finding a statistically significant interaction effect (after Bonferroni correction) between the GRS and ever/never smoking on FEV_1/FVC ($P = 0.00057$). Although replication in independent data was unsuccessful, the study highlighted the

potential for missing interactions when considering individual SNP interaction effects, rather than considering the collective effect of multiple SNPs.

5.2 UK Biobank/SpiroMeta consortium GWAS of lung function

The UK Biobank/SpiroMeta consortium GWAS of lung function was a study led by the University of Leicester genetic epidemiology group utilising both UK Biobank (n = 321,047) and the SpiroMeta consortium (n = 79,055) to identify SNPs associated with lung function. The lung function phenotypes studied were FEV_1, FVC, FEV_1/FVC and PEF (as defined in section 4.2).

5.2.1 Identifying novel signals and replicating previous signals associated with lung function

The 139 novel signals were identified using two approaches, a two stage and a one stage approach, with resulting signals referred to as tier 1 and tier 2 signals respectively. In the two stage approach (tier 1 signals), SNPs were identified firstly using a strict threshold in UK Biobank, before being assessed at a more lenient threshold using data from the SpiroMeta consortium (**Figure 5.1**). In the one stage approach SNPs were identified using a combined meta-analysis of both the UK Biobank and SpiroMeta consortium datasets (n up to 400,102).

Previously reported signals were deemed to be associated in this dataset by adhering to one of three predefined criteria (**Figure 5.1**). The signal was associated if the reported sentinel, a designated proxy (with $r^2 > 0.5$), or a nearby signal passing novel signal criteria (with an $r^2 > 0.1$), were associated with any lung function trait in UK Biobank.

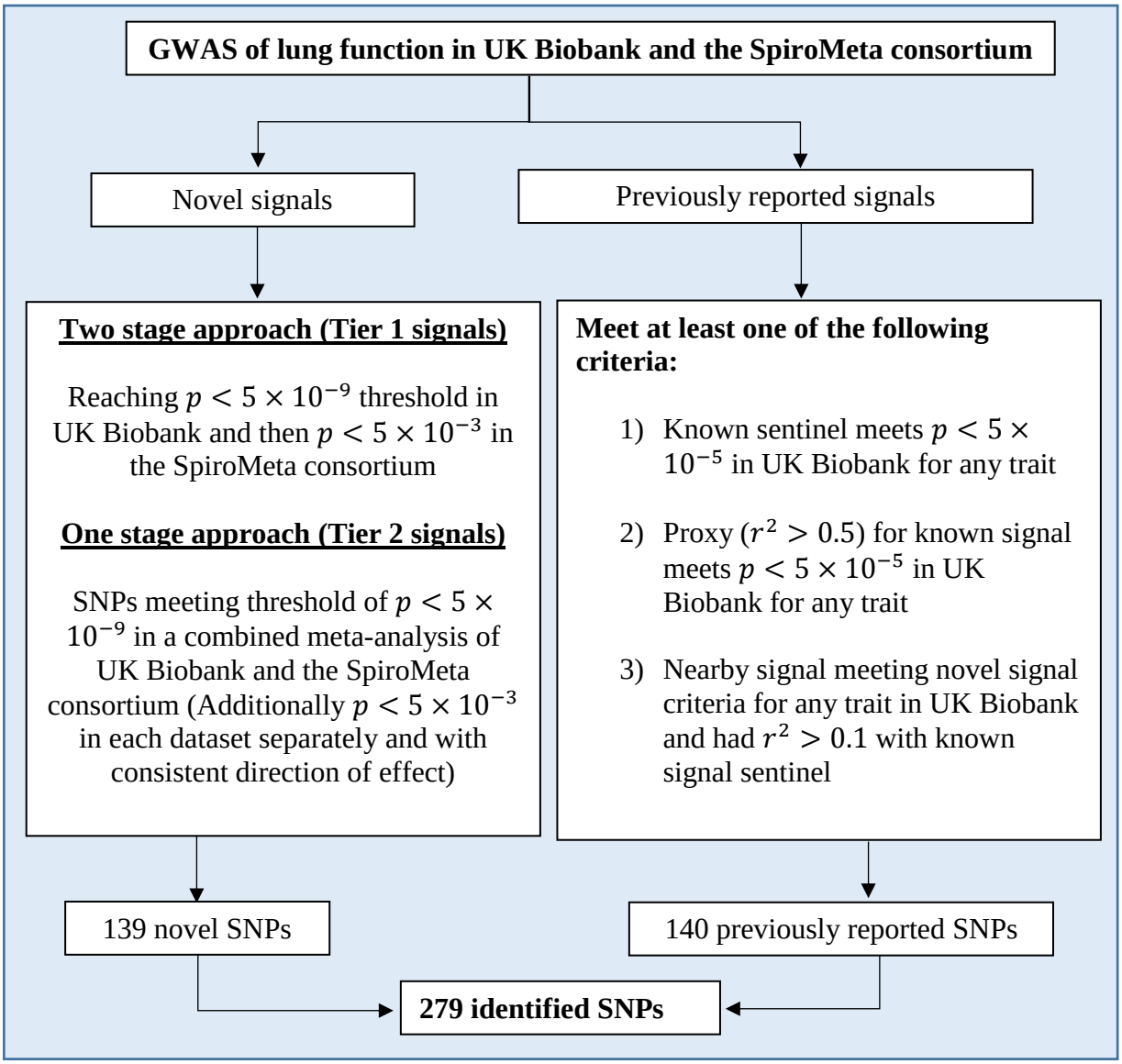


Figure 5.1 - Identification of novel signals and providing corroborative evidence for previously reported signals in UK Biobank and the SpiroMeta consortium

5.3 Phenotype quality control

The analysis undertaken in this chapter utilises the same unrelated European sample from chapter 4 of 303,619 individuals (with phenotype/genotype quality control, relatedness exclusion and ever/never smoker definitions presented in sections 4.2 – 4.5). Quality control of the FVC phenotype is presented in chapter 4 section 4.2. The quality control of PEF values was as follows. In summary, the best PEF value for each individual was derived from an acceptable blow (with acceptability of blows also described in chapter 4 section 4.2) which had the largest sum of FEV_1 and FVC. Although PEF values were provided by UK Biobank, there was an identified

discrepancy between the derived and provided values for a large number of individuals. Further exploration found that for a large subset, the UK Biobank provided PEF values appeared very low given their recorded FEV_1 value, however derived PEF values appeared consistent and plausible (**Figure 5.2**). As a result, rather than making exclusions based on inconsistency between derived and provided values, derived PEF values were used for all individuals and provided PEF values were deemed erroneous.

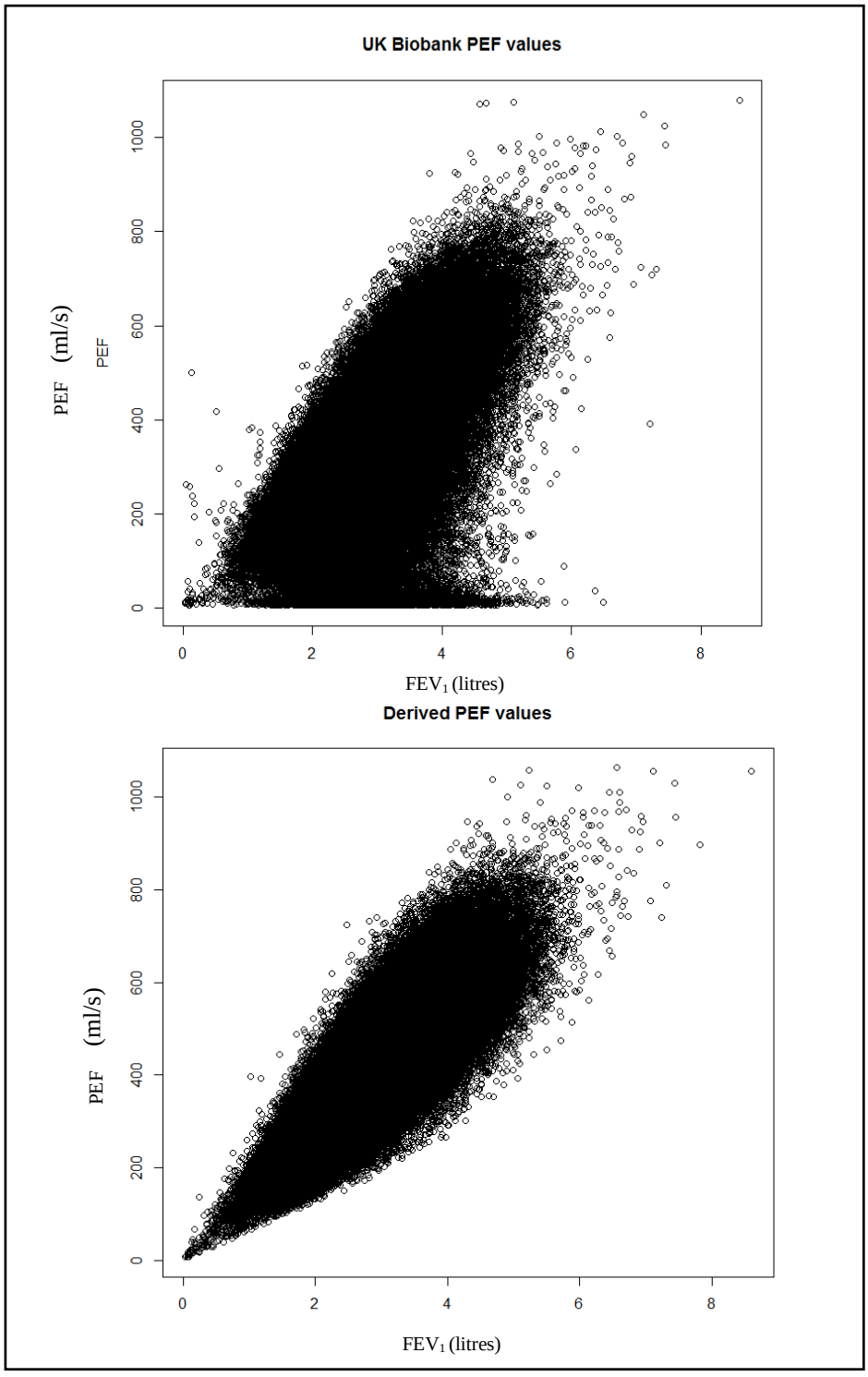


Figure 5.2 - Plots of both the UK provided and derived PEF values against FEV_1 values for the 303,619 individuals resulting from previous QC

5.4 Methods: Interaction and risk score analysis for the 279 SNPs associated with lung function

5.4.1 SNP-smoking interaction analysis

To determine the dependence of each of the 279 previously selected SNPs identified for association with lung function traits on smoking, each SNP was tested for an interaction effect (with ever/never smoking as the exposure). To do this required four stages (**Figure 5.3**).

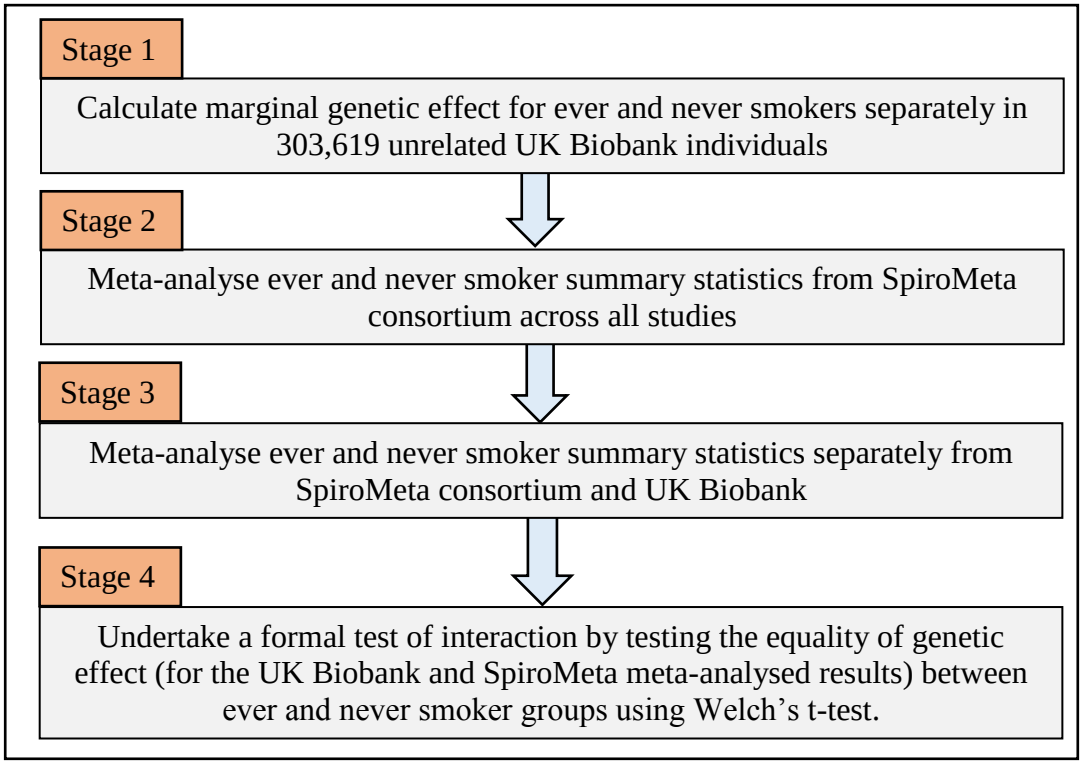


Figure 5.3 - Flow chart of methods to determine the interaction effect for each of the 279 SNPs identified for association with lung function and/or COPD

Firstly, genetic effect size estimates (with corresponding standard errors) were calculated for each SNP in ever and never smokers separately using the unrelated European UK Biobank sample of 303,619 individuals. Phenotypes were adjusted for sex, age, age^2 and height before being inverse normalised. Ten principal components and the genotype array used for each individual was adjusted for during the analysis stage, with the mean genotype method applied in QUICKTEST (171,172) using the following model:

$$Phenotype_i = \beta_0 + \beta_1 G_i + \varepsilon_i$$

where G_i is the genotype for individual i. The genetic effect of β_1 is tested for deviation from 0 (i.e. $H_o: \beta_1 = 0$).

The genetic effect size estimates for the SpiroMeta consortium were available in ever and never smokers separately. Therefore, in the second stage, the genetic effect size estimates were meta-analysed in ever and never smokers across all SpiroMeta consortium studies (as presented in chapter 4 section 4.6.7.1). In a third stage, the resulting genetic effect size estimates from the previous two stages were then meta-analysed (using a fixed effects meta-analysis) across UK Biobank and the SpiroMeta consortium, for the ever and never smoker subgroups. Both stages 2 and 3 were undertaken using R. Finally, in a fourth stage, a test of interaction was undertaken to compare the meta-analysed genetic effect size estimates calculated in stage 3 between ever and never smoker subgroups using Welch's t-test (as described in chapter 4 section 4.6.7.1). The test of interaction for each SNP was undertaken using the reported phenotype, which is the trait which produced the smallest p-value in the UK Biobank and SpiroMeta meta-analysis. The significance threshold was chosen based on a Bonferroni correction for the 279 SNPs tested.

5.4.2 Risk score interaction analysis

In addition, a further aim was to determine whether there was any effect of ever/never smoking on the combined genetic effect of all 279 novel and previous lung function signals. This was undertaken using a weighted genetic risk score (wGRS) and testing for a wGRS by smoking interaction. To do this, each of the 279 SNPs was given a weight based on magnitude of genetic effect, so that SNPs with larger magnitude of effect had a stronger influence on the genetic risk score. This was calculated by dividing the absolute value of each SNP's genetic effect by the total of all the absolute value genetic effects, multiplied by 279, meaning that all weights summed to 279. For each individual the maximum attainable risk score was 558 (possible number of risk alleles) and a unit change in risk score represented an increase or decrease equivalent to that of 1 risk allele. Risk alleles were chosen as the allele which had a deleterious effect on the

lung function trait. Weights were re-calculated for each phenotype to ensure a trait specific genetic risk score.

For each SNP, the source for the contributing genetic effect was dependent on the source used to identify the signal originally. So as not to introduce bias when calculating SNP weights (by influencing the weight using the dataset in which the SNP was identified), for previously reported signals identified using data from the SpiroMeta consortium, the genetic effect was taken from the UK Biobank data. Alternatively, UK Biobank identified signals had genetic effects extracted from the SpiroMeta consortium. Signals identified using data other than these two sources had their weights taken from the UK Biobank (the larger resource). For novel signals, the source of genetic effect was dependent on whether signals were tier 1 or tier 2 (**Figure 5.1**). The SpiroMeta consortium genetic effects were used for tier 1 signals, whilst for tier 2 signals, whichever source contributed the smallest absolute effect size for that signal was used to calculate weights. The discovery data source and weight source for the 279 lung function and COPD associated signals is presented in **Table 5.1**.The external weight method (using effect sizes for each SNP from a source independent of its identification) used here where possible, is suggested as the gold standard for weight calculation, with weights calculated internally in the absence of external weights also reported as a powerful approach for wGRS by environment interaction analysis (179).

Risk scores for each individual were calculated using PLINK (30), with R used to calculate risk score by ever-smoking interaction effect using the following model (genotyping array and 10 principal components were included as additional covariates):

$$Phenotype_i = \beta_0 + \beta_1 RS_i + \beta_2 S_i + \beta_3 RS_i S_i + \varepsilon_i$$

where RS_i and S_i correspond to the risk score and the smoking variable (ever/never smoking) for individual i respectively, and the effect of interest is β_3, the interaction effect between risk score and ever/never smoking.

The significance threshold was chosen based on a Bonferroni correction for the four traits tested for a wGRS by smoking interaction.

Table 5.1- Discovery source and weight source summary for the 279 lung function and COPD signals

Source of signals	No. signals	Source of Weight
UK Biobank References: Wain et al. 2015 (48), Wain et al. 2017 (50)	46	SpiroMeta Consortium
SpiroMeta Consortium References: Soler Artigas et al. 2011 (46), Loth et al. 2014 (47), Soler Artigas et al. 2015 (49), Jackson et al. 2016 (51)	41	UK Biobank
Other source References: Wilk et al. 2009 (43), Repapi et al. 2010 (44), Hancock et al. 2010 (45), Cho et al. 2014 (180), Hobbs et al. 2016 (59), Hobbs et al. 2017 (54), Wyss et al. 2017 (52)	53	UK Biobank
Novel tier 1 References: Shrine et al. 2018 (32,33)	99	SpiroMeta Consortium
Novel tier 2 References: Shrine et al. 2018 (32,33)	40	UK Biobank or SpiroMeta consortium (dependent on smallest magnitude of effect)

5.5 Results: Interaction and risk score analysis for the 279 SNPs associated with lung function

5.5.1 SNP-smoking interaction analysis

SNP-smoking interaction effect size estimates (and corresponding standard errors) were calculated for the 279 SNPs associated with lung function and COPD in up to 374,700 individuals (303,619 from UK Biobank and 71,081 from the SpiroMeta consortium), consisting of 197,999 never smokers and 176,701 ever smokers. None of the 279 previous or novel SNPs identified for having a genetic association with lung function traits produced a statistically significant interaction effect with ever/never smoking ($p < 2 \times 10^{-4}$, Bonferroni corrected significance threshold for 279 SNPs, **Table 5.2**). Only six of the 279 SNPs were nominally statistically significant at $p < 0.05$. Ever/never smoker stratified genetic effect size estimates for UK Biobank and the SpiroMeta consortium are given in **Appendix E**.

Table 5.2 – Meta-analysed genetic effects in ever and never smoker subgroups (and Welch's t-test to test interaction effect)

Results are presented for a combined UK Biobank and the SpiroMeta consortium meta-analysis for the 279 SNPs identified for marginal genetic effect. Fir signals, P = previously reported and N = novel. Betas (β) and standard errors (SE) to 4 d.p and are on the inverse normalised scale.

					Ever smoker			Never smoker			
Trait	**Signal**	**Gene**	**SNP (Chr:Pos)**	**Coded / non-coded allele**	**β**	**SE**	**P**	**β**	**SE**	**P**	**Welch's t-test P**
FEV_1/FVC	N	*PHF13*	rs9661802 (1:6678864)	A/C	0.0245	0.0035	$4.52x10^{-12}$	0.0236	0.0034	$2.61x10^{-12}$	0.874
FEV_1/FVC	P	*MFAP2*	rs9435733 (1:17308254)	T/C	0.0390	0.0033	$1.23x10^{-31}$	0.0377	0.0032	$2.28x10^{-32}$	0.799
FEV_1	N	*MIR4418*	rs12737805 (1:22612690)	A/G	0.0121	0.0040	$2.73x10^{-03}$	0.0281	0.0039	$3.29x10^{-13}$	0.030
FVC	N	*DHDDS*	rs9438626 (1:26775367)	G/C	-0.0158	0.0041	$1.16x10^{-04}$	-0.0187	0.0039	$1.49x10^{-06}$	0.624
FEV_1	N	*DHDDS*	rs12096239 (1:26796922)	G/C	0.0213	0.0038	$2.08x10^{-08}$	0.0160	0.0036	$1.10x10^{-05}$	0.363
FEV_1/FVC	P	*LOC101929516*	rs755249 (1:39995074)	T/C	-0.0273	0.0039	$3.86x10^{-12}$	-0.0213	0.0038	$1.44x10^{-08}$	0.316
FEV_1/FVC	N	*FAF1*	rs1416685 (1:51243374)	G/C	-0.0188	0.0034	$3.00x10^{-08}$	-0.0214	0.0032	$4.82x10^{-11}$	0.617
FEV_1/FVC	N	*LOC101926964*	rs72673461 (1:60966772)	T/G	0.0447	0.0078	$9.60x10^{-09}$	0.0622	0.0074	$4.06x10^{-17}$	0.119
FEV_1/FVC	N	*NEXN*	rs9661687 (1:78387270)	C/T	0.0282	0.0049	$9.38x10^{-09}$	0.0250	0.0047	$9.20x10^{-08}$	0.653
FEV_1/FVC	P	*TGFBR3*	rs1192415 (1:92077097)	G/A	-0.0410	0.0043	$1.64x10^{-21}$	-0.0466	0.0041	$4.06x10^{-30}$	0.375
FEV_1/FVC	N	*TGFBR3*	rs10874851 (1:92106637)	A/C	-0.0136	0.0034	$5.67x10^{-05}$	-0.0167	0.0032	$1.85x10^{-07}$	0.539
FEV_1/FVC	P	*TGFBR3*	rs11165787 (1:92381483)	A/G	0.0223	0.0036	$8.69x10^{-10}$	0.0271	0.0035	$4.46x10^{-15}$	0.391
FEV_1/FVC	N	*DENND2D*	rs9970286 (1:111737398)	G/A	-0.0222	0.0036	$6.76x10^{-10}$	-0.0271	0.0034	$2.43x10^{-15}$	0.373
FVC	P	*SPAG17*	rs35043843 (1:118911295)	T/G	-0.0241	0.0040	$1.12x10^{-09}$	-0.0241	0.0037	$1.08x10^{-10}$	0.999
PEF	N	*C1orf54*	rs11205354 (1:150249101)	C/A	-0.0156	0.0036	$1.73x10^{-05}$	-0.0181	0.0034	$1.03x10^{-07}$	0.634
FVC	P	*MCL1*	rs878471 (1:150547747)	G/A	0.0260	0.0034	$3.31x10^{-14}$	0.0336	0.0032	$4.80x10^{-25}$	0.184
FEV_1/FVC	N	*KRTCAP2*	rs141942982 (1:155137395)	G/T	0.0232	0.0055	$2.13x10^{-05}$	0.0496	0.0052	$1.79x10^{-21}$	0.005
FEV_1	N	*RALGPS2*	rs4651005 (1:178719306)	C/T	-0.0153	0.0036	$2.07x10^{-05}$	-0.0226	0.0034	$3.32x10^{-11}$	0.208
FVC	N	*MIR548F1*	rs2146098 (1:186090370)	A/G	-0.0147	0.0035	$3.52x10^{-05}$	-0.0206	0.0033	$4.81x10^{-10}$	0.282
FEV_1/FVC	N	*MIR548F1*	rs17531405 (1:186113852)	G/C	-0.0308	0.0044	$2.56x10^{-12}$	-0.0299	0.0042	$1.13x10^{-12}$	0.883
FEV_1/FVC	N	*MIR181A1HG*	rs10919604 (1:198898157)	A/G	0.0164	0.0034	$1.85x10^{-06}$	0.0211	0.0033	$1.07x10^{-10}$	0.375
FVC	P	*NR5A2*	rs2816992 (1:200069216)	A/G	-0.0164	0.0034	$1.95x10^{-06}$	-0.0169	0.0033	$2.06x10^{-07}$	0.914
FEV_1/FVC	N	*LMOD1*	rs4309038 (1:201884647)	G/C	-0.0180	0.0034	$1.17x10^{-07}$	-0.0130	0.0032	$5.81x10^{-05}$	0.345
PEF	P	*PIK3C2B*	rs1008833 (1:204426295)	A/G	-0.0244	0.0051	$1.63x10^{-06}$	-0.0397	0.0048	$1.40x10^{-16}$	0.061
FVC	P	*CENPF/KCNK2*	rs556648 (1:215120596)	A/G	0.0160	0.0041	$1.03x10^{-04}$	0.0153	0.0039	$9.12x10^{-05}$	0.902
FEV_1/FVC	N	*TGFB2*	rs2799098 (1:218521609)	G/A	0.0331	0.0044	$6.63x10^{-14}$	0.0218	0.0042	$1.68x10^{-07}$	0.107
PEF	P	*TGFB2*	rs6604614 (1:218631452)	C/G	-0.0165	0.0040	$3.30x10^{-05}$	-0.0155	0.0037	$3.51x10^{-05}$	0.856
FEV_1	P	*MIR548F3/TGFB2*	rs28613267 (1:218855029)	C/G	0.0150	0.0034	$8.51x10^{-06}$	0.0191	0.0032	$2.59x10^{-09}$	0.434
FEV_1/FVC	N	*LYPLAL1*	rs75128958 (1:219483218)	G/A	0.0507	0.0063	$1.11x10^{-15}$	0.0402	0.0060	$2.12x10^{-11}$	0.250
FEV_1/FVC	P	*RNU5F-1*	rs1338227 (1:219853742)	G/T	-0.0245	0.0034	$7.45x10^{-13}$	-0.0263	0.0032	$5.22x10^{-16}$	0.719

Trait	Signal	Gene	SNP (Chr:Pos)	Coded / non-coded allele	Ever smoker			Never smoker			Welch's t-test P
					β	SE	P	β	SE	P	
FVC	N	*HLX*	rs17009288 (1:221204299)	A/C	-0.0262	0.0037	2.19×10^{-12}	-0.0244	0.0035	4.68×10^{-12}	0.731
FVC	P	*C1orf140/DUSP10*	rs12757436 (1:221631938)	A/G	0.0228	0.0036	1.88×10^{-10}	0.0117	0.0034	5.47×10^{-04}	0.081
PEF	P	*CHRM3*	rs2355237 (1:239857524)	A/G	0.0292	0.0036	3.94×10^{-16}	0.0293	0.0034	3.41×10^{-18}	0.998
FEV_1/FVC	N	*LOC101926966*	rs2544536 (2:15906854)	T/C	-0.0181	0.0033	5.17×10^{-08}	-0.0284	0.0032	4.58×10^{-19}	0.091
FEV_1/FVC	P	*KCNS3*	rs55884799 (2:18287623)	T/C	-0.0353	0.0044	1.29×10^{-15}	-0.0467	0.0042	1.58×10^{-28}	0.104
FVC	N	*RDH14*	rs6751968 (2:18570024)	C/A	-0.0254	0.0044	6.11×10^{-09}	-0.0251	0.0041	1.37×10^{-09}	0.964
FVC	N	*RDH14*	rs13430465 (2:18702313)	C/T	-0.0391	0.0062	3.75×10^{-10}	-0.0386	0.0059	6.33×10^{-11}	0.954
FVC	N	*ATAD2B*	rs13009582 (2:24018480)	G/A	-0.0213	0.0034	2.47×10^{-10}	-0.0121	0.0032	1.67×10^{-04}	0.117
FVC	N	*CIB4*	rs732990 (2:26842146)	C/G	-0.0145	0.0034	1.66×10^{-05}	-0.0143	0.0032	8.04×10^{-06}	0.963
FVC	N	*PKDCC*	rs4952564 (2:42243850)	A/G	-0.0167	0.0036	3.22×10^{-06}	-0.0183	0.0034	8.84×10^{-08}	0.774
FVC	P	*EFEMP1*	rs3791679 (2:56096892)	A/G	0.0378	0.0040	6.51×10^{-21}	0.0296	0.0038	6.92×10^{-15}	0.195
FEV_1/FVC	N	*IL1RL1*	rs12470864 (2:102926362)	G/A	0.0149	0.0035	1.51×10^{-05}	0.0246	0.0033	8.18×10^{-14}	0.110
FVC	P	*CCNT2-AS1*	rs62168891 (2:135672187)	C/T	-0.0205	0.0035	3.27×10^{-09}	-0.0161	0.0033	9.98×10^{-07}	0.403
FEV_1/FVC	N	*TEX41*	rs1406225 (2:145797829)	G/T	0.0249	0.0037	2.06×10^{-11}	0.0133	0.0036	1.71×10^{-04}	0.076
FEV_1	P	*LOC101929378*	rs72902177 (2:157016257)	C/T	0.0304	0.0049	7.77×10^{-10}	0.0362	0.0047	1.19×10^{-14}	0.413
FEV_1	N	*RBMS1*	rs7424771 (2:161276378)	G/A	0.0198	0.0034	4.18×10^{-09}	0.0159	0.0032	8.04×10^{-07}	0.443
FEV_1	N	*MIR548N*	rs2304340 (2:179260382)	A/G	-0.0192	0.0034	2.00×10^{-08}	-0.0089	0.0033	6.19×10^{-03}	0.094
FEV_1/FVC	N	*ITGAV*	rs2084448 (2:187530520)	T/C	0.0157	0.0037	2.08×10^{-05}	0.0204	0.0035	6.07×10^{-09}	0.400
FVC	N	*SATB2*	rs1249096 (2:199723365)	G/A	-0.0219	0.0034	1.82×10^{-10}	-0.0215	0.0032	2.85×10^{-11}	0.947
FEV_1/FVC	N	*SPATS2L*	rs985256 (2:201208692)	A/C	0.0195	0.0041	1.76×10^{-06}	0.0181	0.0039	2.78×10^{-06}	0.814
FVC	N	*KIAA2012*	rs12997625 (2:202970250)	C/T	0.0195	0.0034	1.04×10^{-08}	0.0110	0.0032	6.36×10^{-04}	0.142
FEV_1/FVC	N	*IGFBP5*	rs6435952 (2:217614730)	A/T	0.0196	0.0047	3.52×10^{-05}	0.0338	0.0045	3.75×10^{-14}	0.062
FEV_1	N	*DIRC3*	rs4294980 (2:218604356)	G/A	-0.0146	0.0041	4.04×10^{-04}	-0.0186	0.0039	2.03×10^{-06}	0.512
FEV_1	P	*TNS1*	rs2571445 (2:218683154)	A/G	-0.0307	0.0034	3.16×10^{-19}	-0.0264	0.0033	5.85×10^{-16}	0.406
FVC	N	*ASIC4*	rs4674407 (2:220382700)	C/T	0.0070	0.0034	4.17×10^{-02}	0.0191	0.0032	3.64×10^{-09}	0.061
FEV1/FVC	P	*PID1*	rs62201738 (2:229502197)	A/C	-0.0743	0.0063	3.18×10^{-32}	-0.0802	0.0060	4.60×10^{-41}	0.507
FEV_1	P	*TRAF3IP1*	rs6710301 (2:239441308)	A/C	0.0248	0.0047	1.57×10^{-07}	0.0257	0.0045	1.02×10^{-08}	0.892
FVC	N	*LINC01107*	rs6431620 (2:239604970)	T/G	0.0224	0.0042	7.20×10^{-08}	0.0152	0.0039	1.08×10^{-04}	0.254
FEV1/FVC	P	*FLJ43879*	rs4308141 (2:239881309)	C/G	-0.0569	0.0042	2.12×10^{-41}	-0.0446	0.0040	8.13×10^{-29}	0.078
FVC	N	*C2orf54*	rs6437219 (2:241844033)	C/T	-0.0182	0.0036	4.34×10^{-07}	-0.0195	0.0034	9.00×10^{-09}	0.811
FVC	N	*BOK-AS1*	rs6733504 (2:242495953)	A/G	0.0216	0.0034	2.78×10^{-10}	0.0192	0.0032	2.94×10^{-09}	0.635
FEV_1	N	*LINC00620*	rs2974389 (3:13787641)	A/G	0.0143	0.0034	2.15×10^{-05}	0.0166	0.0032	2.33×10^{-07}	0.643
FVC	N	*RARB*	rs73048404 (3:25179533)	T/G	0.0184	0.0047	1.03×10^{-04}	0.0242	0.0045	6.84×10^{-08}	0.404

Trait	Signal	Gene	SNP (Chr:Pos)	Coded / non-coded allele	Ever smoker			Never smoker			Welch's t-test P
					β	SE	P	β	SE	P	
FEV_1/FVC	P	*RARB*	rs1529672 (3:25520582)	C/A	-0.0431	0.0044	$1.71x10^{-22}$	-0.0423	0.0042	$1.04x10^{-23}$	0.898
FEV_1/FVC	P	*RBMS3*	rs17666332 (3:29469675)	T/G	0.0308	0.0037	$1.62x10^{-16}$	0.0216	0.0036	$1.63x10^{-09}$	0.135
FEV_1/FVC	P	*CACNA2D3*	rs12715478 (3:55152319)	A/G	0.0251	0.0034	$2.92x10^{-13}$	0.0246	0.0033	$6.49x10^{-14}$	0.913
FEV_1	P	*SLMAP*	rs6445932 (3:57879611)	T/G	-0.0295	0.0039	$3.60x10^{-14}$	-0.0262	0.0037	$1.21x10^{-12}$	0.563
FEV_1	P	*SUCLG2*	rs4132748 (3:67455803)	T/C	-0.0194	0.0037	$1.28x10^{-07}$	-0.0188	0.0035	$7.29x10^{-08}$	0.910
FVC	N	*FOXP1*	rs35480566 (3:71583177)	A/G	-0.0239	0.0034	$2.85x10^{-12}$	-0.0201	0.0032	$4.58x10^{-10}$	0.464
FEV_1/FVC	N	*PDZRN3-AS1*	rs586936 (3:73862616)	G/A	0.0146	0.0035	$2.72x10^{-05}$	0.0218	0.0033	$3.63x10^{-11}$	0.202
FVC	P	*DCBLD2*	rs12497779 (3:98822050)	G/T	0.0349	0.0040	$2.96x10^{-18}$	0.0277	0.0038	$2.76x10^{-13}$	0.240
FVC	N	*MIR548G*	rs1610265 (3:99420192)	C/T	0.0473	0.0064	$9.75x10^{-14}$	0.0321	0.0060	$8.83x10^{-08}$	0.105
FEV_1/FVC	P	*EEFSEC*	rs2999090 (3:127931340)	A/G	-0.0391	0.0052	$5.75x10^{-14}$	-0.0452	0.0049	$4.91x10^{-20}$	0.414
FVC	P	*RSRC1*	rs12634907 (3:158226886)	A/G	0.0247	0.0036	$3.95x10^{-12}$	0.0271	0.0034	$7.84x10^{-16}$	0.644
FEV1/FVC	N	*BCHE*	rs1799807 (3:165548529)	T/C	0.0565	0.0123	$4.53x10^{-06}$	0.0574	0.0117	$9.72x10^{-07}$	0.960
FEV_1	P	*LOC100507661*	rs879394 (3:168709843)	G/T	0.0274	0.0040	$4.98x10^{-12}$	0.0262	0.0038	$4.32x10^{-12}$	0.831
FEV_1	P	*MECOM*	rs78101726 (3:169295436)	A/G	0.0318	0.0047	$9.38x10^{-12}$	0.0344	0.0044	$7.27x10^{-15}$	0.704
FEV_1	N	*IGF2BP2*	rs6780171 (3:185503456)	T/A	0.0172	0.0036	$1.96x10^{-06}$	0.0186	0.0034	$7.04x10^{-08}$	0.802
FEV_1/FVC	P	*AFAP1*	rs62289340 (4:7879027)	C/T	-0.0159	0.0034	$2.11x10^{-06}$	-0.0169	0.0032	$1.25x10^{-07}$	0.837
FEV_1	N	*KDR*	rs12331869 (4:56012149)	A/G	-0.0137	0.0044	$1.80x10^{-03}$	-0.0187	0.0042	$6.82x10^{-06}$	0.429
FEV_1/FVC	N	*BTC*	rs62316310 (4:75676529)	G/A	-0.0224	0.0038	$5.74x10^{-09}$	-0.0334	0.0037	$8.05x10^{-20}$	0.093
FEV_1/FVC	N	*FRAS1*	rs11098196 (4:79403952)	G/T	0.0158	0.0034	$2.86x10^{-06}$	0.0227	0.0032	$1.50x10^{-12}$	0.212
FEV_1/FVC	P	*FAM13A*	rs2609279 (4:89855495)	T/C	0.0585	0.0041	$3.24x10^{-46}$	0.0454	0.0039	$2.07x10^{-31}$	0.060
FEV_1/FVC	P	*FAM13A*	rs2869966 (4:89869078)	T/C	-0.0435	0.0034	$4.93x10^{-37}$	-0.0385	0.0033	$2.64x10^{-32}$	0.351
FEV_1/FVC	P	*TET2*	rs6533183 (4:106133184)	T/C	-0.0280	0.0035	$2.41x10^{-15}$	-0.0323	0.0034	$7.51x10^{-22}$	0.423
FEV_1	P	*GSTCD*	rs11722225 (4:106766430)	T/C	-0.0705	0.0067	$1.18x10^{-25}$	-0.0750	0.0064	$1.31x10^{-31}$	0.637
FEV_1/FVC	P	*NPNT*	rs34712979 (4:106819053)	G/A	0.0640	0.0039	$6.65x10^{-60}$	0.0737	0.0037	$3.23x10^{-88}$	0.130
FVC	N	*HHIP-AS1*	rs13109426 (4:145330628)	G/A	0.0236	0.0034	$8.52x10^{-12}$	0.0224	0.0032	$5.45x10^{-12}$	0.816
PEF	N	*HHIP-AS1*	rs13116999 (4:145442364)	G/A	-0.0620	0.0036	$8.06x10^{-67}$	-0.0690	0.0034	$1.90x10^{-93}$	0.226
FEV_1/FVC	P	*HHIP-AS1*	rs13141641 (4:145506456)	T/C	-0.0698	0.0034	$6.33x10^{-92}$	-0.0710	0.0033	$5.68x10^{-105}$	0.815
PEF	P	*OTUD4/SMAD1*	rs2353940 (4:145740898)	T/C	0.0365	0.0041	$9.29x10^{-19}$	0.0424	0.0039	$6.61x10^{-28}$	0.338
FEV_1	N	*LOC100996325*	rs11739847 (5:609661)	G/A	0.0214	0.0041	$2.49x10^{-07}$	0.0195	0.0040	$9.07x10^{-07}$	0.751
FVC	P	*TARS*	rs268717 (5:33352738)	T/C	-0.0391	0.0058	$1.52x10^{-11}$	-0.0301	0.0055	$3.94x10^{-08}$	0.286
FEV_1	N	*NNT*	rs4866846 (5:43976162)	A/G	0.0298	0.0047	$1.62x10^{-10}$	0.0250	0.0045	$2.21x10^{-08}$	0.478
FVC	P	*FGF10*	rs6859730 (5:44367221)	A/T	0.0193	0.0036	$5.40x10^{-08}$	0.0213	0.0034	$2.92x10^{-10}$	0.702
FEV_1/FVC	P	*ITGA1*	rs12522114 (5:52187038)	A/C	-0.0352	0.0038	$4.23x10^{-20}$	-0.0399	0.0036	$6.24x10^{-28}$	0.415

Trait	Signal	Gene	SNP (Chr:Pos)	Coded / non-coded allele	Ever smoker			Never smoker			Welch's t-test P
					β	SE	P	β	SE	P	
FVC	P	*ARL15*	rs2441026 (5:53444498)	C/T	-0.0180	0.0034	$1.17x10^{-07}$	-0.0188	0.0032	$5.00x10^{-09}$	0.882
FVC	P	*AP3B1*	rs425102 (5:77396400)	T/G	0.0184	0.0040	$3.49x10^{-06}$	0.0222	0.0037	$2.81x10^{-09}$	0.510
FEV_1/FVC	P	*SPATA9*	rs987068 (5:95025146)	G/C	0.0301	0.0036	$1.23x10^{-16}$	0.0300	0.0035	$3.73x10^{-18}$	0.986
FEV_1/FVC	N	*LOX*	rs10059661 (5:121410529)	C/G	-0.0270	0.0045	$1.42x10^{-09}$	-0.0366	0.0042	$6.38x10^{-18}$	0.163
FEV_1/FVC	N	*ADAMTS19-AS1*	rs17163397 (5:128767384)	A/G	-0.0305	0.0051	$2.83x10^{-09}$	-0.0297	0.0049	$1.15x10^{-09}$	0.915
FVC	P	*P4HA2-AS1*	rs3843503 (5:131466629)	T/A	0.0216	0.0034	$3.82x10^{-10}$	0.0168	0.0033	$2.50x10^{-07}$	0.373
FEV_1/FVC	P	*HTR4*	rs7733410 (5:147856522)	G/A	-0.0553	0.0034	$7.95x10^{-59}$	-0.0467	0.0032	$5.00x10^{-47}$	0.144
FEV_1	N	*ADRB2*	rs1800888 (5:148206885)	C/T	0.0932	0.0140	$2.45x10^{-11}$	0.0793	0.0135	$3.85x10^{-09}$	0.475
FEV_1	P	*ABLIM3*	rs11952673 (5:148652302)	T/G	-0.0224	0.0035	$8.84x10^{-11}$	-0.0158	0.0033	$1.56x10^{-06}$	0.236
FEV_1/FVC	P	*CYFIP2*	rs11134766 (5:156908317)	T/C	-0.0650	0.0069	$3.93x10^{-21}$	-0.0661	0.0066	$6.55x10^{-24}$	0.909
FEV_1/FVC	P	*ADAM19*	rs11134789 (5:156944199)	C/A	0.0344	0.0036	$3.41x10^{-22}$	0.0488	0.0034	$2.04x10^{-47}$	0.038
FEV_1/FVC	N	*FGF18*	rs10059996 (5:170901463)	T/G	-0.0292	0.0036	$9.43x10^{-16}$	-0.0403	0.0035	$1.66x10^{-31}$	0.083
FEV_1/FVC	N	*RASGEF1C*	rs79898473 (5:179598771)	T/C	-0.0303	0.0036	$3.87x10^{-17}$	-0.0324	0.0034	$4.92x10^{-21}$	0.696
FEV_1/FVC	P	*LY86*	rs1294417 (6:6741932)	T/C	-0.0254	0.0033	$2.98x10^{-14}$	-0.0374	0.0032	$9.72x10^{-32}$	0.061
FEV_1/FVC	P	*DSP*	rs2076295 (6:7563232)	T/G	-0.0210	0.0034	$3.41x10^{-10}$	-0.0255	0.0032	$1.75x10^{-15}$	0.393
FVC	N	*BMP6*	rs12198986 (6:7720059)	G/A	0.0213	0.0033	$1.96x10^{-10}$	0.0251	0.0032	$2.51x10^{-15}$	0.450
FVC	P	*BMP6*	rs10498672 (6:7797840)	C/G	0.0317	0.0044	$4.53x10^{-13}$	0.0381	0.0042	$4.46x10^{-20}$	0.322
FEV_1/FVC	P	*CASC15*	rs13198081 (6:22017543)	G/C	-0.0258	0.0035	$1.17x10^{-13}$	-0.0335	0.0033	$7.45x10^{-24}$	0.182
PEF	P	*ZNF184*	rs7752448 (6:28301099)	A/G	0.0531	0.0054	$9.60x10^{-23}$	0.0597	0.0050	$1.13x10^{-32}$	0.394
FEV_1/FVC	P	*AGER*	rs2070600 (6:32151443)	T/C	0.1384	0.0070	$1.17x10^{-86}$	0.1565	0.0066	$1.32x10^{-123}$	0.079
FEV_1/FVC	P	*HLA-DQB1*	rs9274247 (6:32631295)	A/G	-0.0425	0.0041	$5.72x10^{-25}$	-0.0505	0.0039	$1.25x10^{-38}$	0.211
FVC	N	*HMGA1*	rs9689096 (6:34188892)	A/C	-0.0305	0.0070	$1.20x10^{-05}$	-0.0346	0.0067	$2.76x10^{-07}$	0.680
FVC	N	*CDC5L*	rs9357446 (6:44447598)	G/A	0.0147	0.0033	$1.13x10^{-05}$	0.0131	0.0032	$3.80x10^{-05}$	0.739
FEV_1/FVC	N	*RUNX2*	rs12202314 (6:45530471)	T/C	-0.0170	0.0036	$1.92x10^{-06}$	-0.0240	0.0034	$2.13x10^{-12}$	0.221
FVC	N	*RUNX2*	rs9472541 (6:45622748)	T/A	0.0166	0.0037	$8.54x10^{-06}$	0.0127	0.0035	$3.18x10^{-04}$	0.484
FEV_1	N	*RNU6-71P*	rs2894837 (6:56336406)	A/G	0.0181	0.0035	$2.43x10^{-07}$	0.0161	0.0033	$1.47x10^{-06}$	0.689
FEV_1/FVC	P	*KCNQ5*	rs13206405 (6:73663814)	C/A	-0.0344	0.0042	$2.28x10^{-16}$	-0.0380	0.0040	$1.38x10^{-21}$	0.555
FEV_1/FVC	P	*ARMC2*	rs2798641 (6:109268050)	C/T	0.0384	0.0044	$1.17x10^{-18}$	0.0512	0.0041	$2.71x10^{-35}$	0.073
FVC	P	*MIR588*	rs6918725 (6:126990392)	T/G	-0.0172	0.0034	$4.50x10^{-07}$	-0.0218	0.0032	$1.25x10^{-11}$	0.381
FEV_1	N	*SLC2A12*	rs2627237 (6:134339265)	A/G	0.0159	0.0034	$3.22x10^{-06}$	0.0136	0.0033	$3.21x10^{-05}$	0.647
FEV_1	N	*LOC100507477*	rs1102077 (6:140271357)	A/C	0.0106	0.0039	$7.19x10^{-03}$	0.0299	0.0037	$1.44x10^{-15}$	0.015
FEV_1	N	*VTA1*	rs9385988 (6:142560957)	A/G	-0.0312	0.0037	$8.53x10^{-17}$	-0.0257	0.0036	$5.43x10^{-13}$	0.339
FEV_1/FVC	P	*GPR126*	rs17280293 (6:142688969)	A/G	-0.1752	0.0105	$1.72x10^{-62}$	-0.1867	0.0098	$6.52x10^{-81}$	0.432

Trait	Signal	Gene	SNP (Chr:Pos)	Coded / non-coded allele	Ever smoker			Never smoker			Welch's t-test P
					β	SE	P	β	SE	P	
FEV_1/FVC	P	*GPR126*	rs7753012 (6:142745883)	T/G	-0.0689	0.0037	$6.60x10^{-79}$	-0.0756	0.0035	$1.95x10^{-104}$	0.251
FEV_1/FVC	P	*C1GALT1*	rs4318980 (7:7256490)	A/G	-0.0186	0.0034	$3.92x10^{-08}$	-0.0178	0.0032	$3.92x10^{-08}$	0.866
FVC	P	*AGMO*	rs4721442 (7:15506007)	T/G	0.0215	0.0045	$1.98x10^{-06}$	0.0199	0.0043	$3.19x10^{-06}$	0.810
FEV_1/FVC	N	*MEOX2-AS1*	rs4721457 (7:15872324)	T/C	0.0178	0.0047	$1.40x10^{-04}$	0.0245	0.0044	$2.94x10^{-08}$	0.328
FEV_1	N	*SKAP2*	rs559233 (7:26848830)	T/C	0.0163	0.0034	$1.15x10^{-06}$	0.0167	0.0032	$1.81x10^{-07}$	0.933
FVC	N	*HOXA-AS3*	rs62454414 (7:27182329)	T/G	0.0194	0.0049	$8.44x10^{-05}$	0.0176	0.0047	$1.70x10^{-04}$	0.789
FEV_1	N	*JAZF1*	rs1513272 (7:28200097)	C/T	-0.0201	0.0033	$1.35x10^{-09}$	-0.0225	0.0032	$1.25x10^{-12}$	0.633
FVC	N	*IGFBP3*	rs17232687 (7:46448518)	T/C	-0.0210	0.0034	$5.28x10^{-10}$	-0.0161	0.0032	$5.02x10^{-07}$	0.347
FEV_1	N	*SEMA3D*	rs12707691 (7:84569510)	C/G	-0.0173	0.0036	$1.20x10^{-06}$	-0.0252	0.0034	$1.11x10^{-13}$	0.179
FEV_1/FVC	P	*ZKSCAN1*	rs2261360 (7:99692993)	T/G	0.0232	0.0040	$5.56x10^{-09}$	0.0252	0.0038	$2.74x10^{-11}$	0.731
FEV_1/FVC	N	*MET*	rs193686 (7:116431427)	C/T	0.0126	0.0036	$5.30x10^{-04}$	0.0240	0.0034	$3.24x10^{-12}$	0.076
FEV_1	P	*LOC285889*	rs12698403 (7:156127246)	G/A	0.0232	0.0034	$9.04x10^{-12}$	0.0284	0.0032	$1.16x10^{-18}$	0.326
FEV_1/FVC	N	*PPP1R3B*	rs330939 (8:9018590)	T/G	0.0222	0.0035	$1.36x10^{-10}$	0.0267	0.0033	$9.89x10^{-16}$	0.402
FEV_1	N	*DEFB136*	rs4128298 (8:11823332)	T/C	-0.0215	0.0037	$7.52x10^{-09}$	-0.0135	0.0036	$1.43x10^{-04}$	0.184
FEV_1	N	*LOC100505739*	rs7465401 (8:70367248)	T/C	-0.0215	0.0038	$1.24x10^{-08}$	-0.0203	0.0036	$1.52x10^{-08}$	0.828
FVC	N	*BOP1*	rs7838717 (8:145504343)	T/C	-0.0243	0.0036	$1.15x10^{-11}$	-0.0204	0.0034	$1.69x10^{-09}$	0.468
FVC	P	*DMRT2/SMARCA2*	rs771662 (9:1568941)	T/C	-0.0179	0.0035	$3.54x10^{-07}$	-0.0133	0.0033	$6.48x10^{-05}$	0.394
FEV1/FVC	P	*GLIS3*	rs1570203 (9:4120648)	G/A	-0.0231	0.0034	$5.83x10^{-12}$	-0.0250	0.0032	$7.17x10^{-15}$	0.714
FEV_1	N	*SH3GL2*	rs7041139 (9:18013733)	C/T	0.0181	0.0036	$3.86x10^{-07}$	0.0186	0.0034	$4.58x10^{-08}$	0.928
FEV_1/FVC	P	*FLJ35282/ELAVL2*	rs1107677 (9:23587027)	T/C	0.0228	0.0034	$9.90x10^{-12}$	0.0226	0.0032	$1.79x10^{-12}$	0.958
FEV_1/FVC	P	*PTCH1*	rs28446321 (9:98266855)	T/A	0.0552	0.0059	$5.38x10^{-21}$	0.0456	0.0056	$3.51x10^{-16}$	0.260
FEV_1/FVC	N	*LOC158434*	rs72743974 (9:98878881)	A/G	-0.0227	0.0045	$4.50x10^{-07}$	-0.0208	0.0043	$1.11x10^{-06}$	0.770
FEV_1/FVC	N	*GALNT12*	rs57649467 (9:101632854)	G/A	-0.0171	0.0035	$8.81x10^{-07}$	-0.0156	0.0033	$2.26x10^{-06}$	0.770
FEV_1/FVC	P	*TMEM38B/ZNF462*	rs1491106 (9:109483517)	T/G	0.0233	0.0035	$2.04x10^{-11}$	0.0286	0.0033	$3.95x10^{-18}$	0.327
FEV_1/FVC	P	*ASTN2*	rs10983184 (9:119234058)	C/T	-0.0231	0.0035	$6.39x10^{-11}$	-0.0295	0.0034	$1.48x10^{-18}$	0.255
FEV_1	N	*IER5L*	rs967497 (9:131943843)	A/G	0.0180	0.0036	$6.94x10^{-07}$	0.0100	0.0035	$3.67x10^{-03}$	0.178
FVC	P	*QSOX2*	rs7024579 (9:139100413)	C/T	0.0220	0.0037	$1.74x10^{-09}$	0.0215	0.0035	$4.56x10^{-10}$	0.928
FVC	P	*DNLZ*	rs4073153 (9:139259349)	A/G	0.0135	0.0035	$9.32x10^{-05}$	0.0135	0.0033	$3.23x10^{-05}$	0.993
FEV_1/FVC	P	*CDC123*	rs7090277 (10:12278021)	T/A	-0.0358	0.0033	$4.49x10^{-27}$	-0.0444	0.0032	$1.67x10^{-44}$	0.134
PEF	P	*KIAA1462*	rs7914842 (10:30268770)	A/G	0.0156	0.0036	$1.72x10^{-05}$	0.0187	0.0034	$4.20x10^{-08}$	0.566
FEV_1/FVC	N	*PARD3*	rs1274475 (10:34480582)	G/A	-0.0188	0.0035	$6.49x10^{-08}$	-0.0157	0.0033	$2.08x10^{-06}$	0.559
FEV_1	P	*JMJD1C*	rs7082066 (10:64998971)	A/G	0.0232	0.0043	$6.75x10^{-08}$	0.0192	0.0041	$3.17x10^{-06}$	0.520
FVC	P	*MYPN*	rs10998018 (10:69962954)	A/G	-0.0180	0.0034	$1.07x10^{-07}$	-0.0254	0.0032	$2.36x10^{-15}$	0.188

Trait	Signal	Gene	SNP (Chr:Pos)	Coded / non-coded allele	Ever smoker			Never smoker			Welch's t-test P
					β	SE	P	β	SE	P	
FEV_1	P	*CAMK2G*	rs7098573 (10:75580014)	G/A	0.0246	0.0037	$4.19x10^{-11}$	0.0248	0.0035	$2.63x10^{-12}$	0.972
PEF	N	*CAMK2G*	rs60820984 (10:75639578)	C/T	0.0238	0.0046	$2.48x10^{-07}$	0.0192	0.0043	$9.92x10^{-06}$	0.487
FVC	P	*COMTD1/ZNF503-AS1*	rs1259605 (10:77119039)	T/C	-0.0095	0.0039	$1.56x10^{-02}$	-0.0132	0.0037	$3.72x10^{-04}$	0.520
FEV_1	P	*C10orf11*	rs2637254 (10:78312002)	G/A	0.0288	0.0034	$9.88x10^{-18}$	0.0277	0.0032	$4.10x10^{-18}$	0.826
FEV_1/FVC	P	*SFTPD*	rs721917 (10:81706324)	A/G	0.0192	0.0034	$1.89x10^{-08}$	0.0201	0.0032	$6.33x10^{-10}$	0.856
FEV_1	N	*OBFC1*	rs11191841 (10:105639611)	T/C	-0.0122	0.0034	$2.89x10^{-04}$	-0.0201	0.0032	$3.18x10^{-10}$	0.164
FEV_1/FVC	P	*HTRA1*	rs4279944 (10:124297637)	T/C	0.0213	0.0049	$1.14x10^{-05}$	0.0193	0.0046	$2.74x10^{-05}$	0.773
FEV_1/FVC	N	*SLC1A2*	rs10836366 (11:35308988)	T/C	0.0186	0.0038	$1.29x10^{-06}$	0.0163	0.0037	$8.13x10^{-06}$	0.694
FVC	P	*HSD17B12*	rs17596617 (11:43690717)	C/T	0.0249	0.0036	$4.95x10^{-12}$	0.0149	0.0034	$1.25x10^{-05}$	0.105
FEV_1	P	*PRDM11*	rs10838435 (11:45244903)	C/G	0.0197	0.0047	$2.70x10^{-05}$	0.0233	0.0045	$2.43x10^{-07}$	0.596
FEV_1	P	*EML3*	rs71490394 (11:62370155)	G/A	-0.0305	0.0035	$2.57x10^{-18}$	-0.0237	0.0033	$9.11x10^{-13}$	0.227
FEV_1/FVC	P	*ARHGEF17*	rs2027761 (11:73036179)	C/T	-0.0400	0.0053	$5.00x10^{-14}$	-0.0330	0.0051	$6.61x10^{-11}$	0.368
FEV_1/FVC	P	*PRSS23*	rs11234768 (11:86448839)	T/C	0.0265	0.0047	$1.35x10^{-08}$	0.0322	0.0044	$2.93x10^{-13}$	0.399
FEV_1/FVC	P	*RPUSD4*	rs541601 (11:126009500)	T/C	-0.0222	0.0043	$3.05x10^{-07}$	-0.0281	0.0041	$1.08x10^{-11}$	0.361
FVC	N	*FKBP4*	rs56196860 (12:2908330)	C/A	0.0436	0.0100	$1.20x10^{-05}$	0.0634	0.0093	$1.16x10^{-11}$	0.156
FEV_1	N	*CCND2-AS1*	rs12811814 (12:4243749)	T/C	0.0120	0.0034	$3.86x10^{-04}$	0.0185	0.0032	$9.13x10^{-09}$	0.233
FEV_1/FVC	N	*AEBP2*	rs10841302 (12:19808912)	G/C	-0.0180	0.0033	$7.67x10^{-08}$	-0.0171	0.0032	$8.55x10^{-08}$	0.865
FVC	P	*CCDC91*	rs7977418 (12:28588242)	T/C	0.0368	0.0034	$7.17x10^{-28}$	0.0388	0.0032	$5.44x10^{-34}$	0.678
FEV_1	P	*RAB5B*	rs1689510 (12:56396768)	C/G	-0.0106	0.0036	$2.99x10^{-03}$	-0.0193	0.0034	$9.13x10^{-09}$	0.142
FEV_1/FVC	P	*LRP1*	rs11172113 (12:57527283)	T/C	-0.0220	0.0034	$1.20x10^{-10}$	-0.0238	0.0032	$2.35x10^{-13}$	0.733
FEV_1/FVC	N	*RASSF3*	rs1244869 (12:65075332)	T/G	0.0177	0.0035	$4.03x10^{-07}$	0.0130	0.0033	$8.93x10^{-05}$	0.381
FEV_1	P	*MSRB3*	rs12825748 (12:65793153)	G/C	-0.0195	0.0036	$8.01x10^{-08}$	-0.0192	0.0035	$2.68x10^{-08}$	0.954
FEV_1	N	*MIR6074*	rs11176001 (12:66409367)	C/A	0.0333	0.0050	$2.63x10^{-11}$	0.0247	0.0047	$1.81x10^{-07}$	0.245
PEF	P	*ALX1/RASSF9*	rs56390486 (12:85719906)	A/G	0.0168	0.0039	$2.11x10^{-05}$	0.0233	0.0037	$4.01x10^{-10}$	0.286
FVC	P	*CRADD*	rs9788269 (12:94194890)	A/G	-0.0208	0.0038	$5.73x10^{-08}$	-0.0090	0.0036	$1.25x10^{-02}$	0.076
FEV_1/FVC	P	*FGD6*	rs113745635 (12:95554771)	T/C	-0.0308	0.0041	$5.57x10^{-14}$	-0.0257	0.0039	$3.33x10^{-11}$	0.410
FEV_1/FVC	P	*SNRPF*	rs7970544 (12:96242109)	T/G	0.0417	0.0043	$3.62x10^{-22}$	0.0470	0.0041	$2.91x10^{-30}$	0.409
PEF	N	*IGF1*	rs972936 (12:102824921)	T/C	0.0260	0.0041	$1.83x10^{-10}$	0.0332	0.0038	$4.82x10^{-18}$	0.255
FEV_1	N	*TBX5*	rs2701110 (12:114669870)	C/A	-0.0250	0.0045	$2.89x10^{-08}$	-0.0275	0.0043	$1.77x10^{-10}$	0.709
FEV_1	P	*TBX3*	rs10850377 (12:115201436)	G/A	-0.0234	0.0036	$4.90x10^{-11}$	-0.0145	0.0034	$1.83x10^{-05}$	0.138
FVC	P	*TBX3*	rs35505 (12:115501127)	A/G	0.0191	0.0037	$1.84x10^{-07}$	0.0246	0.0035	$1.17x10^{-12}$	0.331
FEV_1/FVC	N	*MIR8079*	rs9533803 (13:44820608)	C/T	0.0249	0.0041	$1.14x10^{-09}$	0.0282	0.0039	$4.68x10^{-13}$	0.579
FEV_1	N	*DLEU1*	rs2812208 (13:50707087)	G/C	-0.0607	0.0116	$1.51x10^{-07}$	-0.0600	0.0111	$5.69x10^{-08}$	0.966

Trait	Signal	Gene	SNP (Chr:Pos)	Coded / non-coded allele	Ever smoker			Never smoker			Welch's t-test P
					β	SE	P	β	SE	P	
FVC	N	*LINC00348*	rs803765 (13:71647588)	C/A	0.0219	0.0035	$4.38x10^{-10}$	0.0290	0.0033	$4.40x10^{-18}$	0.212
FEV_1	N	*LINC00382*	rs4885681 (13:80467235)	C/T	-0.0208	0.0038	$3.47x10^{-08}$	-0.0170	0.0036	$2.37x10^{-06}$	0.500
FEV_1/FVC	N	*DOCK9*	rs11620380 (13:99665512)	C/A	0.0357	0.0055	$1.05x10^{-10}$	0.0210	0.0052	$5.60x10^{-05}$	0.082
FEV_1/FVC	N	*MYO16*	rs9634470 (13:109918493)	T/C	-0.0244	0.0039	$2.75x10^{-10}$	-0.0196	0.0037	$8.18x10^{-08}$	0.410
FEV_1/FVC	N	*HAUS4*	rs1951121 (14:23429729)	T/G	0.0216	0.0034	$2.52x10^{-10}$	0.0158	0.0033	$1.26x10^{-06}$	0.287
FEV_1/FVC	N	*MIR5580*	rs74053129 (14:54346010)	G/A	-0.0421	0.0057	$1.10x10^{-13}$	-0.0362	0.0054	$2.45x10^{-11}$	0.467
FEV_1/FVC	P	*BMP4*	rs35107139 (14:54419106)	A/C	0.0265	0.0035	$6.67x10^{-14}$	0.0350	0.0034	$2.95x10^{-25}$	0.149
FVC	N	*VRTN*	rs10141786 (14:74817418)	A/G	0.0221	0.0035	$1.73x10^{-10}$	0.0223	0.0033	$1.00x10^{-11}$	0.959
FEV_1/FVC	P	*LINC00911*	rs1756281 (14:84338431)	A/G	0.0250	0.0037	$9.58x10^{-12}$	0.0238	0.0035	$1.15x10^{-11}$	0.809
FEV_1	P	*TRIP11*	rs11160037 (14:92512143)	A/G	-0.0187	0.0035	$6.68x10^{-08}$	-0.0184	0.0033	$2.11x10^{-08}$	0.961
FVC	P	*RIN3*	rs11621587 (14:93098339)	G/C	-0.0333	0.0044	$2.71x10^{-14}$	-0.0398	0.0041	$7.19x10^{-22}$	0.323
FVC	N	*BMF*	rs34245505 (15:40397191)	C/G	0.0150	0.0043	$4.92x10^{-04}$	0.0274	0.0041	$1.78x10^{-11}$	0.078
FEV_1	N	*IVD*	rs2304645 (15:40716253)	G/C	0.0170	0.0033	$3.05x10^{-07}$	0.0141	0.0032	$8.95x10^{-06}$	0.563
FVC	N	*CHAC1*	rs4924525 (15:41255396)	C/A	0.0185	0.0034	$4.04x10^{-08}$	0.0185	0.0032	$6.92x10^{-09}$	0.995
FEV_1/FVC	P	*RPAP1*	rs2012453 (15:41840238)	A/G	0.0221	0.0034	$7.32x10^{-11}$	0.0252	0.0032	$7.74x10^{-15}$	0.553
FEV_1/FVC	P	*MGA*	rs56383987 (15:41953211)	T/C	-0.0400	0.0075	$8.17x10^{-08}$	-0.0344	0.0071	$1.45x10^{-06}$	0.596
FEV_1/FVC	N	*COPS2*	rs79234094 (15:49409527)	G/A	-0.0204	0.0038	$8.16x10^{-08}$	-0.0334	0.0036	$3.92x10^{-20}$	0.055
FEV_1/FVC	N	*FAM227B*	rs35251997 (15:49706145)	A/T	-0.0412	0.0066	$3.61x10^{-10}$	-0.0610	0.0063	$2.44x10^{-22}$	0.044
FEV_1/FVC	N	*USP3*	rs62012772 (15:63866877)	T/C	-0.0189	0.0043	$1.34x10^{-05}$	-0.0330	0.0042	$3.07x10^{-15}$	0.053
FVC	P	*AAGAB*	rs12917612 (15:67491274)	C/A	0.0220	0.0040	$3.75x10^{-08}$	0.0220	0.0038	$5.43x10^{-09}$	0.997
FEV_1/FVC	P	*THSD4*	rs1441358 (15:71612514)	T/G	0.0622	0.0035	$3.84x10^{-70}$	0.0689	0.0034	$3.06x10^{-93}$	0.233
FEV_1	P	*THSD4*	rs62015883 (15:71803450)	C/T	0.0190	0.0043	$1.16x10^{-05}$	0.0195	0.0041	$2.50x10^{-06}$	0.936
FEV_1/FVC	N	*REC114*	rs7176074 (15:73833600)	G/T	-0.0260	0.0077	$7.40x10^{-04}$	-0.0407	0.0074	$3.80x10^{-08}$	0.183
FEV_1/FVC	P	*SH3GL3*	rs1896797 (15:84274591)	G/A	-0.0280	0.0034	$9.91x10^{-17}$	-0.0291	0.0032	$8.77x10^{-20}$	0.821
FVC	N	*CLUAP1*	rs3751837 (16:3583173)	C/T	0.0265	0.0041	$6.01x10^{-11}$	0.0376	0.0038	$1.18x10^{-22}$	0.096
FEV_1/FVC	N	*GLIS2-AS1*	rs56104880 (16:4361138)	T/C	0.0247	0.0037	$1.97x10^{-11}$	0.0191	0.0035	$5.79x10^{-08}$	0.321
FVC	N	*GRIN2A*	rs11074547 (16:10136889)	T/G	-0.0207	0.0038	$4.90x10^{-08}$	-0.0113	0.0036	$1.81x10^{-03}$	0.130
FEV_1/FVC	P	*TEKT5*	rs78442819 (16:10740982)	G/C	0.0312	0.0043	$4.04x10^{-13}$	0.0371	0.0041	$1.25x10^{-19}$	0.359
FEV_1	P	*IL27*	rs12446589 (16:28870962)	A/G	-0.0131	0.0034	$1.06x10^{-04}$	-0.0106	0.0032	$1.08x10^{-03}$	0.611
FVC	N	*PAPD5*	rs76219171 (16:50188929)	G/A	0.0410	0.0073	$1.59x10^{-08}$	0.0318	0.0068	$2.50x10^{-06}$	0.366
FEV_1/FVC	N	*FTO*	rs35420030 (16:53935407)	T/C	-0.0484	0.0075	$1.06x10^{-10}$	-0.0445	0.0072	$5.20x10^{-10}$	0.712
FEV_1/FVC	P	*MMP15*	rs11648508 (16:58063513)	G/T	-0.0374	0.0036	$1.93x10^{-25}$	-0.0302	0.0034	$1.16x10^{-18}$	0.211
FEV_1	P	*WWP2*	rs8047194 (16:69891510)	G/T	0.0214	0.0034	$1.87x10^{-10}$	0.0216	0.0032	$1.26x10^{-11}$	0.959

Trait	Signal	Gene	SNP (Chr:Pos)	Coded / non-coded allele	Ever smoker			Never smoker			Welch's t-test P
					β	SE	P	β	SE	P	
FEV_1/FVC	P	*CFDP1*	rs11858992 (16:75411445)	A/C	0.0370	0.0034	$4.76x10^{-27}$	0.0416	0.0033	$1.77x10^{-37}$	0.378
FEV_1	P	*WWOX*	rs2345443 (16:78225633)	A/G	0.0216	0.0036	$2.57x10^{-09}$	0.0212	0.0035	$8.39x10^{-10}$	0.935
FEV_1/FVC	N	*LINC00917*	rs12918140 (16:86403821)	G/C	0.0257	0.0053	$1.22x10^{-06}$	0.0199	0.0050	$7.90x10^{-05}$	0.441
FEV_1	N	*MTHFSD*	rs6539952 (16:86579223)	C/A	0.0178	0.0038	$3.90x10^{-06}$	0.0177	0.0037	$1.44x10^{-06}$	0.989
FEV_1/FVC	N	*ATP2A3*	rs8082036 (17:3882613)	G/C	-0.0251	0.0033	$4.66x10^{-14}$	-0.0217	0.0032	$9.69x10^{-12}$	0.497
FEV_1	N	*PITPNM3*	rs4796334 (17:6469793)	G/A	0.0206	0.0033	$7.21x10^{-10}$	0.0097	0.0032	$2.46x10^{-03}$	0.079
FVC	N	*CLDN7*	rs1215 (17:7163350)	A/G	0.0169	0.0048	$4.30x10^{-04}$	0.0250	0.0046	$4.02x10^{-08}$	0.257
FEV_1	N	*TNFSF12-TNFSF13*	rs4968200 (17:7448457)	C/G	-0.0243	0.0048	$3.41x10^{-07}$	-0.0184	0.0045	$5.25x10^{-05}$	0.393
FVC	N	*NCOR1*	rs34351630 (17:16030520)	T/C	0.0115	0.0034	$6.69x10^{-04}$	0.0177	0.0032	$2.82x10^{-08}$	0.246
FEV_1/FVC	P	*SSH2*	rs2244592 (17:28072327)	A/G	-0.0309	0.0033	$2.57x10^{-20}$	-0.0352	0.0032	$3.50x10^{-28}$	0.407
FVC	P	*SUZ12P1*	rs62070648 (17:29210595)	A/G	0.0253	0.0038	$2.06x10^{-11}$	0.0193	0.0036	$7.65x10^{-08}$	0.298
FEV_1/FVC	P	*PSMB3*	rs35246838 (17:36915540)	T/C	0.0403	0.0050	$8.43x10^{-16}$	0.0344	0.0048	$4.92x10^{-13}$	0.418
FVC	P	*FBXL20*	rs8069451 (17:37504933)	T/C	0.0183	0.0039	$2.25x10^{-06}$	0.0201	0.0037	$5.39x10^{-08}$	0.760
FEV_1	P	*MAPT-AS1*	rs79412431 (17:43940021)	G/A	0.0393	0.0042	$3.98x10^{-21}$	0.0441	0.0039	$2.91x10^{-29}$	0.433
FVC	N	*LOC101927166*	rs12945803 (17:46552229)	T/C	0.0175	0.0041	$1.61x10^{-05}$	0.0226	0.0039	$5.47x10^{-09}$	0.406
FVC	N	*ANKFN1*	rs28519449 (17:54195453)	C/T	-0.0156	0.0034	$4.90x10^{-06}$	-0.0264	0.0032	$3.21x10^{-16}$	0.081
FEV_1/FVC	N	*BCAS3*	rs8068952 (17:59286644)	G/C	0.0319	0.0041	$7.43x10^{-15}$	0.0251	0.0039	$1.09x10^{-10}$	0.278
FVC	N	*DDX5*	rs77672322 (17:62497964)	C/T	0.0389	0.0110	$3.94x10^{-04}$	0.0453	0.0103	$1.15x10^{-05}$	0.672
FEV_1/FVC	N	*SMURF2*	rs11653958 (17:62686730)	G/A	-0.0146	0.0039	$1.90x10^{-04}$	-0.0224	0.0037	$1.18x10^{-09}$	0.200
FVC	P	*CASC17*	rs6501431 (17:68976415)	T/C	0.0236	0.0041	$9.83x10^{-09}$	0.0115	0.0039	$3.12x10^{-03}$	0.077
FEV_1	P	*CASC17*	rs6501455 (17:69201811)	A/G	0.0245	0.0034	$2.75x10^{-13}$	0.0336	0.0032	$7.18x10^{-26}$	0.124
FEV_1/FVC	N	*CASC17*	rs996865 (17:69371318)	C/T	0.0379	0.0064	$3.03x10^{-09}$	0.0560	0.0061	$6.27x10^{-20}$	0.060
FEV_1	P	*TSEN54*	rs9892893 (17:73525670)	T/G	0.0154	0.0039	$6.82x10^{-05}$	0.0255	0.0037	$3.74x10^{-12}$	0.114
FVC	N	*ASPSCR1*	rs59606152 (17:79952944)	C/T	-0.0286	0.0057	$5.02x10^{-07}$	-0.0399	0.0053	$5.47x10^{-14}$	0.175
FEV_1	P	*MTCL1*	rs513953 (18:8801351)	A/G	-0.0203	0.0038	$1.01x10^{-07}$	-0.0325	0.0036	$4.53x10^{-19}$	0.067
FEV_1/FVC	N	*VAPA*	rs8089099 (18:10078071)	G/A	-0.0234	0.0038	$4.87x10^{-10}$	-0.0251	0.0036	$2.54x10^{-12}$	0.749
FEV_1/FVC	N	*GATA6*	rs1985511 (18:19816712)	A/T	0.0151	0.0034	$7.49x10^{-06}$	0.0144	0.0032	$7.15x10^{-06}$	0.896
FEV_1	P	*CTAGE1/RBBP8*	rs11082051 (18:20234336)	A/G	0.0099	0.0033	$3.06x10^{-03}$	0.0203	0.0032	$1.67x10^{-10}$	0.087
FEV_1	P	*CABLES1*	rs9947743 (18:20708321)	A/G	-0.0192	0.0041	$2.50x10^{-06}$	-0.0253	0.0039	$7.64x10^{-11}$	0.320
FVC	N	*C18orf8*	rs303752 (18:21074255)	G/A	0.0195	0.0035	$1.66x10^{-08}$	0.0135	0.0033	$4.14x10^{-05}$	0.268
FVC	N	*LOC729950*	rs1668091 (18:22290711)	T/C	-0.0161	0.0036	$8.12x10^{-06}$	-0.0205	0.0034	$2.25x10^{-09}$	0.425
FEV_1	N	*SLC14A2*	rs9807668 (18:42827898)	C/T	-0.0317	0.0057	$2.44x10^{-08}$	-0.0274	0.0054	$4.23x10^{-07}$	0.597
FVC	P	*DCC*	rs12607758 (18:51022606)	T/C	0.0129	0.0034	$1.65x10^{-04}$	0.0171	0.0032	$1.28x10^{-07}$	0.417

Trait	Signal	Gene	SNP (Chr:Pos)	Coded / non-coded allele	Ever smoker			Never smoker			Welch's t-test P
					β	SE	P	β	SE	P	
FVC	N	*LOC101927273*	rs2202572 (18:53566471)	A/C	0.0166	0.0036	$3.12x10^{-06}$	0.0137	0.0034	$5.12x10^{-05}$	0.583
FEV_1	N	*QTRT1*	rs11085744 (19:10819967)	C/T	0.0174	0.0034	$2.22x10^{-07}$	0.0123	0.0032	$1.24x10^{-04}$	0.336
FEV_1/FVC	P	*TSHZ3*	rs9636166 (19:31829613)	A/C	0.0315	0.0050	$4.12x10^{-10}$	0.0407	0.0048	$2.79x10^{-17}$	0.217
FVC	N	*ZFP82*	rs2967516 (19:36881643)	A/G	-0.0131	0.0037	$3.65x10^{-04}$	-0.0159	0.0035	$5.75x10^{-06}$	0.614
FEV_1/FVC	P	*LTBP4*	rs34093919 (19:41117300)	G/A	-0.1593	0.0149	$1.44x10^{-26}$	-0.1545	0.0141	$8.25x10^{-28}$	0.815
FVC	P	*BMP2*	rs2145272 (20:6626218)	A/G	0.0291	0.0035	$7.34x10^{-17}$	0.0272	0.0033	$1.75x10^{-16}$	0.710
FEV_1	N	*LOC101929395*	rs6032942 (20:10745545)	G/C	-0.0166	0.0039	$2.32x10^{-05}$	-0.0175	0.0038	$3.08x10^{-06}$	0.875
FEV_1	P	*ABHD12*	rs2236180 (20:25282608)	T/C	0.0292	0.0043	$6.92x10^{-12}$	0.0111	0.0041	$6.76x10^{-03}$	0.020
FEV_1/FVC	P	*C20orf112*	rs4413223 (20:30858967)	A/G	-0.0252	0.0044	$1.01x10^{-08}$	-0.0211	0.0042	$5.11x10^{-07}$	0.517
FVC	P	*UQCC1*	rs143384 (20:34025756)	A/G	0.0229	0.0034	$1.94x10^{-11}$	0.0291	0.0032	$2.70x10^{-19}$	0.259
FVC	P	*EYA2*	rs12481092 (20:45486817)	C/T	-0.0251	0.0038	$3.16x10^{-11}$	-0.0272	0.0036	$3.46x10^{-14}$	0.701
FVC	P	*SLC2A4RG*	rs4809221 (20:62372706)	G/A	0.0287	0.0036	$2.78x10^{-15}$	0.0303	0.0034	$7.56x10^{-19}$	0.759
FEV_1/FVC	N	*LINC00649*	rs12627254 (21:35368402)	G/T	-0.0314	0.0050	$2.60x10^{-10}$	-0.0413	0.0048	$4.09x10^{-18}$	0.185
FEV_1/FVC	P	*KCNE2*	rs62213732 (21:35675966)	C/T	-0.0199	0.0034	$7.20x10^{-09}$	-0.0285	0.0033	$3.67x10^{-18}$	0.142
FEV_1	P	*MICAL3*	rs1978968 (22:18448113)	C/T	-0.0249	0.0039	$1.62x10^{-10}$	-0.0316	0.0037	$2.54x10^{-17}$	0.268
FEV_1	P	*SCARF2*	rs9610955 (22:20790723)	C/G	-0.0158	0.0042	$1.73x10^{-04}$	-0.0214	0.0040	$9.44x10^{-08}$	0.369
FEV_1/FVC	P	*MN1*	rs2283847 (22:28181399)	T/C	-0.0239	0.0034	$3.77x10^{-12}$	-0.0206	0.0033	$3.13x10^{-10}$	0.532
FEV_1	N	*PPP6R2*	rs113111175 (22:50867711)	C/T	-0.0204	0.0052	$8.24x10^{-05}$	-0.0215	0.0049	$1.30x10^{-05}$	0.882

5.5.2 Risk score interaction analysis

For the weighted genetic risk score by smoking interaction analysis, for the 279 lung function associated signals, the weighting and source of extraction for SNP effects are presented in **Appendix F.** For each of the four traits, genetic effects to determine suitable weights were sourced from UK Biobank and the SpiroMeta consortium approximately equally, with 148 (~53%), 148 (~53%), 156 (~56%) and 159 (~57%) sourced from the SpiroMeta consortium for traits FEV_1, FEV_1/FVC, FVC, and PEF respectively. The genetic effect size estimates and corresponding standard errors are presented in **Table 5.3**, alongside marginal risk score effects in ever and never smokers separately. The only statistically significant interaction effect reaching Bonferroni threshold was for PEF ($p = 0.002$ with $p < 0.0125$ used as a Bonferroni corrected threshold for four tests). The results suggest a stronger quantitative effect on lung function in ever smokers (beta: -0.147, P = $< 2 \times 10^{-16}$) compared with never smokers (beta: -0.137, P = $< 2 \times 10^{-16}$). FEV_1 was significant at a threshold of $p < 0.05$, which is arguably an acceptable threshold to conclude statistical significance given the correlation between traits. The results suggest a stronger quantitative effect on lung function in ever smokers (beta: -0.174, P = $< 2 \times 10^{-16}$) compared with never smokers (beta: -0.168, P = $< 2 \times 10^{-16}$).

Table 5.3 - Weighted genetic risk score and ever-smoking interaction analysis for the 303,619 unrelated European subset in UK Biobank

Risk score effects are presented as change as a proportion of lung function trait SD for a 1-SD change in risk score. Betas (β) and standard errors (SE) are presented to 3 d.p and are on the inverse-normalised scale.

	β	SE	P
FEV_1			
Interaction effect	-0.008	0.004	0.027
Risk score	-0.167	0.002	$< 2 \times 10^{-16}$
Ever smoker	0.024	0.091	0.792
Ever smoker risk score effect	-0.174	0.003	$< 2 \times 10^{-16}$
Never smoker risk score effect	-0.168	0.002	$< 2 \times 10^{-16}$
FVC			
Interaction effect	-0.003	0.004	0.446
Risk score	-0.147	0.002	$< 2 \times 10^{-16}$
Ever smoker	-0.023	0.087	0.794
Ever smoker risk score effect	-0.149	0.003	$< 2 \times 10^{-16}$
Never smoker risk score effect	-0.147	0.002	$< 2 \times 10^{-16}$
RATIO			
Interaction effect	-0.002	0.003	0.532
Risk score	-0.222	0.002	$< 2 \times 10^{-16}$
Ever smoker	-0.176	0.088	0.047
Ever smoker risk score effect	-0.222	0.003	$< 2 \times 10^{-16}$
Never smoker risk score effect	-0.223	0.002	$< 2 \times 10^{-16}$
PEF			
Interaction effect	-0.011	0.004	0.002
Risk score	-0.136	0.002	$< 2 \times 10^{-16}$
Ever smoker	0.147	0.083	0.077
Ever smoker risk score effect	-0.147	0.003	$< 2 \times 10^{-16}$
Never smoker risk score effect	-0.137	0.002	$< 2 \times 10^{-16}$

5.6 Discussion

This chapter aimed to identify whether any of the 279 signals associated with lung function in the GWAS of UK Biobank and SpiroMeta had genetic effects which were modified by smoking behaviour, thus producing a SNP by smoking interaction. A further aim was to determine whether the combined genetic effect of all 279 SNPs was dependent on smoking behaviour, with contribution of all SNPs represented by the calculation of a genetic risk score for each individual.

For the 279 signals identified for producing a marginal genetic effect, none showed evidence of a statistically significant interaction with ever-smoking ($p < 2 \times 10^{-4}$, Bonferroni corrected threshold for 279 tests, with six signals nominally statistically significant at $p < 0.05$). For the genetic risk score interaction analysis there was however a statistically significant interaction effect with the combined effect of all 279 SNPs (represented by a weighted genetic risk score) on PEF ($p = 0.002$ with $p < 0.0125$ statistically significant), with a larger deleterious effect observed on lung function as risk score increased for ever smokers compared with never smokers. Given the correlation between the four lung function traits, it is arguable whether there needs to be any multiple testing correction. Using a threshold of $p < 0.05$ would also identify FEV_1 as a statistically significant result trait for a wGRS by smoking interaction. The magnitude of both interaction effects were however quite small (≈ 0.01). Evidence of a genetic risk score by smoking interaction affecting lung health has been previously documented, with a study by Aschard et al. (178) finding statistical evidence of an interaction effect between ever/never smoking and an unweighted genetic risk score (calculated using the genetic effects from 26 loci associated with traits FEV_1 and FEV_1/FVC) on lung function trait FEV_1/FVC. In addition, in the same paper which reports the 279 signals associated with lung function (32,33), statistical evidence was found for a genetic risk score by smoking interaction associated with COPD, with a slightly smaller GRS effect on COPD in ever smokers (observed odds ratio of 0.96).

Analysing the 279 SNPs with a strong marginal genetic effect on lung function and COPD for interactions with smoking behaviour, did not identify any novel gene-smoking interactions, of which identification is the main aim of this thesis. That does not however mean that this result holds no significance at all, as it has clinical implication for risk prediction. Specifically, determining the absolute risk of poor lung function and COPD for individuals in the population, can be modelled using the assumption of equal genetic risk in ever and never smokers with confidence. In contrast, had any of these signals produced interaction effects, then it would have suggested that subgroup specific risk score computation and absolute risk prediction would have been more appropriate. This result also suggests that although a SNP may have a strong association with lung function, it is not necessarily a prime suspect for contributing an interaction effect. It is important to note that the primary aim of the analysis undertaken, which reported the 279 lung function associated signals, was to identify SNPs

producing marginal genetic effects for lung function (for which SNPs were heavily scrutinised with the use of a strict genome-wide threshold) in both ever and never smokers. The study was not designed to screen SNPs for an interaction analysis. Therefore, this would not be a fair reflection on the outcome of using marginal screening for interaction analysis, as if this was the aim, implementation would likely make use of a much more relaxed threshold for SNP screening (91). This would allow for the inclusion of SNPs with much smaller and less statistically significant marginal effects, broadening the subset for potential interaction effect producing SNPs in the interaction analysis stage.

The risk score result suggests that although no particular SNP individually interacts with ever/never smoking, there is statistical evidence that the combined effect (presented by a weighted genetic risk score) does. Although it must be noted that the effect was small. The subject of identifying associations between combined SNP effects (which can be collated using different criteria for example by functional relevance, marginal association or due to lying within the same gene) and disease is well discussed in the literature and is particularly documented for the analysis of rare variation (MAF < 1%)(8,112), specifically when detecting small effects for which testing SNPs individually will be largely underpowered. Similarly, in an interaction setting, the use of a combined effect genetic risk score for interaction analysis rather than a univariate interaction analysis which tests each SNP individually has been suggested (35,161). Therefore, with regards to gene-smoking interaction effects on lung function, a potential area for exploration would be to look at whether the combined effects of particular SNPs (with regards to a particular SNP characteristic) differs by ever/never smoking status. Such findings would enable the identification of subgroups in the general population for which their predisposed genetic makeup results in them being more susceptible to the effects of smoking.

5.7 Conclusion

In conclusion, none of the 279 lung function signals identified due to statistical evidence of a marginal genetic effect, showed statistical evidence of an interaction with ever-smoking individually. This allows us to predict the absolute *genetic* risk for ever and never smokers equally, and suggests that current risk prediction models do not

require subgroup specific genetic risk score computation, due to underlying interaction effects between SNPs and smoking behaviour. However, there did appear to be evidence of an ever smoking interaction with the weighted risk score for all of these identified SNPs, specifically for outcomes PEF and FEV_1. Such a result questions whether SNPs with strong marginal effects (using a stringent threshold) are contenders for the contribution of interaction effects, and further work to determine how genetic effects on lung function differ between ever and never smokers may benefit from considering the combined rather than individual effects of SNPs.

The next chapter removes the restrictions applied to the analyses in chapter 4 and chapter 5 regarding only considering a subset of SNPs for analysis, and presents an interaction analyses on a genome-wide scale.

Chapter 6 Genome-wide gene-smoking interaction analysis in UK Biobank

6.1 Introduction

In previous chapters, gene-environment interaction analysis was confined to a subset of variants along the genome, with different hypotheses determining their candidacy. In chapter 4, analysis focussed on regions of the genome known to contain lung function associated variants and thus harbouring genes believed to be relevant for lung function. This involves testing all SNPs within each region, as well as considering the lung function associated sentinels themselves (of which there were 97 at the time of analysis (43-50)). The theory behind this approach was that given these genes have already been implicated for a role in lung function, it is possible that such genes could contain interaction signals (that may be independent of the marginal effect signals which led to the gene's discovery). In chapter 5, the concept of focussing on variants or regions known to be associated with lung function was extended, to consider an increased number of associated variants, representative of the progress made in marginal genetic effect research since the work in chapter 4, including a number of COPD only associated signals (32,33,51,52,54,59,60,180). 279 variants were considered. Additionally, the combined interaction effect of all 279 variants with smoking was explored with the use of a weighted genetic risk score, to determine whether there was an interaction between the combined genetic effect of all lung function associated variants and smoking exposure.

The purpose of this chapter is to extend the previous work discussed by undertaking a genome-wide hypothesis-free analysis of gene-smoking interaction, in the form of a GWAS or GWIS (genome wide interaction study). This approach removes any assumptions or criteria to determine candidacy for analysis, and simply explores the whole (imputed) genome for variants whose effect on lung function is dependent on smoking exposure. This removes the reliance on a previously identified association between the variants and lung function. This is a novel undertaking as there has not yet been a genome-wide analysis of gene-smoking interaction effects on lung function in a resource as large as UK Biobank. By considering all regions of the genome, the aim of this non-restrictive analysis is to identify interaction effects on a genome-wide scale, to

aid in our understanding of how genetics and smoking behaviour interact to affect lung function.

In order to produce a comprehensive analysis of the effect of gene-smoking interactions on lung function, phenotypes FVC and PEF were also analysed in addition to FEV_1 and FEV_1/FVC (the phenotypes analysed in chapter 4, which are the primary lung function traits used for COPD diagnosis). All phenotypes are defined in the lung function quality control section in chapter 4 section 4.2.

6.2 Quality control of phenotype data and resulting sample exclusion

Phenotype (lung function) and genotype sample quality control was undertaken as described in chapter 4 (sections 4.2 – 4.3). To summarise, spirometry for the UK Biobank data was undertaken using a Vitalograph Pneumotrac 6800. Each individual produced a minimum of two blows (unless unable to due to extenuating circumstances), with a third blow performed if the first two were not consistent or reproducible as determined by the spirometry equipment. Problematic blows (and samples as a result) were removed, for which detailed exclusion criteria are presented in section 4.2. For each sample, values of FEV_1 and FVC were chosen as the maximum (reproducible) measurement contributed by that individual. PEF values for each individual were derived from the blow curves (volume-time curves), due to an error in the values provided by UK Biobank (see chapter 5 section 5.3 for more detailed information). Samples were then removed on the basis of problematic genotype data and K-means clustering to determine a European sample, before related individuals were removed to give a final unrelated European data set (as in chapter 4 section 4.4). The sample quality control previously undertaken in chapter 4 with regards to phenotype and genotype QC is summarised in **Figure 6.1**.

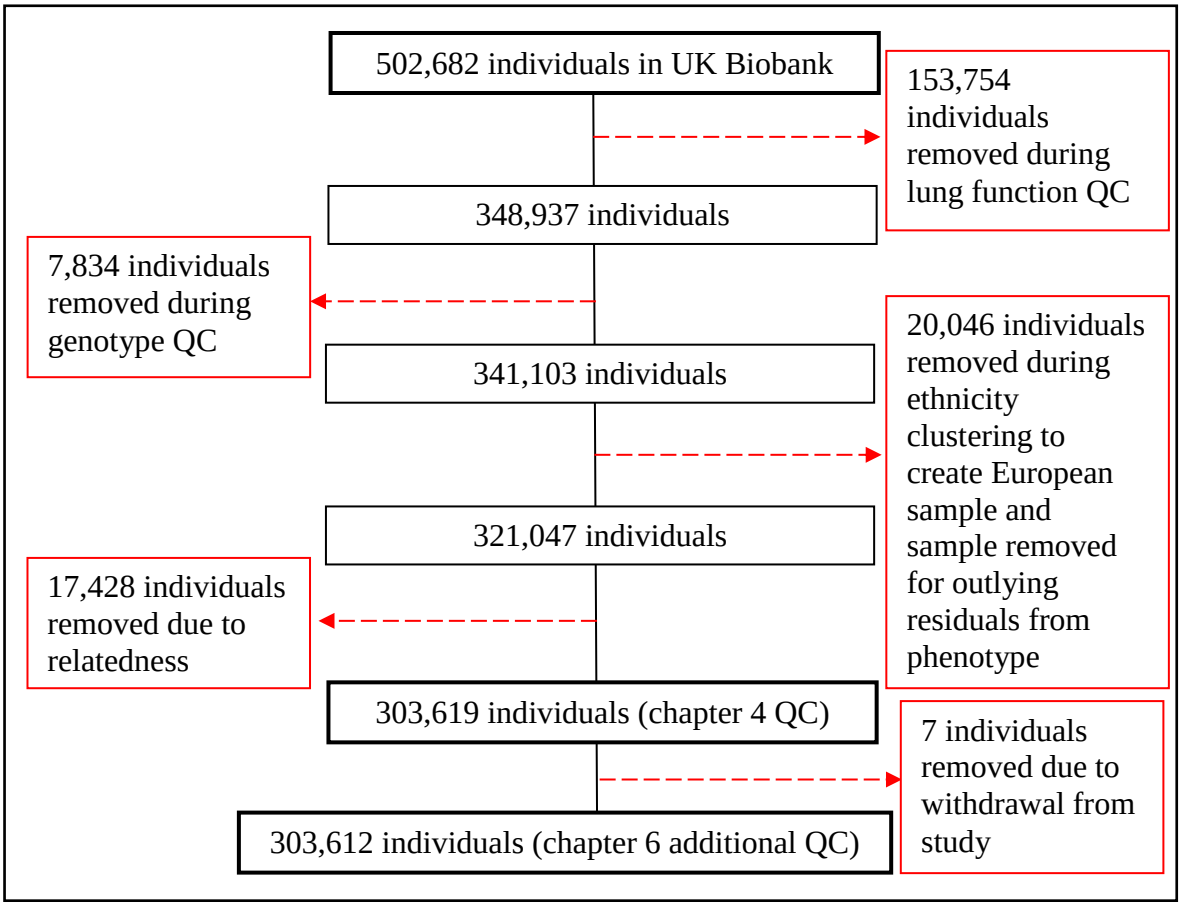


Figure 6.1- Summary of phenotype QC, genotype QC and relatedness sample exclusion
Exclusion to produce sample of 303,619 individuals is presented in full in chapter 4, sections 4.2 - 4.4)

As a result of the sample quality control previously undertaken, 303,619 unrelated European individuals remained.

Upon being recruited to the UK Biobank resource, individuals were informed that should they no longer want their data to be used, they can withdraw at any time. Since the work undertaken in chapter 4, UK Biobank provided a list of 81 participant IDs corresponding to individuals that have since withdrawn from the study. Of these 81 individuals, seven contributed to the unrelated European sample of 303,619 individuals and thus were removed prior to genome-wide interaction analysis (**Figure 6.1**). Therefore, the final sample size for the genome-wide interaction analysis was 303,612.

6.3 Methods: Genome-wide interaction analysis

6.3.1 Quality control of genotype data

Quality control was also applied at the SNP level. The following criteria were used to remove SNPs:

- Minor allele count (MAC) ≤ 3
- Multi-allelic (had more than 2 alleles at SNP locus)
- INFO (imputation quality score) value less than 0.5
    - A more stringent filter was applied for low frequency variants (defined as MAF < 5%) with INFO score less than 0.8 removed
- MAF less than 0.05%
    - To reduce the "noise" or over inflation from extremely rare variants in the data, which can affect the ability to identify independent signals using signal selection software (explained further in section 6.3.4).

6.3.2 Phenotype adjustment and analysis

The four phenotypes (FEV_1, FEV_1/FVC, FVC and PEF) were adjusted for sex, age, age^2 and height before inverse normalisation of traits (the residuals from the regression model were ranked and converted into normal z-scores). Genotype array and ten principal components were adjusted for during the analysis stage (to account for the differing genotype array used for a subset of individuals and population structure respectively), with analysis undertaken using the mean genotype (expected allele dosage) test in QUICKTEST (171,172) using the following model:

$$Phenotype_i = \beta_0 + \beta_1 G_i + \beta_2 S_i + \beta_3 G_i S_i + \varepsilon_i$$

where G_i is the genotype for individual i, and S_i is the ever/never smoking status. The p-value returned corresponds to a test of $\beta_3 = 0$, the interaction effect (GS) between SNP (G) and ever/never smoking (S) respectively. Genotypes were input in dosage format (taking a continuous value from 0 to 2).

For each trait genome-wide inflation factors were calculated using equation (1.3) presented in section 1.1.3.3 and p-values were adjusted (re-scaled) if calculated lambdas were extreme (> 1.1).

6.3.3 Threshold to determine statistical significance of interaction effect

To determine statistical significance for the interaction test for a particular SNP, a threshold of $p < 5 \times 10^{-6}$ was used. This is more relaxed than the threshold typical of a GWAS of marginal effect, however this is chosen deliberately due to there being lower power (thus a larger sample size requirement) to detect and observe interaction effects, in comparison to a GWAS of marginal genetic effects (71). This was therefore chosen as a suitable threshold for the identification of interaction signals which could then be taken forward into a replication stage.

6.3.4 Selection of independent interaction signals

The term signal here refers to a group of associated SNPs (correlated with each other due to patterns of LD), where the sentinel SNP is defined as the SNP with the smallest p-value in that LD group for the association test.

To determine the number of independent signals associated with the four lung function phenotypes, the following process was applied. Firstly, all genome-wide SNPs associated with each trait at threshold $p < 5 \times 10^{-6}$ were identified. The first sentinel was then chosen as the most statistically significant SNP. All SNPs $\pm$1Mb of this sentinel SNP were then assigned to a 2Mb region, and the next most significant SNP outside of this 2Mb region was chosen as the next sentinel SNP. This was repeated until all SNPs meeting the threshold of $p < 5 \times 10^{-6}$ had been accounted for. This produced a set of sentinels (and corresponding 2Mb regions) which were taken forward to the next signal selection step.

Due to the possibility of observing more than one independent signal within each 2Mb region, further exploration was undertaken to identify possible secondary (or tertiary) independent signals. Within each 2Mb region, a stepwise model selection process was

applied using GCTA-cojo (181,182), with a threshold of $P < 5 \times 10^{-6}$. This approach identifies multiple signals by choosing a sentinel SNP, then re-calculating p-values for the remaining SNPs conditional on this sentinel. Independence of SNPs was determined through calculation of LD using the genotype data for the full sample of 303,612 individuals. Region plots were also produced using LocusZoom (29) for each defined region, to check the plausibility of results returned by GCTA-cojo.

Finally, after producing a set of *within-trait* independent signals, SNPs were then assessed for independence *across-traits*. Due to correlation of lung function phenotypes, the same SNP or signal could be associated with more than one lung function phenotype. In this scenario, the trait taken forward corresponds to the one with the strongest association (the trait which has the smallest p-value for the interaction test) with the given signal or SNP.

6.3.5 Follow-up of identified interaction signals

6.3.5.1 Association of interaction signals with smoking

The final set of independent (both within trait and across trait) interaction signals were then explored for association with smoking to ensure smoking was not the driving mechanism behind the observed interaction signal. As described before in section 4.6.7.2, the smoking phenotypes that were available were smoking initiation (SI) with N = 275,596 (123,890 ever smokers and 151,706 never smokers), smoking cessation (SC) with N = 123,851 (25,905 current smokers and 97,946 former smokers) and number of cigarettes smoked per day (CPD) with N = 80,015. Phenotypes were inverse normalised after adjusting for age, age^2, sex and the first 15 PC's. Analysis was undertaken using a mixed model in BOLT-LMM version 2.3 (166). Statistical significance for an association with smoking behaviour was determined with the use of a Bonferroni corrected threshold for the number of SNPs tested.

6.3.5.2 Replication of signals using the SpiroMeta consortium

Interaction signals passing all previous steps were then tested for replication as before using the meta-analysed summary statistics from the SpiroMeta consortium (see **Table**

4.9 in chapter 4 for contributing studies and imputation panels), with a maximum sample size of 71,067 individuals (37,411 ever smokers and 33,656 never smokers).

As before, summary statistics from the ever smoker and never smoker stratified meta-analysis were tested for a statistical difference using Welch's t-test with test statistic and degrees of freedom calculated as follows:

$$t = \frac{\beta_1 - \beta_2}{\sqrt{{se_1}^2 + {se_2}^2}} \qquad d.f. = \frac{({se_1}^2 + {se_2}^2)^2}{\frac{{se_1}^2}{n_1 - 1} + \frac{{se_2}^2}{n_2 - 1}}$$

where β_1 and β_2 are the effect sizes for each smoking group, se_1 and se_2 are the standard errors and n_1 and n_2 are the respective sample sizes (N effective sample size for each SNP). Statistical significance was determined with the use of a Bonferroni corrected threshold for the number of SNPs tested.

A flow chart summarising the methods for the genome-wide interaction analysis is given in **Figure 6.2**.

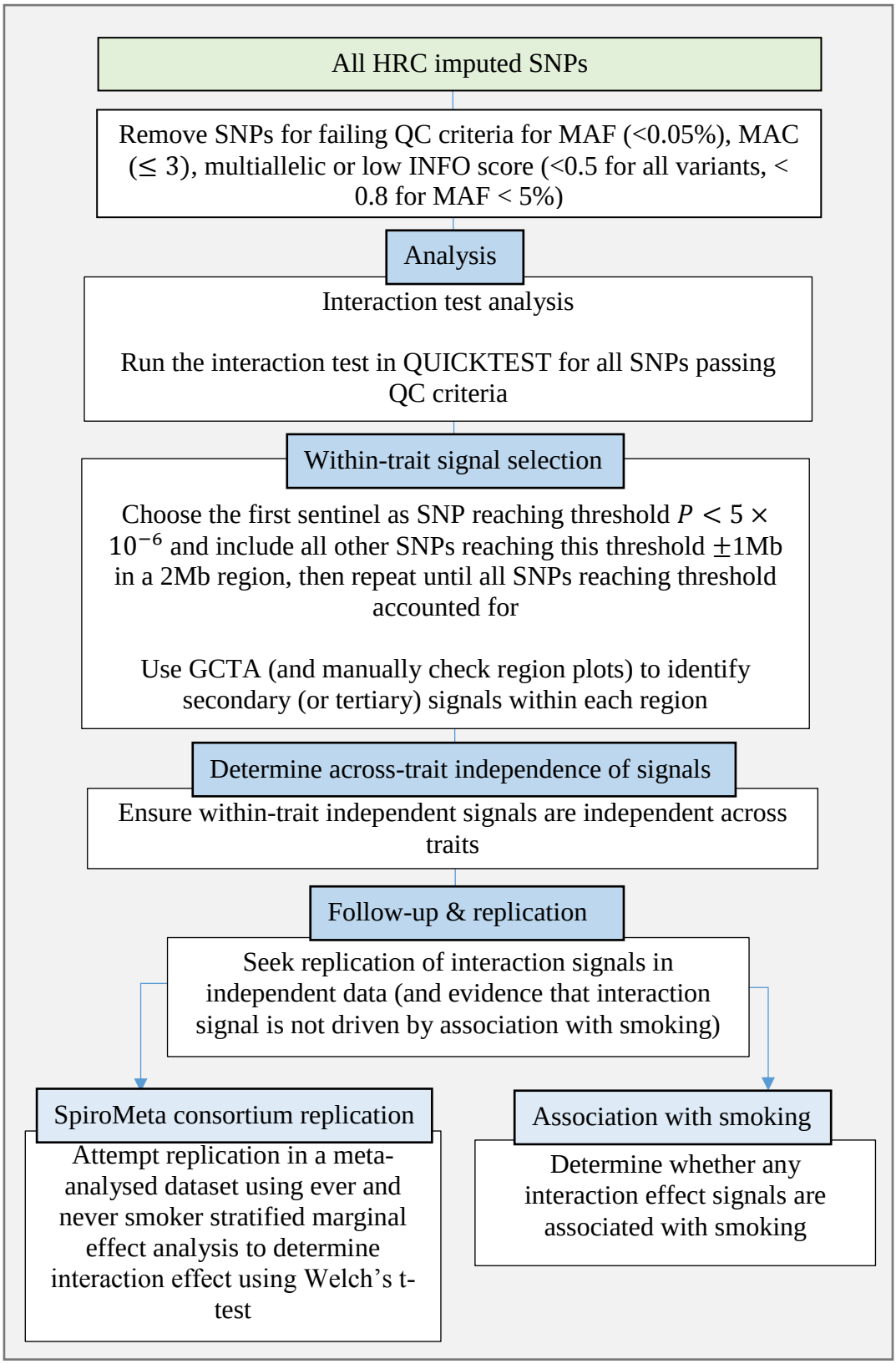


Figure 6.2 - Flowchart to present the summary of methods for the genome-wide gene-smoking interaction analysis

6.4 Results: Genome-wide interaction analysis

6.4.1 Summary of lung function phenotypes and covariates

The summaries for the analysed phenotypes and the adjusted covariates for the total sample, and also stratified by ever and never smoker subgroups, are presented in **Table 6.1**. There were less males (~45%) in the analysed sample, and this was more extreme in the never smoker subgroup (~39% male). The average age was 56.5. Average values for lung function phenotypes were marginally lower for the ever smoker subgroup when compared with the never smoker subgroup for FEV_1 and FEV_1/FVC (with $p = 3.69 \times 10^{-29}$ and $p = 0$ observed when using a two sample t test in R). For FVC and PEF however, values were higher for ever smokers with p values for a difference in means of $p = 1.55 \times 10^{-42}$ and $p = 8.99 \times 10^{-8}$ respectively.

Table 6.1- Phenotype and covariate summaries for the full sample of 303,612 individuals, and also stratified by ever smoking

Variable	Total (n = 303,612)	Ever smokers (n = 139,285)	Never smokers (n = 164,327)
Sex - n(%) male	135,475 (44.6)	71,523 (51.4)	63,952 (38.9)
Age - mean(sd)	56.47 (7.98)	57.29 (7.85)	55.78 (8.01)
Standing height – cm, mean(sd)	168.61 (9.13)	169.40 (9.05)	167.93 (9.14)
FEV_1 – Litres, mean(sd)	2.844 (0.759)	2.828 (0.767)	2.859 (0.752)
FEV_1/FVC – mean(sd)	0.760 (0.064)	0.750 (0.069)	0.769 (0.058)
FVC – Litres, mean(sd)	3.742 (0.963)	3.768 (0.961)	3.720 (0.964)
PEF – Litres/min, mean(sd)	406.716 (117.427)	407.958 (120.172)	405.664(115.038)

6.4.2 SNP removal and SNP summary

For the sample of 303,612 unrelated European individuals, there was a total of 37,310,006 HRC panel (67) imputed variants available for genome-wide interaction analysis. After removing variants on the basis of quality control criteria presented in section 6.3.1, 8,647,748 variants remained. The variant QC process which presents the number of variants removed for each QC criteria is shown in **Figure 6.3**.

Of the remaining 8,647,748 variants, 990,249 were rare (MAF < 1%), 2,256,033 were low frequency (1% ≤ MAF < 5%) and 5,401,463 were common (MAF > 5%), with the

distribution represented in **Figure 6.4.** Imputation quality was high for the analysed variants with 8,627,671 of the 8,647,748 (i.e. 99.8%) variants having an imputation quality of at least 0.8 (**Figure 6.5**).

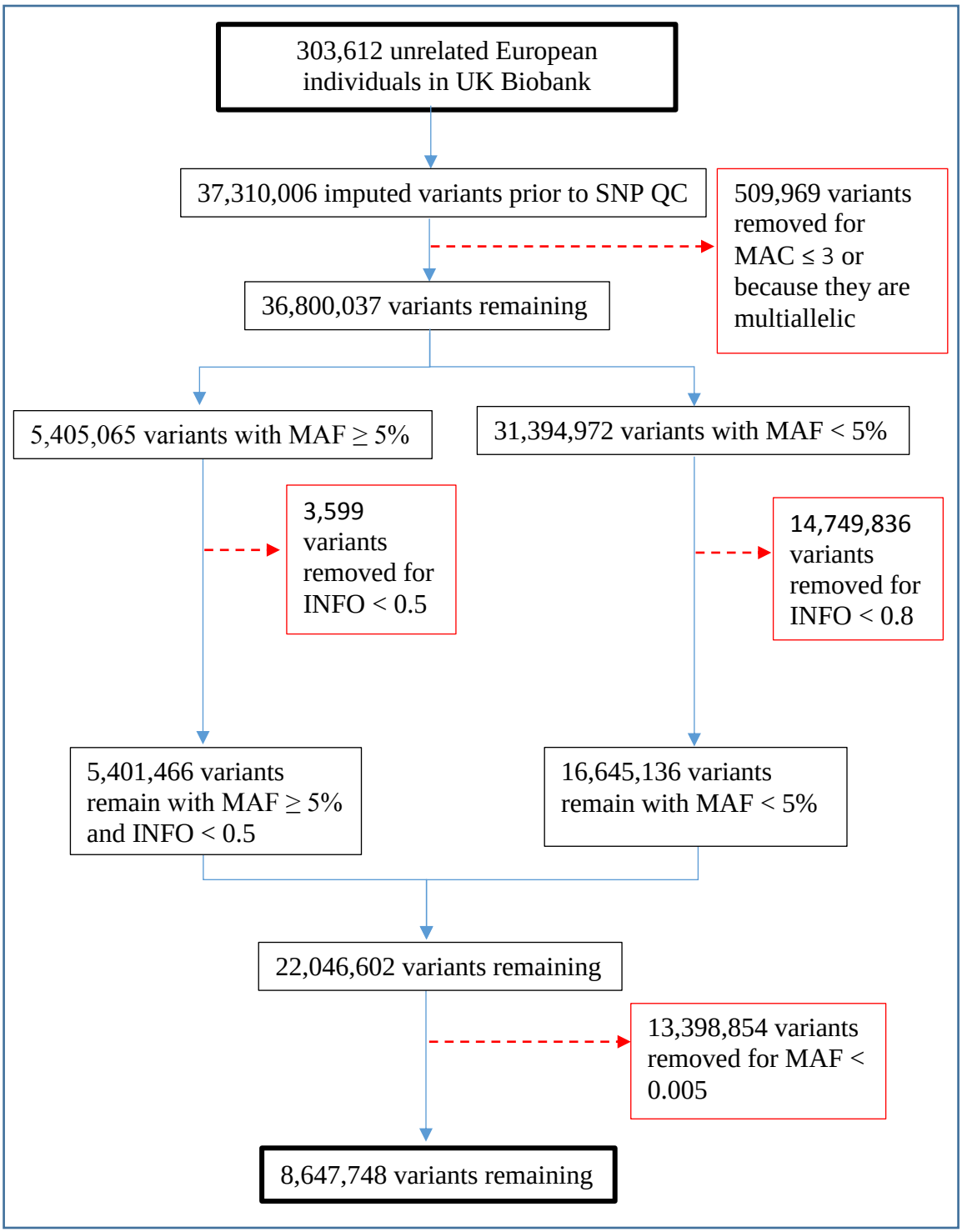


Figure 6.3 - SNP exclusion for HRC imputed UK Biobank data
MAC – minor allele count, MAF – minor allele frequency, and INFO – imputation score

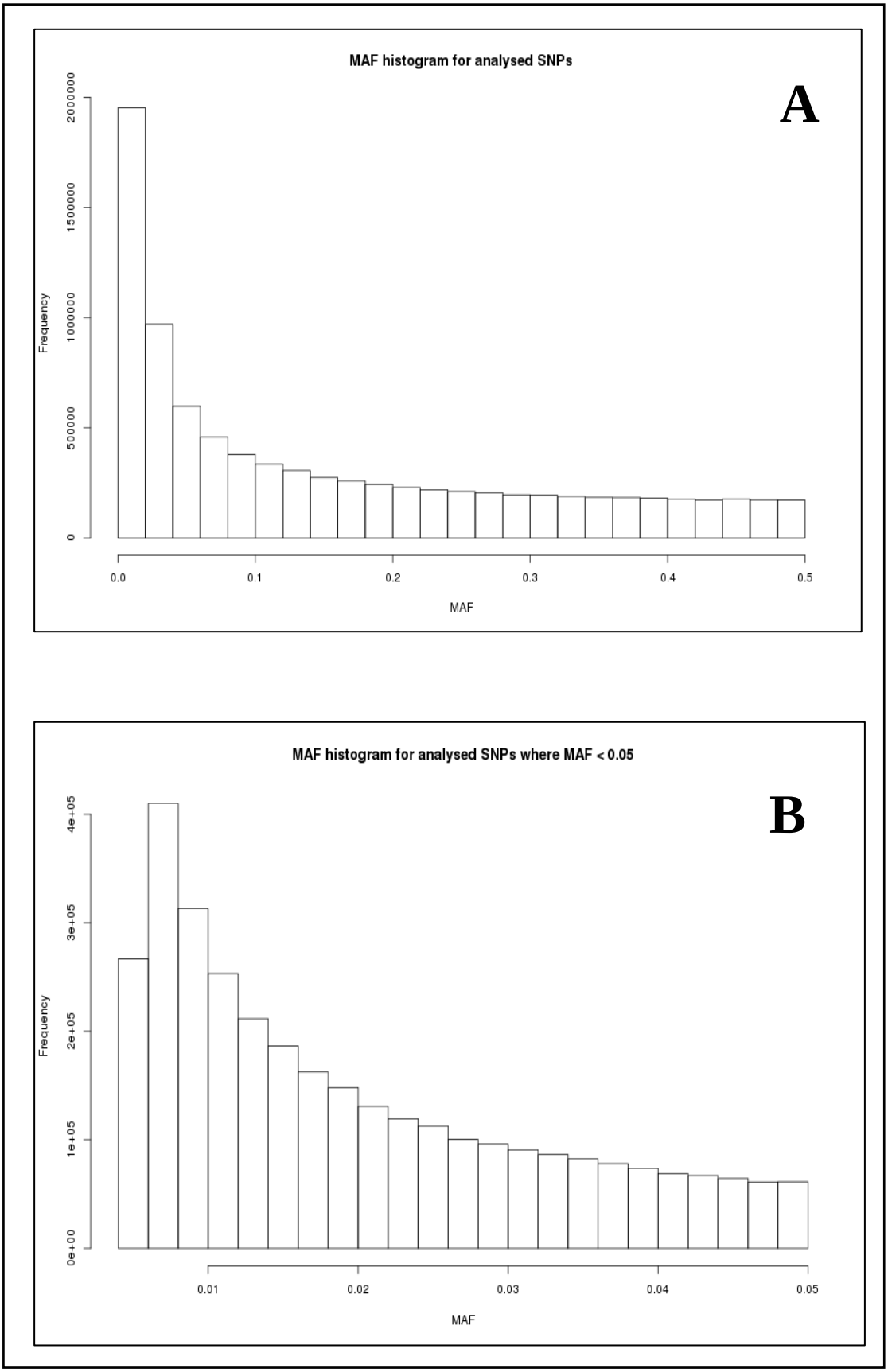


Figure 6.4 - MAF distribution for the 8,647,748 HRC imputed variants passing SNP QC, analysed as part of the genome-wide interaction analysis

For (A) the full MAF range and (B) MAF < 5%)

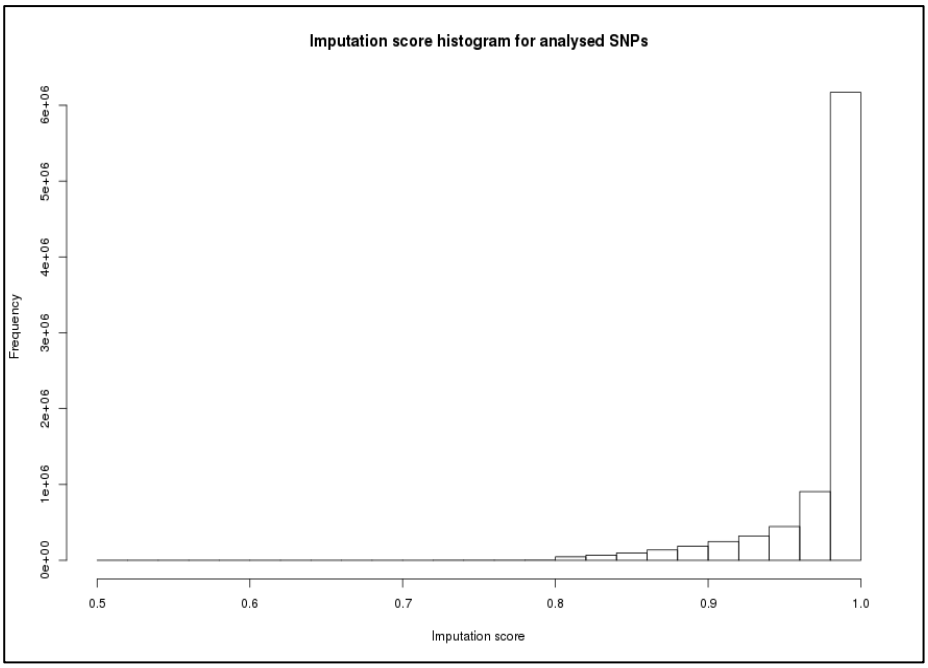


Figure 6.5 - Imputation distribution for the 8,647,748 HRC imputed variants passing SNP QC, analysed as part of the genome-wide interaction analysis

6.4.3 Interaction analysis results

Genome-wide interaction analysis was undertaken for the 8,647,748 variants and 303,612 unrelated European individuals that passed quality control steps.

The genomic control inflation factor, lambda (λ), was calculated for each trait and all analysed MAFs (MAF > 0.005), and did not indicate over inflation of observed test statistics (compared with the expected test statistics), beyond the inflation observed for potentially true associated signals. Lambda values of 1.0495, 1.0385, 1.0265 and 1.0458 were observed for traits FEV_1, FEV_1/FVC, FVC and PEF respectively. QQ plot inflation due to potentially associated interaction signals was particularly evident for traits FEV_1, FEV_1/FVC and PEF, due to a large deviation away from the constant expected gradient illustrated by a red line in **Figure 6.6**. The Manhattan plots for all traits are presented in **Figures 6.7 – 6.10**. It appeared as though the prominent observable inflation for traits FEV_1, FEV_1/FVC and PEF were driven by strongly associated variants on chromosome 15, which although appeared present for FVC, was not as extreme (illustrated by the smaller peak and smaller –log(p-value) observed in the chromosome 15 region of the FVC Manhattan plot, when compared with the plots for the other traits).

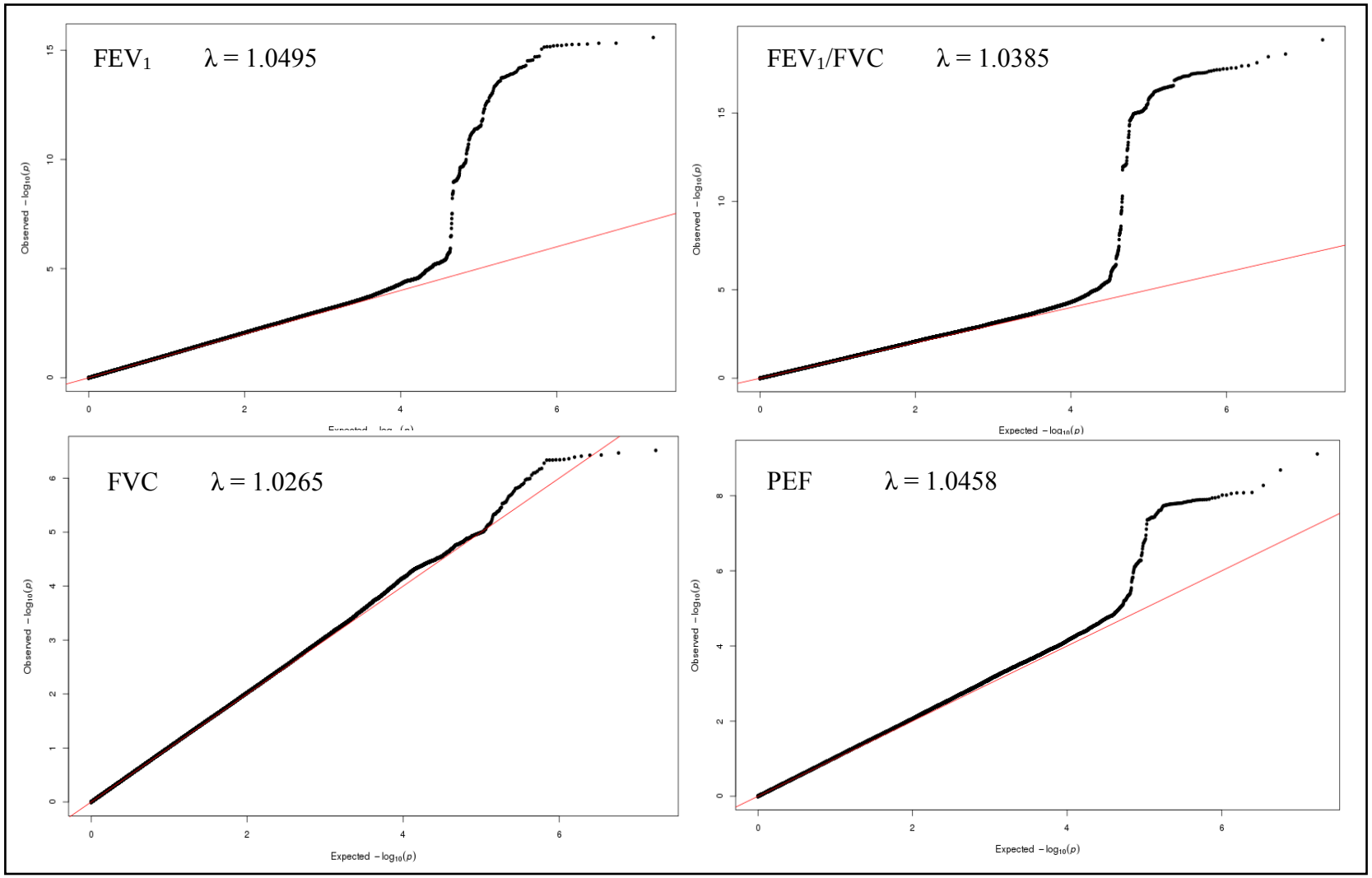


Figure 6.6 - QQ plots for each trait for the genome-wide interaction analysis results

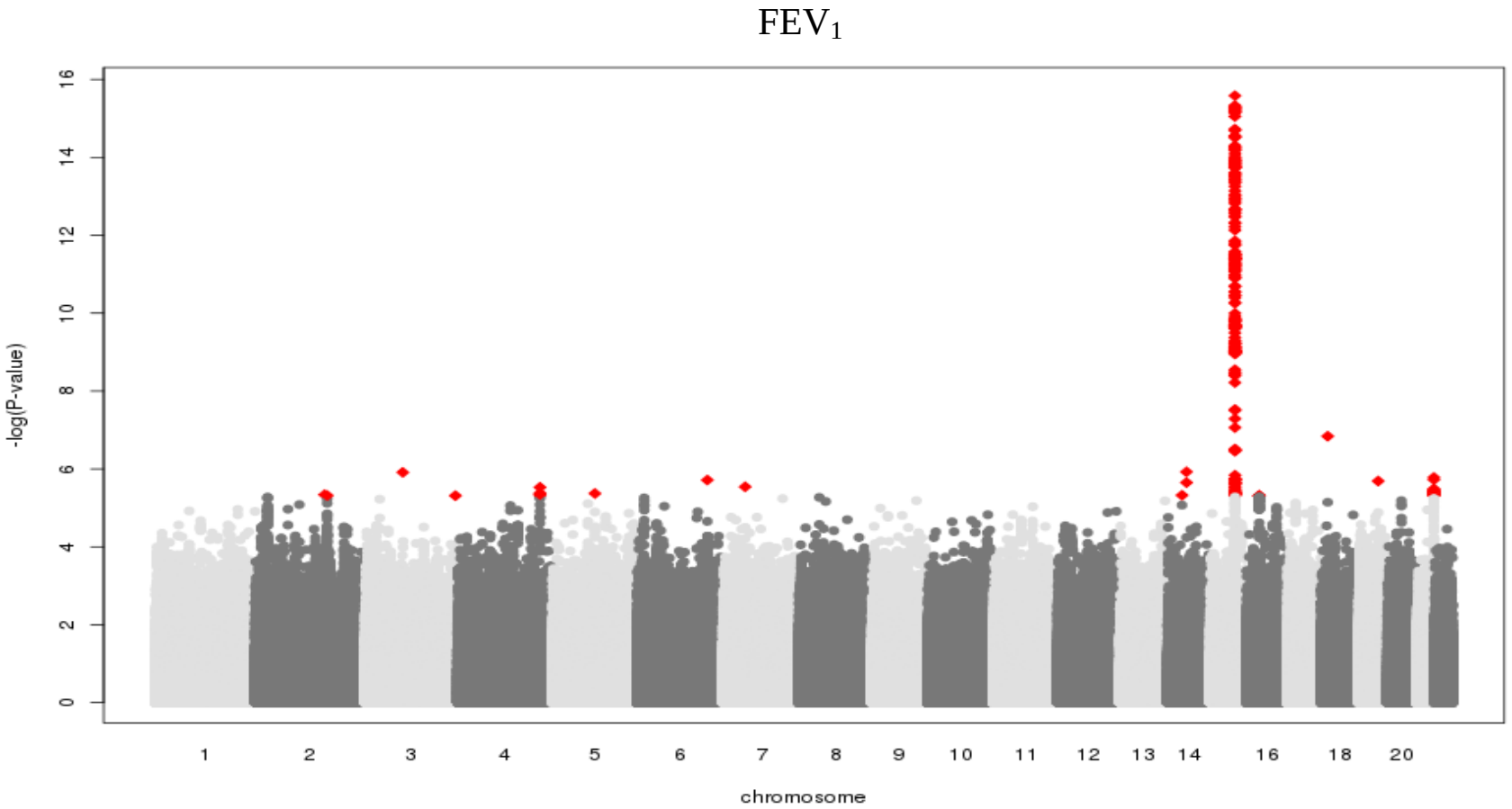


Figure 6.7 - Manhattan plot for FEV_1 lung function phenotype for the genome-wide interaction analysis. Points in red are those which meet the threshold of $p < 5 \times 10^{-6}$

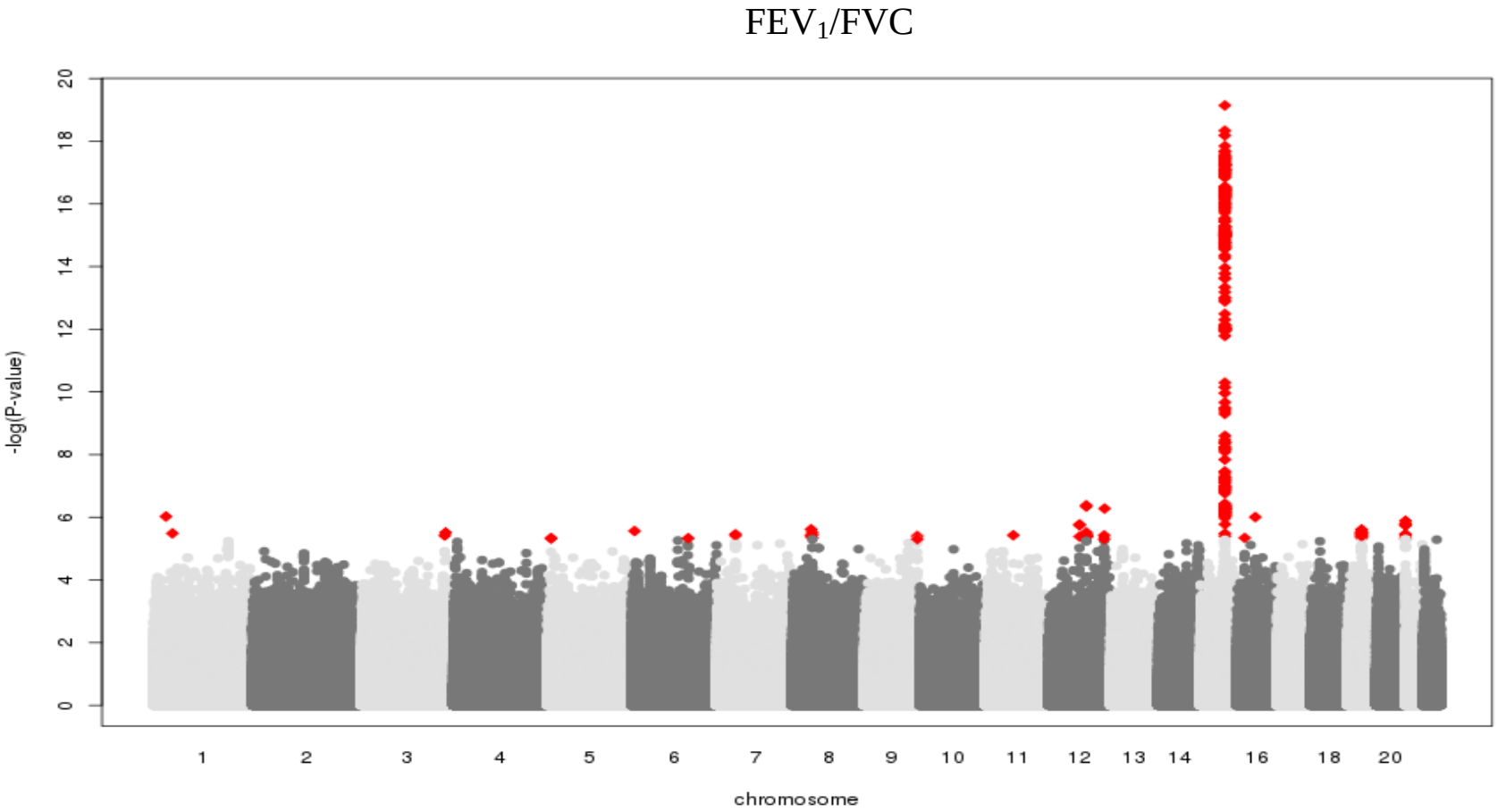


Figure 6.8 - Manhattan plot for FEV_1/FVC lung function phenotype for the genome-wide interaction analysis. Points in red are those which meet the threshold of $p < 5 \times 10^{-6}$

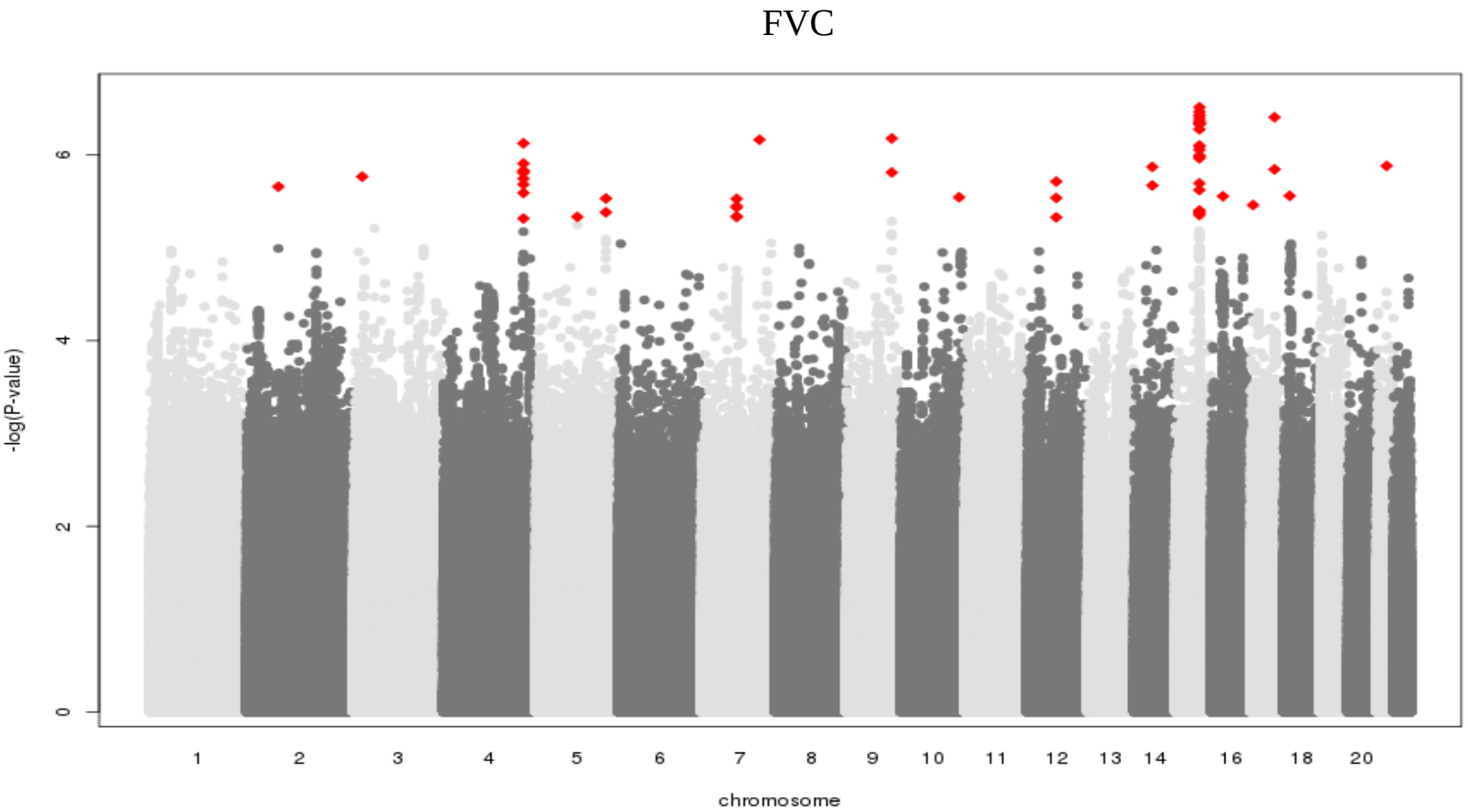


Figure 6.9 - Manhattan plot for FVC lung function phenotype for the genome-wide interaction analysis. Points in red are those which meet the threshold of $p < 5 \times 10^{-6}$

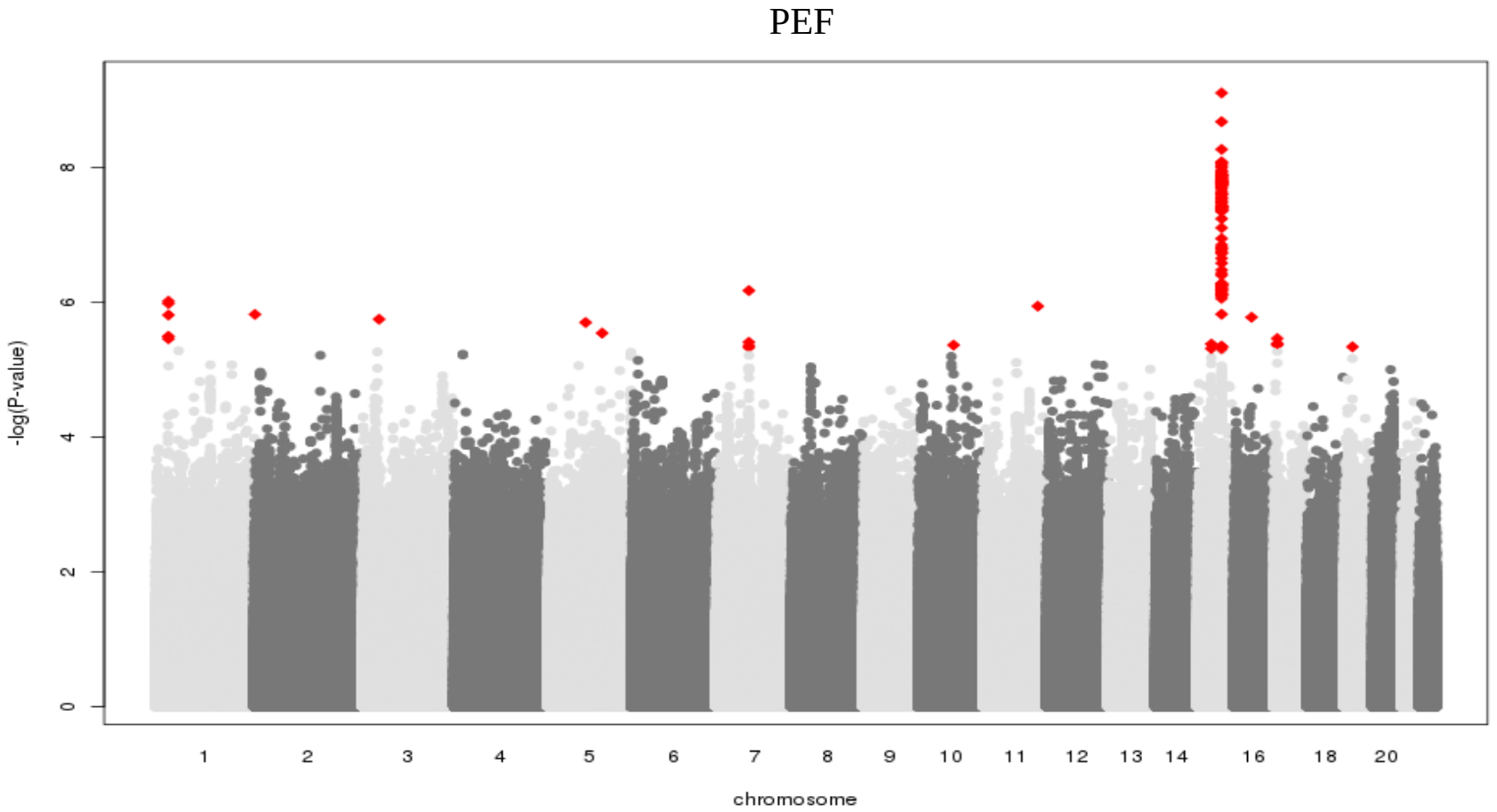


Figure 6.10 - Manhattan plot for PEF lung function phenotype for the genome-wide interaction analysis. Points in red are those which meet the threshold of $p < 5 \times 10^{-6}$

6.4.4 Selection of independent interaction signals

Preliminary signal selection using the process outlined in section 6.3.4, identified 64 sentinels with statistically significant smoking interaction effects at $p < 5 \times 10^{-6}$, thus resulting in the identification of 64 2Mb regions for secondary signal exploration. Of these 64 sentinels identified, 15, 19, 17 and 13 (within-trait) regions corresponded to traits FEV_1, FEV_1/FVC, FVC and PEF respectively.

Utilising both GCTA and a manual check of region plots to identify further potential independent signals within each 2Mb region, highlighted two secondary signals, both in the chromosome 15 region for traits FEV_1 and FEV_1/FVC. This resulted in an increase in the number of identified signals from 15 to 16, and 19 to 20, for FEV_1 and FEV_1/FVC respectively. Therefore, a total of 66 independent signals were identified with interaction results. The process is summarised in **Figure 6.11**.

The 66 within-trait independent smoking-interaction signals were then checked for across trait independence, with dependency possible due to the correlation between phenotypes as presented in **Figures 6.12** and **6.13**. This was explored using genome position, region plots and LD calculation between SNPs (to avoid retaining duplicate signals across traits). 14 of the within-trait independent interaction signals were identified for an association at $p < 5 \times 10^{-6}$ with more than one trait (either the same SNP arising more than once or SNPs associated as part of the same signal). These 14 signals were consolidated into six independent signals across traits, when selecting the trait with the most statistically significant interaction effect (smallest p-value for the interaction test) (**Figure 6.14**). The SNPs identified due to being associated with more than one phenotype, and the independent SNPs chosen to represent them are presented in **Appendix G**. This left 58 gene-smoking interaction signals that were independent both *within* each trait and also *across* the four analysed traits (12 for FEV_1, 20 for FEV_1/FVC, 14 for FVC, and 12 for PEF).

Figures 6.15 and **6.16** present examples of signal selection in two separate regions, for a region where more than one signal was identified, and for a region where only one signal was identified respectively.

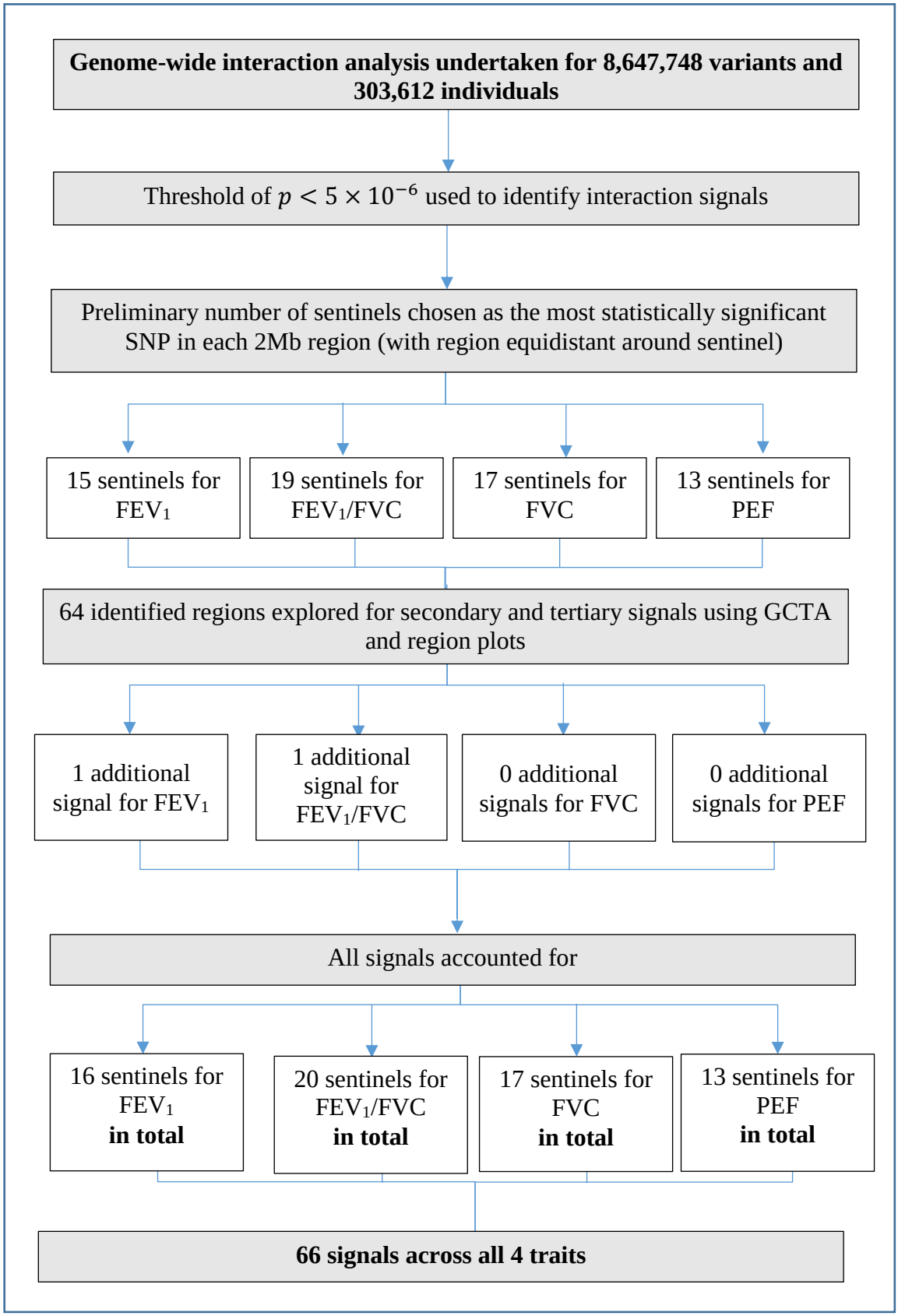
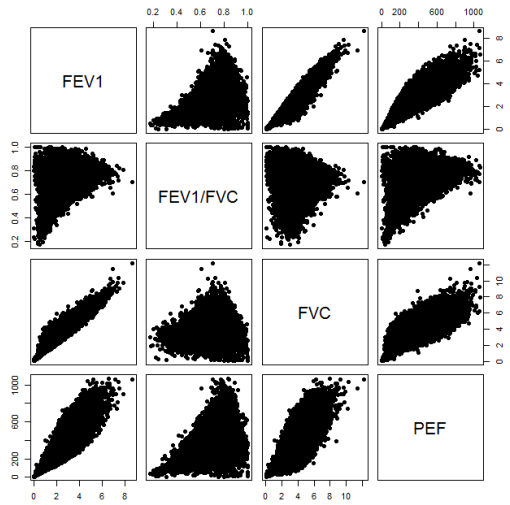


Figure 6.11 - Signal selection process for the genome-wide interaction analysis

Untransformed phenotypes



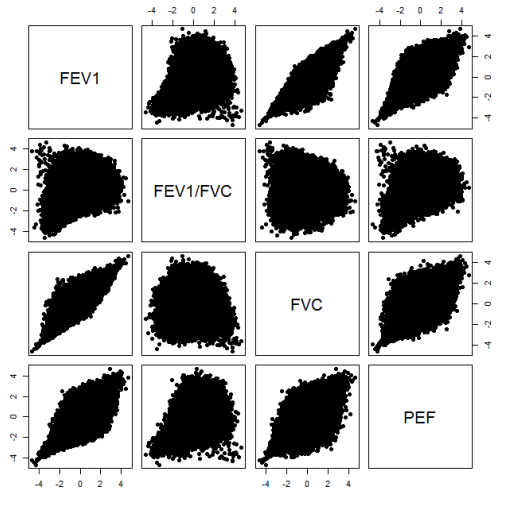
Correlation matrix

	FEV_1	**FEV_1/FVC**	**FVC**	**PEF**
FEV_1		0.28	0.95	0.85
FEV_1/FVC	0.28		-0.02	0.31
FVC	0.95	-0.02		0.78
PEF	0.85	0.31	0.78	

Figure 6.12 - Correlations between untransformed phenotypes

Correlations are calculated using Pearson's correlation coefficient method. The plot on the left shows the pairwise plots of all four analysed phenotypes The table on the right shows the correlation matrix for all four phenotypes, taking the value -1 to 1 for perfectly negatively correlated to perfectly positively correlated respectively.

Transformed phenotypes



Correlation matrix

	FEV_1	**FEV_1/FVC**	**FVC**	**PEF**
FEV_1		0.40	0.89	0.68
FEV_1/FVC	0.40		-0.03	0.44
FVC	0.89	-0.03		0.52
PEF	0.68	0.44	0.52	

Figure 6.13 - Correlation between inverse normalised phenotypes

Correlations are calculated using Pearson's correlation coefficient method. The plot on the left shows the pairwise plots of all four analysed phenotypes The table on the right shows the correlation matrix for all four phenotypes, taking the value -1 to 1 for perfectly negatively correlated to perfectly positively correlated respectively.

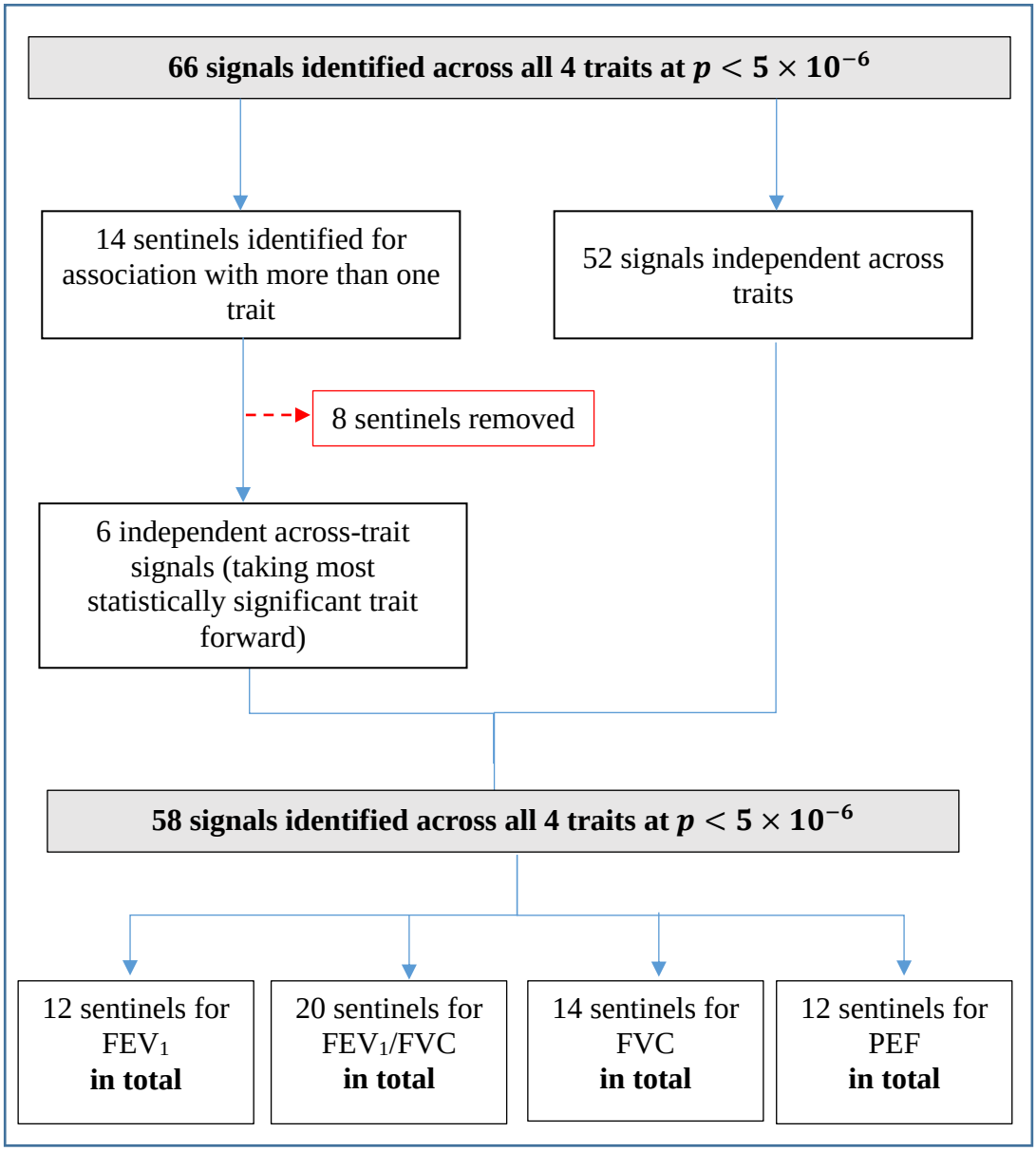
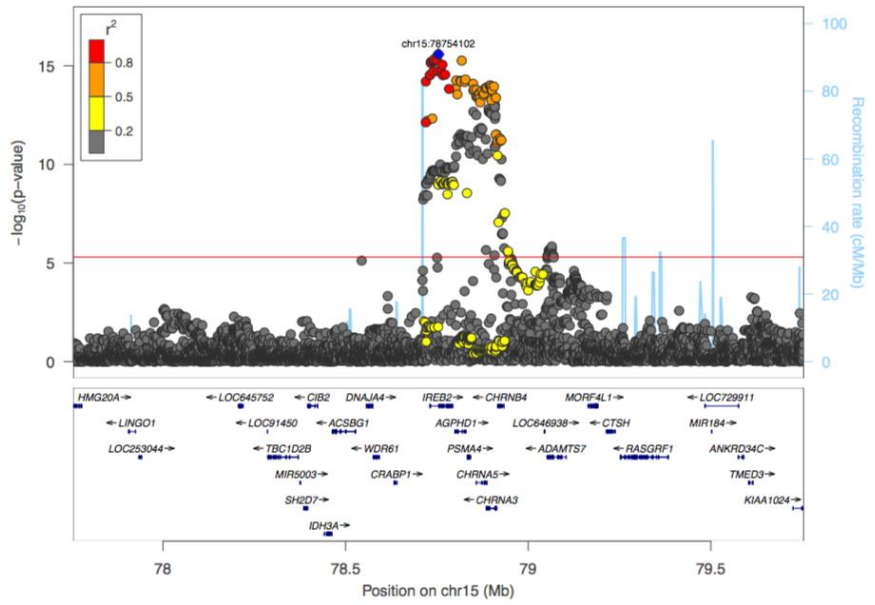


Figure 6.14 - Across trait signal independence process for the genome-wide interaction signals

rs2009746 (chr15:78754102) chosen as preliminary sentinel and thus 2Mb region is chosen equidistant around sentinel from 77754102 – 79754102



Use of GCTA identifies a second signal with sentinel rs8042059 (chr15: 78907859). All SNPs reaching threshold ($p < 5 \times 10^{-6}$) accounted for thus no more signals in region

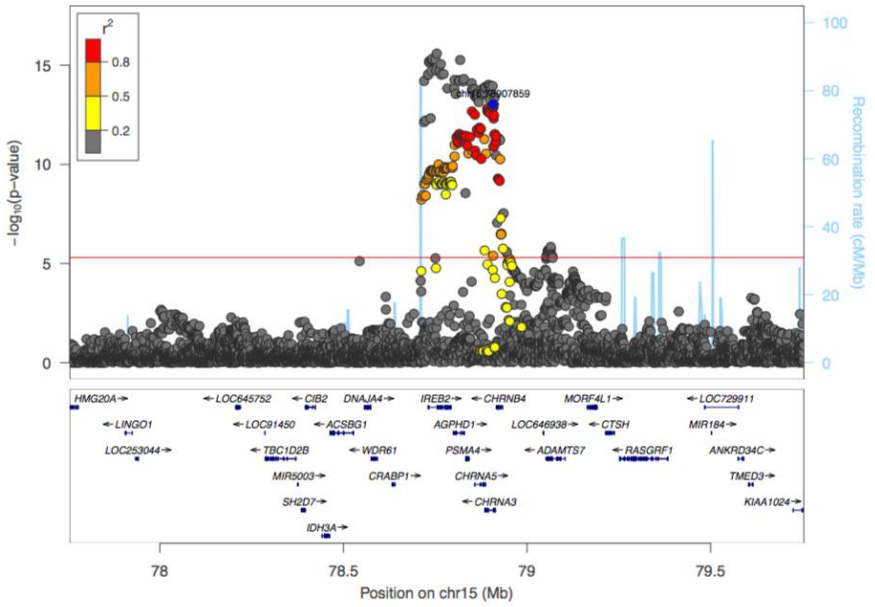


Figure 6.15- Example of secondary signal found during signal selection process using GCTA for phenotype FEV_1

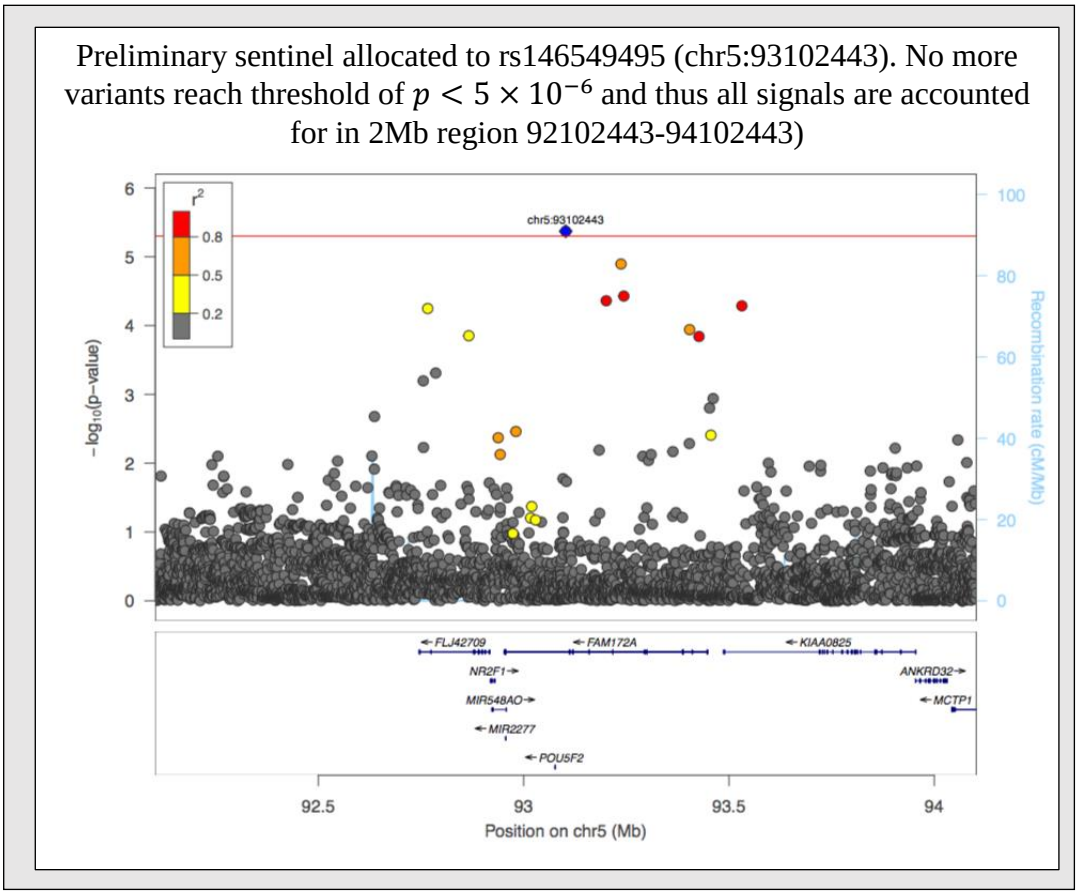


Figure 6.16 - A scenario where only one signal was found (rs146549495) within a region and no further signals identified, for phenotype FEV_1

6.4.5 Interaction signals resulting from the signal selection process

The 58 interaction signals identified as part of the signal selection process are presented in **Tables 6.2**, **6.3**, **6.4** and **6.5** for traits FEV_1 (12 signals), FEV_1/FVC (20 signals), FVC (14 signals) and PEF (12 signals) respectively. Across the four traits, the 58 SNPs presented were well imputed with imputation quality ranging from 0.81 to 1. Of the 58 SNPs identified, 29 were common (MAF ≥ 5%), 18 were low frequency (1% ≤ MAF < 5%) and 11 were rare (MAF < 1%).

The most strongly associated signals were for the trait FEV_1/FVC, both of which were common SNPs (MAF > 5%) at chromosomal location 15q25, specifically rs8042849 in *HYKK* (MAF = 34.1%, Interaction beta = 0.049, P = 7.25×10^{-20}) and rs7173514

downstream of *CHRNA3* (MAF = 22.9%, Interaction beta = 0.054, P = 4.62×10^{-19}). Both signals were in the 15q25 region, a region well known for its association with smoking behaviour (containing the nicotine receptor genes *CHRNA3 – 5*), specifically identified for its reported association with cigarettes per day (CPD), with reported sentinel SNP rs1051730 (64). For both SNPs the interaction effect was driven by a strong association in ever smokers (**Appendix H**), with an opposite direction of effect (which was not statistically significant) in never smokers.

Five of the 56 signals were near (with regards to genomic position) to previously reported lung function signals (defined as within 1Mb), specifically variants rs113246660, rs147414811, rs4796410, rs12452505 and rs2865035 in or nearest to *RBMS3*, *MIR4465, ACAP1, AXIN2* and *ITSN1*, with interaction effects identified for PEF, FEV_1, PEF, FVC and FVC respectively (**Table 6.6**). Observed values of LD between the interaction signal and the previously reported lung function signal were all $r^2 < 0.024$, suggesting that none of the interaction signals identified were already known to be associated with lung function or COPD.

For each signal which was identified at $P < 5 \times 10^{-6}$ for one trait, suggestive evidence of an association with any of the other three traits was determined by a threshold of 1×10^{-3}. Of the 58 signals, 20 were unique to one trait (**Figure 6.17**). Specifically, there were 15 signals that were unique to the ratio of FEV_1/FVC, and 5 for PEF. There were 21 signals with an interaction effect association with both FEV_1 and FVC only. Only eight signals were associated with three or more traits (six associated with FEV_1, FVC and PEF and two associated with all four traits).

Of the variants identified, one is a missense variant, namely rs148424048 (MAF = 1.2%, Interaction beta = -0.107, P = 3.46×10^{-6}), in *ZNF594*, associated with FVC**.** Of the remaining variants, 28 were intronic, 28 were intergenic, and one was a three prime untranslated region variant (3'-UTR) in *MPEG1* (rs142254414).

The marginal genetic effects in both the ever and never smoking subgroups separately are presented for all SNPs in **Appendix H**.

Table 6.2 - FEV_1 gene-smoking interaction signals.

Betas (β) and standard errors (SE) are on the inverse normalised scale. Ever smokers are coded as 1 (and never smokers coded as 0), such that a positive or negative β means a larger or smaller genetic effect in ever smokers respectively for the coded allele. Loci are build 37. CAF = coded allele frequency, MAF = Minor allele frequency and INFO = Imputation score.

SNP (chr:pos)	Coded / Non coded allele	CAF	MAF	INFO	β	SE	P value	Gene/closest gene
rs142312019 (2:153914485)	T / C	0.006	0.006	0.895	-0.156	0.034	4.58×10^{-6}	*ARL6IP6 (downstream)*
rs10497204 (2:159950898)	T / C	0.485	0.485	0.987	0.023	0.005	4.83×10^{-6}	*TANC1*
rs139977403 (3:76755996)	A / C	0.009	0.009	0.874	0.138	0.028	1.23×10^{-6}	*ROBO2*
rs74823357 (3:193713123)	T / C	0.041	0.041	1.000	0.058	0.013	4.85×10^{-6}	*LOC647323* (upstream of DPPA2P3)
rs146549495 (5:93102443)	T / C	0.016	0.016	0.967	0.093	0.020	4.26×10^{-6}	*FAM172A*
rs147414811 (6:141243046)	C / T	0.025	0.025	0.979	0.079	0.017	1.93×10^{-6}	*MIR4465* (downstream)
rs34154123 (7:37786836)	A / T	0.104	0.104	0.809	-0.043	0.009	2.87×10^{-6}	*GPR141*
rs17120700 (14:49349224)	A / G	0.647	0.353	0.987	-0.024	0.005	4.74×10^{-6}	*RPS29* (downstream)
rs117585696 (14:58706512)	C / T	0.018	0.018	1.000	0.093	0.019	1.18×10^{-6}	*ACTR10*
rs35524777 (18: 11424551)	T / C	0.062	0.062	0.953	-0.056	0.011	1.44×10^{-7}	*PIEZO2* (upstream)
rs2604894 (19:41292404)	G / A	0.549	0.451	0.983	-0.024	0.005	2.04×10^{-6}	*RAB4B*
rs2225434 (21:46634146)	C / T	0.568	0.432	0.995	-0.025	0.005	1.66×10^{-6}	*ADARB1*

Table 6.3 - FEV_1/FVC gene-smoking interaction signals.

Betas (β) and standard errors (SE) are on the inverse normalised scale. Ever smokers are coded as 1 (and never smokers coded as 0), such that a positive or negative β means a larger or smaller genetic effect in ever smokers respectively for the coded allele. Loci are build 37. CAF = coded allele frequency, MAF = Minor allele frequency and INFO = Imputation score.

SNP (chr:pos)	Coded / Non coded allele	CAF	MAF	INFO	β	SE	P value	Gene/closest gene
rs61787074 (1:27851315)	*T / C*	0.026	0.026	0.848	0.085	0.017	9.46×10^{-7}	*AHDC1* (downstream)
rs116757305 (1:43876305)	T / C	0.023	0.023	0.954	-0.080	0.017	3.26×10^{-6}	*SZT2* (upstream)
rs116799787 (3:180022491)	T / C	0.009	0.009	0.941	0.129	0.028	3.80×10^{-6}	*PEX5L* (upstream)
rs74376726 (3:182694981)	A / G	0.060	0.060	0.969	0.050	0.011	3.09×10^{-6}	*DCUN1D1*
rs115906789 (5:5809051)	A / G	0.029	0.029	0.911	0.072	0.016	4.62×10^{-6}	*KIAA0947* (downstream)
rs11969624 (6:5460085)	C / G	0.117	0.117	0.918	0.038	0.008	2.73×10^{-6}	*FARS2*
rs76004091 (6:111716001)	C / G	0.014	0.014	0.928	0.102	0.022	4.68×10^{-6}	*REV3L*
rs35535406 (7:31427263)	C / T	0.238	0.238	0.997	0.028	0.006	3.49×10^{-6}	*NEUROD6* (upstream)
rs7817569 (8:25461315)	T / C	0.714	0.286	0.972	-0.027	0.006	2.42×10^{-6}	*CDCA2* (downstream)
rs11780592 (8:27418747)	G / A	0.180	0.180	0.997	0.031	0.007	3.07×10^{-6}	*EPHX2* (downstream)
rs73555789 (9:136456673)	A / G	0.040	0.040	0.983	0.060	0.013	3.94×10^{-6}	*ADAMTSL2* (downstream)
rs142254414 (11:58976080)	A / G	0.011	0.011	0.910	-0.120	0.026	3.75×10^{-6}	*MPEG1, DTX4* (downstream)
rs117367754 (12:67751257)	A / G	0.017	0.017	0.993	-0.093	0.019	1.69×10^{-6}	*CAND1* (downstream)
rs10862408 (12:82359619)	C / G	0.275	0.275	0.992	0.029	0.006	4.16×10^{-7}	*PPFIA2* (upstream)
rs142704172 (12:122496894)	T / C	0.008	0.008	0.919	-0.145	0.029	5.26×10^{-7}	*BCL7A*
rs8042849 (15:78817929)	T / C	0.659	0.341	0.999	0.049	0.005	7.25×10^{-20}	*HYKK*
rs7173514 (15:78849918)	T / C	0.229	0.229	0.991	0.054	0.006	4.62×10^{-19}	*CHRNA3* (downstream)
rs62023825 (16:11506666)	T / C	0.161	0.161	0.959	-0.032	0.007	4.53×10^{-6}	*RMI2*
rs80277243 (16:52793919)	A / G	0.011	0.011	0.902	0.123	0.025	9.85×10^{-7}	*LOC643714* (upstream)
rs67134151 (19:29510958)	G / T	0.070	0.070	0.991	0.047	0.010	2.47×10^{-6}	*LOC100505835* (downstream)

Table 6.4 - FVC interaction gene-smoking interaction signals.
Betas (β) and standard errors (SE) are on the inverse normalised scale. Ever smokers are coded as 1 (and never smokers coded as 0), such that a positive or negative β means a larger or smaller genetic effect in ever smokers respectively for the coded allele. Loci are build 37. CAF = coded allele frequency, MAF = Minor allele frequency and INFO = Imputation score.

SNP (chr:pos)	Coded / Non coded allele	CAF	MAF	INFO	β	SE	P value	Gene/closest gene
rs146541032 (2:62358665)	C / T	0.006	0.006	0.885	-0.164	0.035	2.19×10^{-6}	*COMMD1*
rs6773439 (3:15530595)	A / G	0.173	0.173	0.975	-0.033	0.007	1.71×10^{-6}	*COLQ*
rs17704183 (4:168390470)	A / G	0.362	0.362	0.989	0.026	0.005	7.48×10^{-7}	*SPOCK3* (upstream)
rs1593464 (5:94360291)	A / G	0.879	0.121	0.995	-0.036	0.008	4.63×10^{-6}	*MCTP1*
rs7728169 (5:155340750)	T / A	0.036	0.036	0.971	-0.065	0.014	2.94×10^{-6}	*SGCD*
rs61244245 (7:68595239)	T / C	0.405	0.405	0.991	-0.024	0.005	2.96×10^{-6}	*AUTS2* (upstream)
rs186074884 (7:125325472)	A / G	0.006	0.006	0.840	0.174	0.035	6.84×10^{-7}	*GRM8* (downstream)
rs77608508 (9:120844163)	G / A	0.076	0.076	0.983	-0.048	0.010	6.62×10^{-7}	*TLR4* (downstream)
rs4962379 (10:126228742)	C / T	0.873	0.127	0.989	-0.036	0.008	2.85×10^{-6}	*LHPP*
rs117575177 (12:65696432)	G / A	0.022	0.022	0.966	0.084	0.018	1.93×10^{-6}	*MSRB3*
rs4783512 (16:20937848)	C / A	0.625	0.375	0.982	-0.025	0.005	2.78×10^{-6}	LYRM1 (downstream)
rs148424048 (17:5086198)	A / G	0.012	0.012	1.000	-0.107	0.023	3.46×10^{-6}	*ZNF594*
rs12452505 (17:63556402)	G / C	0.142	0.142	0.992	0.037	0.007	3.91×10^{-7}	*AXIN2*
rs28650353 (21:35166202)	A / G	0.014	0.014	0.958	0.107	0.022	1.31×10^{-6}	*ITSN1*

Table 6.5 - PEF gene-smoking interaction signals

Betas (β) and standard errors (SE) are on the inverse normalised scale. Ever smokers are coded as 1 (and never smokers coded as 0), such that a positive or negative β means a larger or smaller genetic effect in ever smokers respectively for the coded allele. Loci are build 37. CAF = coded allele frequency, MAF = Minor allele frequency and INFO = Imputation score.

SNP (chr:pos)	Coded / Non coded allele	CAF	MAF	INFO	β	SE	P value	Gene/closest gene
rs12077425 (1:29744567)	T / C	0.006	0.006	0.971	-0.165	0.034	9.58×10^{-7}	*PTPRU* (downstream)
rs72772245 (2:1359174)	C / T	0.008	0.008	0.911	-0.140	0.029	1.50×10^{-6}	*SNTG2*
rs113246660 (3:30189600)	C / T	0.008	0.008	0.837	-0.144	0.030	1.77×10^{-6}	*RBMS3* (downstream)
rs55730263 (5:83166632)	G / A	0.154	0.154	0.965	0.034	0.007	1.98×10^{-6}	*EDIL3* (downstream)
rs76412370 (5:118115062)	T / G	0.118	0.118	0.963	0.038	0.008	2.85×10^{-6}	*DTWD2* (downstream)
rs73139542 (7:63325787)	A / G	0.045	0.045	0.972	-0.062	0.012	6.66×10^{-7}	*LOC100506050* (downstream)
rs61853871 (10:73325318)	T / G	0.075	0.075	0.984	0.045	0.010	4.29×10^{-6}	*CDH23*
rs553187851 (11:117448580)	T / C	0.006	0.006	0.875	0.165	0.034	1.14×10^{-6}	*DSCAML1*
rs61183515 (15:55175622)	C / A	0.197	0.197	0.995	-0.030	0.006	4.14×10^{-6}	*UNC13C* (downstream), RSL24D1 (downstream)
rs539865765 (16:51905937)	G / C	0.005	0.005	0.896	0.178	0.037	1.66×10^{-6}	*C16orf97* (downstream)
rs4796410 (17:7273247)	G / A	0.119	0.119	0.988	0.037	0.008	3.42×10^{-6}	*ACAP1* (downstream)
rs151310656 (19:16777665)	A / G	0.029	0.029	0.878	0.074	0.016	4.55×10^{-6}	*TMEM38A*

Table 6.6 - SNPs in close proximity of previous lung function or COPD signals
Loci are build 37.

Interaction signal SNP (chr:pos)	Trait	Previously reported SNP	Previously reported trait	Ref.	Distance (bases)	LD (r^2)	Gene
rs113246660 (3:30189600)	PEF	rs17666332 (3:29469675)	FEV_1/FVC	(52)	719925	0.0002	*RBMS3*
rs147414811 (6:141243046)	FEV_1	rs1102077 (6:140271357)	FEV_1	(32,33)	971689	0.0093	*MIR4465*
rs4796410 (17:7273247)	PEF	rs1215 (17:7163350)	FVC	(32,33)	109897	0.0040	*ACAP1*
		rs4968200 (17:7448457)	FEV_1	(32,33)	175210	0.0240	*ACAP1*
rs12452505 (17:63556402)	FVC	rs11653958 (17:62686730)	FEV_1/FVC	(32,33)	869672	0.0002	*AXIN2*
rs28650353 (21:35166202)	FVC	rs12627254 (21:35368402)	FEV_1/FVC	(32,33)	202200	0.0004	*ITSN1*

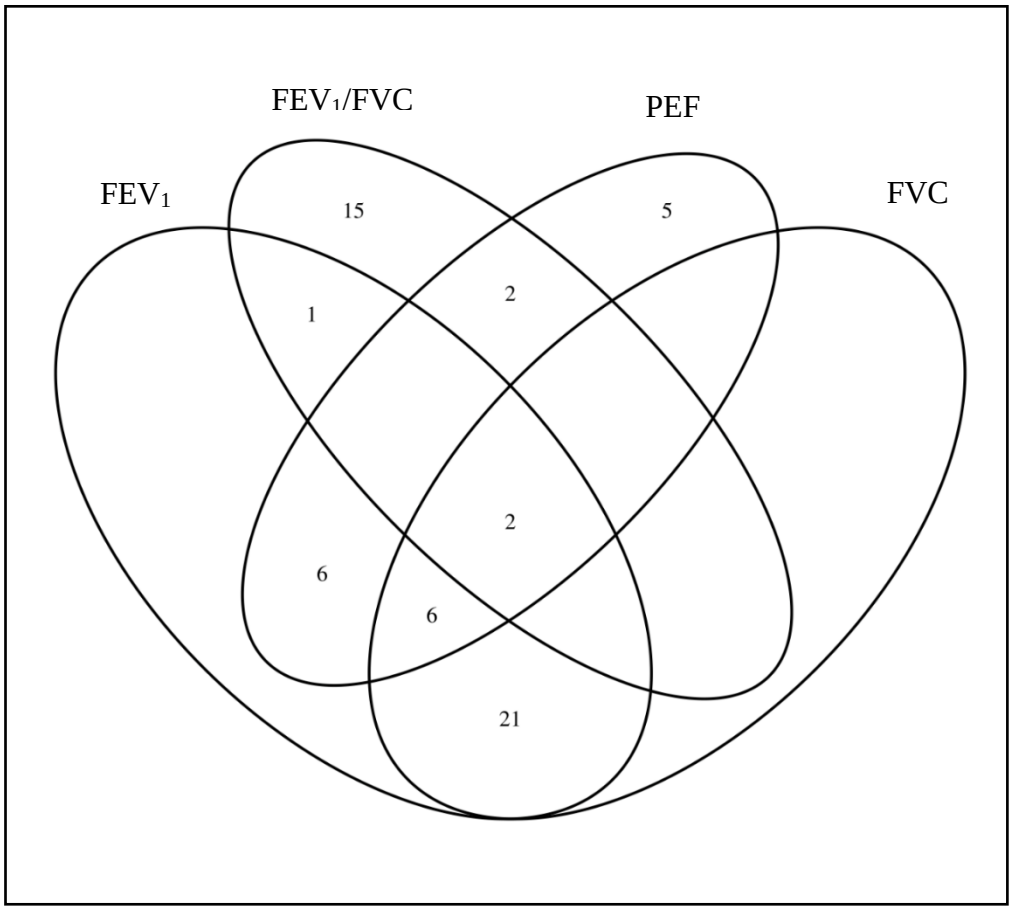


Figure 6.17- Overlap of interaction effects between the four analysed traits using a suggestive threshold of $P < 1 \times 10^{-3}$

6.4.6 Association of interaction signals with smoking

Five of the 58 interaction effect signals were found to have a statistically significant association with smoking behaviour at the Bonferroni adjusted threshold for 58 tests (i.e. $p < 8.6 \times 10^{-4}$). **Table 6.7** presents the association results that were statistically significant.

As expected, the most statistically significant associations were for the 15q25 SNPs (with 15q25 a widely reported smoking locus) rs8042849 (MAF = 0.34, P = 1.1×10^{-78}) in *HYKK* and rs7173514 (MAF = 0.22, P = 1.5×10^{-36}) near *CHRNA3* with cigarettes per day (CPD), whilst rs8042849 (MAF = 0.34, P = 9.6×10^{-5}) was also associated with smoking initiation. Both SNPs were checked for independence from the reported CPD signal, rs1051730 (64). The r^2 values of 0.33 and 0.11 observed for SNPs rs8042849 and rs7173514 respectively, suggested they were not independent of the previously reported CPD signal.

There were also statistically significant associations for rs11780592 (MAF = 0.18, P = 7×10^{-9}) in *EPHX2*, rs73555789 (MAF = 0.04, P = 7.9×10^{-4}) near *ADAMTSL2*, and rs2604894 (MAF = 0.45, P = 3×10^{-12}) in *RAB4B* with smoking initiation (SI), smoking cessation (SC) and CPD respectively. The latter SNP, rs2604894, is in high LD with another previously reported sentinel SNP for association with CPD, namely rs7937 (62), with an r^2 value of 0.55. For SNPs rs11780592 and rs73555789, there was no previous reports to suggest that they are known smoking signals. Only the latter was in close proximity of a previously reported smoking association signal (~22Kb), namely rs3025343 (64), reported for its association with SI. LD between the two SNPs was however low (r^2 = 0.0036) suggesting they were independent.

The full smoking association results for all 58 SNPs are given in **Appendix I**. The five signals identified were removed from further follow-up.

Table 6.7 - SNPs with a statistically significant association with any of the smoking traits SI, SC or CPD ($p < 8.6 \times 10^{-4}$).
CAF – coded allele frequency, MAF – Minor allele frequency and INFO = Imputation score. Betas (β) and standard errors (SE) are on the inverse normalised scale.

SNP (Chr:Pos)	Trait	Coded / Non-coded allele	CAF	MAF	INFO	β	SE	Gene	P value
rs11780592 (8:27418747)	SI	A/G	0.819	0.181	0.997	0.020	0.003	*EPHX2*	7.8×10^{-9}
rs73555789 (9:136456673)	SC	G/A	0.960	0.040	0.983	0.034	0.010	*ADAMTSL2*	7.9×10^{-4}
rs8042849 (15:78817929)	SI	C/T	0.339	0.339	0.999	-0.011	0.003	*HYKK*	9.6×10^{-5}
	CPD	C/T	0.343	0.343	0.999	0.098	0.005	HYKK	1.1×10^{-78}
rs7173514 (15:78849918)	CPD	C/T	0.778	0.222	0.991	0.076	0.006	CHRNA3	1.5×10^{-36}
rs2604894 (19:41292404)	CPD	A/G	0.453	0.453	0.983	-0.035	0.005	*RAB4B*	3.0×10^{-12}

6.4.7 SpiroMeta replication of interaction signals

For the 53 SNPs that remained after removing those which were associated with smoking traits, replication was sought using the SpiroMeta consortium. Association testing results were available for all 53 SNPs in the SpiroMeta consortium meta-analysis. Although the maximum attainable sample size in the SpiroMeta consortium was 71,067 with 37,411 ever smokers and 33,656 never smokers, N effective sample size estimates for each SNP (which takes into account imputation quality) were on the whole much lower (**Figure 6.18**). 27 of the 53 signals had an N effective below 50,000 individuals (~70% of the maximum attainable sample size), with the lowest N effective as low as 19,575 individuals, which is less than 30% of the maximum attainable sample size in the SpiroMeta consortium (10,455 ever smokers and 9,120 never smokers). The largest N effective was 70,626 (37,166 smokers and 33,460 never smokers). None of the 53 SNPs had a statistically significant interaction effect with the respective trait when using a Bonferroni corrected threshold for 53 tests of $p < 9.4 \times 10^{-4}$ (**Table 6.8**). P-values ranged from 0.0937 to 0.9813, thus the most statistically significant SNP rs553187851 only reached a significance threshold of $p < 0.1$ (P = 0.0937), but did have consistent direction of effect in ever and never smokers. For the 53 SNPs, 11 had direction of effect in ever and never smokers consistent with the UK Biobank analysis. There did not appear to be any link between effective sample size and consistent

direction of effect for these 11 SNPs, with consistent directions of effect observed for both extremities of effective sample sizes.

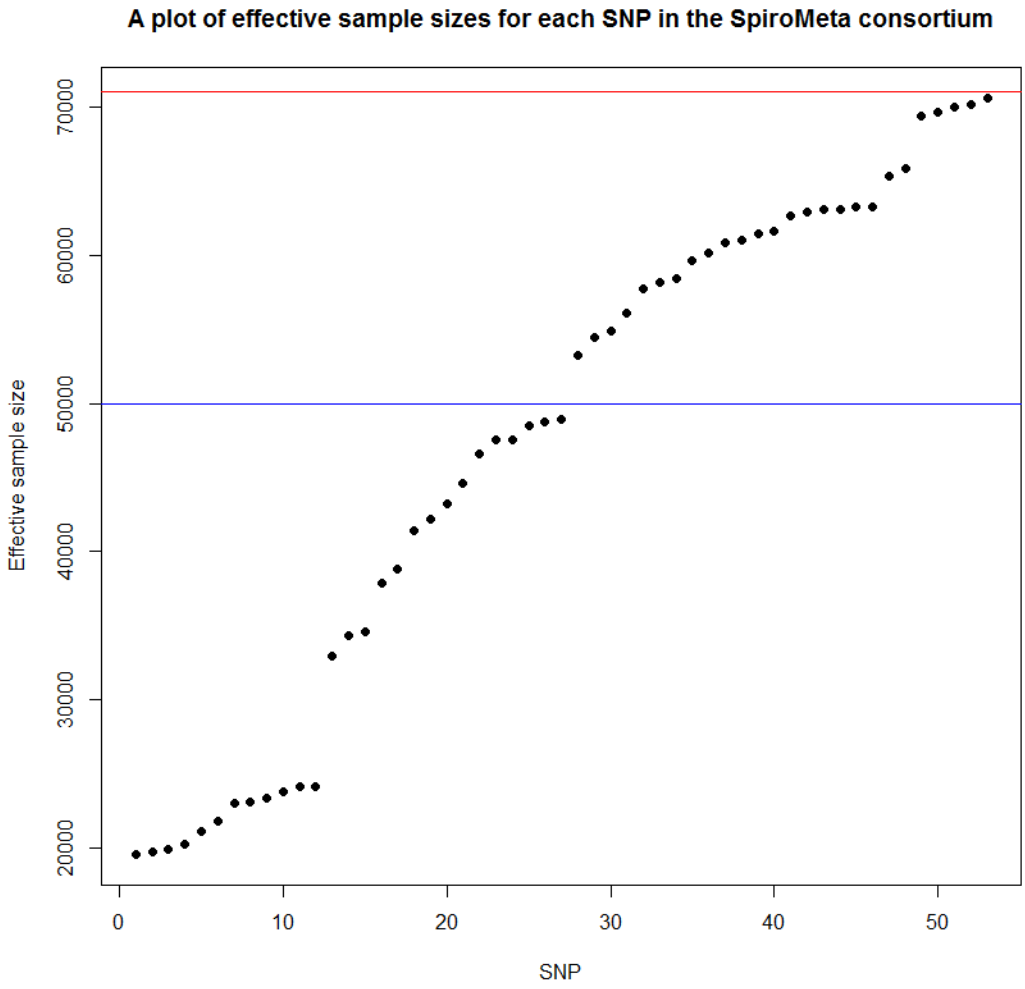


Figure 6.18 - A plot of effective sample sizes for each of the 53 variants which were included in the SpiroMeta consortium replication stage

The red line indicates the maximum attainable sample size and the blue line represents 50,000 individuals (~ 70% of the maximum available sample size)

Table 6.8 – Welch's t-test results comparing genetic effect in ever and never smokers using meta-analysed data from the SpiroMeta consortium for the 53 SNPs producing statically significant interaction effects in UK Biobank

CAF = coded allele frequency, MAF = minor allele frequency, TS = test statistic, DF = degrees of freedom and N_{eff} = effective sample size (considers imputation quality). Betas (β) and standard errors (SE) are on the inverse normalised scale. The maximum attainable sample sizes for ever smokers and never smokers were 37,411 and 33,656 respectively.

			Ever Smoker						Never Smoker								
SNP (chr:pos)	**Trait**	**Coded / Non-coded allele**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**Welch TS**	**Welch DF**	**Welch P**
rs61787074 (1:27851315)	FEV_1/FVC	T / C	0.020	0.020	21549	0.0394	0.0341	0.2474	0.020	0.0197	19848	0.0246	0.0356	0.4894	0.30	50.10	0.7656
rs12077425 (1:29744567)	PEF	T / C	0.006	0.006	11828	-0.0302	0.0852	0.7228	0.005	0.0052	9279	0.0433	0.1040	0.6767	-0.55	183.55	0.5848
rs116757305 (1:43876305)	FEV_1/FVC	T / C	0.022	0.022	30124	-0.0287	0.0276	0.3000	0.023	0.0228	28084	-0.0114	0.0281	0.6840	-0.44	45.08	0.6639
rs72772245 (2:1359174)	PEF	C / T	0.009	0.009	11655	-0.0851	0.0692	0.2184	0.009	0.0092	10173	0.0663	0.0755	0.3804	-1.48	113.27	0.1422
rs146541032 (2:62358665)	FVC	C / T	0.006	0.006	22743	-0.0286	0.0611	0.6401	0.005	0.0054	20469	0.0560	0.0691	0.4177	-0.92	182.15	0.3604
rs142312019 (2:153914485)	FEV_1	T / C	0.005	0.005	22835	0.0639	0.0690	0.3546	0.004	0.0042	21825	0.0439	0.0734	0.5495	0.20	226.11	0.8432
rs10497204 (2:159950898)	FEV_1	T / C	0.467	0.467	32698	-0.0010	0.0080	0.8973	0.456	0.4562	30557	-0.0147	0.0083	0.0755	1.19	4.16	0.2971
rs6773439 (3:15530595)	FVC	A / G	0.163	0.163	36806	-0.0004	0.0102	0.9707	0.159	0.1589	33374	-0.0093	0.0108	0.3901	0.60	7.64	0.5657
rs113246660 (3:30189600)	PEF	C / T	0.009	0.009	10630	0.0108	0.0767	0.8885	0.008	0.0083	9246	0.0322	0.0816	0.6931	-0.19	123.61	0.8484
rs139977403 (3:76755996)	FEV_1	A / C	0.005	0.005	16716	-0.0374	0.0767	0.6260	0.004	0.0045	17588	0.0232	0.0782	0.7671	-0.55	205.66	0.5811
rs116799787 (3:180022491)	FEV_1/FVC	T / C	0.008	0.008	24872	0.0102	0.0496	0.8372	0.008	0.0077	23592	0.0267	0.0524	0.6096	-0.23	125.75	0.8187
rs74376726 (3:182694981)	FEV_1/FVC	A / G	0.065	0.065	30747	0.0138	0.0164	0.4027	0.067	0.0668	28915	-0.0112	0.0169	0.5092	1.06	16.59	0.3058

			Ever Smoker						Never Smoker								
SNP (chr:pos)	**Trait**	**Coded / Non-coded allele**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**Welch TS**	**Welch DF**	**Welch P**
rs74823357 (3:193713123)	FEV_1	T / C	0.033	0.033	21231	-0.0049	0.0273	0.8559	0.031	0.0312	20934	0.0170	0.0281	0.5464	-0.56	32.33	0.5797
rs17704183 (4:168390470)	FVC	A / G	0.362	0.362	31948	0.0083	0.0084	0.3249	0.364	0.3639	29525	0.0025	0.0086	0.7751	0.48	4.44	0.6523
rs115906789 (5:5809051)	FEV_1/FVC	A / G	0.026	0.026	24304	-0.0129	0.0285	0.6506	0.026	0.0260	23268	0.0216	0.0287	0.4524	-0.85	38.99	0.3993
rs55730263 (5:83166632)	PEF	G / A	0.154	0.154	12455	0.0009	0.0179	0.9582	0.155	0.1546	10872	-0.0063	0.0191	0.7418	0.28	7.88	0.7895
rs146549495 (5:93102443)	FEV_1	T / C	0.018	0.018	28771	0.0305	0.0318	0.3375	0.022	0.0216	27334	0.0018	0.0295	0.9516	0.66	52.90	0.5112
rs1593464 (5:94360291)	FVC	A / G	0.859	0.141	32483	0.0163	0.0116	0.1592	0.847	0.8474	30591	0.0220	0.0115	0.0561	-0.35	8.41	0.7339
rs76412370 (5:118115062)	PEF	T / G	0.120	0.120	12295	-0.0088	0.0201	0.6617	0.127	0.1268	10723	0.0043	0.0207	0.8359	-0.45	9.51	0.6606
rs7728169 (5:155340750)	FVC	T / A	0.034	0.034	27080	-0.0149	0.0241	0.5362	0.031	0.0309	26163	-0.0228	0.0258	0.3768	0.22	33.23	0.8243
rs11969624 (6:5460085)	FEV_1/FVC	C / G	0.115	0.115	32234	0.0009	0.0125	0.9455	0.113	0.1130	28756	-0.0080	0.0134	0.5522	0.48	10.19	0.6410
rs76004091 (6:111716001)	FEV_1/FVC	C / G	0.016	0.016	24702	-0.0382	0.0354	0.2797	0.017	0.0171	24034	-0.0593	0.0347	0.0871	0.43	59.77	0.6720
rs147414811 (6:141243046)	FEV_1	C / T	0.017	0.017	24656	-0.0184	0.0345	0.5933	0.017	0.0166	22857	0.0337	0.0361	0.3496	-1.04	59.02	0.3003
rs35535406 (7:31427263)	FEV_1/FVC	C / T	0.243	0.243	36801	0.0040	0.0087	0.6430	0.244	0.2444	33191	0.0037	0.0092	0.6843	0.02	5.57	0.9813
rs34154123 (7:37786836)	FEV_1	A / T	0.098	0.098	17090	-0.0020	0.0182	0.9112	0.093	0.0927	15854	-0.0359	0.0195	0.0650	1.27	11.65	0.2284
rs73139542 (7:63325787)	PEF	A / G	0.044	0.044	12368	-0.0124	0.0316	0.6953	0.046	0.0458	10784	-0.0691	0.0331	0.0369	1.24	24.04	0.2272
rs61244245 (7:68595239)	FVC	T / C	0.408	0.408	32266	0.0023	0.0082	0.7795	0.414	0.4143	30390	-0.0009	0.0084	0.9150	0.27	4.28	0.7983

			Ever Smoker						Never Smoker								
SNP (chr:pos)	**Trait**	**Coded / Non-coded allele**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**Welch TS**	**Welch DF**	**Welch P**
rs186074884 (7:125325472)	FVC	A / G	0.005	0.005	18067	0.0693	0.0719	0.3355	0.004	0.0042	16559	0.0509	0.0815	0.5324	0.17	203.06	0.8656
rs7817569 (8:25461315)	FEV_1/FVC	T / C	0.708	0.292	33068	0.0010	0.0086	0.9121	0.704	0.7040	29824	-0.0095	0.0091	0.2948	0.84	4.89	0.4421
rs77608508 (9:120844163)	FVC	G / A	0.076	0.076	28109	0.0116	0.0162	0.4735	0.080	0.0803	26322	-0.0026	0.0162	0.8739	0.62	14.20	0.5458
rs61853871 (10:73325318)	PEF	T / G	0.076	0.076	12720	0.0159	0.0239	0.5068	0.073	0.0728	11105	-0.0184	0.0261	0.4815	0.97	14.76	0.3489
rs4962379 (10:126228742)	FVC	C / T	0.865	0.135	31716	0.0064	0.0118	0.5859	0.860	0.8601	29949	-0.0092	0.0120	0.4400	0.93	8.74	0.3763
rs142254414 (11:58976080)	FEV_1/FVC	A / G	0.008	0.008	18805	0.0192	0.0561	0.7318	0.008	0.0078	19053	0.0355	0.0579	0.5399	-0.20	122.91	0.8404
rs553187851 (11:117448580)	PEF	T / C	0.007	0.007	10829	0.1993	0.0821	0.0152	0.006	0.0061	9416	-0.0138	0.0960	0.8857	1.69	159.09	0.0937
rs117575177 (12:65696432)	FVC	G / A	0.024	0.024	27972	-0.0170	0.0282	0.5470	0.025	0.0245	26888	-0.0263	0.0279	0.3449	0.24	43.19	0.8149
rs117367754 (12:67751257)	FEV_1/FVC	A / G	0.016	0.016	31456	-0.0344	0.0319	0.2821	0.017	0.0165	29433	-0.0493	0.0330	0.1347	0.33	63.99	0.7456
rs10862408 (12:82359619)	FEV_1/FVC	C / G	0.263	0.263	32624	-0.0043	0.0090	0.6296	0.260	0.2600	30477	-0.0126	0.0094	0.1787	0.63	5.31	0.5519
rs142704172 (12:122496894)	FEV_1/FVC	T / C	0.010	0.010	19584	-0.0222	0.0486	0.6485	0.010	0.0102	19251	0.0212	0.0496	0.6694	-0.62	93.69	0.5341
rs17120700 (14:49349224)	FEV_1	A / G	0.642	0.358	36672	0.0039	0.0078	0.6195	0.647	0.6471	33004	0.0058	0.0082	0.4800	-0.17	4.45	0.8706
rs117585696 (14:58706512)	FEV_1	C / T	0.015	0.015	23994	0.0167	0.0377	0.6568	0.014	0.0139	22569	0.0143	0.0410	0.7277	0.04	71.81	0.9646
rs61183515 (15:55175622)	PEF	C / A	0.196	0.196	12880	-0.0031	0.0159	0.8468	0.197	0.1968	11242	-0.0045	0.0169	0.7910	0.06	6.44	0.9533
rs62023825 (16:11506666)	FEV_1/FVC	T / C	0.166	0.166	34696	-0.0145	0.0103	0.1572	0.171	0.1714	30689	0.0099	0.0108	0.3617	-1.64	7.24	0.1445

			Ever Smoker						Never Smoker								
SNP (chr:pos)	**Trait**	**Coded / Non-coded allele**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**CAF**	**MAF**	**N_{eff}**	**β**	**SE**	**P**	**Welch TS**	**Welch DF**	**Welch P**
rs4783512 (16:20937848)	FVC	C / A	0.623	0.377	34730	0.0020	0.0080	0.8030	0.624	0.6243	31117	0.0058	0.0084	0.4896	-0.33	4.37	0.7578
rs539865765 (16:51905937)	PEF	G / C	0.004	0.004	10676	0.0380	0.1064	0.7207	0.005	0.0052	9073	0.0440	0.1039	0.6717	-0.04	217.38	0.9679
rs80277243 (16:52793919)	FEV_1/FVC	A / G	0.011	0.011	25785	0.0162	0.0432	0.7070	0.010	0.0105	23191	0.0052	0.0475	0.9125	0.17	100.16	0.8640
rs148424048 (17:5086198)	FVC	A / G	0.014	0.014	32186	-0.0115	0.0343	0.7376	0.013	0.0135	27982	-0.0184	0.0368	0.6175	0.14	75.43	0.8910
rs4796410 (17:7273247)	PEF	G / A	0.123	0.123	12894	-0.0252	0.0194	0.1930	0.124	0.1241	11252	-0.0147	0.0206	0.4759	-0.37	9.57	0.7171
rs12452505 (17:63556402)	FVC	G / C	0.141	0.141	30371	-0.0052	0.0119	0.6643	0.139	0.1388	28056	-0.0122	0.0123	0.3226	0.41	8.52	0.6923
rs35524777 (18:11424551)	FEV_1	T / C	0.058	0.058	33341	-0.0107	0.0169	0.5257	0.056	0.0560	29935	-0.0206	0.0179	0.2490	0.40	19.00	0.6918
rs151310656 (19:16777665)	PEF	A / G	0.028	0.028	10455	0.0078	0.0433	0.8567	0.029	0.0290	9120	0.0196	0.0447	0.6606	-0.19	37.58	0.8505
rs67134151 (19:29510958)	FEV_1/FVC	G / T	0.073	0.073	36462	-0.0076	0.0144	0.5974	0.075	0.0750	32930	0.0043	0.0149	0.7732	-0.57	14.89	0.5744
rs28650353 (21:35166202)	FVC	A / G	0.010	0.010	30479	-0.0641	0.0415	0.1229	0.010	0.0097	27270	-0.1053	0.0439	0.0164	0.68	104.74	0.4962
rs2225434 (21:46634146)	FEV_1	C / T	0.570	0.430	37166	-0.0110	0.0075	0.1433	0.569	0.5690	33460	0.0061	0.0079	0.4393	-1.57	4.17	0.1888

6.4.8 Discussion

This chapter presented the largest genome-wide interaction analysis exploring gene-smoking interaction effects on lung function traits FEV_1, FEV_1/FVC, FVC and PEF undertaken to date. The purpose was to identify novel signals whose genetic effect on lung function is dependent on smoking behaviour. In a genome-wide interaction analysis of 303,612 individuals and 8,647,748 variants, a total of 58 signals were identified for producing an interaction effect with ever/never smoking (at a threshold of $p < 5 \times 10^{-6}$). 53 of these signals showed no statistically significant association with smoking traits in UK Biobank, and thus were not driven by an association with smoking behaviour. The identified signals did not replicate in a meta-analysis of 71,067 individuals from the SpiroMeta consortium, thus further independent replication efforts will be needed to determine the validity of the identified signals.

Of the 53 interaction signals, none were for signals previously identified for association with lung function. Five of the 58 signals were within 1Mb of a previously identified lung function signal, however r^2 values of 0.0240 and below suggested independence between the interaction signal, and the previously reported SNPs. This result, which is consistent with the analysis undertaken in chapter 5 (where none of the known marginal effect SNPs for lung function showed evidence of interaction effect) suggests that interaction effects may be produced by SNPs that do not contribute marginal effects.

The interaction signals identified in the discovery analysis in UK Biobank failed to replicate in the independent dataset used i.e. the SpiroMeta consortium meta-analysis. There is however a limitation here in that the maximum sample available (~ 71k individuals) is much smaller than the discovery sample in UK Biobank (~ 300k individuals). Furthermore, when considering N effective estimates which account for imputation quality of the SNPs from the contributing studies, 27 of the 53 signals had a follow up effective sample size of less than 50,000 individuals (~70% of the maximum attainable) and 15 had an effective sample size less than 50% of the maximum attainable sample size. The smallest effective sample size observed was as low as 19,575 individuals. This is less than 7% of the discovery sample size which was used to identify the interaction signal and less than 30% of the maximum attainable sample size in the SpiroMeta consortium. Therefore, the replication attempt does not necessarily

deem the interaction signals to be chance findings, but rather inconclusive findings, as power will be limited by the smaller replication sample size, and even more severely by the low effective sample sizes observed. Ideally, a replication sample size would resemble the discovery sample size when seeking replication, with independent replication considered the gold standard to confirm GWAS results (183). However, with resources such as the UK Biobank offering such large sample sizes, locating suitable independent datasets is currently difficult and will largely rely on meta-analysis efforts, with sample sizes that can be boosted by the availability of data from additional individual smaller studies. For example, more studies have since contributed to the SpiroMeta consortium which will increase the sample size from that presented here. Furthermore, future replication efforts will benefit from the availability of new imputation panels with greater variant density, which will increase imputation quality, and thus effective sample sizes will more closely resemble the maximum attainable sample size, to avoid large losses in power.

Five of the identified interaction signals in the genome-wide analysis were associated with smoking behaviour. This suggests that these signals are not observed due to differing genetic effect by smoking group, but are instead signals which have been identified due to an association between the signal and smoking. These signals are therefore acting on lung function via an addiction pathway, rather than being directly associated with lung function. Two of the five SNPs provided further evidence for the strong smoking signal at the 15q25 locus corresponding to SNP rs1051730 (64) and an association with CPD. Another signal rs2604894 on chromosome 19 was in high LD with a previous smoking signal rs7937 (62). The remaining two signals, rs11780592 and rs73555789 did not appear to be correlated with previous smoking signals, despite strong associations with smoking initiation and smoking cessation respectively ($p = 7.8 \times 10^{-9}$ and $p = 7.9 \times 10^{-4}$ respectively). Alternatively, other approaches could have been taken into account to address possible associations with smoking behaviour. For example, one could have included further smoking covariates in the regression model used for analysis, for example smoking cessation and pack-years. In this instance however, pack-years data was not available for the full sample, thus adjusting for this covariate would have penalised the sample size. Therefore, with the availability of a large well-powered GWAS of smoking behaviour, determining the association of the

identified interaction signals with smoking behaviour post analysis was the approach taken.

The significance threshold used to identify interaction signals here is more relaxed than that often applied in studies of this type, with $P < 5 \times 10^{-8}$ generally the threshold of choice for GWAS, based on a Bonferroni multiple testing correction for the number of independent SNPs. This threshold, although broadly used, is however suggested for GWAS of marginal genetic effects rather than interaction effects. With this in mind, and with the suggested loss of power to identify interaction signals in comparison to marginal effect signals, the decision made was to perform signal selection with a threshold of $P < 5 \times 10^{-6}$. Further adjustment was not made for the consideration of 4 lung function phenotypes due to the correlation between them (see **Figure 6.12** and **Figure 6.13**).

6.5 Conclusion

In conclusion, although the hypothesis free genome-wide interaction analysis presented in this chapter did identify a number of gene-smoking interaction signals, further research will need to be undertaken to conclude such signals with confidence. This is mainly due to the fact that resources large enough to provide replication of results are in demand, particularly when using sample sizes as large as UK Biobank for the discovery stage. Therefore, future research into the effect of gene-smoking interactions on lung function will rely on datasets or meta-analyses with comparable sample size to that of UK Biobank. This will ensure that power is not lost between discovery and replication stages, increasing the likelihood of replicating discovered interaction effects.

The following chapter will present a discussion of the work undertaken as part of this thesis, addressing the challenges, limitations and potential for future research.

Chapter 7 Discussion

Chronic obstructive pulmonary disease (COPD) which encompasses diseases emphysema and chronic bronchitis, is characterised in an individual by a severe airflow obstruction. It is currently one of the top causes of worldwide mortality and thus is a major concern in public health. The biggest risk factor for COPD is smoking, although there is also a reported genetic component, with shared genetic loci identified between COPD and heritable lung function traits, which are used in COPD definition. Although significant progress has been made in understanding the genetic contribution to poor lung function and COPD in recent years, much of the phenotypic variance is still unexplained and we still do not really understand why there are many smokers who do not develop the disease. To attempt to provide some clarity on these issues, the primary aim of this thesis was to identify gene by smoking interactions associated with lung function (thus determining whether genetic effects on lung function are dependent on smoking behaviour), for which there has been little robust evidence to date, likely due to power limitations. To combat this issue this thesis utilises the large UK Biobank resource, providing the best opportunity to date in terms of power, to discover interaction effects, with the hope to provide direction for drug targets and treatment development for COPD and poor lung function.

This final chapter presents a summary of what is known about COPD and lung function to date, before discussing the work undertaken in this thesis and the clinical implications of gene-environment interactions. The challenges and limitations of the work undertaken will then be presented before concluding with direction and ideas for future work in this area.

7.1 Summary of previous genetic studies of lung function and COPD

GWAS efforts in recent years have rapidly increased our understanding of the role of genetics in lung function, significantly increasing the number of loci known to be associated with lung function and COPD. Since the first reported association from a GWAS in 2009 (43), there are now nearly 300 identified loci reported in the literature for association with lung function and/or COPD (32,33,44-52,54,56-58,69,180). Approximately half of these signals were reported as novel findings in a recent 2018 GWAS of lung function utilising UK Biobank (32,33), which is the largest GWAS of

lung function to date. Loci identified to date for marginal genetic effect on lung function have predominantly been due to the effects of common variation (MAF > 5%), with individually small contributions to the observed variance of lung function traits.

So far, analyses undertaken to uncover potential gene-environment interactions (specifically considering smoking as the environmental variable) affecting lung function and COPD have been largely unsuccessful. This is most likely due to limitations in power, with large sample sizes needed to observe interaction effects (70,71). Studies undertaken have considered both a candidate gene (72-74) and genome-wide (75-77) approach, with the largest sample size considered approximately 50,000 individuals collated with the use of meta-analysis (75).

The UK Biobank dataset used in this thesis eclipses previously attainable sample sizes for the analysis of gene-smoking interaction effects with data available for more than 500,000 individuals, ten times the largest sample size considered in previous genome-wide gene-smoking interaction analyses. This resource therefore provides our best opportunity to date in terms of power, to identify genetic effects on lung function which are altered by smoking behaviour.

7.2 Summary of work undertaken as part of this thesis

In chapter 2, the concept of an interaction was introduced, before presenting how this could be incorporated and analysed using a regression framework. With gene-environment and gene-gene interactions gaining popularity due to the aforementioned reasons, so too has the demand for interaction analysis methods. The remaining sections of chapter 2 aimed to provide a comprehensive literature review of these methods, the various caveats to their application and the benefits and drawbacks of each approach. Methods were broadly grouped into 6 categories; joint tests, the case-only approach, screening approaches, data mining, gene-based methods and pathway-level analysis. The chapter highlights the fact that there are numerous methods available to efficiently capture interaction effects, but in application choice becomes limited by study design, research question, and computational complexity, which given the substantial sample sizes from resources such as UK Biobank, must be considered. The joint test of Kraft et al. (83) and the standard interaction test incorporated within a regression framework

were identified as the most suitable methods for gene-smoking interaction analysis in proceeding chapters. This decision was based on the criteria that these methods are computationally simple (in anticipation of genome-wide application on imputed data), are able to incorporate quantitative traits, and are methods suitable for the analysis of individual SNPs (rather than SNP-sets).

Chapter 3 presented a simulation with the primary aim to determine the power available for gene-smoking (ever/never smoking) interaction analysis in UK Biobank. Secondary aims were to determine power dependency on method choice, sample size (300,000 individuals chosen to represent the sample size in UK Biobank post quality control, and 100,000 individuals to represent previously attainable sample sizes), minor allele frequency and interaction effect size. The simulation considered the two methods chosen in chapter 2, the standard interaction test and the joint test of Kraft et al. (83) using two main simulation scenarios; main and interaction effect present or interaction effect present only. Assuming rare variants produce larger effect sizes, the interaction and joint tests could detect effects of 1.5 times and 1.05 times larger respectively in exposed individuals than unexposed individuals, when main and interaction effects were present. When no main effect was present, the interaction test performed similarly, however the joint test required 1.5 – 2 times larger effect in exposed individuals. These power estimates were largely dependent on the assumption of increasing effect size for smaller MAF, with the scenario where effect size was constant across MAF requiring much larger interaction effects to achieve the same power. Furthermore, the detectable effects were smaller when considering 300,000 individuals in comparison to 100,000 individuals as expected. The joint test was marginally more powerful in the majority of scenarios, including when only the interaction effect was present, highlighting its potential robustness when characteristics of the SNPs analysed are unknown. However, the joint test and interaction test do ask distinct statistical and biological questions, thus this result should be interpreted with care.

Chapter 4 presented a candidate region gene-smoking interaction analysis, using ever/never smoking as the smoking variable (as used for all interaction analyses throughout this thesis). The analysis focussed on 70 regions of the genome which have already been implicated in lung function through previous marginal effect studies (97 signals to date of analysis), under the assumption that we may be more likely to see

interaction effects in regions already associated with lung function. The analysis incorporated a two stage design. The joint test was first applied to all SNPs individually within a region equidistant around previously reported associated lung function loci, given its robust performance in simulation in chapter 3. In a second stage, the interaction test was applied to determine whether the interaction (rather than the main genetic effect) was the driving mechanism behind the association. The analysis included 303,619 unrelated Europeans from UK Biobank and 1,831,014 SNPs imputed to the Haplotype Reference Consortium (HRC) panel (15), and included two lung function traits, FEV_1 and FEV_1/FVC. 632 independent joint test signals were identified and two of these produced a statistically significant interaction with ever/never smoking in the second analysis stage. The analysis identified a SNP downstream of *LTBP4* associated with FEV_1 and a SNP in *GNB1L* (Guanine nucleotide-binding protein subunit beta-like protein 1) associated with FEV_1/FVC, a gene with a reported role in autism and schizophrenia (176). Both SNPs were independent of the known lung function signal in their defined candidate region, and were not associated with smoking traits, indicating that the larger effect in ever smokers was not driven by smoking behaviour. Neither of the SNPs replicated however in a sample size of ~71k in the SpiroMeta consortium, possibly due to power limitations due to a smaller sample than that used for discovery. Therefore these signals warrant further research.

Chapter 5 presented an interaction analysis of the 279 marginal effect signals identified for lung function to date, to determine whether signals that show strong marginal effect are dependent on ever/never smoking status. Furthermore, the interaction of all 279 SNPs with ever/never smoking was analysed using a weighted genetic risk score (wGRS), to determine whether the combined effect of all marginal effect SNPs was dependent on smoking behaviour (wGRS by smoking interaction analysis). The analysis used the same sample of 303,619 unrelated European individuals from UK Biobank from chapter 4. None of the 279 marginal effect signals produced a statistically significant interaction with ever/never smoking (when tested for association with the phenotype for which the signal was initially identified). These results are presented in Shrine et al. (32,33) (in press – Nature Genetics, the bioRxiv version is presented in **Appendix J**). There were statistically significant results for both FEV_1 ($p = 0.027$) and PEF ($p = 0.002$) at a threshold of $P < 0.05$ for the wGRS by smoking interaction analysis. For both traits, the interaction effect suggested a larger deleterious effect on

lung function for ever smokers compared with never smokers as the risk score increased, however the observed effects were small (~0.01). This provides further statistical evidence of interaction effects between risk scores (calculated using lung function associated SNPs) and smoking behaviour, with Aschard et al. reporting a differing risk score effect on lung function depending on ever/never smoking status (184). The work I undertook in this chapter provides direction for current risk prediction estimates, under the premise that ever and never smokers carry the same *genetic* risk of poor lung function and COPD, when utilising the most recent set of identified risk SNPs.

The aim of chapter 6 was to remove any restriction with regards the search space for gene-smoking analysis (with analysis restricted to regions and/or SNPs in chapter 4 and chapter 5), and presents the largest genome-wide interaction study (GWIS) with smoking behaviour to date for lung function, specifically analysing traits FEV_1, FEV_1/FVC, FVC and PEF. The analysis included 303,612 individuals (7 removed from previous analyses due to withdrawal from UK Biobank) and 8,647,748 variants imputed to the Haplotype Reference Consortium (HRC) panel. Using a threshold of $P < 5 \times 10^{-6}$, the analysis identified 58 signals that interacted with smoking behaviour (ever/never smoking), 53 of which were not associated with smoking behaviour directly. Five of the 53 interaction signals were in close proximity of previously reported signals associated with lung function (< 1Mb), however none were correlated with their respective previously reported signals, with LD less than 0.0240. None of the 53 signals replicated in the SpiroMeta consortium (~71,000 individuals) however the replication attempt was penalised by small effective sample sizes, with over half of the signals having effective sample sizes of less than 50,000, with the smallest as little as 19,575 (~28% of the maximum attainable sample size). Therefore, further follow up of these signals is needed with more power in a larger independent sample size to determine whether these interaction signals are genuine.

7.3 Clinical implications of gene-environment interactions and the work undertaken

Although identifying gene-smoking interactions for lung function has so far been unsuccessful in the literature, and signals identified here require further follow up

efforts, uncovering such effects could have important clinical implications. Firstly, should loci be found for which genetic effect on lung function is dependent on smoking behaviour, then this could uncover biological pathways that lead to the development of personalised medicine for individuals, based on their exposure to smoking. For example, if an interaction signal identifies a biological process which leads to COPD in ever-smokers only, this could inform the development of specific treatment in ever-smokers that might differ from therapies suitable for those who have never smoked. Secondly, such results would aid in prediction modelling, to forecast the lung function and COPD outcomes of individuals with the relevant genetic variants, given their exposure to smoking. High risk individuals for COPD can already be identified without the requirement of an interaction between genes and the environment (31,32). We already know for example that individuals are at higher risk if they smoke and are genetically predisposed to the disease. However, should we discover a gene by ever-smoking interaction signal in the future, this could identify a subgroup at particularly high risk, that are not currently defined when assuming independence between genetic and smoking effects. This information could benefit potential genetic screening methods to target individuals for exposure reduction therapies and medication (i.e. smoking cessation services), although ethics issues which arise with genetic screening would need to be considered (33). Additionally, if individuals are informed that they carry the interacting genetic variant, they may be more likely to address their smoking habits. A study by Sparks et al. (34) for example found a link between the disclosure of personalised genetic risk for rheumatoid arthritis (RA) and an improvement of behaviours which increase the risk of the disease. Similarly, the Finnish GeneRISK study found that informing individuals of their genetic risk for cardiovascular disease increased beneficial lifestyle changes such as weight loss and giving up smoking (35).

Ultimately, the clinical message remains the same, quitting smoking is vital in reducing the risk not only of respiratory disease but also of other diseases. Identifying genetic signals for which the effect is more severe or only present in smokers however, is still extremely beneficial, particularly because the effects of prolonged exposure to smoking may be irreversible and there are individuals that are more genetically susceptible to nicotine addiction (7,36-38).

Specific to the work undertaken in this thesis, I have shown that the 279 marginal effect signals associated with lung function traits, do not depend on ever/never smoking status. Given this result, we can determine relative and absolute risks for all individuals, based on all 279 variants, collated using a risk score (17,18). For example, we can produce infographics such as that in **Figure 7.1,** which presents the absolute risk of COPD for smokers with high and low genetic risk (comparing the highest and lowest deciles of risk score), and the effect quitting smoking has on absolute risk estimates.

The work to identify novel SNP-smoking interactions could have further clinical implication, such that, if SNPs are identified for producing an interaction effect with smoking, it may suggest that separate risk scores for ever and never smokers would be more appropriate and infographics such as that presented in **Figure 7.1** would need to account for this. Although there is statistical evidence of a GRS by smoking interaction, the effects were very small. Care would have to be taken however when presenting these updated infographics to the media. If smoking risk on COPD is higher given the individual has a certain genetic variant, then individuals without it may conclude that smoking is not as harmful for them.

With such clinical implications, efforts should be made to maximise power and in turn, our ability to detect their effects. Suggestions for how to do this are discussed in section 7.5, which considers future directions for gene-smoking interaction research.

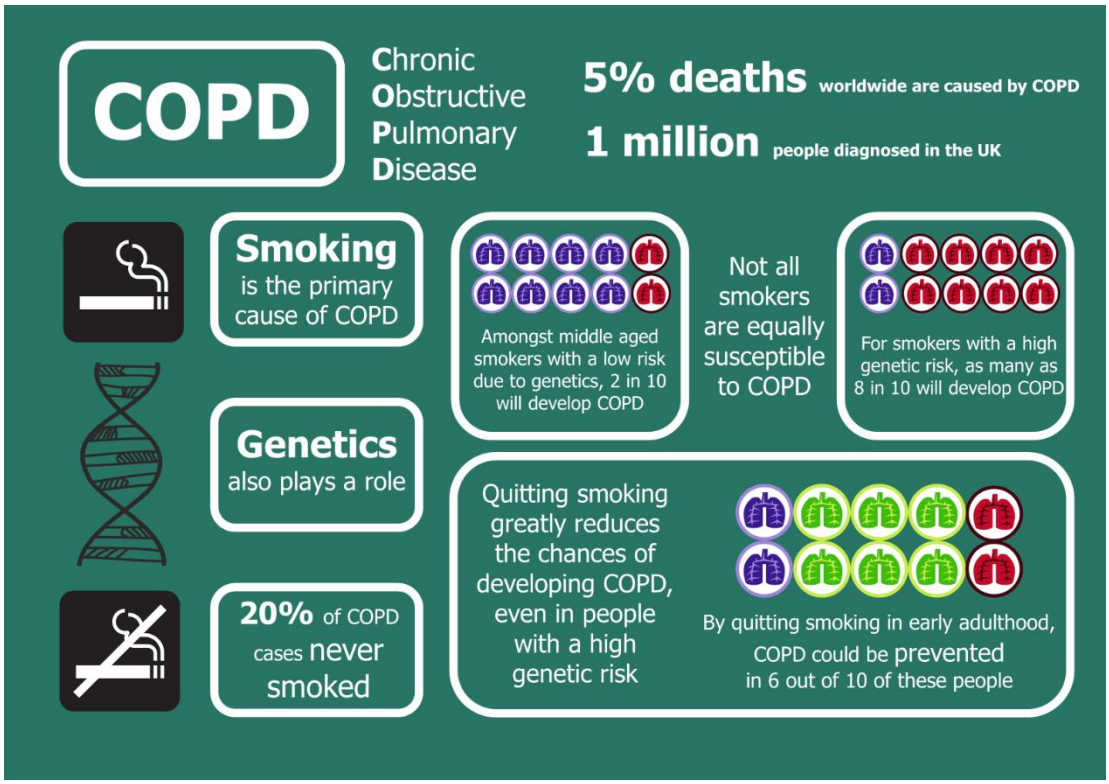


Figure 7.1 - Absolute risk of developing COPD for smokers with low and high genetic risk (infographic created with the use of risk score analysis for the 279 lung function and COPD associated loci from Shrine et al. (17,18))

7.4 Challenges and limitations

As is the case with all research, this thesis does have limitations. Firstly, for the SNPs showing putative SNP-smoking interactions in chapters 4 and 6, potential replication was severely limited by the available sample size. Although this analysis benefited from the availability of in-house individual level data for replication, there were also disadvantages to the replication data available in the SpiroMeta consortium. Specifically, although the consortium was large, the sample size in comparison to that of UK Biobank used for the discovery of gene-smoking interaction signals was much smaller (~71,000 compared with ~300,000). Additionally, there were further penalisations to sample size for two reasons. Firstly, not all studies contributed to the replication sample for each SNP, possibly due to quality control or not being imputed. Secondly, further reductions were evident for effective sample sizes, which account for imputation quality. Given the size of UK Biobank, locating independent data with comparable sample size is difficult, thus the approach was to seek replication in as large a sample as was available, whilst also taking into consideration the time constraints on

this thesis and the analyses undertaken. Moving forward, sample size with respect to the SpiroMeta consortium in particular, has since been improved with the inclusion of additional studies and imputation can be improved by the denser imputation panels available e.g. imputation to HRC panel (39) for all studies.

The smoking variable chosen to represent smoking behaviour may not provide the best representation of smoking behaviour for each individual. The pragmatic approach taken was to consider ever and never smoking as a categorical environmental exposure, in order to assign individuals to smoking groups. However, this could be a generalised representation and does not take into account characteristics such as exposure duration, age of exposure, or the magnitude of the exposure. An alternative option here would have been to consider a smoking trait such as pack-years, which takes into account both length and magnitude of exposure, calculated using information on the number of cigarettes smoked per day and the number of years smoked. This approach was considered when designing the analyses undertaken in this thesis, but was later reconsidered in favour of ever/never smoking due to a large number of individuals not having pack-years data available, which would impact on sample size. Specifically, for the 303,619 individuals with ever/never smoking information available analysed in chapter 4, 44,536 of the 139,288 ever-smokers did not have pack-years information available, thus resulting in the loss of approximately 15% of the maximum sample size. This would have resulted in a loss of power, limiting our ability to identify gene-smoking interactions. Additionally, there are documented effects of exposure misclassification on analysis results (40-42) and issues could arise if individuals have been assigned to the wrong smoking group (never smokers to the ever smoking group and vice versa). This could be the case specifically when sub setting individuals based on questionnaire data which relies on honesty and accuracy from participating individuals, as the only way to record smoking behaviour (see chapter 4 section 4.5 for ever/never smoker criteria), with individuals often understating smoking habits. An example of this in application is the emergence of the well reported smoking locus, 15q25, which appears in COPD case-control analysis, even after adjustment for smoking behaviour (43). Bias could be introduced most severely when allocating occasional smokers for example, which are hard to allocate to ever/never smoking groups in their own right, but additionally difficult due to definitions relying on the number of cigarettes smoked in a lifetime. It is possible that the answer to this question

could be under or over estimated by an individual, thus assigning them to the wrong category. Defining criteria used for ever and never smokers was however the subject of much discussion and was defined as accurately as possible here. Furthermore, a strength of this thesis was having access to GWAS results for smoking behaviour, which ensured that interaction signals driven by association with smoking behaviour were identified. For example, the 15q25 locus for smoking behaviour was identified as an interaction signal as part of this thesis, but access to the smoking behaviour GWAS allowed us to conclude that this was acting via an addiction pathway rather than genetic effect differences on lung function between ever and never smokers.

A further concept to consider is the dependency of the presence of gene-environment interaction on the phenotype transformation used (or not used in the case of analysing raw phenotype data) as explored by Murray et al. (44). It is suggested that the method used to transform the phenotype could determine whether an interaction is observed and additionally whether it is observed with the correct direction of effect. The approach taken here was to inverse normalise phenotypes, but it could be possible that leaving the phenotype data untransformed may have produced different results. Transforming phenotypes in this way however ensures that effect sizes produced as a result of analysis undertaken are generalisable and not specific to the data being analysed, making them accessible and interpretable for others exploring the same phenotype, for example, in different populations. This is the approach taken in many marginal effect genome-wide association studies and was suggested as a solution to avoid unbiased results by Murray et al. (44).

An additional challenge with this thesis was the limitations of available software. Although genomic analysis benefits from large sample sizes, the software available to analyse sample sizes of several 100,000 individuals, using methods other than those implemented in marginal effect association analysis, is restricted. The first challenge is choosing an appropriate interaction method, the second challenge is identifying software which will apply it. Finding software that would run the joint test for example, was difficult, and further complications were encountered due to the limitations of the chosen software with regards the inclusion of related individuals, which due to the substantial sample size being used, had to be excluded due to difficulties in computing a kinship matrix. This resulted in a loss of more than 17,000 individuals. Given more

time, it may have been possible to develop my own software, which could apply the methods desired whilst retaining related samples. This was however not the focus of this thesis and could be a possible avenue for exploration in the future.

Finally, as discussed in chapter 1, it is important to reiterate that the UK Biobank sample used for this analysis is one which is healthier than the general population and thus does raise issues when putting results into context. Although this may create problems with generalisability in epidemiological studies with regards incidence of disease between sample and population, this issue will however be less problematic for genetic association studies. The main issues in this context will be a loss in power for genetic association analysis due to restricted distributions of analysed phenotypes and environmental exposures that may not be representative of the population as a whole i.e. spirometry and smoking behaviour. Additionally, generalisability of results for other populations may be difficult given that UK Biobank is predominantly European. Nevertheless, the substantial sample size available and scope for the exploration of genetic contributions for multiple diseases and traits makes it a very important resource for genetic research and allowed for the largest gene-smoking interaction analysis in lung function to date.

7.5 Future work

The work I have undertaken here has shown that to date, the SNPs used for risk prediction of COPD do not show differences in genetic effect between ever and never smokers, removing the need for subgroup specific risk prediction based on smoking behaviour. With regards to identifying novel gene-smoking interactions, this work has identified SNPs with potential interaction effects, but lack of replication results are inconclusive and warrant further research. The pertinent question with regards to the work undertaken here, and previous work regarding gene-smoking interactions associated with lung function and COPD is; Are gene-smoking interactions important in lung function and COPD, and should time and resources be spent researching them? Given the lack of findings to date, one could argue the case that research efforts may be better directed towards marginal effects, which as discussed have proven extremely successful. However, although successful in terms of number of identified genetic loci for various complex diseases, contributions towards heritability estimates have been

small, with current contributions for lung function traits estimated to be approximately 10% (18) with heritability estimates of 40% (45). It is unclear whether the contributors of this missing heritability are marginal genetic effects yet to be detected or due to more complex biological pathways such as interactions between pairs of genes or genes and the environment. For example rare variants may explain some of this missing heritability and although identifying such effects has been relatively unsuccessful, this does not necessarily mean they do not play a role in lung function, but more likely that we are still underpowered to detect them. In addition, there are still a number of areas, particularly around the effect of gene-smoking interactions on lung function that need further research, before there can be any final conclusions made regarding the significance of their role, and whether or not they are contributors to the heritability of lung function traits. With regards to the work undertaken in this thesis, there is so far no evidence that interactions are the contributors of the missing heritability for lung function traits.

Firstly, although resources such as UK Biobank provide a fantastic resource for genetic epidemiological research for a number of phenotypes and research questions, replication efforts will likely fall short due to contrasts in sample size, specifically due to power limitations. Therefore, particularly for the work undertaken in this thesis where results provide direction but not conclusive evidence, we should seek larger independent samples with lung function data for which powerful replication can be achieved. We are already in the process of developing a more powerful replication resource with regards the SpiroMeta consortium, with current estimates of sample size now approaching 100,000 individuals due to the addition of new studies, and the use of new imputation panels, such as the HRC panel (28), will mean better coverage to aid in replication of discovery signals.

As discussed in section 7.4 as a limitation, future research into the effect of gene-smoking interactions on lung function and COPD could benefit from the consideration of a smoking behaviour variable which takes into account magnitude of the exposure, the length of time the individual was exposed for and at what stage of life the exposure took place. As discussed, pack-years takes these criteria into account. Currently the best option we have is to undertake gene-environment interaction analysis in the individuals

for which we have pack-years data available, but we could look for additional samples with this information available elsewhere for use in a meta-analysis framework.

This thesis has considered the interaction effect of individual SNPs along the same theoretical pipeline as single variant marginal effect analysis. The work undertaken here did not consider the prospect of genes or SNP-sets (for which SNPs are grouped using other criteria such as functionality). Given the well documented loss in power in rare variant marginal effect analysis (46) and when considering interactions compared with marginal effects (19,20), it is then plausible to believe that considering rare variants individually in gene-environment interaction analysis will be underpowered. The approach to combat the loss in power of rare variant marginal effect analysis is to simultaneously test multiple SNPs together using gene-based approaches (for example burden tests). Therefore, future work in this area could benefit from applying the same approach to SNPs in an interaction setting.

The work in this thesis only considered variants imputed to the Haplotype Reference Consortium (HRC) imputation panel (39). UK Biobank imputed variants using two different imputation panels, the HRC panel as previously mentioned and the UK10K + 1000 Genomes panel (47). However, an error in allocating genome position to UK10K + 1000 Genomes imputed variants, meant that these data could not be used. This has been corrected by UK Biobank since the interaction analysis undertaken in this thesis. Additionally, future interaction analysis would benefit from the imminent release of more dense imputation panels than that offered by the Haplotype Reference Consortium (28) and the UK10K + 1000 Genomes panel (47), such as that provided by TOPMed (48). Furthermore, chromosome X was not analysed as part of this thesis but is available in UK Biobank. Analysis of the X chromosome brings new complexities beyond autosomal analysis however, particularly with regards interaction analysis. This is due to being mainly haploid in males and diploid in females, thus there would need to be consideration of how to model this in a statistical framework.

Although imputation is a useful and cost effective tool for determining genotypes for SNPs not included on the direct genotyping array, the gold standard for observing variation along the genome (particularly rare variation) is the use of whole genome sequencing (WGS), which captures and records the full nucleotide sequence of each

individual. This would hugely benefit the analysis of gene-environment interactions, providing a much denser and accurate genetic profile for which to identify associations. Although WGS data is not available for UK Biobank individuals, whole exome sequencing (WES) is underway for the full UK Biobank dataset, which will more accurately capture rare and low frequency genetic variation within protein-coding areas of the genome. This will allow for a more comprehensive interaction analysis of low frequency and rare variants within these regions. The genetic data is expected to be available to analysts this year (2019). Variants showing gene-smoking interactions within these regions could impact on protein structures, and may assist in understanding the biological pathways and possibly guide the development of new treatments.

If it is the case that even with sample sizes as large as that provided by UK Biobank we are still underpowered for gene-environment interaction analysis, given that this sample size is unlikely to be superseded in the near future, other options need to be considered to maximise power. One way to increase power would be to further explore the options available for screening the full search space of variants in order to reduce the penalisation of multiple testing. For example, a recent study by Wang et al. (49) used the theory of variation prioritisation suggested by Pare et al. (50) (discussed in chapter 2 section 2.2.4.1), in which presence of an interaction effect (with an unmeasured exposure variable) could be inferred by an observed difference in phenotypic variance across genotype groups at a particular SNP. The study identified 75 signals associated with 9 traits using 348,501 unrelated Europeans from UK Biobank. The benefit of this approach is that the environmental variable need not be quantified and thus sample size can be increased, in turn boosting power. Specific to this work, this approach could be used to screen variants that potentially interact with an environment variable to affect lung function, using a second analysis stage to determine whether the responsible environmental variable is smoking behaviour. Other screening options based on functionality could also be considered for gene-smoking interaction analysis moving forward. For example, restricting analysis to SNPs based on their proximity to CpG (cytosine-guanine dinucleotides) sites associated with smoking behaviour (51,52), areas of the genome where DNA methylation predominantly occurs, which contribute towards the observed variation of gene expression levels between individuals (53,54). These areas could be of interest for interaction analysis due to the resulting functional relevance of any identified interaction. For example, should a SNP within or near a CpG

site have a genetic effect on lung function which is dependent on smoking behaviour, then this may be due to the effect of smoking behaviour on methylation levels, providing direction for treatment development.

7.6 Conclusion

In conclusion, the work undertaken as part of this thesis has aimed to identify gene-smoking interactions that affect lung function. To do so, this thesis presents a comprehensive review of gene-environment interaction methods and an understanding of the power available for interaction analysis in UK Biobank. The analysis undertaken focussed on regions known to be associated with lung function, SNPs with strong marginal effect on lung function, and presents the largest genome-wide gene-smoking interaction analysis to date, using data for over 300,000 individuals. Analysis informs the most up to date estimates of relative and absolute risk of developing COPD, according to genetic variation and smoking status. The analyses also identified 55 loci whose effects could be dependent on smoking behaviour, with increased replication sample sizes with better imputation quality needed in the future to confirm these results. This work does however provide direction for future interaction analysis and results for further research, which will aid in uncovering how genetics affects lung function, and specifically its dependency on exposure to smoking. This contributes towards achieving a common goal in lung function and COPD genetic research, which is to develop effective treatment and improve risk prediction.

Appendix

A. The 139 novel signals identified in Shrine et al. (1,2) from a GWAS of marginal genetic effect on lung function using both UK Biobank and the SpiroMeta consortium.

Reported RSID	CHR	POS	TRAIT	GENE
rs9661802	1	6678864	FEV_1/FVC	*PHF13*
rs12737805	1	22612690	FEV_1	*MIR4418*
rs9438626	1	26775367	FVC	*DHDDS*
rs12096239	1	26796922	FEV_1	*DHDDS*
rs1416685	1	51243374	FEV_1/FVC	*FAF1*
rs72673461	1	60966772	FEV_1/FVC	*LOC101926964*
rs9661687	1	78387270	FEV_1/FVC	*NEXN*
rs10874851	1	92106637	FEV_1/FVC	*TGFBR3*
rs9970286	1	111737398	FEV_1/FVC	*DENND2D*
rs11205354	1	150249101	PEF	*C1orf54*
rs141942982	1	155137395	FEV_1/FVC	*KRTCAP2*
rs4651005	1	178719306	FEV_1	*RALGPS2*
rs2146098	1	186090370	FVC	*MIR548F1*
rs17531405	1	186113852	FEV_1/FVC	*MIR548F1*
rs10919604	1	198898157	FEV_1/FVC	*MIR181A1HG*
rs4309038	1	201884647	FEV_1/FVC	*LMOD1*
rs2799098	1	218521609	FEV_1/FVC	*TGFB2*
rs75128958	1	219483218	FEV_1/FVC	*LYPLAL1*
rs17009288	1	221204299	FVC	*HLX*
rs2544536	2	15906854	FEV_1/FVC	*LOC101926966*
rs6751968	2	18570024	FVC	*RDH14*
rs13430465	2	18702313	FVC	*RDH14*
rs13009582	2	24018480	FVC	*ATAD2B*
rs732990	2	26842146	FVC	*CIB4*
rs4952564	2	42243850	FVC	*PKDCC*
rs12470864	2	102926362	FEV_1/FVC	*IL1RL1*
rs1406225	2	145797829	FEV_1/FVC	*TEX41*
rs7424771	2	161276378	FEV_1	*RBMS1*
rs2304340	2	179260382	FEV_1	*MIR548N*
rs2084448	2	187530520	FEV_1/FVC	*ITGAV*
rs1249096	2	199723365	FVC	*SATB2*
rs985256	2	201208692	FEV_1/FVC	*SPATS2L*
rs12997625	2	202970250	FVC	*KIAA2012*
rs6435952	2	217614730	FEV_1/FVC	*IGFBP5*
rs4294980	2	218604356	FEV_1	*DIRC3*
rs4674407	2	220382700	FVC	*ASIC4*
rs6431620	2	239604970	FVC	*LINC01107*
rs6437219	2	241844033	FVC	*C2orf54*
rs6733504	2	242495953	FVC	*BOK-AS1*
rs2974389	3	13787641	FEV_1	*LINC00620*
rs73048404	3	25179533	FVC	*RARB*
rs35480566	3	71583177	FVC	*FOXP1*
rs586936	3	73862616	FEV_1/FVC	*PDZRN3-AS1*
rs1610265	3	99420192	FVC	*MIR548G*
rs1799807	3	165548529	FEV_1/FVC	*BCHE*
rs6780171	3	185503456	FEV_1	*IGF2BP2*
rs12331869	4	56012149	FEV_1	*KDR*

Reported RSID	CHR	POS	TRAIT	GENE
rs62316310	4	75676529	FEV_1/FVC	*BTC*
rs11098196	4	79403952	FEV_1/FVC	*FRAS1*
rs13109426	4	145330628	FVC	*HHIP-AS1*
rs13116999*	4	145442364	PEF	*HHIP-AS1*
rs11739847	5	609661	FEV_1	*LOC100996325*
rs4866846	5	43976162	FEV_1	*NNT*
rs10059661	5	121410529	FEV_1/FVC	*LOX*
rs17163397	5	128767384	FEV_1/FVC	*ADAMTS19-AS1*
rs1800888	5	148206885	FEV_1	*ADRB2*
rs10059996	5	170901463	FEV_1/FVC	*FGF18*
rs79898473	5	179598771	FEV_1/FVC	*RASGEF1C*
rs12198986	6	7720059	FVC	*BMP6*
rs9689096	6	34188892	FVC	*HMGA1*
rs9357446	6	44447598	FVC	*CDC5L*
rs12202314	6	45530471	FEV1/FVC	*RUNX2*
rs9472541	6	45622748	FVC	*RUNX2*
rs2894837	6	56336406	FEV_1	*RNU6-71P*
rs2627237	6	134339265	FEV_1	*SLC2A12*
rs1102077	6	140271357	FEV_1	*LOC100507477*
rs9385988	6	142560957	FEV_1	*VTA1*
rs4721457	7	15872324	FEV_1/FVC	*MEOX2-AS1*
rs559233	7	26848830	FEV_1	*SKAP2*
rs62454414	7	27182329	FVC	*HOXA-AS3*
rs1513272	7	28200097	FEV_1	*JAZF1*
rs17232687	7	46448518	FVC	*IGFBP3*
rs12707691	7	84569510	FEV_1	*SEMA3D*
rs193686	7	116431427	FEV_1/FVC	*MET*
rs330939	8	9018590	FEV_1/FVC	*PPP1R3B*
rs4128298	8	11823332	FEV_1	*DEFB136*
rs7465401	8	70367248	FEV_1	*LOC100505739*
rs7838717	8	145504343	FVC	*BOP1*
rs7041139	9	18013733	FEV_1	*SH3GL2*
rs72743974	9	98878881	FEV_1/FVC	*LOC158434*
rs57649467	9	101632854	FEV_1/FVC	*GALNT12*
rs967497	9	131943843	FEV_1	*IER5L*
rs1274475	10	34480582	FEV_1/FVC	*PARD3*
rs60820984	10	75639578	PEF	*CAMK2G*
rs11191841	10	105639611	FEV_1	*OBFC1*
rs10836366	11	35308988	FEV_1/FVC	*SLC1A2*
rs56196860	12	2908330	FVC	*FKBP4*
rs12811814	12	4243749	FEV_1	*CCND2-AS1*
rs10841302	12	19808912	FEV_1/FVC	*AEBP2*
rs1244869	12	65075332	FEV_1/FVC	*RASSF3*
rs11176001	12	66409367	FEV_1	*MIR6074*
rs972936	12	102824921	PEF	*IGF1*
rs2701110	12	114669870	FEV_1	*TBX5*
rs9533803	13	44820608	FEV_1/FVC	*MIR8079*
rs2812208	13	50707087	FEV_1	*DLEU1*
rs803765	13	71647588	FVC	*LINC00348*
rs4885681	13	80467235	FEV_1	*LINC00382*
rs11620380	13	99665512	FEV_1/FVC	*DOCK9*
rs9634470	13	109918493	FEV_1/FVC	*MYO16*
rs1951121	14	23429729	FEV_1/FVC	*HAUS4*
rs74053129	14	54346010	FEV_1/FVC	*MIR5580*
rs10141786	14	74817418	FVC	*VRTN*
rs34245505	15	40397191	FVC	*BMF*
rs2304645	15	40716253	FEV_1	*IVD*

Reported RSID	CHR	POS	TRAIT	GENE
rs4924525	15	41255396	FVC	*CHAC1*
rs79234094	15	49409527	FEV_1/FVC	*COPS2*
rs35251997	15	49706145	FEV_1/FVC	*FAM227B*
rs62012772	15	63866877	FEV_1/FVC	*USP3*
rs7176074	15	73833600	FEV_1/FVC	*REC114*
rs3751837	16	3583173	FVC	*CLUAP1*
rs56104880	16	4361138	FEV_1/FVC	*GLIS2-AS1*
rs11074547	16	10136889	FVC	*GRIN2A*
rs76219171	16	50188929	FVC	*PAPD5*
rs35420030	16	53935407	FEV_1/FVC	*FTO*
rs12918140	16	86403821	FEV_1/FVC	*LINC00917*
rs6539952	16	86579223	FEV_1	*MTHFSD*
rs8082036	17	3882613	FEV_1/FVC	*ATP2A3*
rs4796334	17	6469793	FEV_1	*PITPNM3*
rs1215	17	7163350	FVC	*CLDN7*
rs4968200	17	7448457	FEV_1	*TNFSF12-TNFSF13*
rs34351630	17	16030520	FVC	*NCOR1*
rs12945803	17	46552229	FVC	*LOC101927166*
rs28519449	17	54195453	FVC	*ANKFN1*
rs8068952	17	59286644	FEV_1/FVC	*BCAS3*
rs77672322	17	62497964	FVC	*DDX5*
rs11653958	17	62686730	FEV_1/FVC	*SMURF2*
rs996865	17	69371318	FEV_1/FVC	*CASC17*
rs59606152	17	79952944	FVC	*ASPSCR1*
rs8089099	18	10078071	FEV_1/FVC	*VAPA*
rs1985511	18	19816712	FEV_1/FVC	*GATA6*
rs303752	18	21074255	FVC	*C18orf8*
rs1668091	18	22290711	FVC	*LOC729950*
rs9807668	18	42827898	FEV_1	*SLC14A2*
rs2202572	18	53566471	FVC	*LOC101927273*
rs11085744	19	10819967	FEV_1	*QTRT1*
rs2967516	19	36881643	FVC	*ZFP82*
rs6032942	20	10745545	FEV_1	*LOC101929395*
rs12627254	21	35368402	FEV_1/FVC	*LINC00649*
rs113111175	22	50867711	FEV_1	*PPP6R2*

B. Code for simulation in chapter 3. The code includes the function created to simulate the data and an example scenario (300,000 individuals, MAF=0.4, main effect present, varying effect sizes for rarer MAF)

```
# Set up a function for the simulation with n individuals, minor allele frequency (maf) and indicators for whether effect sizes vary or not, and whether main effect is present or not

     unitest<-
     function(n=300000,maf=0.4,consteffsize=TRUE,maineffect
     =TRUE){

     repeat {

# Sampling 1 SNP for 300000 individuals

     G<-rbinom(n,2,maf)

# Simulate smoking data (prevalence based on biobank)

     sm<-rbinom(n,1,0.44)

     if
     (length(which((G==1|G==2)&sm==0))!=0&length(which((G==
     1|G==2)&sm==1))!=0) break

     }

# Create interaction variable

     int<-sm*G

# Set b0=0

     b0<-0

# Choose effect for SNP

     if(maineffect&consteffsize){b1<-
sqrt(0.002/(0.4*0.6*2))
     } else if (maineffect&consteffsize==F){b1<-
     sqrt(0.002/(maf*(1-maf)*2))
     } else {
     b1<-0
     }

# Choose effect for ever/never smoking (based on contribution to variation of FEV1 in UK Biobank)
```

```
     b2<-0.1

# Choose interaction effect

# Create fifteen interaction effect sizes

     ratio <- c(0.05,seq(0.1,0.5,0.1),1,1.5,seq(2,8,1))


# Create interaction effect based on primary and secondary
scenarios e.g. main or no main effect, varying or constant
effects


     if(maineffect){beta3<-ratio*b1}
     if(maineffect==F&consteffsize){beta3 <-
     ratio*sqrt(0.002/(0.4*0.6*2))}

     if(maineffect==F&consteffsize==F){beta3 <-
     ratio*sqrt(0.002/(maf*(1-maf)*2))}

# Run simulation for each b3 value

     jtestp<-numeric(15)
     inttestp<-numeric(15)
     conttovarint<-numeric(15)


     for (b3 in beta3){

# Create error variance

     simfev1<-b0+b1*G+b2*sm+b3*int
     sigma<-sqrt(1-var(simfev1))
          err <- rnorm(n,0,sigma)

# Simulate phenotype

     simfev1<-b0+b1*G+b2*sm+b3*int+err

# Run required tests (joint and interaction)

# Interaction test

     fit<-lm(simfev1~G+sm+int)
     sumfit1<-summary(fit)

# Joint test (required second model)

     fit2<-lm(simfev1~sm)
```

```
     sumfit2<-summary(fit2)
     jointtest<-anova(fit,fit2)


# Pull out p-values for each test

     inttest<-sumfit1$coefficients[4,4]
     jtest<-jointtest$"Pr(>F)"[2]


# Pull out contribution of interaction to variance of
phenotype

     fitwoint<-summary(lm(simfev1~G+sm))
     conttovar<-sumfit1$r.squared-fitwoint$r.squared

#Store in results in vectors

     intsize<-which(beta3==b3)
     jtestp[intsize]<-jtest
     inttestp[intsize]<-inttest
     conttovarint[intsize]<-conttovar

     }

#return the three vectors of p-values

     list(int=inttestp,joint=jtestp,twomod=twomodtestp,cont
     tovar=conttovarint)
     }

# Simulate one scenario: 300,000 individuals, main and
interaction effect, common variant with maf=0.4, varying
effect sizes (so n=300000, maf=0.4, consteffsize=FALSE,
maineffect=TRUE)

itestcvarymandi<-data.frame(matrix(0,1000,15))
jtestcvarymandi<-data.frame(matrix(0,1000,15))
conttovarcvarymandi<-data.frame(matrix(0,1000,15))

for (i in 1:1000){
dat<-unitest(300000,0.4,FALSE,TRUE)
itestcvarymandi[i,]<-dat$int
jtestcvarymandi[i,]<-dat$joint
conttovarcvarymandi[i,]<-dat$conttovar
print(i)
}
```

C. Joint test results for previous known signals of association with lung function. For SNPs which were identified for a marginal association with FEV_1 or FEV_1/FVC, the corresponding p-values are highlighted in red. *Signal originally identified with the use of UK Biobank. MAF = Minor allele frequency.

SNP (Chr:Pos)	GENE	UK Biobank RSID	MAF	FEV_1 p-value	FEV_1/FVC p-value	Identified trait
rs2284746 (1:17306675)	*MFAP2*	rs2284746	0.49	1.99×10^{-01}	1.37×10^{-46}	FEV_1/FVC
*rs17513135 (1:40035686)	*LOC101929516*	rs17513135	0.22	9.35×10^{-11}	1.87×10^{-15}	FEV_1/FVC
*rs1192404 (1:92068967)	*TGFBR3*	rs1192404	0.16	1.04×10^{-01}	2.39×10^{-41}	FEV_1/FVC
*rs12140637 (1:92374517)	*TGFBR3*	rs12140637	0.31	7.55×10^{-02}	5.85×10^{-22}	FEV_1/FVC
*rs200154334 (1:118862070)	*SPAG17*	rs60804050	0.26	6.46×10^{-12}	5.81×10^{-01}	FVC
*rs6681426 (1:150586971)	*ENSA*	rs6681426	0.35	4.06×10^{-10}	3.44×10^{-03}	FEV_1
rs993925 (1:218860068)	*TGFB2*	rs993925	0.35	9.22×10^{-04}	3.93×10^{-07}	FEV_1/FVC
rs4328080 (1:219963088)	*RNU5F-1*	rs4328080	0.39	7.73×10^{-03}	3.30×10^{-16}	FEV_1/FVC
rs6688537 (1:239850588)	*CHRM3*	rs6688537	0.50	7.83×10^{-13}	8.58×10^{-33}	FEV_1/FVC
rs62126408 (2:18309132)	*KCNS3*	rs62126408	0.14	1.35×10^{-17}	7.01×10^{-33}	FEV_1/FVC
rs1430193 (2:56120853)	*EFEMP1*	rs1430193	0.36	2.57×10^{-08}	5.86×10^{-07}	FVC
rs2571445 (2:218683154)	*TNS1*	rs2571445	0.39	3.29×10^{-28}	1.18×10^{-17}	FEV_1
rs10498230 (2:229502503)	*PID1*	rs10498230	0.08	9.38×10^{-07}	1.34×10^{-67}	FEV_1/FVC
*rs61332075 (2:239316560)	*TRAF3IP1*	rs61332075	0.12	7.97×10^{-09}	3.03×10^{-08}	FEV_1/FVC
rs12477314 (2:239877148)	*FLJ43879*	rs12477314	0.19	4.48×10^{-20}	1.01×10^{-55}	FEV_1/FVC
rs1529672 (3:25479090)	*RARB*	rs1286664	0.17	3.73×10^{-05}	2.34×10^{-37}	FEV_1/FVC
*rs1458979 (3:55150677)	*CACNA2D3*	rs1458979	0.50	5.69×10^{-07}	4.55×10^{-19}	FEV_1/FVC
*rs1490265 (3:67452043)	*SUCLG2*	rs1490265	0.29	1.13×10^{-11}	2.53×10^{-01}	FVC
*rs2811415 (3:127991527)	*EEFSEC*	rs2811415	0.16	3.99×10^{-03}	1.77×10^{-19}	FEV_1/FVC
rs1595029 (3:158241767)	*RSRC1/RP11-538P18.2*	rs1595029	0.48	6.79×10^{-14}	6.99×10^{-02}	FVC
*esv2660202 (3:168738454)	*MECOM*	rs56341938	0.50	2.24×10^{-19}	4.72×10^{-21}	FEV_1/FVC
rs1344555 (3:169300219)	*MECOM*	rs1344555	0.21	4.78×10^{-12}	7.12×10^{-04}	FEV_1
*rs13110699 (4:89815695)	*FAM13A*	rs13110699	0.18	1.05×10^{-02}	1.74×10^{-65}	FEV_1/FVC
rs2045517 (4:89870964)	*FAM13A*	rs2045517	0.41	3.11×10^{-02}	5.86×10^{-65}	FEV_1/FVC
*rs2047409 (tag for rs34480284) (4:105215875)	*TET2*	rs2007403	0.37	1.71×10^{-18}	3.70×10^{-36}	FEV_1

SNP (Chr:Pos)	GENE	UK Biobank RSID	MAF	FEV_1 p-value	FEV_1/FVC p-value	Identified trait
rs10516526 (4:106688904)	*GSTCD*	rs10516526	0.06	5.89×10^{-44}	5.94×10^{-20}	FEV_1
*rs34712979 (4:106819053)	*NPNT*	rs34712979	0.25	2.10×10^{-83}	2.75×10^{-139}	FEV_1/FVC
rs138641402 (4:145445779)	*HHIP-AS1*	rs12504628	0.39	2.45×10^{-43}	1.10×10^{-174}	FEV_1/FVC
*rs91731 (5:33334312)	*TARS*	rs91731	0.10	1.18×10^{-09}	7.32×10^{-01}	FVC
*rs1551943 (5:52195033)	*ITGA1*	rs1551943	0.23	9.12×10^{-08}	6.29×10^{-39}	FEV_1/FVC
*rs2441026 (5:53444498)	*ARL15*	rs2441026	0.46	2.39×10^{-10}	8.82×10^{-02}	FVC
rs153916 (5:95036700)	*SPATA9*	rs153916	0.45	1.38×10^{-01}	7.07×10^{-27}	FEV_1/FVC
*rs7713065 (5:131788334)	*C5orf56*	rs7713065	0.26	7.60×10^{-04}	2.53×10^{-29}	FEV_1/FVC
rs7715901 (5:147856392)	*HTR4*	rs7715901	0.40	4.84×10^{-38}	2.71×10^{-93}	FEV_1
*rs3839234 (5:148596693)	*ABLIM3*	rs2014787	0.47	2.22×10^{-13}	1.20×10^{-02}	FEV_1
*rs10515750 (5:156810072)	*CYFIP2*	rs10515750	0.07	2.43×10^{-11}	5.29×10^{-32}	FEV_1/FVC
rs1990950 (5:156920756)	*ADAM19*	rs1990950	0.40	3.06×10^{-19}	2.81×10^{-51}	FEV_1/FVC
rs6924424 (6:7801611)	*BMP6*	rs6924424	0.16	2.87×10^{-16}	2.83×10^{-01}	FVC
*rs141651520 (6:73670095)	*KCNQ5*	rs16883089	0.20	1.97×10^{-01}	6.01×10^{-32}	FEV_1/FVC
rs2768551 (6:109270656)	*ARMC2*	rs2768551	0.19	6.40×10^{-23}	4.44×10^{-50}	FEV_1/FVC
rs7753012 (6:142745883)	*GPR126/LOC153910*	rs7753012	0.32	2.24×10^{-22}	2.10×10^{-176}	FEV_1/FVC
rs148274477 (6:142838173)	*GPR126/LOC153910*	rs148274477	0.03	1.47×10^{-02}	2.06×10^{-121}	FEV_1/FVC
*rs10246303 (7:7286445)	*C1GALT1*	rs10246303	0.43	4.08×10^{-08}	5.09×10^{-13}	FEV_1/FVC
*rs72615157 (7:99635967)	*ZKSCAN1*	rs72615157	0.17	1.08×10^{-01}	2.61×10^{-11}	FEV_1/FVC
*rs12698403 (7:156127246)	*LOC285889*	rs12698403	0.44	1.38×10^{-24}	3.23×10^{-21}	FEV_1
*rs7872188 (9:4124377)	*GLIS3*	rs7872188	0.40	1.74×10^{-16}	1.32×10^{-20}	FEV_1
rs16909859 (9:98204792)	*PTCH1*	rs16909859	0.09	3.63×10^{-01}	3.36×10^{-20}	FEV_1/FVC
rs803923 (9:119401650)	*ASTN2*	rs803923	0.46	3.73×10^{-14}	9.44×10^{-18}	FEV_1/FVC
rs10858246 (9:139102831)	*QSOX2/LHX3*	rs10858246	0.31	3.26×10^{-09}	7.76×10^{-02}	FVC
*rs10870202 (9:139257411)	*DNLZ*	rs10870202	0.50	2.53×10^{-05}	8.40×10^{-01}	FVC
rs7090277 (10:12278021)	*CDC123*	rs7090277	0.49	1.58×10^{-36}	9.85×10^{-60}	FEV_1/FVC
*rs3847402 (10:30267810)	*KIAA1462*	rs3847402	0.42	1.99×10^{-04}	8.04×10^{-08}	FEV_1/FVC
*rs7095607 (10:69957350)	*MYPN*	rs7095607	0.49	2.72×10^{-17}	3.16×10^{-01}	FVC
rs2637254 (10:78312002)	*C10orf11*	rs2637254	0.50	3.23×10^{-30}	3.06×10^{-17}	FEV_1
rs4237643 (11:43648368)	*MIR129-2*	rs4237643	0.31	2.68×10^{-09}	6.76×10^{-01}	FVC
rs2863171 (11:45250732)	*PRDM11*	rs2863171	0.16	3.07×10^{-05}	2.30×10^{-03}	FVC
*rs2509961 (11:62310909)	*AHNAK*	rs2509961	0.38	4.38×10^{-21}	2.31×10^{-05}	FEV_1
*rs11234757 (11:86443072)	*PRSS23*	rs7108254	0.16	4.79×10^{-11}	4.42×10^{-17}	FEV_1

SNP (Chr:Pos)	GENE	UK Biobank RSID	MAF	FEV_1 p-value	FEV_1/FVC p-value	Identified trait
*rs567508 (11:126008910)	*RPUSD4*	rs567508	0.17	6.66×10^{-07}	8.05×10^{-10}	FEV_1
rs2348418 (12:28689514)	*CCDC91*	rs2348418	0.45	4.12×10^{-31}	4.94×10^{-03}	FVC
rs11172113 (12:57527283)	*LRP1*	rs11172113	0.41	1.19×10^{-02}	2.04×10^{-17}	FEV_1/FVC
*rs1494502 (12:65824670)	*MSRB3*	rs1494502	0.37	6.00×10^{-11}	7.92×10^{-02}	FEV_1
*rs113745635 (12:95554771)	*FGD6*	rs113745635	0.21	2.14×10^{-03}	1.58×10^{-20}	FEV_1/FVC
rs12820313 (12:96255704)	*SNRPF/CCDC38*	rs12820313	0.21	6.60×10^{-03}	1.82×10^{-41}	FEV_1/FVC
*chr12_114743533 (12:114743533)	*TBX5*	rs569058293	0.0000761	6.44×10^{-01}	6.85×10^{-01}	FEV_1
rs10850377 (12:115201436)	*TBX3*	rs10850377	0.33	2.00×10^{-09}	8.53×10^{-04}	FEV_1
rs35506 (12:115500691)	*TBX3*	rs35506	0.28	1.49×10^{-12}	9.72×10^{-01}	FVC
*rs1698268 (14:84309664)	*LINC00911*	rs1698268	0.31	5.89×10^{-07}	7.35×10^{-22}	FEV_1/FVC
rs7155279 (14:92485881)	*TRIP11*	rs7155279	0.37	2.12×10^{-08}	6.81×10^{-02}	FEV_1
rs117068593 (14:93118229)	*RIN3*	rs117068593	0.18	7.57×10^{-24}	6.83×10^{-03}	FEV_1
*rs72724130 (15:41977690)	*MGA*	rs72724130	0.05	5.03×10^{-01}	6.42×10^{-11}	FEV_1/FVC
rs10851839 (15:71628370)	*THSD4*	rs10851839	0.34	1.97×10^{-17}	4.28×10^{-140}	FEV_1/FVC
*rs12591467 (15:71788387)	*THSD4*	rs12591467	0.33	1.62×10^{-06}	8.47×10^{-16}	FEV_1/FVC
*rs66650179 (15:84261689)	*SH3GL3*	rs12438269	0.21	8.60×10^{-09}	7.35×10^{-23}	FEV_1/FVC
rs12149828 (16:10706328)	*TEKT5*	rs12149593	0.17	3.05×10^{-01}	8.20×10^{-08}	FEV_1/FVC
rs12447804 (16:58075282)	*MMP15*	rs12447804	0.22	1.37×10^{-02}	8.95×10^{-36}	FEV_1/FVC
rs3743609 (16:75467021)	*CFDP1*	rs3743609	0.42	5.18×10^{-20}	6.53×10^{-52}	FEV_1/FVC
rs1079572 (16:78187138)	*WWOX*	rs1079572	0.42	1.06×10^{-03}	7.82×10^{-01}	FVC
*rs59835752 (17:28265330)	*EFCAB5*	rs62070270	0.45	2.16×10^{-02}	2.09×10^{-38}	FEV_1/FVC
*rs11658500 (17:36886828)	*CISD3*	rs11658500	0.13	5.57×10^{-10}	1.91×10^{-24}	FEV_1/FVC
*rs35524223 (17:44192590)	*KANSL1*	rs17577877	0.22	2.29×10^{-48}	3.46×10^{-04}	FEV_1
rs6501431 (17:68976415)	*CASC17*	rs6501431	0.21	3.22×10^{-03}	1.23×10^{-03}	FVC
*rs7218675 (17:73513185)	*TSEN54*	rs7218675	0.27	5.83×10^{-10}	8.59×10^{-06}	FEV_1
rs113473882 (19:41124155)	*LTBP4*	rs113473882	0.01	1.63×10^{-02}	3.09×10^{-37}	FEV_1/FVC
*rs6140050 (20:6632901)	*BMP2*	rs6140050	0.37	1.55×10^{-18}	4.74×10^{-04}	FVC
*rs72448466 (20:62363640)	*ZGPAT*	rs6062304	0.32	1.94×10^{-23}	2.40×10^{-02}	FEV_1
rs2834440 (21:35690499)	*KCNE2*	rs2834440	0.39	7.64×10^{-01}	8.05×10^{-20}	FEV_1/FVC
*rs11704827 (22:18450287)	*MICAL3*	rs11704827	0.23	1.15×10^{-20}	1.97×10^{-13}	FEV_1
rs134041 (22:28056338)	*MN1*	rs134041	0.44	5.31×10^{-03}	4.75×10^{-03}	FEV_1
*rs2283847 (22:28181399)	*MN1*	rs2283847	0.45	1.66×10^{-14}	9.06×10^{-18}	FEV_1

D. Joint test p-values and interaction test p-values, betas (β), and standard errors (SE) for the 632 SNPs which were screened using the joint test and tested for interaction effect. CAF = coded allele frequency, MAF = minor allele frequency, and INFO = imputation score. Results are on the inverse-normalised scale.

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1	rs3768321(1:40035928)	T/G	0.190	0.190	1.00	1.59×10^{-11}	2.05×10^{-1}	-0.008177	0.006455
FEV_1	rs1188935(1:91305057)	C/A	0.345	0.345	0.99	3.46×10^{-8}	6.79×10^{-2}	-0.009934	0.005442
FEV_1	rs2359463(1:118363223)	A/G	0.084	0.084	0.97	5.19×10^{-8}	9.16×10^{-1}	0.000964	0.009152
FEV_1	rs12735613(1:118883973)	A/G	0.244	0.244	1.00	1.33×10^{-12}	5.57×10^{-1}	0.003498	0.005960
FEV_1	rs9428295(1:119674087)	G/T	0.458	0.458	0.96	3.86×10^{-9}	5.87×10^{-1}	-0.002848	0.005248
FEV_1	rs11806879(1:150103261)	T/C	0.091	0.091	1.00	2.18×10^{-9}	4.56×10^{-1}	0.006524	0.008751
FEV_1	rs12133238(1:150165860)	G/A	0.303	0.303	0.99	3.27×10^{-11}	6.18×10^{-1}	0.002828	0.005677
FEV_1	rs66773125(1:150174366)	A/G	0.134	0.134	0.99	2.75×10^{-12}	4.24×10^{-2}	-0.015281	0.007529
FEV_1	rs878471(1:150547747)	A/G	0.566	0.434	1.00	1.33×10^{-16}	9.51×10^{-1}	0.000321	0.005199
FEV_1	rs28736888(1:218857609)	T/C	0.480	0.480	1.00	3.21×10^{-11}	1.80×10^{-1}	-0.006876	0.005134
FEV_1	rs2738755(1:221057646)	T/C	0.331	0.331	1.00	3.05×10^{-9}	9.23×10^{-1}	-0.000526	0.005459
FEV_1	rs17009288(1:221204299)	C/A	0.291	0.291	0.99	3.68×10^{-12}	9.51×10^{-1}	0.000350	0.005660
FEV_1	rs7354867(1:221268907)	T/C	0.324	0.324	0.99	8.78×10^{-9}	5.26×10^{-1}	-0.003487	0.005495
FEV_1	rs6680689(1:221462038)	A/G	0.429	0.429	1.00	2.00×10^{-9}	3.97×10^{-2}	0.010665	0.005185
FEV_1	rs1155612(1:239897705)	C/T	0.511	0.489	1.00	8.87×10^{-15}	2.97×10^{-1}	0.005352	0.005133
FEV_1	rs55884799(2:18287623)	C/T	0.178	0.178	0.99	9.36×10^{-23}	2.79×10^{-1}	0.007334	0.006779
FEV_1	rs76771253(2:18291613)	T/C	0.072	0.072	1.00	1.41×10^{-9}	8.31×10^{-1}	-0.002088	0.009776
FEV_1	rs11900434(2:18293827)	A/G	0.437	0.437	0.99	3.64×10^{-10}	1.55×10^{-1}	0.007432	0.005226
FEV_1	rs4335948(2:18577104)	T/C	0.192	0.192	0.99	1.91×10^{-16}	5.52×10^{-1}	-0.003967	0.006666
FEV_1	rs12613344(2:18587516)	C/G	0.307	0.307	0.99	8.94×10^{-9}	6.01×10^{-1}	-0.002909	0.005568
FEV_1	rs55937737(2:18688581)	G/A	0.229	0.229	1.00	6.88×10^{-10}	2.34×10^{-1}	0.007288	0.006120
FEV_1	rs11890443(2:18725739)	A/G	0.660	0.340	1.00	6.63×10^{-11}	4.97×10^{-1}	-0.003671	0.005403
FEV_1	rs2346133(2:18728276)	T/C	0.083	0.083	1.00	4.96×10^{-17}	2.55×10^{-1}	0.010817	0.009496
FEV_1	rs34811804(2:18886374)	T/C	0.306	0.306	1.00	1.81×10^{-8}	1.59×10^{-3}	0.017439	0.005521
FEV_1	rs1430190(2:56031305)	G/C	0.270	0.270	0.99	5.56×10^{-14}	1.93×10^{-2}	-0.013769	0.005887

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1	rs11125611(2:56173907)	T/A	0.368	0.368	0.99	4.62×10^{-12}	4.35×10^{-2}	-0.010748	0.005323
FEV_1	rs2571445(2:218683154)	G/A	0.609	0.391	1.00	3.29×10^{-28}	9.99×10^{-2}	0.008619	0.005239
FEV_1	rs2552529(2:218695798)	G/A	0.652	0.348	0.98	2.01×10^{-10}	2.35×10^{-1}	0.006417	0.005403
FEV_1	rs62202385(2:229591965)	T/C	0.082	0.082	0.97	4.00×10^{-8}	2.19×10^{-1}	-0.011741	0.009554
FEV_1	rs6710301(2:239441308)	A/C	0.149	0.149	0.99	9.32×10^{-13}	6.68×10^{-1}	0.003101	0.007221
FEV_1	rs13410025(2:239614258)	A/C	0.238	0.238	0.99	4.33×10^{-13}	3.13×10^{-1}	-0.006102	0.006042
FEV_1	rs62191107(2:239876527)	C/T	0.195	0.195	1.00	3.11×10^{-20}	2.42×10^{-1}	0.007533	0.006440
FEV_1	rs9819463(3:53672471)	C/T	0.205	0.205	0.99	5.43×10^{-13}	9.75×10^{-1}	0.000198	0.006360
FEV_1	rs358493(3:55072187)	T/C	0.336	0.336	0.99	3.68×10^{-8}	5.16×10^{-1}	0.003549	0.005466
FEV_1	rs6795984(3:55130867)	A/G	0.270	0.270	1.00	5.12×10^{-11}	8.40×10^{-1}	0.001153	0.005726
FEV_1	rs3923256(3:67454630)	G/A	0.708	0.292	1.00	7.89×10^{-12}	9.07×10^{-1}	0.000656	0.005628
FEV_1	rs73154313(3:158225485)	C/T	0.264	0.264	1.00	4.43×10^{-16}	4.76×10^{-1}	0.004116	0.005773
FEV_1	rs11925651(3:168645335)	A/T	0.272	0.272	1.00	7.10×10^{-9}	8.50×10^{-1}	-0.001090	0.005747
FEV_1	rs114902833(3:168689756)	A/C	0.077	0.077	1.00	4.51×10^{-8}	3.99×10^{-1}	-0.008062	0.009557
FEV_1	rs9883125(3:168748911)	A/T	0.372	0.372	1.00	1.25×10^{-21}	2.54×10^{-2}	-0.011852	0.005302
FEV_1	rs6444837(3:168781277)	C/T	0.725	0.275	0.99	2.92×10^{-17}	4.18×10^{-1}	0.004696	0.005792
FEV_1	rs998749(3:168972802)	G/A	0.473	0.473	0.99	8.80×10^{-14}	8.68×10^{-1}	0.000853	0.005144
FEV_1	rs78101726(3:169295436)	G/A	0.155	0.155	1.00	2.65×10^{-18}	6.56×10^{-1}	-0.003171	0.007126
FEV_1	rs4955665(3:169355019)	C/T	0.396	0.396	0.95	1.14×10^{-8}	5.35×10^{-1}	-0.003321	0.005349
FEV_1	rs7661349(4:106066982)	C/T	0.642	0.358	1.00	1.41×10^{-19}	2.05×10^{-1}	0.006726	0.005308
FEV_1	rs75321784(4:106125022)	G/A	0.122	0.122	1.00	1.76×10^{-12}	4.88×10^{-1}	0.005437	0.007840
FEV_1	rs59428412(4:106524455)	C/T	0.221	0.221	1.00	2.04×10^{-16}	8.98×10^{-2}	-0.010525	0.006204
FEV_1	rs12512339(4:106592617)	C/G	0.786	0.214	0.99	3.35×10^{-14}	5.15×10^{-1}	0.004158	0.006384
FEV_1	rs80126340(4:106690466)	G/C	0.044	0.044	1.00	1.36×10^{-8}	5.92×10^{-1}	-0.006573	0.012275
FEV_1	rs143396422(4:106697395)	A/G	0.007	0.007	0.95	5.43×10^{-9}	4.34×10^{-1}	-0.024514	0.031362
FEV_1	rs112819759(4:106711416)	A/G	0.043	0.043	1.00	4.40×10^{-29}	6.57×10^{-1}	-0.005471	0.012330
FEV_1	rs79263635(4:106791573)	C/A	0.094	0.094	1.00	1.46×10^{-8}	6.89×10^{-1}	0.003437	0.008578
FEV_1	rs182400279(4:106812298)	C/G	0.006	0.006	0.82	5.38×10^{-8}	5.94×10^{-1}	-0.019444	0.036471
FEV_1	rs67277942(4:106813309)	G/A	0.072	0.072	0.99	1.35×10^{-45}	8.22×10^{-1}	-0.002296	0.010233
FEV_1	rs34712979(4:106819053)	A/G	0.245	0.245	1.00	2.10×10^{-83}	5.02×10^{-1}	-0.003934	0.005860
FEV_1	rs6852099(4:106819889)	C/T	0.337	0.337	0.97	1.33×10^{-23}	9.28×10^{-2}	0.009259	0.005510
FEV_1	rs4513575(4:106862181)	G/T	0.920	0.080	1.00	1.02×10^{-8}	3.29×10^{-1}	-0.009224	0.009455
FEV_1	rs141966336(4:106893391)	G/A	0.026	0.026	0.92	8.53×10^{-10}	2.03×10^{-1}	-0.021014	0.016496
FEV_1	rs11945032(4:106897596)	A/G	0.357	0.357	0.99	2.87×10^{-11}	2.29×10^{-2}	0.012302	0.005406
FEV_1	rs62320063(4:106901301)	C/T	0.062	0.062	0.96	4.20×10^{-12}	6.90×10^{-2}	0.019594	0.010774

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1	rs72660503(4:107162069)	A/G	0.031	0.031	0.92	3.68×10^{-10}	2.75×10^{-2}	0.033620	0.015252
FEV_1	rs3018056(4:107329974)	T/C	0.251	0.251	0.98	3.40×10^{-8}	9.89×10^{-1}	-0.000079	0.005936
FEV_1	rs4634292(4:107473889)	T/A	0.951	0.049	0.95	3.59×10^{-13}	4.52×10^{-1}	0.008920	0.011864
FEV_1	rs151025055(4:144674986)	T/C	0.122	0.122	0.98	8.62×10^{-12}	9.19×10^{-2}	0.012998	0.007713
FEV_1	rs4434191(4:145024452)	T/G	0.271	0.271	0.94	2.70×10^{-10}	1.95×10^{-2}	0.013797	0.005908
FEV_1	rs62334650(4:145038318)	T/C	0.303	0.303	0.98	9.16×10^{-9}	5.18×10^{-1}	0.003601	0.005574
FEV_1	rs72729558(4:145133493)	C/T	0.310	0.310	0.98	4.40×10^{-13}	4.50×10^{-1}	0.004201	0.005560
FEV_1	rs79709819(4:145185212)	A/G	0.043	0.043	1.00	3.39×10^{-8}	5.48×10^{-1}	0.007424	0.012365
FEV_1	rs11100844(4:145212394)	A/G	0.365	0.365	0.86	4.76×10^{-11}	1.87×10^{-1}	-0.007502	0.005688
FEV_1	rs2719335(4:145289459)	C/T	0.200	0.200	0.99	1.28×10^{-17}	9.59×10^{-1}	-0.000333	0.006395
FEV_1	rs75674934(4:145383334)	A/G	0.080	0.080	0.95	3.67×10^{-17}	3.58×10^{-1}	-0.008757	0.009534
FEV_1	rs13116999(4:145442364)	A/G	0.549	0.451	1.00	1.60×10^{-55}	3.77×10^{-1}	0.004540	0.005141
FEV_1	rs62345400(4:145455765)	A/C	0.051	0.051	1.00	1.22×10^{-11}	3.84×10^{-1}	-0.009899	0.011368
FEV_1	rs12648786(4:145521703)	A/G	0.393	0.393	0.95	1.68×10^{-9}	3.85×10^{-1}	0.004673	0.005384
FEV_1	rs2639576(4:145658429)	C/T	0.464	0.464	0.98	2.42×10^{-19}	5.56×10^{-1}	-0.003047	0.005179
FEV_1	rs11727676(4:145659064)	C/T	0.093	0.093	1.00	1.77×10^{-26}	7.73×10^{-1}	0.002498	0.008676
FEV_1	rs11933087(4:145722862)	T/A	0.372	0.372	0.99	1.74×10^{-10}	5.77×10^{-1}	0.002976	0.005335
FEV_1	rs268717(5:33352738)	T/C	0.903	0.097	0.99	1.01×10^{-9}	7.75×10^{-1}	-0.002528	0.008859
FEV_1	rs35597318(5:52274101)	A/C	0.205	0.205	0.99	2.26×10^{-13}	5.68×10^{-1}	0.003602	0.006312
FEV_1	rs3212627(5:52380923)	T/A	0.189	0.189	0.99	1.59×10^{-8}	2.06×10^{-1}	0.008287	0.006552
FEV_1	rs1644814(5:53365452)	T/C	0.629	0.371	0.98	2.17×10^{-10}	2.09×10^{-1}	-0.006701	0.005335
FEV_1	rs35006(5:53442725)	G/C	0.304	0.304	0.99	3.98×10^{-8}	7.68×10^{-1}	0.001644	0.005567
FEV_1	rs1345815(5:53490737)	C/T	0.619	0.381	0.99	6.43×10^{-9}	4.01×10^{-1}	0.004447	0.005292
FEV_1	rs35466090(5:131622836)	T/C	0.091	0.091	0.99	7.91×10^{-9}	4.33×10^{-4}	0.031656	0.008995
FEV_1	rs57079115(5:132382272)	T/C	0.257	0.257	1.00	7.34×10^{-9}	2.59×10^{-1}	-0.006621	0.005870
FEV_1	rs10476892(5:147720723)	T/C	0.259	0.259	1.00	9.62×10^{-15}	2.67×10^{-1}	0.006520	0.005870
FEV_1	rs13190336(5:147823134)	T/C	0.039	0.039	1.00	1.44×10^{-9}	8.98×10^{-1}	0.001660	0.012930
FEV_1	rs7733410(5:147856522)	A/G	0.442	0.442	1.00	1.33×10^{-43}	5.69×10^{-3}	0.014277	0.005163
FEV_1	rs35684381(5:148203236)	C/T	0.152	0.152	0.98	1.06×10^{-10}	7.23×10^{-1}	0.002601	0.007339
FEV_1	rs1800888(5:148206885)	T/C	0.014	0.014	1.00	9.64×10^{-16}	1.30×10^{-1}	-0.031984	0.021142
FEV_1	rs11952673(5:148652302)	T/G	0.395	0.395	0.98	5.64×10^{-16}	6.43×10^{-2}	-0.009754	0.005273
FEV_1	rs4704863(5:156879727)	T/C	0.514	0.486	1.00	1.36×10^{-9}	8.39×10^{-1}	0.001045	0.005133
FEV_1	rs77273565(5:156916555)	G/A	0.061	0.061	1.00	7.45×10^{-12}	1.91×10^{-1}	-0.013723	0.010501
FEV_1	rs11741723(5:156916987)	A/C	0.129	0.129	1.00	5.35×10^{-15}	4.61×10^{-1}	0.005549	0.007530
FEV_1	rs13361953(5:156926442)	C/T	0.344	0.344	1.00	2.70×10^{-19}	9.96×10^{-1}	-0.000024	0.005431

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1	rs11948668(5:157109069)	T/C	0.251	0.251	0.98	8.43×10^{-12}	8.84×10^{-1}	0.000867	0.005964
FEV_1	rs62385362(5:158279381)	C/T	0.222	0.222	0.99	5.85×10^{-10}	8.49×10^{-2}	0.010559	0.006129
FEV_1	rs72643433(5:158364449)	A/G	0.254	0.254	0.99	2.56×10^{-13}	7.63×10^{-1}	0.001791	0.005937
FEV_1	rs9392172(6:7723962)	G/C	0.465	0.465	1.00	9.50×10^{-17}	9.04×10^{-1}	-0.000621	0.005141
FEV_1	rs270411(6:7740660)	A/G	0.456	0.456	0.98	5.00×10^{-8}	4.50×10^{-1}	0.003924	0.005193
FEV_1	rs10498672(6:7797840)	G/C	0.170	0.170	1.00	5.98×10^{-17}	4.36×10^{-1}	0.005238	0.006721
FEV_1	rs948518(6:109167087)	C/G	0.594	0.406	0.99	3.03×10^{-8}	5.13×10^{-1}	-0.003450	0.005268
FEV_1	rs2806356(6:109266255)	C/T	0.183	0.183	1.00	3.24×10^{-24}	2.06×10^{-1}	-0.008337	0.006590
FEV_1	rs9496274(6:142424051)	A/T	0.081	0.081	1.00	6.70×10^{-9}	4.22×10^{-1}	0.007584	0.009455
FEV_1	rs9376680(6:142553740)	T/C	0.261	0.261	1.00	6.32×10^{-23}	2.64×10^{-2}	0.013103	0.005901
FEV_1	rs7753012(6:142745883)	G/T	0.322	0.322	1.00	2.24×10^{-22}	7.77×10^{-1}	0.001580	0.005586
FEV_1	rs4724960(7:7241299)	G/A	0.340	0.340	1.00	1.39×10^{-10}	3.80×10^{-1}	-0.004784	0.005451
FEV_1	rs6972975(7:156119581)	C/A	0.443	0.443	1.00	8.28×10^{-9}	8.79×10^{-1}	-0.000789	0.005165
FEV_1	rs12698403(7:156127246)	A/G	0.437	0.437	1.00	1.38×10^{-24}	4.24×10^{-1}	0.004146	0.005185
FEV_1	rs4741893(9:4123284)	C/G	0.414	0.414	0.98	6.50×10^{-18}	8.35×10^{-1}	0.001090	0.005245
FEV_1	rs62524071(9:4153283)	A/G	0.086	0.086	0.99	2.04×10^{-10}	2.23×10^{-1}	-0.011013	0.009039
FEV_1	rs12115436(9:98879734)	A/G	0.167	0.167	1.00	2.40×10^{-10}	1.62×10^{-1}	0.009638	0.006893
FEV_1	rs4837565(9:119258820)	A/G	0.140	0.140	0.99	1.04×10^{-9}	8.70×10^{-1}	-0.001217	0.007411
FEV_1	rs803912(9:119414823)	G/A	0.541	0.459	1.00	6.07×10^{-15}	8.89×10^{-1}	0.000718	0.005136
FEV_1	rs3860179(9:119603526)	C/T	0.798	0.202	0.99	3.07×10^{-9}	1.84×10^{-1}	0.008449	0.006367
FEV_1	rs10858246(9:139102831)	C/G	0.310	0.310	1.00	3.26×10^{-9}	6.34×10^{-1}	-0.002618	0.005507
FEV_1	rs2271804(10:12252217)	A/G	0.522	0.478	1.00	1.57×10^{-37}	5.58×10^{-1}	-0.003016	0.005145
FEV_1	rs12779790(10:12328010)	G/A	0.185	0.185	0.97	3.63×10^{-14}	8.58×10^{-1}	-0.001195	0.006663
FEV_1	rs7079481(10:69959242)	A/C	0.487	0.487	1.00	6.70×10^{-18}	2.56×10^{-1}	0.005822	0.005129
FEV_1	rs11594905(10:77659733)	A/G	0.132	0.132	1.00	3.03×10^{-10}	1.66×10^{-3}	-0.023473	0.007465
FEV_1	rs2637254(10:78312002)	A/G	0.505	0.495	0.99	3.23×10^{-30}	2.33×10^{-1}	-0.006127	0.005136
FEV_1	rs11001894(10:78593502)	T/C	0.465	0.465	1.00	1.53×10^{-9}	3.09×10^{-1}	0.005229	0.005135
FEV_1	rs17596617(11:43690717)	T/C	0.318	0.318	0.99	1.86×10^{-9}	9.94×10^{-2}	-0.009102	0.005524
FEV_1	rs7928792(11:61698488)	G/C	0.337	0.337	1.00	7.38×10^{-9}	7.74×10^{-1}	0.001570	0.005464
FEV_1	rs1801144(11:62381808)	C/G	0.356	0.356	0.99	1.33×10^{-25}	4.19×10^{-1}	0.004315	0.005343
FEV_1	rs117510149(11:62394297)	C/G	0.052	0.052	0.99	4.09×10^{-8}	7.84×10^{-1}	0.003104	0.011310
FEV_1	rs2512561(11:62453952)	A/G	0.751	0.249	1.00	3.91×10^{-8}	7.96×10^{-2}	-0.010402	0.005933
FEV_1	rs113554344(11:62610758)	A/C	0.064	0.064	0.95	2.26×10^{-10}	2.07×10^{-1}	0.013296	0.010543
FEV_1	rs117261012(11:86444761)	G/A	0.152	0.152	0.98	1.65×10^{-11}	2.15×10^{-1}	0.008824	0.007120
FEV_1	rs79864933(11:125904580)	C/T	0.067	0.067	0.98	3.30×10^{-8}	2.00×10^{-1}	-0.013237	0.010327

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1	rs560813(11:126045822)	C/T	0.228	0.228	1.00	2.10×10^{-10}	7.23×10^{-1}	-0.002187	0.006174
FEV_1	rs57380671(12:28010407)	C/T	0.183	0.183	0.99	3.22×10^{-9}	2.88×10^{-1}	-0.006975	0.006562
FEV_1	rs1314085(12:28140253)	G/T	0.880	0.120	1.00	3.33×10^{-9}	1.89×10^{-1}	0.010349	0.007877
FEV_1	rs476703(12:28247250)	A/G	0.130	0.130	0.97	7.06×10^{-9}	1.21×10^{-1}	0.011839	0.007641
FEV_1	rs12146812(12:28261546)	A/T	0.336	0.336	0.99	1.09×10^{-14}	4.86×10^{-1}	-0.003771	0.005408
FEV_1	rs10843109(12:28288716)	T/C	0.099	0.099	1.00	4.09×10^{-11}	9.62×10^{-2}	-0.014029	0.008433
FEV_1	rs7977418(12:28588242)	C/T	0.454	0.454	1.00	1.09×10^{-32}	4.86×10^{-1}	0.003588	0.005151
FEV_1	rs12368897(12:28782170)	G/A	0.143	0.143	0.99	1.37×10^{-8}	4.61×10^{-1}	0.005438	0.007383
FEV_1	rs1689510(12:56396768)	C/G	0.331	0.331	1.00	3.46×10^{-9}	2.41×10^{-1}	0.006349	0.005418
FEV_1	rs61921171(12:58662712)	T/C	0.212	0.212	1.00	3.59×10^{-8}	7.74×10^{-4}	0.021058	0.006264
FEV_1	rs12825748(12:65793153)	C/G	0.307	0.307	0.99	5.23×10^{-11}	9.19×10^{-1}	-0.000564	0.005571
FEV_1	rs79487293(12:65905126)	T/C	0.311	0.311	1.00	4.56×10^{-11}	9.61×10^{-2}	-0.009121	0.005482
FEV_1	rs4581496(12:66022450)	A/T	0.790	0.210	1.00	7.59×10^{-10}	2.39×10^{-1}	0.007616	0.006464
FEV_1	rs76900422(12:66166141)	A/G	0.050	0.050	0.97	8.50×10^{-10}	1.20×10^{-1}	-0.018248	0.011734
FEV_1	rs74097857(12:66393756)	C/T	0.128	0.128	0.99	4.54×10^{-13}	2.07×10^{-1}	-0.009584	0.007591
FEV_1	rs2555009(12:114666099)	G/A	0.544	0.456	1.00	8.40×10^{-12}	7.58×10^{-1}	-0.001580	0.005136
FEV_1	rs2701110(12:114669870)	A/C	0.169	0.169	0.99	3.12×10^{-13}	6.73×10^{-1}	0.002892	0.006857
FEV_1	rs10850377(12:115201436)	A/G	0.335	0.335	0.99	2.00×10^{-9}	5.36×10^{-2}	0.010470	0.005423
FEV_1	rs35505(12:115501127)	A/G	0.688	0.312	0.99	6.33×10^{-13}	9.03×10^{-2}	-0.009376	0.005535
FEV_1	rs1874903(12:115950227)	T/C	0.471	0.471	0.97	4.23×10^{-8}	1.75×10^{-1}	0.007046	0.005190
FEV_1	rs1511318(14:84298899)	C/T	0.158	0.158	1.00	1.89×10^{-8}	3.40×10^{-2}	-0.015217	0.007178
FEV_1	rs11160037(14:92512143)	G/A	0.388	0.388	1.00	9.25×10^{-9}	7.78×10^{-1}	-0.001494	0.005296
FEV_1	rs11621587(14:93098339)	C/G	0.179	0.179	1.00	2.38×10^{-24}	9.91×10^{-1}	0.000076	0.006608
FEV_1	rs942064(14:93117429)	C/T	0.681	0.319	1.00	7.77×10^{-12}	5.71×10^{-1}	0.003104	0.005478
FEV_1	rs1956028(14:93507197)	C/T	0.138	0.138	0.99	6.24×10^{-10}	9.05×10^{-1}	0.000924	0.007761
FEV_1	rs1898882(15:40655873)	C/G	0.452	0.452	1.00	4.47×10^{-8}	9.26×10^{-1}	-0.000478	0.005161
FEV_1	rs12910520(15:41481420)	T/A	0.619	0.381	0.99	2.40×10^{-9}	3.89×10^{-2}	-0.011011	0.005332
FEV_1	rs1441356(15:71631173)	A/G	0.396	0.396	0.99	5.88×10^{-13}	9.23×10^{-1}	0.000509	0.005260
FEV_1	rs2119568(15:71665824)	C/T	0.179	0.179	1.00	5.08×10^{-23}	5.45×10^{-1}	-0.004036	0.006674
FEV_1	rs8041231(15:83498795)	G/T	0.522	0.478	1.00	1.54×10^{-8}	9.57×10^{-1}	0.000279	0.005133
FEV_1	rs12903359(15:83502787)	G/A	0.546	0.454	0.99	9.24×10^{-15}	7.03×10^{-1}	-0.001970	0.005161
FEV_1	rs872598(15:83584579)	C/A	0.468	0.468	1.00	4.94×10^{-9}	2.84×10^{-1}	-0.005512	0.005141
FEV_1	rs7162082(15:83896608)	T/C	0.196	0.196	1.00	1.02×10^{-8}	5.57×10^{-1}	-0.003725	0.006347
FEV_1	rs12441893(15:84001389)	G/A	0.244	0.244	0.99	1.06×10^{-12}	4.09×10^{-1}	0.004893	0.005926
FEV_1	rs17841201(15:84065381)	G/A	0.236	0.236	1.00	2.86×10^{-10}	1.71×10^{-1}	-0.008207	0.006001

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1	rs301847(15:84092825)	G/C	0.383	0.383	0.99	3.76×10^{-12}	6.94×10^{-1}	-0.002091	0.005310
FEV_1	rs1491579(15:84178679)	C/T	0.552	0.448	0.99	6.62×10^{-9}	8.32×10^{-1}	0.001102	0.005182
FEV_1	rs1896797(15:84274591)	A/G	0.499	0.499	0.99	2.34×10^{-13}	7.39×10^{-1}	0.001711	0.005136
FEV_1	rs77203805(15:84286544)	A/G	0.110	0.110	1.00	4.17×10^{-15}	4.89×10^{-1}	-0.005624	0.008128
FEV_1	rs979543(15:84352564)	A/C	0.791	0.209	0.99	3.71×10^{-9}	5.63×10^{-1}	0.003620	0.006266
FEV_1	rs4843154(15:84419524)	T/G	0.450	0.450	1.00	3.62×10^{-9}	6.79×10^{-1}	0.002130	0.005154
FEV_1	rs118134155(15:84508221)	C/A	0.021	0.021	0.96	2.39×10^{-8}	6.06×10^{-1}	-0.009220	0.017880
FEV_1	rs4414460(15:84562740)	G/A	0.336	0.336	1.00	4.78×10^{-8}	4.72×10^{-1}	0.003956	0.005494
FEV_1	rs12932007(16:75428556)	T/C	0.131	0.131	0.78	1.28×10^{-11}	1.35×10^{-2}	0.021251	0.008604
FEV_1	rs11642572(16:75452782)	T/C	0.590	0.410	1.00	1.18×10^{-20}	3.08×10^{-1}	-0.005334	0.005229
FEV_1	rs247454(16:75519390)	C/G	0.393	0.393	0.89	2.26×10^{-9}	4.40×10^{-1}	0.004309	0.005574
FEV_1	rs2345443(16:78225633)	G/A	0.694	0.306	1.00	9.39×10^{-14}	5.06×10^{-1}	-0.003694	0.005549
FEV_1	rs11859414(16:79189453)	G/C	0.078	0.078	0.98	2.89×10^{-10}	6.08×10^{-2}	0.018731	0.009991
FEV_1	rs62070637(17:29118181)	T/C	0.168	0.168	1.00	2.83×10^{-12}	6.65×10^{-1}	0.002926	0.006750
FEV_1	rs11657029(17:36864953)	A/G	0.191	0.191	0.99	6.92×10^{-14}	3.66×10^{-1}	0.005920	0.006547
FEV_1	rs8069451(17:37504933)	C/T	0.257	0.257	0.99	1.36×10^{-9}	8.65×10^{-1}	-0.001009	0.005935
FEV_1	rs9303280(17:38074031)	C/T	0.492	0.492	1.00	4.62×10^{-8}	4.93×10^{-1}	-0.003524	0.005136
FEV_1	rs11871217(17:43057585)	T/C	0.339	0.339	0.98	2.83×10^{-8}	9.21×10^{-1}	0.000542	0.005450
FEV_1	rs10853050(17:43445792)	G/A	0.671	0.329	0.98	1.49×10^{-9}	4.38×10^{-1}	-0.004266	0.005501
FEV_1	rs9905348(17:43571815)	A/G	0.574	0.426	1.00	7.54×10^{-15}	8.46×10^{-1}	-0.001012	0.005209
FEV_1	rs35116560(17:43804186)	C/T	0.417	0.417	0.98	1.90×10^{-16}	9.78×10^{-1}	0.000143	0.005240
FEV_1	rs1158660(17:43945288)	A/G	0.311	0.311	0.96	8.91×10^{-13}	9.21×10^{-1}	-0.000560	0.005633
FEV_1	rs117499775(17:44078618)	C/T	0.038	0.038	1.00	3.54×10^{-8}	3.06×10^{-1}	-0.013328	0.013020
FEV_1	rs112333322(17:44126673)	G/A	0.215	0.215	0.99	6.05×10^{-49}	9.42×10^{-1}	0.000451	0.006187
FEV_1	rs1863115(17:44625928)	A/C	0.732	0.268	0.90	1.94×10^{-16}	7.38×10^{-1}	-0.002062	0.006170
FEV_1	rs34724124(17:44902516)	A/G	0.473	0.473	0.99	5.82×10^{-10}	1.94×10^{-2}	-0.012060	0.005161
FEV_1	rs2190693(17:69103758)	A/G	0.588	0.412	0.97	1.38×10^{-9}	6.14×10^{-1}	0.002656	0.005265
FEV_1	rs7214488(17:69190974)	A/T	0.814	0.186	0.99	1.14×10^{-12}	7.71×10^{-1}	0.001908	0.006567
FEV_1	rs17178377(17:69198133)	A/G	0.472	0.472	1.00	3.99×10^{-32}	6.96×10^{-2}	0.009266	0.005106
FEV_1	rs9892893(17:73525670)	G/T	0.739	0.261	0.99	1.27×10^{-13}	2.75×10^{-1}	0.006414	0.005882
FEV_1	rs10410606(19:41108975)	C/A	0.446	0.446	1.00	1.83×10^{-8}	4.40×10^{-3}	0.014733	0.005173
FEV_1	rs2561562(19:41150086)	A/C	0.509	0.491	1.00	4.16×10^{-9}	3.98×10^{-5}	0.021090	0.005133
FEV_1	rs6085574(20:6466574)	A/T	0.271	0.271	0.98	4.55×10^{-9}	5.48×10^{-1}	0.003466	0.005768
FEV_1	rs979011(20:6623833)	A/C	0.635	0.365	1.00	4.68×10^{-19}	1.55×10^{-1}	0.007582	0.005333
FEV_1	rs200383755(20:61050522)	C/G	0.006	0.006	0.88	5.33×10^{-14}	4.85×10^{-1}	0.024629	0.035258

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1	rs4809548(20:62001148)	G/A	0.066	0.066	0.99	1.52×10^{-10}	1.16×10^{-3}	-0.033579	0.010338
FEV_1	rs71325435(20:62011379)	T/C	0.271	0.271	0.98	4.20×10^{-9}	1.59×10^{-3}	0.018362	0.005814
FEV_1	rs3865527(20:62047925)	C/T	0.567	0.433	0.98	3.64×10^{-8}	7.52×10^{-5}	0.020634	0.005212
FEV_1	rs77993536(20:62296349)	T/C	0.082	0.082	0.97	2.06×10^{-10}	9.54×10^{-1}	-0.000536	0.009386
FEV_1	rs6011026(20:62308486)	G/A	0.483	0.483	0.96	2.55×10^{-8}	1.30×10^{-1}	-0.007933	0.005240
FEV_1	rs6062506(20:62352389)	G/T	0.675	0.325	1.00	1.87×10^{-23}	2.29×10^{-1}	-0.006573	0.005462
FEV_1	rs56057703(20:62363512)	T/C	0.161	0.161	0.98	2.91×10^{-8}	9.16×10^{-2}	-0.011884	0.007044
FEV_1	rs6011067(20:62364376)	C/G	0.918	0.082	1.00	2.65×10^{-11}	3.69×10^{-1}	-0.008399	0.009353
FEV_1	rs4819639(22:18347127)	T/C	0.251	0.251	0.99	1.03×10^{-16}	4.31×10^{-1}	-0.004646	0.005901
FEV_1	rs5992136(22:18453103)	C/G	0.265	0.265	0.99	4.96×10^{-21}	2.44×10^{-1}	-0.006791	0.005829
FEV_1	rs2283847(22:28181399)	T/C	0.548	0.452	0.95	1.66×10^{-14}	5.87×10^{-1}	-0.002874	0.005292
FEV_1/FVC	rs7413433(1:17213369)	T/C	0.221	0.221	0.82	6.60×10^{-14}	9.34×10^{-1}	-0.000566	0.006783
FEV_1/FVC	rs9435731(1:17306029)	A/C	0.509	0.491	1.00	2.84×10^{-47}	9.92×10^{-1}	-0.000049	0.005118
FEV_1/FVC	rs9787230(1:17370970)	T/C	0.222	0.222	1.00	6.72×10^{-9}	1.59×10^{-1}	-0.008688	0.006176
FEV_1/FVC	rs2647145(1:17377890)	C/T	0.183	0.183	0.99	1.84×10^{-9}	1.57×10^{-1}	0.009476	0.006700
FEV_1/FVC	rs2235909(1:17429774)	T/G	0.333	0.333	0.95	5.66×10^{-15}	8.02×10^{-1}	-0.001404	0.005587
FEV_1/FVC	rs755249(1:39995074)	T/C	0.225	0.225	1.00	4.52×10^{-16}	4.81×10^{-1}	-0.004263	0.006056
FEV_1/FVC	rs3131689(1:40359485)	T/C	0.152	0.152	0.99	9.19×10^{-10}	1.89×10^{-1}	-0.009493	0.007221
FEV_1/FVC	rs9287148(1:91090125)	G/A	0.641	0.359	0.99	2.73×10^{-8}	1.51×10^{-1}	-0.007639	0.005321
FEV_1/FVC	rs13447514(1:91980282)	A/G	0.037	0.037	0.97	1.40×10^{-13}	1.31×10^{-1}	-0.020336	0.013483
FEV_1/FVC	rs72730042(1:92026941)	A/C	0.230	0.230	0.96	6.78×10^{-9}	8.03×10^{-1}	0.001545	0.006184
FEV_1/FVC	rs6681573(1:92053802)	G/A	0.344	0.344	0.99	2.50×10^{-14}	7.99×10^{-1}	-0.001382	0.005435
FEV_1/FVC	rs10874824(1:92068011)	A/G	0.472	0.472	0.99	2.77×10^{-14}	2.99×10^{-1}	0.005343	0.005142
FEV_1/FVC	rs140209273(1:92072000)	T/C	0.021	0.021	0.95	5.51×10^{-11}	5.42×10^{-1}	-0.011057	0.018117
FEV_1/FVC	rs1192415(1:92077097)	A/G	0.810	0.190	1.00	2.47×10^{-47}	9.21×10^{-1}	0.000647	0.006533
FEV_1/FVC	rs4658102(1:92077511)	T/C	0.164	0.164	0.99	1.19×10^{-20}	9.60×10^{-1}	-0.000352	0.007101
FEV_1/FVC	rs2478165(1:92097508)	G/A	0.519	0.481	0.85	2.27×10^{-8}	5.42×10^{-1}	0.003389	0.005552
FEV_1/FVC	rs1555891(1:92102009)	C/T	0.469	0.469	0.98	1.74×10^{-12}	1.29×10^{-1}	-0.007843	0.005164
FEV_1/FVC	rs12082710(1:92155337)	C/T	0.377	0.377	0.99	5.60×10^{-10}	1.47×10^{-2}	-0.012921	0.005297
FEV_1/FVC	rs17878454(1:92164876)	C/A	0.183	0.183	1.00	2.99×10^{-22}	5.45×10^{-2}	0.012665	0.006586
FEV_1/FVC	rs2391067(1:92234287)	G/T	0.273	0.273	0.97	2.73×10^{-10}	1.64×10^{-1}	0.007996	0.005752
FEV_1/FVC	rs78517377(1:92335232)	C/A	0.069	0.069	0.98	3.68×10^{-15}	4.17×10^{-1}	0.008037	0.009894
FEV_1/FVC	rs35075866(1:92384777)	T/C	0.307	0.307	0.99	5.71×10^{-23}	9.58×10^{-1}	0.000291	0.005544
FEV_1/FVC	rs11165906(1:92440939)	A/T	0.232	0.232	0.99	4.79×10^{-9}	8.20×10^{-2}	-0.010598	0.006094
FEV_1/FVC	rs17131607(1:92670783)	A/G	0.025	0.025	1.00	2.44×10^{-12}	6.63×10^{-2}	0.029430	0.016025

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs139795227(1:92842367)	C/A	0.014	0.014	0.94	6.48×10^{-9}	8.31×10^{-1}	0.004614	0.021644
FEV_1/FVC	rs6676141(1:92939217)	C/T	0.203	0.203	0.99	8.02×10^{-18}	2.15×10^{-1}	-0.007858	0.006338
FEV_1/FVC	rs12030613(1:93135195)	A/G	0.780	0.220	0.99	6.83×10^{-9}	9.65×10^{-1}	0.000269	0.006117
FEV_1/FVC	rs6658835(1:218520995)	G/A	0.279	0.279	0.99	2.43×10^{-18}	5.01×10^{-1}	-0.003895	0.005789
FEV_1/FVC	rs2799097(1:218524632)	G/A	0.849	0.151	0.99	1.31×10^{-17}	2.57×10^{-2}	-0.015994	0.007170
FEV_1/FVC	rs10482810(1:218607532)	C/G	0.006	0.006	1.00	5.66×10^{-12}	8.41×10^{-1}	-0.006331	0.031521
FEV_1/FVC	rs3009947(1:218689155)	C/T	0.501	0.499	0.99	5.29×10^{-8}	1.37×10^{-1}	-0.007628	0.005128
FEV_1/FVC	rs62817(1:218746863)	C/G	0.435	0.435	0.99	6.26×10^{-13}	1.98×10^{-2}	-0.012011	0.005156
FEV_1/FVC	rs28459747(1:218763810)	G/A	0.115	0.115	0.97	5.99×10^{-11}	8.40×10^{-2}	0.013858	0.008019
FEV_1/FVC	rs17048367(1:218833890)	G/A	0.581	0.419	1.00	3.61×10^{-9}	2.38×10^{-1}	-0.006134	0.005194
FEV_1/FVC	rs6698478(1:219056000)	T/C	0.228	0.228	0.91	1.24×10^{-9}	2.46×10^{-1}	-0.007364	0.006353
FEV_1/FVC	rs79274749(1:219500754)	T/C	0.075	0.075	0.99	8.75×10^{-19}	3.05×10^{-2}	-0.020904	0.009664
FEV_1/FVC	rs2802544(1:219572573)	C/T	0.717	0.283	0.98	1.38×10^{-9}	4.76×10^{-1}	0.004116	0.005771
FEV_1/FVC	rs1338227(1:219853742)	T/G	0.565	0.435	0.99	1.14×10^{-22}	7.23×10^{-1}	-0.001846	0.005199
FEV_1/FVC	rs4846595(1:219961253)	C/T	0.817	0.183	1.00	3.81×10^{-8}	7.60×10^{-1}	0.002071	0.006767
FEV_1/FVC	rs74651079(1:239781310)	G/C	0.076	0.076	0.95	3.24×10^{-8}	3.45×10^{-1}	0.009102	0.009636
FEV_1/FVC	rs111461590(1:239838152)	T/G	0.137	0.137	0.93	3.40×10^{-10}	2.41×10^{-1}	-0.008950	0.007633
FEV_1/FVC	rs1155612(1:239897705)	C/T	0.511	0.489	1.00	3.17×10^{-35}	4.74×10^{-1}	0.003661	0.005117
FEV_1/FVC	rs2881046(2:18283531)	A/C	0.386	0.386	0.99	3.13×10^{-14}	1.37×10^{-1}	-0.007914	0.005326
FEV_1/FVC	rs55884799(2:18287623)	C/T	0.178	0.178	0.99	1.25×10^{-34}	4.20×10^{-1}	-0.005445	0.006759
FEV_1/FVC	rs79889842(2:18306412)	T/A	0.054	0.054	0.99	5.39×10^{-16}	2.24×10^{-1}	-0.013598	0.011178
FEV_1/FVC	rs2345501(2:18336829)	G/C	0.143	0.143	0.99	4.42×10^{-11}	3.97×10^{-1}	0.006302	0.007448
FEV_1/FVC	rs12998488(2:18558525)	T/A	0.318	0.318	0.99	1.27×10^{-9}	3.48×10^{-1}	-0.005138	0.005474
FEV_1/FVC	rs3791675(2:56111309)	T/C	0.235	0.235	1.00	1.06×10^{-15}	1.53×10^{-1}	-0.008637	0.006037
FEV_1/FVC	rs8179795(2:217613285)	T/C	0.849	0.151	0.99	1.02×10^{-11}	4.90×10^{-2}	0.014212	0.007219
FEV_1/FVC	rs2571445(2:218683154)	G/A	0.609	0.391	1.00	1.18×10^{-17}	3.03×10^{-1}	0.005382	0.005224
FEV_1/FVC	rs2552529(2:218695798)	G/A	0.652	0.348	0.98	2.48×10^{-9}	1.76×10^{-1}	0.007284	0.005388
FEV_1/FVC	rs146074526(2:229385145)	T/C	0.013	0.013	0.92	6.10×10^{-9}	4.14×10^{-1}	0.018562	0.022704
FEV_1/FVC	rs190203082(2:229517553)	T/C	0.012	0.012	0.95	2.58×10^{-12}	6.80×10^{-1}	-0.009901	0.024035
FEV_1/FVC	rs62202379(2:229561921)	C/G	0.081	0.081	0.99	5.19×10^{-70}	4.08×10^{-1}	-0.007843	0.009484
FEV_1/FVC	rs10184640(2:229590473)	C/T	0.285	0.285	0.98	1.49×10^{-8}	3.53×10^{-1}	-0.005330	0.005734
FEV_1/FVC	rs1358370(2:229633134)	T/C	0.174	0.174	1.00	8.24×10^{-18}	2.71×10^{-1}	-0.007509	0.006817
FEV_1/FVC	rs4487082(2:229723961)	G/A	0.081	0.081	1.00	4.18×10^{-11}	3.45×10^{-1}	-0.008694	0.009214
FEV_1/FVC	rs6733289(2:230264691)	C/G	0.656	0.344	0.99	2.10×10^{-15}	7.73×10^{-1}	0.001567	0.005423
FEV_1/FVC	rs1050785(2:238232752)	A/C	0.609	0.391	0.96	4.23×10^{-13}	8.54×10^{-1}	-0.000983	0.005334

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs56365531(2:239458456)	T/C	0.160	0.160	0.99	3.26×10^{-10}	9.91×10^{-1}	-0.000082	0.007185
FEV_1/FVC	rs13410025(2:239614258)	A/C	0.238	0.238	0.99	1.23×10^{-8}	2.96×10^{-1}	0.006295	0.006024
FEV_1/FVC	rs74001718(2:239625596)	T/A	0.047	0.047	0.99	5.28×10^{-8}	8.05×10^{-2}	0.021709	0.012420
FEV_1/FVC	rs7571467(2:239849075)	G/A	0.631	0.369	0.98	1.13×10^{-9}	2.33×10^{-1}	-0.006356	0.005333
FEV_1/FVC	rs111583047(2:239859029)	C/T	0.382	0.382	0.96	4.46×10^{-13}	3.20×10^{-1}	-0.005317	0.005345
FEV_1/FVC	rs112948457(2:239859052)	T/A	0.053	0.053	0.93	1.87×10^{-11}	2.19×10^{-1}	0.014369	0.011702
FEV_1/FVC	rs4308141(2:239881309)	G/C	0.195	0.195	1.00	2.04×10^{-56}	5.68×10^{-2}	0.012243	0.006426
FEV_1/FVC	rs11124207(2:239897650)	A/C	0.465	0.465	0.98	1.37×10^{-15}	4.12×10^{-1}	0.004247	0.005175
FEV_1/FVC	rs4334516(2:239918121)	G/A	0.430	0.430	0.90	4.21×10^{-9}	5.24×10^{-2}	0.010502	0.005413
FEV_1/FVC	rs75857886(3:25508253)	A/G	0.109	0.109	0.95	2.43×10^{-10}	7.23×10^{-1}	0.002991	0.008425
FEV_1/FVC	rs7629478(3:25513628)	G/T	0.249	0.249	1.00	2.94×10^{-9}	1.78×10^{-1}	-0.008036	0.005969
FEV_1/FVC	rs13087022(3:25517646)	T/C	0.043	0.043	1.00	4.08×10^{-10}	6.56×10^{-1}	-0.005759	0.012931
FEV_1/FVC	rs1529672(3:25520582)	A/C	0.175	0.175	0.99	2.34×10^{-37}	7.54×10^{-1}	0.002131	0.006797
FEV_1/FVC	rs1286772(3:25580776)	G/C	0.600	0.400	0.97	1.12×10^{-10}	9.57×10^{-3}	0.013658	0.005271
FEV_1/FVC	rs17016894(3:25672778)	T/C	0.296	0.296	1.00	3.72×10^{-8}	2.45×10^{-1}	0.006542	0.005630
FEV_1/FVC	rs76490431(3:25684571)	A/G	0.055	0.055	0.95	8.89×10^{-19}	6.32×10^{-1}	-0.005410	0.011309
FEV_1/FVC	rs6809164(3:53712910)	G/T	0.270	0.270	1.00	1.02×10^{-17}	7.68×10^{-1}	0.001694	0.005752
FEV_1/FVC	rs358079(3:55110307)	G/T	0.314	0.314	0.98	1.56×10^{-8}	4.76×10^{-2}	0.010989	0.005547
FEV_1/FVC	rs9865871(3:55120948)	T/C	0.554	0.446	0.99	3.09×10^{-13}	2.92×10^{-2}	-0.011265	0.005164
FEV_1/FVC	rs1380118(3:55137259)	G/C	0.623	0.377	1.00	1.07×10^{-8}	9.73×10^{-2}	-0.008797	0.005306
FEV_1/FVC	rs17759204(3:55158224)	G/A	0.274	0.274	0.99	8.28×10^{-28}	8.54×10^{-1}	0.001053	0.005728
FEV_1/FVC	rs59719061(3:55163401)	C/T	0.382	0.382	0.99	2.09×10^{-12}	2.40×10^{-1}	-0.006206	0.005286
FEV_1/FVC	rs62254563(3:55167815)	T/C	0.052	0.052	1.00	3.09×10^{-8}	1.24×10^{-1}	-0.017293	0.011246
FEV_1/FVC	rs17216573(3:56282978)	C/T	0.169	0.169	0.98	7.14×10^{-12}	3.28×10^{-1}	-0.006882	0.007042
FEV_1/FVC	rs9849853(3:127713993)	T/C	0.409	0.409	0.98	3.52×10^{-8}	9.83×10^{-1}	0.000113	0.005249
FEV_1/FVC	rs77996940(3:127956148)	G/A	0.024	0.024	0.96	2.93×10^{-12}	2.30×10^{-1}	-0.020029	0.016689
FEV_1/FVC	rs2955083(3:127961178)	A/T	0.881	0.119	1.00	1.92×10^{-29}	4.59×10^{-1}	0.005838	0.007890
FEV_1/FVC	rs71329988(3:127976456)	T/A	0.025	0.025	0.99	2.55×10^{-9}	6.32×10^{-1}	0.007863	0.016412
FEV_1/FVC	rs2955103(3:128015236)	T/C	0.598	0.402	0.99	5.86×10^{-13}	2.94×10^{-1}	-0.005535	0.005275
FEV_1/FVC	rs9871556(3:168671733)	C/T	0.468	0.468	0.99	3.25×10^{-22}	2.32×10^{-1}	-0.006164	0.005160
FEV_1/FVC	rs56017063(3:168707007)	T/C	0.008	0.008	0.91	5.31×10^{-11}	4.75×10^{-1}	0.021457	0.030004
FEV_1/FVC	rs9846832(3:168911663)	T/C	0.663	0.337	0.99	5.30×10^{-8}	1.15×10^{-1}	0.008570	0.005431
FEV_1/FVC	rs784286(3:168970043)	A/G	0.593	0.407	1.00	1.09×10^{-12}	3.60×10^{-1}	0.004772	0.005213
FEV_1/FVC	rs6444848(3:168982572)	A/C	0.337	0.337	1.00	3.11×10^{-8}	5.78×10^{-1}	-0.003048	0.005472
FEV_1/FVC	rs710834(4:89618837)	C/T	0.379	0.379	1.00	9.66×10^{-11}	4.04×10^{-2}	0.010858	0.005297

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs3775373(4:89743821)	T/G	0.205	0.205	1.00	2.86×10^{-9}	7.46×10^{-2}	0.011481	0.006439
FEV_1/FVC	rs11737515(4:89777379)	G/C	0.132	0.132	1.00	6.77×10^{-15}	7.61×10^{-1}	-0.002281	0.007498
FEV_1/FVC	rs75106620(4:89807613)	T/C	0.046	0.046	0.98	1.14×10^{-22}	2.53×10^{-3}	0.037646	0.012467
FEV_1/FVC	rs2609279(4:89855495)	C/T	0.781	0.219	1.00	3.98×10^{-68}	1.68×10^{-4}	-0.023553	0.006259
FEV_1/FVC	rs185574798(4:89863626)	A/G	0.005	0.005	0.82	2.03×10^{-9}	3.96×10^{-3}	0.110231	0.038254
FEV_1/FVC	rs2869966(4:89869078)	T/C	0.407	0.407	1.00	4.83×10^{-65}	2.18×10^{-2}	-0.011938	0.005205
FEV_1/FVC	rs139584036(4:89870099)	T/C	0.010	0.010	0.92	2.79×10^{-8}	2.47×10^{-1}	0.030283	0.026169
FEV_1/FVC	rs78952727(4:89876109)	G/A	0.038	0.038	1.00	1.18×10^{-10}	5.34×10^{-1}	-0.008185	0.013165
FEV_1/FVC	rs78681184(4:89906598)	T/C	0.094	0.094	1.00	1.75×10^{-14}	9.25×10^{-2}	-0.014461	0.008595
FEV_1/FVC	rs2904262(4:89917060)	T/C	0.475	0.475	1.00	1.62×10^{-24}	2.34×10^{-2}	0.011615	0.005123
FEV_1/FVC	rs1708661(4:89938564)	T/C	0.435	0.435	1.00	4.21×10^{-10}	2.60×10^{-1}	-0.005828	0.005171
FEV_1/FVC	rs115491270(4:90004053)	T/G	0.040	0.040	0.95	7.81×10^{-9}	4.72×10^{-1}	0.009446	0.013131
FEV_1/FVC	rs1533292(4:90044048)	A/G	0.234	0.234	0.99	1.17×10^{-14}	1.70×10^{-1}	0.008419	0.006140
FEV_1/FVC	rs17770341(4:90048177)	A/C	0.088	0.088	0.98	4.52×10^{-10}	3.73×10^{-1}	0.008030	0.009012
FEV_1/FVC	rs76935189(4:90104537)	A/G	0.136	0.136	1.00	5.12×10^{-12}	3.70×10^{-1}	0.006722	0.007491
FEV_1/FVC	rs636895(4:90106613)	C/G	0.730	0.270	0.99	3.83×10^{-8}	3.08×10^{-1}	0.005844	0.005732
FEV_1/FVC	rs6824908(4:106046872)	A/G	0.815	0.185	0.95	2.50×10^{-10}	3.03×10^{-1}	-0.006906	0.006708
FEV_1/FVC	rs9884482(4:106081636)	C/T	0.368	0.368	1.00	2.41×10^{-15}	1.63×10^{-2}	0.012708	0.005289
FEV_1/FVC	rs6533183(4:106133184)	T/C	0.649	0.351	1.00	2.50×10^{-37}	4.67×10^{-1}	0.003922	0.005392
FEV_1/FVC	rs2726486(4:106264681)	C/T	0.921	0.079	1.00	2.24×10^{-15}	7.88×10^{-1}	-0.002631	0.009788
FEV_1/FVC	rs61117510(4:106329868)	A/C	0.316	0.316	1.00	1.81×10^{-8}	8.51×10^{-1}	0.001033	0.005488
FEV_1/FVC	rs28478829(4:106379808)	A/G	0.297	0.297	0.99	1.62×10^{-11}	3.72×10^{-1}	-0.004979	0.005576
FEV_1/FVC	rs13129703(4:106389337)	C/G	0.705	0.295	1.00	3.54×10^{-8}	4.87×10^{-1}	0.003963	0.005704
FEV_1/FVC	rs59428412(4:106524455)	C/T	0.221	0.221	1.00	2.15×10^{-11}	8.89×10^{-1}	-0.000861	0.006187
FEV_1/FVC	rs112038653(4:106564682)	T/A	0.007	0.007	0.86	2.46×10^{-8}	8.64×10^{-1}	0.005469	0.031946
FEV_1/FVC	rs17036098(4:106597452)	C/T	0.236	0.236	1.00	1.84×10^{-14}	7.69×10^{-1}	0.001799	0.006138
FEV_1/FVC	rs112819759(4:106711416)	A/G	0.043	0.043	1.00	1.88×10^{-29}	7.29×10^{-1}	-0.004262	0.012295
FEV_1/FVC	rs148331258(4:106784171)	T/C	0.044	0.044	0.97	2.66×10^{-9}	1.88×10^{-1}	0.016352	0.012427
FEV_1/FVC	rs79263635(4:106791573)	C/A	0.094	0.094	1.00	9.41×10^{-25}	6.95×10^{-1}	-0.003356	0.008553
FEV_1/FVC	rs10516529(4:106799316)	G/T	0.071	0.071	0.99	7.83×10^{-23}	4.73×10^{-1}	-0.007310	0.010187
FEV_1/FVC	rs182400279(4:106812298)	C/G	0.006	0.006	0.82	2.79×10^{-9}	7.19×10^{-1}	0.013088	0.036366
FEV_1/FVC	rs34712979(4:106819053)	A/G	0.245	0.245	1.00	2.75×10^{-139}	8.99×10^{-1}	-0.000742	0.005841
FEV_1/FVC	rs6852099(4:106819889)	C/T	0.337	0.337	0.97	6.17×10^{-30}	4.66×10^{-1}	0.004004	0.005493
FEV_1/FVC	rs4597836(4:106888263)	C/A	0.346	0.346	1.00	4.56×10^{-16}	2.00×10^{-1}	0.006921	0.005397
FEV_1/FVC	rs62320063(4:106901301)	C/T	0.062	0.062	0.96	3.99×10^{-11}	9.27×10^{-1}	0.000979	0.010743

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs62320065(4:106905609)	G/A	0.107	0.107	1.00	2.12×10^{-12}	1.24×10^{-1}	-0.012513	0.008132
FEV_1/FVC	rs3018056(4:107329974)	T/C	0.251	0.251	0.98	1.45×10^{-11}	4.81×10^{-1}	-0.004172	0.005919
FEV_1/FVC	rs4634292(4:107473889)	T/A	0.951	0.049	0.95	2.10×10^{-19}	4.26×10^{-1}	-0.009416	0.011830
FEV_1/FVC	rs1553067(4:144555318)	G/A	0.128	0.128	0.99	2.08×10^{-8}	8.65×10^{-1}	-0.001320	0.007760
FEV_1/FVC	rs151025055(4:144674986)	T/C	0.122	0.122	0.98	1.80×10^{-14}	1.14×10^{-1}	0.012164	0.007690
FEV_1/FVC	rs7677044(4:144801779)	A/G	0.458	0.458	0.99	2.61×10^{-9}	6.94×10^{-1}	0.002022	0.005149
FEV_1/FVC	rs4434191(4:145024452)	T/G	0.271	0.271	0.94	1.30×10^{-9}	2.89×10^{-1}	0.006245	0.005891
FEV_1/FVC	rs150585087(4:145024913)	G/T	0.012	0.012	0.81	4.69×10^{-10}	7.18×10^{-1}	0.009370	0.025958
FEV_1/FVC	rs62334650(4:145038318)	T/C	0.303	0.303	0.98	3.10×10^{-15}	9.22×10^{-1}	-0.000547	0.005558
FEV_1/FVC	rs4835622(4:145125913)	G/C	0.621	0.379	0.99	8.22×10^{-9}	2.95×10^{-1}	-0.005505	0.005261
FEV_1/FVC	rs72729558(4:145133493)	C/T	0.310	0.310	0.98	7.19×10^{-23}	2.73×10^{-1}	-0.006080	0.005544
FEV_1/FVC	rs7669987(4:145161554)	T/G	0.086	0.086	0.98	7.34×10^{-17}	2.17×10^{-2}	-0.020813	0.009068
FEV_1/FVC	rs77991580(4:145177949)	G/A	0.028	0.028	1.00	9.51×10^{-9}	9.76×10^{-1}	-0.000459	0.015438
FEV_1/FVC	rs13112701(4:145208104)	C/T	0.371	0.371	0.86	5.53×10^{-38}	5.39×10^{-1}	-0.003474	0.005655
FEV_1/FVC	rs72731504(4:145252763)	T/C	0.043	0.043	0.92	1.93×10^{-8}	2.26×10^{-1}	0.015508	0.012813
FEV_1/FVC	rs116825877(4:145277105)	G/A	0.027	0.027	0.90	3.67×10^{-10}	5.90×10^{-1}	-0.008730	0.016184
FEV_1/FVC	rs72731541(4:145310689)	A/C	0.065	0.065	0.90	2.09×10^{-25}	7.85×10^{-1}	-0.002935	0.010744
FEV_1/FVC	rs62334727(4:145360445)	T/C	0.074	0.074	0.99	5.01×10^{-28}	7.33×10^{-1}	0.003296	0.009673
FEV_1/FVC	rs1602238(4:145363247)	T/C	0.804	0.196	0.99	2.33×10^{-28}	8.60×10^{-2}	-0.011009	0.006413
FEV_1/FVC	rs189268532(4:145375696)	T/C	0.009	0.009	0.96	2.06×10^{-8}	1.93×10^{-2}	-0.064574	0.027609
FEV_1/FVC	rs79074298(4:145376474)	G/A	0.037	0.037	1.00	2.63×10^{-13}	3.54×10^{-1}	-0.012224	0.013189
FEV_1/FVC	rs149941228(4:145409696)	G/A	0.042	0.042	1.00	1.20×10^{-19}	9.69×10^{-1}	-0.000480	0.012342
FEV_1/FVC	rs115004137(4:145434756)	A/C	0.037	0.037	1.00	3.59×10^{-14}	4.85×10^{-2}	0.026353	0.013360
FEV_1/FVC	rs62346061(4:145472490)	A/G	0.020	0.020	0.84	1.04×10^{-14}	5.09×10^{-1}	-0.012945	0.019610
FEV_1/FVC	rs62346062(4:145472666)	T/C	0.012	0.012	0.87	1.96×10^{-11}	8.81×10^{-2}	-0.042515	0.024930
FEV_1/FVC	rs2175586(4:145496941)	G/A	0.942	0.058	0.94	2.51×10^{-10}	4.33×10^{-1}	-0.008689	0.011090
FEV_1/FVC	rs10013495(4:145505638)	T/C	0.164	0.164	1.00	8.28×10^{-20}	1.48×10^{-1}	0.010255	0.007092
FEV_1/FVC	rs13141641(4:145506456)	C/T	0.390	0.390	1.00	5.78×10^{-180}	5.44×10^{-1}	0.003170	0.005226
FEV_1/FVC	rs11724319(4:145511040)	G/A	0.228	0.228	0.97	2.83×10^{-74}	2.81×10^{-3}	-0.018187	0.006088
FEV_1/FVC	rs115555717(4:145513704)	T/C	0.025	0.025	1.00	2.20×10^{-15}	1.79×10^{-1}	-0.021481	0.015984
FEV_1/FVC	rs2353397(4:145517578)	T/C	0.555	0.445	1.00	1.20×10^{-27}	3.00×10^{-3}	0.015231	0.005132
FEV_1/FVC	rs75686861(4:145621328)	A/G	0.089	0.089	0.99	2.50×10^{-16}	9.30×10^{-1}	0.000781	0.008885
FEV_1/FVC	rs62343111(4:145851452)	T/C	0.122	0.122	1.00	3.98×10^{-8}	4.35×10^{-1}	0.006368	0.008165
FEV_1/FVC	rs10020593(4:145899099)	A/T	0.503	0.497	1.00	4.60×10^{-29}	7.34×10^{-1}	-0.001742	0.005118
FEV_1/FVC	rs62343144(4:145928473)	A/G	0.150	0.150	0.98	5.80×10^{-11}	7.70×10^{-1}	-0.002076	0.007113

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs12189242(5:34627666)	A/C	0.242	0.242	1.00	3.64×10^{-11}	2.68×10^{-2}	-0.013104	0.005918
FEV_1/FVC	rs73086816(5:51298697)	T/G	0.127	0.127	1.00	2.72×10^{-8}	1.61×10^{-1}	-0.010905	0.007772
FEV_1/FVC	rs12522114(5:52187038)	A/C	0.265	0.265	0.99	6.69×10^{-44}	7.74×10^{-1}	0.001673	0.005822
FEV_1/FVC	rs74379306(5:52192552)	G/A	0.034	0.034	0.99	1.19×10^{-9}	2.80×10^{-1}	-0.015164	0.014034
FEV_1/FVC	rs4074793(5:52193125)	G/A	0.074	0.074	1.00	3.68×10^{-17}	2.38×10^{-1}	0.011466	0.009707
FEV_1/FVC	rs1054085(5:52248291)	C/T	0.623	0.377	1.00	1.24×10^{-11}	6.49×10^{-1}	0.002406	0.005284
FEV_1/FVC	rs410064(5:52280670)	C/T	0.755	0.245	0.99	5.93×10^{-9}	6.17×10^{-1}	-0.002981	0.005956
FEV_1/FVC	rs3212666(5:52322369)	A/G	0.083	0.083	0.97	6.02×10^{-11}	1.00×10^{-1}	-0.015157	0.009225
FEV_1/FVC	rs984966(5:52368922)	A/T	0.398	0.398	1.00	1.76×10^{-8}	2.77×10^{-1}	-0.005670	0.005216
FEV_1/FVC	rs28719840(5:94952842)	A/G	0.283	0.283	0.94	2.06×10^{-9}	9.00×10^{-2}	-0.009912	0.005847
FEV_1/FVC	rs987068(5:95025146)	C/G	0.696	0.304	1.00	4.17×10^{-29}	4.70×10^{-1}	-0.003981	0.005506
FEV_1/FVC	rs10069376(5:95040803)	C/T	0.216	0.216	0.98	5.00×10^{-10}	8.68×10^{-1}	0.001049	0.006336
FEV_1/FVC	rs2548128(5:95071832)	G/A	0.304	0.304	0.96	1.67×10^{-10}	5.12×10^{-1}	-0.003722	0.005673
FEV_1/FVC	rs162892(5:131623250)	G/A	0.670	0.330	0.98	3.77×10^{-9}	4.29×10^{-1}	0.004342	0.005495
FEV_1/FVC	rs1045020(5:131730011)	T/C	0.122	0.122	1.00	4.64×10^{-13}	9.90×10^{-1}	-0.000094	0.007659
FEV_1/FVC	rs13190001(5:131744482)	T/C	0.433	0.433	1.00	1.27×10^{-11}	7.47×10^{-1}	0.001662	0.005147
FEV_1/FVC	rs7713065(5:131788334)	C/A	0.735	0.265	0.99	2.53×10^{-29}	7.62×10^{-1}	-0.001766	0.005823
FEV_1/FVC	rs11747722(5:132381617)	C/T	0.267	0.267	1.00	2.68×10^{-12}	9.64×10^{-1}	-0.000264	0.005848
FEV_1/FVC	rs986494(5:147449067)	G/C	0.675	0.325	0.98	7.31×10^{-11}	2.74×10^{-2}	0.012051	0.005464
FEV_1/FVC	rs17719662(5:147693202)	G/A	0.274	0.274	1.00	1.77×10^{-33}	3.41×10^{-1}	0.005527	0.005806
FEV_1/FVC	rs17705710(5:147702516)	G/A	0.022	0.022	1.00	1.28×10^{-10}	5.95×10^{-2}	0.033355	0.017702
FEV_1/FVC	rs114340813(5:147725831)	A/G	0.031	0.031	1.00	3.21×10^{-8}	7.22×10^{-1}	0.005151	0.014457
FEV_1/FVC	rs10515609(5:147752991)	C/T	0.105	0.105	1.00	4.37×10^{-12}	5.02×10^{-2}	-0.015979	0.008159
FEV_1/FVC	rs17638781(5:147783710)	G/A	0.139	0.139	1.00	3.31×10^{-11}	7.24×10^{-1}	0.002586	0.007328
FEV_1/FVC	rs7721661(5:147827921)	G/A	0.846	0.154	1.00	1.13×10^{-13}	8.00×10^{-2}	-0.012354	0.007057
FEV_1/FVC	rs7733410(5:147856522)	A/G	0.442	0.442	1.00	9.65×10^{-100}	1.70×10^{-2}	0.012284	0.005146
FEV_1/FVC	rs35684381(5:148203236)	C/T	0.152	0.152	0.98	1.15×10^{-11}	1.19×10^{-1}	0.011422	0.007318
FEV_1/FVC	rs1800888(5:148206885)	T/C	0.014	0.014	1.00	3.20×10^{-25}	1.86×10^{-1}	-0.027904	0.021080
FEV_1/FVC	rs13174179(5:149150671)	A/G	0.331	0.331	0.99	6.09×10^{-9}	3.39×10^{-1}	-0.005252	0.005487
FEV_1/FVC	rs2421471(5:156631551)	G/A	0.409	0.409	0.99	4.06×10^{-9}	1.80×10^{-3}	0.016324	0.005229
FEV_1/FVC	rs1862874(5:156693958)	A/T	0.335	0.335	1.00	2.85×10^{-15}	8.86×10^{-1}	0.000781	0.005445
FEV_1/FVC	rs2591460(5:156796319)	A/G	0.725	0.275	1.00	6.81×10^{-10}	8.44×10^{-1}	0.001127	0.005730
FEV_1/FVC	rs10076465(5:156801180)	C/T	0.419	0.419	0.99	9.52×10^{-16}	1.26×10^{-2}	0.012973	0.005198
FEV_1/FVC	rs112292708(5:156823447)	T/C	0.101	0.101	0.99	5.24×10^{-10}	3.45×10^{-1}	0.008043	0.008519
FEV_1/FVC	rs4704863(5:156879727)	T/C	0.514	0.486	1.00	5.17×10^{-22}	8.37×10^{-2}	-0.008853	0.005118

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs77058781(5:156886850)	G/A	0.096	0.096	0.99	9.27×10^{-9}	2.40×10^{-1}	0.010245	0.008711
FEV_1/FVC	rs11134766(5:156908317)	T/C	0.061	0.061	1.00	1.24×10^{-39}	8.99×10^{-1}	-0.001327	0.010502
FEV_1/FVC	rs35413759(5:156918147)	T/A	0.135	0.135	1.00	9.40×10^{-26}	9.24×10^{-1}	-0.000704	0.007401
FEV_1/FVC	rs13361953(5:156926442)	C/T	0.344	0.344	1.00	2.76×10^{-61}	2.05×10^{-2}	0.012542	0.005413
FEV_1/FVC	rs11465226(5:157003651)	C/A	0.140	0.140	1.00	6.36×10^{-29}	5.00×10^{-1}	0.005069	0.007512
FEV_1/FVC	rs80213842(5:157027552)	A/T	0.038	0.038	1.00	7.18×10^{-10}	6.47×10^{-1}	-0.005973	0.013027
FEV_1/FVC	rs7708495(5:157044880)	C/G	0.126	0.126	0.99	5.46×10^{-9}	1.47×10^{-1}	0.011672	0.008050
FEV_1/FVC	rs7713029(5:157075445)	C/G	0.281	0.281	0.99	5.37×10^{-13}	2.99×10^{-1}	-0.005977	0.005750
FEV_1/FVC	rs4704751(5:157193624)	C/T	0.926	0.074	0.87	6.85×10^{-11}	8.73×10^{-1}	0.001639	0.010248
FEV_1/FVC	rs139346430(6:6602844)	T/C	0.000	0.000	0.83	1.31×10^{-8}	7.60×10^{-1}	0.065348	0.213916
FEV_1/FVC	rs1294448(6:6727363)	G/C	0.507	0.493	0.99	3.34×10^{-10}	9.84×10^{-1}	-0.000101	0.005136
FEV_1/FVC	rs1294417(6:6741932)	C/T	0.534	0.466	0.99	2.68×10^{-36}	1.50×10^{-1}	-0.007403	0.005143
FEV_1/FVC	rs6938081(6:6759105)	C/G	0.402	0.402	0.97	2.58×10^{-13}	5.94×10^{-1}	0.002814	0.005280
FEV_1/FVC	rs943613(6:6832811)	T/C	0.445	0.445	0.99	3.47×10^{-9}	3.80×10^{-2}	-0.010710	0.005161
FEV_1/FVC	rs9968963(6:6850658)	T/C	0.121	0.121	0.98	1.06×10^{-11}	4.73×10^{-1}	0.005843	0.008138
FEV_1/FVC	rs7356991(6:6998130)	T/C	0.634	0.366	1.00	3.01×10^{-10}	9.51×10^{-1}	0.000328	0.005341
FEV_1/FVC	rs9505001(6:7027384)	T/C	0.697	0.303	0.99	1.24×10^{-9}	1.36×10^{-1}	-0.008442	0.005666
FEV_1/FVC	rs6904346(6:7091385)	C/A	0.923	0.077	0.99	2.76×10^{-11}	4.98×10^{-1}	0.006699	0.009881
FEV_1/FVC	rs2842895(6:7106316)	C/G	0.538	0.462	1.00	6.94×10^{-37}	8.98×10^{-1}	-0.000660	0.005151
FEV_1/FVC	rs116345157(6:7126534)	T/C	0.031	0.031	0.89	1.12×10^{-11}	9.89×10^{-1}	-0.000213	0.015296
FEV_1/FVC	rs1285884(6:7143075)	C/T	0.122	0.122	0.99	8.91×10^{-17}	3.99×10^{-1}	-0.006623	0.007845
FEV_1/FVC	rs59431963(6:7189858)	A/G	0.385	0.385	0.97	4.81×10^{-14}	2.23×10^{-1}	-0.006477	0.005312
FEV_1/FVC	rs9392863(6:7195009)	A/G	0.249	0.249	0.99	4.69×10^{-20}	6.48×10^{-1}	-0.002685	0.005881
FEV_1/FVC	rs12192672(6:7229619)	A/G	0.291	0.291	1.00	4.10×10^{-10}	8.66×10^{-1}	0.000945	0.005596
FEV_1/FVC	rs17672837(6:7257023)	C/G	0.038	0.038	1.00	3.81×10^{-8}	2.59×10^{-2}	0.029320	0.013161
FEV_1/FVC	rs9505168(6:7327846)	C/T	0.556	0.444	1.00	1.30×10^{-10}	2.85×10^{-1}	-0.005495	0.005142
FEV_1/FVC	rs2076295(6:7563232)	G/T	0.449	0.449	0.99	2.63×10^{-18}	5.91×10^{-1}	-0.002773	0.005155
FEV_1/FVC	rs9351958(6:73661010)	C/G	0.197	0.197	1.00	5.75×10^{-32}	4.61×10^{-1}	-0.004704	0.006386
FEV_1/FVC	rs2153960(6:108988184)	A/G	0.696	0.304	1.00	1.90×10^{-8}	5.36×10^{-1}	0.003488	0.005638
FEV_1/FVC	rs35711192(6:109104798)	A/G	0.030	0.030	1.00	2.98×10^{-8}	5.49×10^{-1}	-0.008724	0.014558
FEV_1/FVC	rs911477(6:109262907)	T/C	0.579	0.421	0.99	1.00×10^{-15}	3.03×10^{-1}	0.005378	0.005216
FEV_1/FVC	rs2798641(6:109268050)	T/C	0.181	0.181	1.00	4.80×10^{-51}	1.27×10^{-1}	0.010081	0.006599
FEV_1/FVC	rs79774757(6:109276400)	G/A	0.026	0.026	1.00	5.39×10^{-11}	6.50×10^{-1}	0.007277	0.016032
FEV_1/FVC	rs147590351(6:142368007)	C/G	0.005	0.005	0.89	2.20×10^{-16}	6.86×10^{-1}	-0.015086	0.037269
FEV_1/FVC	rs9496267(6:142419667)	A/G	0.083	0.083	1.00	5.27×10^{-20}	9.33×10^{-1}	-0.000785	0.009285

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs9385987(6:142426244)	T/C	0.248	0.248	0.96	1.81×10^{-17}	9.24×10^{-1}	0.000571	0.005974
FEV_1/FVC	rs6928024(6:142551082)	A/G	0.262	0.262	1.00	6.59×10^{-39}	6.23×10^{-1}	-0.002889	0.005882
FEV_1/FVC	rs2782530(6:142582541)	A/G	0.889	0.111	0.99	5.45×10^{-8}	6.30×10^{-1}	0.004065	0.008434
FEV_1/FVC	rs17790314(6:142624280)	C/T	0.089	0.089	0.93	1.08×10^{-13}	5.18×10^{-1}	0.005958	0.009220
FEV_1/FVC	rs12204775(6:142634142)	T/C	0.072	0.072	0.93	5.13×10^{-42}	9.34×10^{-1}	-0.000835	0.010017
FEV_1/FVC	rs146088795(6:142640832)	G/A	0.015	0.015	0.98	1.18×10^{-12}	6.15×10^{-1}	-0.010337	0.020557
FEV_1/FVC	rs182686136(6:142659908)	C/T	0.012	0.012	0.93	4.41×10^{-10}	8.99×10^{-1}	0.003001	0.023693
FEV_1/FVC	rs142016511(6:142678675)	T/G	0.003	0.003	0.92	1.70×10^{-16}	7.81×10^{-1}	-0.014141	0.050886
FEV_1/FVC	rs73780221(6:142725182)	C/G	0.033	0.033	1.00	1.54×10^{-129}	8.77×10^{-1}	-0.002345	0.015185
FEV_1/FVC	rs75697856(6:142733170)	A/G	0.003	0.003	0.95	8.48×10^{-16}	6.16×10^{-1}	-0.025809	0.051503
FEV_1/FVC	rs7753012(6:142745883)	G/T	0.322	0.322	1.00	2.10×10^{-176}	3.52×10^{-1}	-0.005175	0.005563
FEV_1/FVC	rs12529186(6:142773504)	T/G	0.269	0.269	0.94	9.34×10^{-26}	1.99×10^{-1}	0.007572	0.005893
FEV_1/FVC	rs79309679(6:142823393)	T/C	0.102	0.102	1.00	8.33×10^{-33}	4.70×10^{-1}	-0.006110	0.008460
FEV_1/FVC	rs77319647(6:142826871)	G/C	0.035	0.035	1.00	1.67×10^{-10}	3.38×10^{-1}	-0.013229	0.013805
FEV_1/FVC	rs191453644(6:142827131)	G/A	0.004	0.004	0.93	3.18×10^{-22}	7.73×10^{-1}	-0.011717	0.040621
FEV_1/FVC	rs145743542(6:142836854)	G/C	0.017	0.017	0.83	4.71×10^{-8}	9.56×10^{-1}	-0.001197	0.021553
FEV_1/FVC	rs55771139(6:142842841)	A/G	0.006	0.006	1.00	1.64×10^{-20}	9.56×10^{-1}	0.001868	0.033739
FEV_1/FVC	rs537713083(6:142846459)	G/A	0.002	0.002	0.91	1.21×10^{-11}	8.94×10^{-1}	0.007327	0.055001
FEV_1/FVC	rs263169(6:142873650)	C/A	0.153	0.153	1.00	8.03×10^{-9}	9.63×10^{-1}	-0.000332	0.007190
FEV_1/FVC	rs139461848(6:143029318)	A/G	0.012	0.012	0.82	6.97×10^{-9}	4.56×10^{-1}	-0.018841	0.025255
FEV_1/FVC	rs117779266(6:143043970)	T/C	0.031	0.031	1.00	7.40×10^{-20}	2.86×10^{-1}	-0.015397	0.014440
FEV_1/FVC	rs79217575(6:143047865)	A/G	0.044	0.044	0.90	5.18×10^{-8}	9.56×10^{-1}	0.000702	0.012816
FEV_1/FVC	rs11771259(7:7277215)	G/C	0.116	0.116	1.00	2.71×10^{-11}	4.64×10^{-1}	0.005811	0.007941
FEV_1/FVC	rs10246303(7:7286445)	T/A	0.430	0.430	0.99	5.09×10^{-13}	7.09×10^{-1}	-0.001948	0.005227
FEV_1/FVC	rs12531809(7:99620473)	T/G	0.364	0.364	1.00	6.21×10^{-22}	4.55×10^{-1}	0.003957	0.005301
FEV_1/FVC	rs56252897(7:99684271)	A/C	0.078	0.078	0.98	3.49×10^{-8}	1.09×10^{-1}	-0.015338	0.009579
FEV_1/FVC	rs35305377(7:99938955)	A/G	0.549	0.451	0.99	2.02×10^{-16}	1.92×10^{-2}	0.012067	0.005152
FEV_1/FVC	rs17855473(7:100160264)	C/T	0.181	0.181	1.00	3.68×10^{-8}	3.44×10^{-1}	0.006300	0.006662
FEV_1/FVC	rs1233553(7:155598145)	A/G	0.936	0.064	0.88	5.39×10^{-8}	3.27×10^{-1}	-0.010722	0.010940
FEV_1/FVC	rs1233556(7:155600417)	C/T	0.844	0.156	0.94	3.70×10^{-12}	1.95×10^{-1}	0.009367	0.007236
FEV_1/FVC	rs4458740(7:156089537)	C/A	0.526	0.474	0.99	3.72×10^{-9}	2.39×10^{-1}	0.006038	0.005128
FEV_1/FVC	rs12698403(7:156127246)	A/G	0.437	0.437	1.00	3.23×10^{-21}	2.97×10^{-1}	0.005390	0.005170
FEV_1/FVC	rs73492422(7:156128566)	C/T	0.176	0.176	0.99	2.88×10^{-9}	8.68×10^{-1}	0.001140	0.006868
FEV_1/FVC	rs3925025(9:4080864)	G/T	0.267	0.267	1.00	1.13×10^{-11}	8.30×10^{-2}	-0.009968	0.005750
FEV_1/FVC	rs806038(9:4092700)	T/C	0.598	0.402	0.99	6.05×10^{-9}	7.11×10^{-1}	0.001935	0.005223

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs10974345(9:4144203)	G/C	0.436	0.436	0.99	8.66×10^{-25}	6.92×10^{-1}	-0.002045	0.005168
FEV_1/FVC	rs62524071(9:4153283)	A/G	0.086	0.086	0.99	4.02×10^{-10}	7.70×10^{-2}	-0.015940	0.009013
FEV_1/FVC	rs357564(9:98209594)	A/G	0.340	0.340	1.00	1.69×10^{-9}	5.77×10^{-1}	0.003014	0.005401
FEV_1/FVC	rs60417486(9:98262178)	A/G	0.090	0.090	0.99	2.51×10^{-34}	2.13×10^{-1}	-0.011190	0.008976
FEV_1/FVC	rs28620668(9:98267746)	G/T	0.332	0.332	0.99	1.65×10^{-24}	6.54×10^{-2}	-0.010067	0.005464
FEV_1/FVC	rs111066154(9:98330226)	T/C	0.228	0.228	0.99	4.98×10^{-9}	2.43×10^{-1}	-0.007195	0.006159
FEV_1/FVC	rs72743974(9:98878881)	G/A	0.167	0.167	1.00	8.38×10^{-9}	2.70×10^{-1}	0.007584	0.006872
FEV_1/FVC	rs1054402(9:119163509)	C/T	0.748	0.252	0.99	9.88×10^{-13}	9.90×10^{-1}	-0.000072	0.005947
FEV_1/FVC	rs10983184(9:119234058)	T/C	0.633	0.367	0.99	2.11×10^{-22}	5.98×10^{-1}	-0.002826	0.005360
FEV_1/FVC	rs4837580(9:119273819)	T/C	0.381	0.381	0.99	6.37×10^{-14}	6.37×10^{-1}	0.002481	0.005262
FEV_1/FVC	rs3758379(10:12236773)	A/G	0.128	0.128	1.00	8.23×10^{-9}	5.10×10^{-1}	-0.005004	0.007590
FEV_1/FVC	rs2399794(10:12248800)	A/G	0.556	0.444	0.95	1.09×10^{-21}	1.37×10^{-1}	-0.007836	0.005271
FEV_1/FVC	rs7090277(10:12278021)	A/T	0.509	0.491	1.00	9.85×10^{-60}	1.54×10^{-1}	-0.007295	0.005115
FEV_1/FVC	rs11597664(10:12314197)	T/A	0.233	0.233	1.00	5.96×10^{-13}	1.82×10^{-1}	-0.008084	0.006064
FEV_1/FVC	rs6602570(10:12320571)	C/A	0.812	0.188	0.99	1.10×10^{-26}	4.42×10^{-2}	-0.013185	0.006551
FEV_1/FVC	rs12358571(10:12327724)	T/C	0.593	0.407	0.97	1.60×10^{-11}	5.67×10^{-1}	0.003028	0.005294
FEV_1/FVC	rs3995695(10:12712070)	A/G	0.359	0.359	1.00	4.84×10^{-8}	8.26×10^{-2}	0.009222	0.005314
FEV_1/FVC	rs10906448(10:13733141)	T/C	0.251	0.251	1.00	1.66×10^{-8}	2.21×10^{-1}	-0.007142	0.005834
FEV_1/FVC	rs35495115(10:30253250)	T/A	0.132	0.132	0.99	7.20×10^{-9}	8.29×10^{-1}	-0.001623	0.007523
FEV_1/FVC	rs4749511(10:30266579)	C/T	0.449	0.449	0.99	9.84×10^{-9}	9.92×10^{-1}	0.000053	0.005152
FEV_1/FVC	rs4749652(10:31270278)	T/C	0.523	0.477	1.00	2.78×10^{-9}	1.68×10^{-2}	-0.012238	0.005117
FEV_1/FVC	rs11594905(10:77659733)	A/G	0.132	0.132	1.00	3.02×10^{-9}	1.11×10^{-2}	-0.018900	0.007444
FEV_1/FVC	rs2579762(10:78318879)	C/A	0.464	0.464	1.00	5.77×10^{-19}	2.28×10^{-2}	0.011653	0.005117
FEV_1/FVC	rs11234768(11:86448839)	C/T	0.149	0.149	1.00	5.18×10^{-18}	6.76×10^{-1}	0.002961	0.007075
FEV_1/FVC	rs3096718(11:86458705)	C/T	0.454	0.454	0.99	4.77×10^{-8}	3.23×10^{-1}	0.005097	0.005162
FEV_1/FVC	rs662265(11:126010971)	G/A	0.809	0.191	0.99	1.48×10^{-13}	2.19×10^{-1}	-0.008145	0.006633
FEV_1/FVC	rs7976850(12:27882291)	A/G	0.085	0.085	1.00	5.16×10^{-9}	8.02×10^{-1}	-0.002291	0.009131
FEV_1/FVC	rs1701704(12:56412487)	G/T	0.334	0.334	1.00	2.85×10^{-9}	8.05×10^{-1}	-0.001333	0.005384
FEV_1/FVC	rs11172113(12:57527283)	C/T	0.411	0.411	1.00	2.04×10^{-17}	8.84×10^{-1}	-0.000755	0.005193
FEV_1/FVC	rs186767801(12:57727682)	A/G	0.003	0.003	0.89	2.95×10^{-11}	2.41×10^{-2}	0.103414	0.045863
FEV_1/FVC	rs4762080(12:65796310)	A/T	0.214	0.214	0.99	2.33×10^{-8}	5.27×10^{-1}	0.003925	0.006207
FEV_1/FVC	rs7306256(12:95348178)	A/G	0.427	0.427	1.00	3.95×10^{-10}	3.83×10^{-1}	-0.004505	0.005159
FEV_1/FVC	rs12814712(12:95360709)	A/G	0.103	0.103	0.99	1.79×10^{-8}	1.03×10^{-1}	0.013945	0.008546
FEV_1/FVC	rs117624293(12:95438785)	C/G	0.051	0.051	0.96	1.76×10^{-11}	1.58×10^{-1}	-0.016462	0.011665
FEV_1/FVC	rs7954260(12:95518223)	C/T	0.190	0.190	1.00	2.18×10^{-14}	9.24×10^{-1}	0.000635	0.006628

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs55703445(12:95541202)	T/C	0.211	0.211	1.00	1.52×10^{-20}	9.25×10^{-1}	-0.000588	0.006224
FEV_1/FVC	rs6538615(12:95587856)	C/T	0.639	0.361	1.00	6.65×10^{-9}	8.48×10^{-1}	-0.001022	0.005339
FEV_1/FVC	rs11107973(12:95702385)	C/T	0.369	0.369	1.00	3.87×10^{-13}	5.10×10^{-1}	-0.003515	0.005332
FEV_1/FVC	rs34528271(12:96072516)	T/A	0.108	0.108	0.98	4.74×10^{-10}	9.24×10^{-1}	-0.000808	0.008434
FEV_1/FVC	rs7309423(12:96097147)	C/T	0.218	0.218	1.00	8.55×10^{-9}	3.51×10^{-1}	0.005734	0.006147
FEV_1/FVC	rs78477125(12:96120350)	G/A	0.020	0.020	0.91	6.10×10^{-9}	2.83×10^{-1}	0.020033	0.018647
FEV_1/FVC	rs4274262(12:96136458)	G/A	0.540	0.460	0.99	3.95×10^{-16}	6.70×10^{-1}	0.002195	0.005147
FEV_1/FVC	rs10859932(12:96138327)	T/G	0.847	0.153	1.00	1.51×10^{-10}	9.07×10^{-1}	0.000850	0.007279
FEV_1/FVC	rs78711843(12:96192776)	G/A	0.067	0.067	0.99	3.12×10^{-8}	7.12×10^{-1}	-0.003850	0.010412
FEV_1/FVC	rs35211087(12:96244244)	T/C	0.316	0.316	1.00	1.95×10^{-23}	3.84×10^{-1}	0.004750	0.005462
FEV_1/FVC	rs4762630(12:96244409)	A/G	0.812	0.188	1.00	1.48×10^{-44}	9.15×10^{-1}	-0.000696	0.006519
FEV_1/FVC	rs11108351(12:96348536)	A/C	0.204	0.204	0.99	8.07×10^{-16}	5.92×10^{-1}	0.003416	0.006377
FEV_1/FVC	rs61938793(12:96481462)	A/G	0.134	0.134	1.00	2.13×10^{-13}	8.09×10^{-1}	0.001797	0.007441
FEV_1/FVC	rs11067278(12:115184053)	C/G	0.393	0.393	0.98	4.62×10^{-8}	4.16×10^{-1}	0.004285	0.005271
FEV_1/FVC	rs1874903(12:115950227)	T/C	0.471	0.471	0.97	1.98×10^{-12}	3.27×10^{-1}	0.005076	0.005175
FEV_1/FVC	rs2261760(14:84285998)	T/C	0.311	0.311	1.00	1.57×10^{-22}	2.54×10^{-1}	-0.006364	0.005584
FEV_1/FVC	rs111920458(14:84293632)	A/G	0.062	0.062	0.99	1.36×10^{-10}	2.90×10^{-1}	-0.011236	0.010616
FEV_1/FVC	rs2003253(14:91130763)	A/G	0.279	0.279	0.99	4.42×10^{-9}	3.41×10^{-1}	-0.005414	0.005689
FEV_1/FVC	rs12894780(14:93503386)	C/T	0.138	0.138	0.99	3.50×10^{-20}	9.64×10^{-1}	-0.000345	0.007744
FEV_1/FVC	rs4900195(14:94245652)	T/C	0.488	0.488	1.00	2.62×10^{-10}	2.59×10^{-2}	-0.011395	0.005114
FEV_1/FVC	rs17129265(14:94341752)	T/C	0.172	0.172	0.99	7.97×10^{-10}	4.85×10^{-1}	0.004797	0.006873
FEV_1/FVC	rs2012453(15:41840238)	G/A	0.590	0.410	0.99	1.18×10^{-19}	5.74×10^{-1}	-0.002938	0.005231
FEV_1/FVC	rs62003872(15:41886714)	T/C	0.294	0.294	1.00	5.00×10^{-22}	4.48×10^{-1}	0.004252	0.005605
FEV_1/FVC	rs56383987(15:41953211)	T/C	0.054	0.054	0.97	3.75×10^{-11}	3.63×10^{-1}	-0.010253	0.011267
FEV_1/FVC	rs62002122(15:42146189)	T/C	0.334	0.334	0.98	1.51×10^{-11}	8.61×10^{-1}	-0.000955	0.005437
FEV_1/FVC	rs11853555(15:71528435)	T/G	0.322	0.322	0.97	3.82×10^{-15}	8.33×10^{-1}	0.001167	0.005549
FEV_1/FVC	rs35683356(15:71539772)	G/A	0.154	0.154	0.99	1.34×10^{-14}	1.60×10^{-1}	0.009893	0.007040
FEV_1/FVC	rs79745150(15:71543487)	C/T	0.026	0.026	1.00	7.54×10^{-14}	9.44×10^{-1}	0.001095	0.015627
FEV_1/FVC	rs62015473(15:71555837)	A/G	0.241	0.241	0.98	3.56×10^{-26}	2.11×10^{-1}	0.007497	0.005992
FEV_1/FVC	rs80019083(15:71572350)	T/C	0.045	0.045	1.00	3.61×10^{-12}	1.55×10^{-2}	-0.029219	0.012069
FEV_1/FVC	rs2063826(15:71583413)	C/T	0.320	0.320	1.00	2.60×10^{-13}	8.68×10^{-1}	-0.000912	0.005484
FEV_1/FVC	rs11072270(15:71606121)	T/C	0.271	0.271	0.98	3.30×10^{-28}	1.81×10^{-1}	0.007731	0.005780
FEV_1/FVC	rs1441358(15:71612514)	G/T	0.344	0.344	1.00	1.95×10^{-142}	6.99×10^{-1}	0.002090	0.005411
FEV_1/FVC	rs62015510(15:71621835)	T/C	0.015	0.015	0.95	9.70×10^{-11}	9.48×10^{-1}	0.001362	0.020807
FEV_1/FVC	rs75153206(15:71623906)	T/C	0.064	0.064	1.00	9.26×10^{-11}	6.70×10^{-1}	-0.004420	0.010360

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs12911203(15:71640214)	A/G	0.021	0.021	1.00	4.97×10^{-11}	9.91×10^{-1}	0.000201	0.017497
FEV_1/FVC	rs12441628(15:71655093)	G/A	0.213	0.213	0.99	4.17×10^{-96}	4.71×10^{-1}	0.004501	0.006249
FEV_1/FVC	rs547095830(15:71674337)	A/T	0.001	0.001	0.94	6.24×10^{-9}	8.27×10^{-1}	-0.018147	0.082928
FEV_1/FVC	rs2119570(15:71677788)	G/A	0.026	0.026	1.00	1.21×10^{-29}	7.62×10^{-1}	0.004695	0.015522
FEV_1/FVC	rs4531689(15:71688979)	T/C	0.425	0.425	1.00	2.27×10^{-64}	1.14×10^{-1}	0.008168	0.005164
FEV_1/FVC	rs188056730(15:71692619)	A/G	0.008	0.008	0.91	2.72×10^{-12}	9.75×10^{-1}	0.000949	0.030379
FEV_1/FVC	rs117193529(15:71701610)	A/G	0.025	0.025	1.00	7.46×10^{-17}	5.39×10^{-1}	0.009976	0.016257
FEV_1/FVC	rs34719283(15:71706624)	C/T	0.362	0.362	0.98	7.75×10^{-56}	1.78×10^{-1}	-0.007179	0.005334
FEV_1/FVC	rs77494339(15:71714873)	G/T	0.015	0.015	0.87	3.37×10^{-10}	8.19×10^{-2}	0.039937	0.022955
FEV_1/FVC	rs113850582(15:71721981)	A/G	0.003	0.003	0.85	2.91×10^{-8}	9.02×10^{-1}	0.006341	0.051357
FEV_1/FVC	rs57589422(15:71727706)	G/T	0.144	0.144	0.98	3.61×10^{-15}	6.29×10^{-1}	-0.003561	0.007374
FEV_1/FVC	rs12909439(15:71739708)	G/T	0.286	0.286	0.97	1.70×10^{-17}	4.52×10^{-1}	-0.004264	0.005670
FEV_1/FVC	rs4238437(15:71741779)	C/T	0.764	0.236	0.95	1.46×10^{-12}	4.03×10^{-2}	0.012690	0.006189
FEV_1/FVC	rs7183859(15:71805006)	T/C	0.174	0.174	0.99	1.65×10^{-20}	9.82×10^{-1}	0.000147	0.006708
FEV_1/FVC	rs28650139(15:71812163)	C/T	0.599	0.401	1.00	4.21×10^{-8}	4.60×10^{-2}	0.010400	0.005213
FEV_1/FVC	rs34660045(15:71816660)	T/C	0.130	0.130	0.99	2.85×10^{-10}	2.47×10^{-2}	0.017175	0.007644
FEV_1/FVC	rs56081433(15:84011144)	A/G	0.244	0.244	1.00	8.73×10^{-9}	2.57×10^{-1}	0.006701	0.005911
FEV_1/FVC	rs17584591(15:84191336)	C/T	0.208	0.208	0.97	3.81×10^{-8}	1.11×10^{-1}	0.010071	0.006319
FEV_1/FVC	rs1491577(15:84202348)	G/A	0.811	0.189	0.99	7.89×10^{-10}	8.94×10^{-1}	-0.000865	0.006506
FEV_1/FVC	rs2585071(15:84264404)	A/G	0.659	0.341	0.98	8.77×10^{-27}	5.02×10^{-1}	0.003641	0.005425
FEV_1/FVC	rs7162245(15:84300727)	G/A	0.659	0.341	1.00	1.30×10^{-10}	3.55×10^{-1}	-0.004955	0.005357
FEV_1/FVC	rs117540214(15:84338642)	G/A	0.062	0.062	0.98	4.62×10^{-11}	8.94×10^{-1}	-0.001399	0.010505
FEV_1/FVC	rs73437211(15:84418173)	A/T	0.154	0.154	1.00	2.52×10^{-11}	9.78×10^{-1}	0.000203	0.007193
FEV_1/FVC	rs56226101(15:84426284)	T/C	0.204	0.204	1.00	9.57×10^{-24}	6.64×10^{-1}	0.002767	0.006369
FEV_1/FVC	rs7181926(15:84549428)	C/G	0.520	0.480	1.00	2.83×10^{-30}	9.04×10^{-2}	0.008664	0.005117
FEV_1/FVC	rs10520572(15:84591675)	C/T	0.234	0.234	0.99	8.06×10^{-22}	2.22×10^{-1}	-0.007543	0.006174
FEV_1/FVC	rs17370572(15:84616713)	T/C	0.156	0.156	1.00	4.47×10^{-13}	2.15×10^{-1}	-0.008592	0.006929
FEV_1/FVC	rs366717(15:85088079)	A/G	0.221	0.221	0.99	7.18×10^{-15}	8.17×10^{-1}	-0.001416	0.006130
FEV_1/FVC	rs8057607(16:10711783)	G/C	0.387	0.387	0.99	2.81×10^{-11}	9.31×10^{-1}	0.000459	0.005293
FEV_1/FVC	rs78442819(16:10740982)	C/G	0.190	0.190	0.95	7.32×10^{-25}	6.82×10^{-1}	0.002689	0.006569
FEV_1/FVC	rs62025795(16:10752141)	A/G	0.368	0.368	0.98	6.34×10^{-9}	7.18×10^{-1}	-0.001921	0.005316
FEV_1/FVC	rs12935657(16:11219041)	A/G	0.242	0.242	0.99	1.83×10^{-8}	6.27×10^{-1}	-0.002890	0.005950
FEV_1/FVC	rs3743552(16:58029954)	G/C	0.083	0.083	0.98	4.43×10^{-25}	8.03×10^{-2}	-0.016688	0.009540
FEV_1/FVC	rs11648508(16:58063513)	T/G	0.681	0.319	0.99	1.75×10^{-39}	4.79×10^{-2}	0.010929	0.005523
FEV_1/FVC	rs41390948(16:58075235)	T/C	0.013	0.013	0.94	6.62×10^{-10}	2.88×10^{-1}	0.023883	0.022466

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs74470468(16:58132535)	A/G	0.059	0.059	1.00	1.71×10^{-11}	3.74×10^{-2}	-0.022615	0.010865
FEV_1/FVC	rs34687030(16:58140687)	A/G	0.148	0.148	0.97	6.27×10^{-14}	7.31×10^{-1}	-0.002478	0.007216
FEV_1/FVC	rs61688272(16:58336100)	G/A	0.121	0.121	0.99	5.64×10^{-9}	2.32×10^{-1}	0.009405	0.007873
FEV_1/FVC	rs73605154(16:75303810)	T/C	0.215	0.215	0.89	2.81×10^{-16}	1.07×10^{-1}	0.010497	0.006510
FEV_1/FVC	rs72787129(16:75407682)	G/A	0.076	0.076	0.99	1.98×10^{-10}	7.81×10^{-1}	0.002715	0.009760
FEV_1/FVC	rs12932007(16:75428556)	T/C	0.131	0.131	0.78	8.47×10^{-20}	6.38×10^{-1}	0.004034	0.008579
FEV_1/FVC	rs247440(16:75434992)	C/G	0.037	0.037	1.00	2.54×10^{-8}	5.15×10^{-1}	-0.008607	0.013227
FEV_1/FVC	rs11149828(16:75439489)	C/T	0.589	0.411	1.00	1.64×10^{-54}	4.33×10^{-1}	0.004088	0.005213
FEV_1/FVC	rs247454(16:75519390)	C/G	0.393	0.393	0.89	3.62×10^{-19}	8.78×10^{-1}	-0.000852	0.005558
FEV_1/FVC	rs636000(17:27891864)	G/A	0.276	0.276	0.98	1.35×10^{-9}	2.01×10^{-1}	0.007341	0.005735
FEV_1/FVC	rs3115094(17:27913807)	G/C	0.808	0.192	0.98	2.36×10^{-9}	5.36×10^{-1}	0.004020	0.006488
FEV_1/FVC	rs1808923(17:27980885)	T/C	0.467	0.467	1.00	2.84×10^{-42}	6.39×10^{-1}	0.002405	0.005129
FEV_1/FVC	rs34675417(17:28079167)	G/A	0.246	0.246	0.99	1.32×10^{-16}	7.22×10^{-2}	0.010581	0.005886
FEV_1/FVC	rs62070347(17:28482366)	C/T	0.112	0.112	0.96	4.68×10^{-8}	2.90×10^{-1}	0.008559	0.008093
FEV_1/FVC	rs1872924(17:28546346)	T/C	0.775	0.225	0.99	4.50×10^{-8}	4.80×10^{-1}	0.004411	0.006239
FEV_1/FVC	rs8072345(17:28604289)	T/C	0.445	0.445	1.00	6.53×10^{-15}	4.99×10^{-2}	-0.010098	0.005151
FEV_1/FVC	rs216450(17:28878817)	G/A	0.581	0.419	1.00	4.54×10^{-9}	1.58×10^{-2}	0.012497	0.005179
FEV_1/FVC	rs11079059(17:36813346)	A/G	0.251	0.251	0.98	1.13×10^{-13}	8.93×10^{-1}	-0.000795	0.005914
FEV_1/FVC	rs11543289(17:36882595)	A/T	0.383	0.383	0.96	2.51×10^{-10}	3.76×10^{-1}	0.004763	0.005385
FEV_1/FVC	rs35246838(17:36915540)	C/T	0.127	0.127	0.97	8.90×10^{-25}	1.66×10^{-1}	-0.010609	0.007661
FEV_1/FVC	rs2338115(17:36929578)	T/C	0.542	0.458	1.00	3.97×10^{-9}	4.62×10^{-1}	-0.003792	0.005149
FEV_1/FVC	rs9303283(17:38192633)	T/C	0.575	0.425	0.99	1.18×10^{-9}	3.70×10^{-5}	0.021340	0.005173
FEV_1/FVC	rs56304903(17:38346802)	G/C	0.156	0.156	0.99	2.65×10^{-15}	1.11×10^{-1}	-0.011317	0.007097
FEV_1/FVC	rs983085(17:69212061)	G/A	0.494	0.494	1.00	1.31×10^{-15}	5.87×10^{-1}	0.002776	0.005112
FEV_1/FVC	rs996865(17:69371318)	T/C	0.080	0.080	0.98	4.44×10^{-23}	4.23×10^{-1}	0.007740	0.009662
FEV_1/FVC	rs8077434(17:69843473)	C/T	0.450	0.450	1.00	4.98×10^{-8}	8.65×10^{-1}	-0.000872	0.005142
FEV_1/FVC	rs8067763(17:70012939)	A/G	0.601	0.399	0.99	2.36×10^{-8}	2.14×10^{-2}	-0.012054	0.005240
FEV_1/FVC	rs9913936(17:70074148)	C/T	0.670	0.330	0.99	8.34×10^{-11}	1.10×10^{-2}	-0.013903	0.005466
FEV_1/FVC	rs34093919(19:41117300)	A/G	0.012	0.012	1.00	1.61×10^{-42}	4.59×10^{-1}	0.017133	0.023150
FEV_1/FVC	rs62107911(19:41126447)	G/A	0.294	0.294	0.99	7.06×10^{-11}	1.42×10^{-1}	0.008330	0.005678
FEV_1/FVC	rs2604894(19:41292404)	G/A	0.546	0.454	0.98	3.58×10^{-8}	5.10×10^{-2}	-0.010103	0.005177
FEV_1/FVC	rs12151878(20:60997330)	G/T	0.243	0.243	0.98	1.17×10^{-9}	4.32×10^{-1}	0.004769	0.006076
FEV_1/FVC	rs2427330(20:60999562)	A/C	0.525	0.475	0.98	1.78×10^{-9}	7.93×10^{-1}	-0.001355	0.005166
FEV_1/FVC	rs6089386(20:61038900)	C/T	0.922	0.078	0.94	4.51×10^{-11}	5.60×10^{-2}	-0.018999	0.009943
FEV_1/FVC	rs6421437(20:61040313)	T/G	0.366	0.366	0.98	7.43×10^{-15}	3.68×10^{-1}	-0.004831	0.005367

Trait	SNP (Chr:Pos)	Coded / Non coded allele	CAF	MAF	INFO	Joint test P	Interaction test P	Interaction test β	Interaction test SE
FEV_1/FVC	rs13041686(20:61049831)	C/T	0.212	0.212	1.00	1.06×10^{-9}	1.89×10^{-1}	-0.008235	0.006274
FEV_1/FVC	rs4514955(20:61060934)	G/C	0.569	0.431	0.96	4.20×10^{-9}	8.86×10^{-1}	-0.000748	0.005240
FEV_1/FVC	rs11702760(21:35294638)	A/G	0.026	0.026	0.98	1.10×10^{-9}	9.65×10^{-1}	-0.000695	0.015808
FEV_1/FVC	rs12627254(21:35368402)	T/G	0.128	0.128	1.00	9.51×10^{-21}	3.82×10^{-1}	-0.006690	0.007658
FEV_1/FVC	rs2834435(21:35647853)	A/G	0.150	0.150	1.00	5.35×10^{-15}	9.21×10^{-1}	-0.000720	0.007261
FEV_1/FVC	rs2834440(21:35690499)	A/G	0.613	0.387	1.00	8.05×10^{-20}	1.12×10^{-2}	-0.013339	0.005262
FEV_1/FVC	rs4819639(22:18347127)	T/C	0.251	0.251	0.99	3.42×10^{-9}	2.20×10^{-1}	-0.007223	0.005884
FEV_1/FVC	rs415651(22:18436944)	A/G	0.476	0.476	1.00	1.12×10^{-15}	1.07×10^{-1}	-0.008271	0.005132
FEV_1/FVC	rs9605473(22:18463266)	G/A	0.253	0.253	0.98	7.54×10^{-12}	7.88×10^{-1}	-0.001596	0.005935
FEV_1/FVC	rs72646967(22:19754091)	C/A	0.215	0.215	0.99	2.47×10^{-11}	2.21×10^{-1}	-0.007615	0.006226
FEV_1/FVC	rs9618700(22:19803382)	A/C	0.224	0.224	0.98	1.72×10^{-12}	3.70×10^{-5}	0.025754	0.006243
FEV_1/FVC	rs2283847(22:28181399)	T/C	0.548	0.452	0.95	9.06×10^{-18}	1.22×10^{-1}	-0.008155	0.005277

E. Ever and never smoker genetic effect sizes (β) and standard errors (SE) for both UK Biobank (UKB) and the SpiroMeta consortium (SpM). Results are on the inverse normalised scale.

Trait	SNP (Chr:Pos)	Coded / Non-coded allele	UKB ever smoker			UKB never smoker			SpM ever smoker			SpM never smoker		
			β	SE	P	β	SE	P	β	SE	P	β	SE	P
FEV_1/FVC	rs9661802(1:6678864)	A/C	0.0267	0.0039	1.16×10^{-11}	0.0242	0.0037	4.80×10^{-11}	0.0151	0.0080	6.05×10^{-2}	0.0207	0.0086	1.56×10^{-2}
FEV_1/FVC	rs9435733(1:17308254)	T/C	0.0381	0.0037	1.10×10^{-24}	0.0374	0.0035	6.19×10^{-27}	0.0427	0.0075	1.49×10^{-8}	0.0396	0.0079	6.06×10^{-7}
FEV_1	rs12737805(1:22612690)	A/G	0.0116	0.0045	1.05×10^{-2}	0.0272	0.0042	1.33×10^{-10}	0.0140	0.0089	1.14×10^{-1}	0.0328	0.0094	5.15×10^{-4}
FVC	rs9438626(1:26775367)	G/C	-0.0142	0.0046	1.91×10^{-3}	-0.0179	0.0043	2.57×10^{-5}	-0.0220	0.0091	1.58×10^{-2}	-0.0225	0.0095	1.75×10^{-2}
FEV_1	rs12096239(1:26796922)	G/C	0.0215	0.0042	4.08×10^{-7}	0.0153	0.0040	1.29×10^{-4}	0.0204	0.0085	1.62×10^{-2}	0.0201	0.0091	2.66×10^{-2}
FEV_1/FVC	rs755249(1:39995074)	T/C	-0.0266	0.0044	1.56×10^{-9}	-0.0221	0.0041	7.47×10^{-8}	-0.0304	0.0088	5.65×10^{-4}	-0.0172	0.0093	6.41×10^{-2}
FEV_1/FVC	rs1416685(1:51243374)	G/C	-0.0206	0.0038	5.53×10^{-8}	-0.0214	0.0035	1.45×10^{-9}	-0.0116	0.0076	1.29×10^{-1}	-0.0210	0.0081	9.88×10^{-3}
FEV_1/FVC	rs72673461(1:60966772)	T/G	0.0478	0.0086	2.52×10^{-8}	0.0654	0.0080	2.63×10^{-16}	0.0301	0.0186	1.06×10^{-1}	0.0429	0.0195	2.81×10^{-2}
FEV_1/FVC	rs9661687(1:78387270)	C/T	0.0276	0.0054	4.03×10^{-7}	0.0249	0.0051	1.15×10^{-6}	0.0305	0.0113	6.75×10^{-3}	0.0258	0.0117	2.72×10^{-2}
FEV_1/FVC	rs1192415(1:92077097)	G/A	-0.0438	0.0047	2.53×10^{-20}	-0.0494	0.0044	7.49×10^{-29}	-0.0278	0.0102	6.39×10^{-3}	-0.0307	0.0106	3.66×10^{-3}
FEV_1/FVC	rs10874851(1:92106637)	A/C	-0.0121	0.0037	1.17×10^{-3}	-0.0178	0.0035	2.89×10^{-7}	-0.0202	0.0079	1.07×10^{-2}	-0.0103	0.0082	2.09×10^{-1}
FEV_1/FVC	rs11165787(1:92381483)	A/G	0.0266	0.0040	4.08×10^{-11}	0.0283	0.0038	4.15×10^{-14}	0.0032	0.0085	7.04×10^{-1}	0.0201	0.0089	2.35×10^{-2}
FEV_1/FVC	rs9970286(1:111737398)	G/A	-0.0206	0.0040	2.05×10^{-7}	-0.0270	0.0037	3.10×10^{-13}	-0.0292	0.0085	5.46×10^{-4}	-0.0274	0.0089	2.01×10^{-3}
FVC	rs35043843(1:118911295)	T/G	-0.0263	0.0044	1.63×10^{-9}	-0.0245	0.0040	1.48×10^{-9}	-0.0138	0.0094	1.40×10^{-1}	-0.0218	0.0096	2.29×10^{-2}
PEF	rs11205354(1:150249101)	C/A	-0.0141	0.0038	1.98×10^{-4}	-0.0173	0.0035	8.53×10^{-7}	-0.0328	0.0128	1.02×10^{-2}	-0.0308	0.0138	2.57×10^{-2}
FVC	rs878471(1:150547747)	G/A	0.0238	0.0038	3.00×10^{-10}	0.0307	0.0035	2.87×10^{-18}	0.0360	0.0081	8.96×10^{-6}	0.0498	0.0084	2.84×10^{-9}
FEV_1/FVC	rs141942982(1:155137395)	G/T	0.0246	0.0060	3.87×10^{-5}	0.0515	0.0056	2.99×10^{-20}	0.0162	0.0134	2.26×10^{-1}	0.0368	0.0145	1.10×10^{-2}
FEV_1	rs4651005(1:178719306)	C/T	-0.0150	0.0040	1.60×10^{-4}	-0.0227	0.0037	9.54×10^{-10}	-0.0166	0.0084	4.82×10^{-2}	-0.0222	0.0086	1.04×10^{-2}
FVC	rs2146098(1:186090370)	A/G	-0.0163	0.0039	3.05×10^{-5}	-0.0180	0.0036	7.34×10^{-7}	-0.0071	0.0083	3.93×10^{-1}	-0.0342	0.0082	2.81×10^{-5}
FEV_1/FVC	rs17531405(1:186113852)	G/C	-0.0303	0.0049	4.11×10^{-10}	-0.0319	0.0046	2.53×10^{-12}	-0.0330	0.0104	1.57×10^{-3}	-0.0184	0.0108	8.91×10^{-2}
FEV_1/FVC	rs10919604(1:198898157)	A/G	0.0154	0.0038	5.00×10^{-5}	0.0210	0.0036	3.40×10^{-9}	0.0209	0.0081	9.73×10^{-3}	0.0218	0.0084	9.30×10^{-3}
FVC	rs2816992(1:200069216)	A/G	-0.0172	0.0038	5.92×10^{-6}	-0.0171	0.0035	1.49×10^{-6}	-0.0126	0.0081	1.21×10^{-1}	-0.0163	0.0083	5.06×10^{-2}
FEV_1/FVC	rs4309038(1:201884647)	G/C	-0.0177	0.0038	2.23×10^{-6}	-0.0121	0.0035	5.36×10^{-4}	-0.0191	0.0080	1.68×10^{-2}	-0.0177	0.0083	3.27×10^{-2}
PEF	rs1008833(1:204426295)	A/G	-0.0234	0.0053	1.10×10^{-5}	-0.0396	0.0050	1.41×10^{-15}	-0.0362	0.0178	4.19×10^{-2}	-0.0407	0.0191	3.29×10^{-2}
FVC	rs556648(1:215120596)	A/G	0.0159	0.0045	4.75×10^{-4}	0.0161	0.0042	1.43×10^{-4}	0.0166	0.0098	9.05×10^{-2}	0.0106	0.0102	2.95×10^{-1}
FEV_1/FVC	rs2799098(1:218521609)	G/A	0.0354	0.0049	3.54×10^{-13}	0.0214	0.0045	2.28×10^{-6}	0.0224	0.0105	3.30×10^{-2}	0.0242	0.0107	2.41×10^{-2}
PEF	rs6604614(1:218631452)	C/G	-0.0180	0.0041	1.40×10^{-5}	-0.0173	0.0039	7.93×10^{-6}	0.0006	0.0140	9.63×10^{-1}	0.0111	0.0149	4.56×10^{-1}

Trait	SNP (Chr:Pos)	Coded / Non-coded allele	UKB ever smoker			UKB never smoker			SpM ever smoker			SpM never smoker		
			β	SE	P	β	SE	P	β	SE	P	β	SE	P
FEV_1	rs28613267(1:218855029)	C/G	0.0157	0.0037	2.42×10^{-5}	0.0189	0.0035	4.56×10^{-8}	0.0120	0.0081	1.41×10^{-1}	0.0200	0.0084	1.81×10^{-2}
FEV_1/FVC	rs75128958(1:219483218)	G/A	0.0516	0.0070	1.63×10^{-13}	0.0370	0.0065	1.52×10^{-8}	0.0465	0.0148	1.64×10^{-3}	0.0570	0.0151	1.55×10^{-4}
FEV_1/FVC	rs1338227(1:219853742)	G/T	-0.0252	0.0038	2.60×10^{-11}	-0.0266	0.0035	4.86×10^{-14}	-0.0213	0.0080	7.48×10^{-3}	-0.0249	0.0083	2.76×10^{-3}
FVC	rs17009288(1:221204299)	A/C	-0.0261	0.0041	2.45×10^{-10}	-0.0263	0.0038	6.79×10^{-12}	-0.0267	0.0088	2.37×10^{-3}	-0.0138	0.0091	1.27×10^{-1}
FVC	rs12757436(1:221631938)	A/G	0.0252	0.0040	2.09×10^{-10}	0.0132	0.0037	3.36×10^{-4}	0.0123	0.0085	1.47×10^{-1}	0.0034	0.0088	7.00×10^{-1}
PEF	rs2355237(1:239857524)	A/G	0.0289	0.0037	1.20×10^{-14}	0.0292	0.0035	4.72×10^{-17}	0.0331	0.0127	8.95×10^{-3}	0.0307	0.0135	2.26×10^{-2}
FEV_1/FVC	rs2544536(2:15906854)	T/C	-0.0189	0.0037	3.70×10^{-7}	-0.0277	0.0035	1.64×10^{-15}	-0.0149	0.0074	4.43×10^{-2}	-0.0319	0.0079	5.17×10^{-5}
FEV_1/FVC	rs55884799(2:18287623)	T/C	-0.0367	0.0049	7.62×10^{-14}	-0.0471	0.0046	1.12×10^{-24}	-0.0293	0.0101	3.61×10^{-3}	-0.0448	0.0107	2.70×10^{-5}
FVC	rs6751968(2:18570024)	C/A	-0.0225	0.0049	4.44×10^{-6}	-0.0255	0.0045	2.10×10^{-8}	-0.0367	0.0097	1.44×10^{-4}	-0.0234	0.0101	2.07×10^{-2}
FVC	rs13430465(2:18702313)	C/T	-0.0398	0.0070	1.17×10^{-8}	-0.0369	0.0065	1.22×10^{-8}	-0.0361	0.0139	9.28×10^{-3}	-0.0467	0.0143	1.10×10^{-3}
FVC	rs13009582(2:24018480)	G/A	-0.0214	0.0038	1.19×10^{-8}	-0.0123	0.0035	4.48×10^{-4}	-0.0209	0.0076	5.95×10^{-3}	-0.0110	0.0080	1.71×10^{-1}
FVC	rs732990(2:26842146)	C/G	-0.0125	0.0038	9.00×10^{-4}	-0.0131	0.0035	1.78×10^{-4}	-0.0226	0.0076	2.78×10^{-3}	-0.0202	0.0079	1.05×10^{-2}
FVC	rs4952564(2:42243850)	A/G	-0.0165	0.0040	3.78×10^{-5}	-0.0182	0.0037	1.05×10^{-6}	-0.0177	0.0081	2.99×10^{-2}	-0.0187	0.0086	2.89×10^{-2}
FVC	rs3791679(2:56096892)	A/G	0.0377	0.0045	2.61×10^{-17}	0.0309	0.0041	9.04×10^{-14}	0.0379	0.0094	5.15×10^{-5}	0.0228	0.0096	1.74×10^{-2}
FEV_1/FVC	rs12470864(2:102926362)	G/A	0.0157	0.0038	3.82×10^{-5}	0.0250	0.0036	2.72×10^{-12}	0.0116	0.0082	1.60×10^{-1}	0.0226	0.0086	8.39×10^{-3}
FVC	rs62168891(2:135672187)	C/T	-0.0199	0.0038	2.01×10^{-7}	-0.0143	0.0036	5.96×10^{-5}	-0.0236	0.0083	4.28×10^{-3}	-0.0265	0.0086	2.01×10^{-3}
FEV_1/FVC	rs1406225(2:145797829)	G/T	0.0249	0.0041	1.56×10^{-9}	0.0125	0.0039	1.22×10^{-3}	0.0253	0.0087	3.65×10^{-3}	0.0180	0.0090	4.58×10^{-2}
FEV_1	rs72902177(2:157016257)	C/T	0.0319	0.0054	4.23×10^{-9}	0.0361	0.0051	1.08×10^{-12}	0.0231	0.0119	5.25×10^{-2}	0.0370	0.0124	2.88×10^{-3}
FEV_1	rs7424771(2:161276378)	G/A	0.0190	0.0037	3.47×10^{-7}	0.0150	0.0035	1.62×10^{-5}	0.0238	0.0080	2.89×10^{-3}	0.0204	0.0083	1.34×10^{-2}
FEV_1	rs2304340(2:179260382)	A/G	-0.0173	0.0038	4.66×10^{-6}	-0.0082	0.0035	2.01×10^{-2}	-0.0279	0.0081	5.51×10^{-4}	-0.0128	0.0084	1.25×10^{-1}
FEV_1/FVC	rs2084448(2:187530520)	T/C	0.0171	0.0041	2.66×10^{-5}	0.0195	0.0038	3.09×10^{-7}	0.0093	0.0087	2.86×10^{-1}	0.0257	0.0091	4.63×10^{-3}
FVC	rs1249096(2:199723365)	G/A	-0.0210	0.0038	3.07×10^{-8}	-0.0218	0.0035	5.30×10^{-10}	-0.0262	0.0082	1.31×10^{-3}	-0.0201	0.0084	1.65×10^{-2}
FEV_1/FVC	rs985256(2:201208692)	A/C	0.0179	0.0045	7.52×10^{-5}	0.0172	0.0042	4.34×10^{-5}	0.0265	0.0095	5.10×10^{-3}	0.0231	0.0098	1.85×10^{-2}
FVC	rs12997625(2:202970250)	C/T	0.0172	0.0038	5.05×10^{-6}	0.0099	0.0035	4.67×10^{-3}	0.0304	0.0081	1.70×10^{-4}	0.0172	0.0083	3.72×10^{-2}
FEV_1/FVC	rs6435952(2:217614730)	A/T	0.0182	0.0053	5.46×10^{-4}	0.0323	0.0049	3.73×10^{-11}	0.0256	0.0109	1.86×10^{-2}	0.0416	0.0111	1.70×10^{-4}
FEV_1	rs4294980(2:218604356)	G/A	-0.0126	0.0045	5.40×10^{-3}	-0.0175	0.0042	3.37×10^{-5}	-0.0242	0.0100	1.52×10^{-2}	-0.0248	0.0103	1.61×10^{-2}
FEV_1	rs2571445(2:218683154)	A/G	-0.0305	0.0038	7.57×10^{-16}	-0.0261	0.0035	1.61×10^{-13}	-0.0321	0.0082	8.76×10^{-5}	-0.0280	0.0084	8.63×10^{-4}
FVC	rs4674407(2:220382700)	C/T	0.0067	0.0038	7.68×10^{-2}	0.0171	0.0035	9.79×10^{-7}	0.0083	0.0081	3.06×10^{-1}	0.0301	0.0084	3.32×10^{-4}
FEV_1/FVC	rs62201738(2:229502197)	A/C	-0.0813	0.0069	4.69×10^{-32}	-0.0890	0.0065	2.58×10^{-43}	-0.0400	0.0152	8.74×10^{-3}	-0.0269	0.0159	9.04×10^{-2}
FEV_1	rs6710301(2:239441308)	A/C	0.0281	0.0052	7.10×10^{-8}	0.0267	0.0049	4.39×10^{-8}	0.0096	0.0112	3.92×10^{-1}	0.0201	0.0114	7.77×10^{-2}
FVC	rs6431620(2:239604970)	T/G	0.0233	0.0046	4.41×10^{-7}	0.0132	0.0043	1.98×10^{-3}	0.0186	0.0097	5.46×10^{-2}	0.0258	0.0099	9.17×10^{-3}
FEV_1/FVC	rs4308141(2:239881309)	C/G	-0.0552	0.0047	3.20×10^{-32}	-0.0467	0.0044	7.81×10^{-27}	-0.0644	0.0098	6.35×10^{-11}	-0.0332	0.0102	1.17×10^{-3}

Trait	SNP (Chr:Pos)	Coded / Non-coded allele	UKB ever smoker			UKB never smoker			SpM ever smoker			SpM never smoker		
			β	SE	P	β	SE	P	β	SE	P	β	SE	P
FVC	rs6437219(2:241844033)	C/T	-0.0172	0.0040	1.34×10^{-5}	-0.0189	0.0037	2.30×10^{-7}	-0.0234	0.0088	8.19×10^{-3}	-0.0230	0.0090	1.10×10^{-2}
FVC	rs6733504(2:242495953)	A/G	0.0210	0.0038	2.13×10^{-8}	0.0191	0.0035	4.57×10^{-8}	0.0241	0.0082	3.44×10^{-3}	0.0196	0.0085	2.09×10^{-2}
FEV_1	rs2974389(3:13787641)	A/G	0.0123	0.0038	1.01×10^{-3}	0.0141	0.0035	5.46×10^{-5}	0.0220	0.0075	3.42×10^{-3}	0.0291	0.0079	2.49×10^{-4}
FVC	rs73048404(3:25179533)	T/G	0.0122	0.0053	2.14×10^{-2}	0.0210	0.0049	1.68×10^{-5}	0.0443	0.0108	3.76×10^{-5}	0.0408	0.0112	2.82×10^{-4}
FEV_1/FVC	rs1529672(3:25520582)	C/A	-0.0435	0.0049	1.35×10^{-18}	-0.0444	0.0046	5.70×10^{-22}	-0.0415	0.0098	2.48×10^{-5}	-0.0316	0.0104	2.45×10^{-3}
FEV_1/FVC	rs17666332(3:29469675)	T/G	0.0336	0.0042	6.98×10^{-16}	0.0249	0.0039	1.53×10^{-10}	0.0193	0.0084	2.26×10^{-2}	0.0034	0.0091	7.04×10^{-1}
FEV_1/FVC	rs12715478(3:55152319)	A/G	0.0293	0.0038	1.21×10^{-14}	0.0263	0.0035	1.17×10^{-13}	0.0059	0.0082	4.72×10^{-1}	0.0144	0.0086	9.39×10^{-2}
FEV_1	rs6445932(3:57879611)	T/G	-0.0304	0.0043	1.67×10^{-12}	-0.0243	0.0040	1.37×10^{-9}	-0.0255	0.0091	5.28×10^{-3}	-0.0367	0.0094	9.49×10^{-5}
FEV_1	rs4132748(3:67455803)	T/C	-0.0191	0.0040	2.54×10^{-6}	-0.0202	0.0038	9.20×10^{-8}	-0.0210	0.0087	1.60×10^{-2}	-0.0107	0.0091	2.40×10^{-1}
FVC	rs35480566(3:71583177)	A/G	-0.0250	0.0038	3.69×10^{-11}	-0.0207	0.0035	3.36×10^{-9}	-0.0189	0.0081	1.92×10^{-2}	-0.0167	0.0083	4.33×10^{-2}
FEV_1/FVC	rs586936(3:73862616)	G/A	0.0131	0.0038	6.15×10^{-4}	0.0221	0.0036	5.65×10^{-10}	0.0214	0.0083	9.63×10^{-3}	0.0199	0.0086	1.99×10^{-2}
FVC	rs12497779(3:98822050)	G/T	0.0359	0.0044	5.11×10^{-16}	0.0293	0.0041	1.00×10^{-12}	0.0304	0.0094	1.23×10^{-3}	0.0184	0.0097	5.72×10^{-2}
FVC	rs1610265(3:99420192)	C/T	0.0468	0.0070	2.17×10^{-11}	0.0316	0.0065	1.25×10^{-6}	0.0495	0.0152	1.11×10^{-3}	0.0352	0.0155	2.31×10^{-2}
FEV_1/FVC	rs2999090(3:127931340)	A/G	-0.0416	0.0057	4.23×10^{-13}	-0.0502	0.0054	6.33×10^{-21}	-0.0275	0.0123	2.53×10^{-2}	-0.0170	0.0127	1.81×10^{-1}
FVC	rs12634907(3:158226886)	A/G	0.0263	0.0039	2.06×10^{-11}	0.0270	0.0036	1.22×10^{-13}	0.0172	0.0084	4.07×10^{-2}	0.0273	0.0087	1.63×10^{-3}
FEV_1/FVC	rs1799807(3:165548529)	T/C	0.0508	0.0134	1.43×10^{-4}	0.0537	0.0126	1.99×10^{-5}	0.0898	0.0321	5.16×10^{-3}	0.0819	0.0322	1.11×10^{-2}
FEV_1	rs879394(3:168709843)	G/T	0.0302	0.0044	5.30×10^{-12}	0.0238	0.0041	6.95×10^{-9}	0.0146	0.0094	1.20×10^{-1}	0.0394	0.0097	4.52×10^{-5}
FEV_1	rs78101726(3:169295436)	A/G	0.0305	0.0052	3.20×10^{-9}	0.0327	0.0048	9.39×10^{-12}	0.0379	0.0110	6.04×10^{-4}	0.0438	0.0114	1.14×10^{-4}
FEV_1	rs6780171(3:185503456)	T/A	0.0151	0.0040	1.62×10^{-4}	0.0176	0.0037	2.50×10^{-6}	0.0272	0.0086	1.52×10^{-3}	0.0241	0.0089	6.72×10^{-3}
FEV_1/FVC	rs62289340(4:7879027)	C/T	-0.0129	0.0038	5.80×10^{-4}	-0.0188	0.0035	7.89×10^{-8}	-0.0280	0.0075	1.96×10^{-4}	-0.0074	0.0079	3.53×10^{-1}
FEV_1	rs12331869(4:56012149)	A/G	-0.0120	0.0048	1.27×10^{-2}	-0.0169	0.0045	1.82×10^{-4}	-0.0209	0.0103	4.20×10^{-2}	-0.0289	0.0107	6.93×10^{-3}
FEV_1/FVC	rs62316310(4:75676529)	G/A	-0.0245	0.0042	7.26×10^{-9}	-0.0349	0.0040	1.11×10^{-18}	-0.0125	0.0092	1.73×10^{-1}	-0.0240	0.0096	1.22×10^{-2}
FEV_1/FVC	rs11098196(4:79403952)	G/T	0.0154	0.0037	3.46×10^{-5}	0.0242	0.0035	3.31×10^{-12}	0.0174	0.0079	2.82×10^{-2}	0.0140	0.0083	9.16×10^{-2}
FEV_1/FVC	rs2609279(4:89855495)	T/C	0.0627	0.0045	3.06×10-43	0.0445	0.0042	1.05×10^{-25}	0.0405	0.0095	2.14×10^{-5}	0.0496	0.0097	2.94×10^{-7}
FEV_1/FVC	rs2869966(4:89869078)	T/C	-0.0467	0.0038	4.47×10^{-35}	-0.0401	0.0035	7.21×10^{-30}	-0.0289	0.0080	3.03×10^{-4}	-0.0295	0.0083	4.03×10^{-4}
FEV_1/FVC	rs6533183(4:106133184)	T/C	-0.0301	0.0039	1.41×10^{-14}	-0.0375	0.0037	1.31×10^{-24}	-0.0186	0.0082	2.42×10^{-2}	-0.0042	0.0085	6.21×10^{-1}
FEV_1	rs11722225(4:106766430)	T/C	-0.0651	0.0074	1.45×10^{-18}	-0.0771	0.0069	8.57×10^{-29}	-0.0966	0.0162	2.65×10^{-9}	-0.0622	0.0169	2.33×10^{-4}
FEV_1/FVC	rs34712979(4:106819053)	G/A	0.0699	0.0042	6.46×10^{-61}	0.0747	0.0040	1.68×10^{-79}	0.0298	0.0102	3.56×10^{-3}	0.0663	0.0104	1.86×10^{-10}
FVC	rs13109426(4:145330628)	G/A	0.0238	0.0038	3.86×10^{-10}	0.0217	0.0035	7.62×10^{-10}	0.0222	0.0081	6.20×10^{-3}	0.0261	0.0083	1.65×10^{-3}
PEF	rs13116999(4:145442364)	G/A	-0.0634	0.0037	2.01×10^{-64}	-0.0718	0.0035	6.71×10^{-95}	-0.0453	0.0126	3.32×10^{-4}	-0.0264	0.0136	5.28×10^{-2}
FEV_1/FVC	rs13141641(4:145506456)	T/C	-0.0736	0.0038	1.06×10^{-83}	-0.0761	0.0035	2.85×10^{-102}	-0.0529	0.0081	5.01×10^{-11}	-0.0426	0.0084	3.31×10^{-7}
PEF	rs2353940(4:145740898)	T/C	0.0365	0.0043	1.93×10^{-17}	0.0447	0.0040	5.29×10^{-29}	0.0362	0.0148	1.44×10^{-2}	0.0075	0.0158	6.36×10^{-1}
FEV_1	rs11739847(5:609661)	G/A	0.0221	0.0046	1.73×10^{-6}	0.0173	0.0043	6.02×10^{-5}	0.0185	0.0094	4.97×10^{-2}	0.0310	0.0100	1.96×10^{-3}

Trait	SNP (Chr:Pos)	Coded / Non-coded allele	UKB ever smoker			UKB never smoker			SpM ever smoker			SpM never smoker		
			β	SE	P	β	SE	P	β	SE	P	β	SE	P
FVC	rs268717(5:33352738)	T/C	-0.0363	0.0065	1.85×10^{-8}	-0.0308	0.0060	2.77×10^{-7}	-0.0504	0.0131	1.20×10^{-4}	-0.0267	0.0136	4.97×10^{-2}
FEV_1	rs4866846(5:43976162)	A/G	0.0275	0.0052	1.26×10^{-7}	0.0262	0.0049	7.30×10^{-8}	0.0393	0.0105	1.86×10^{-4}	0.0186	0.0113	1.00×10^{-1}
FVC	rs6859730(5:44367221)	A/T	0.0172	0.0040	1.52×10^{-5}	0.0242	0.0037	6.55×10^{-11}	0.0280	0.0080	4.50×10^{-4}	0.0068	0.0084	4.20×10^{-1}
FEV_1/FVC	rs12522114(5:52187038)	A/C	-0.0384	0.0042	1.10×10^{-19}	-0.0438	0.0039	1.38×10^{-28}	-0.0205	0.0091	2.39×10^{-2}	-0.0177	0.0095	6.15×10^{-2}
FVC	rs2441026(5:53444498)	C/T	-0.0181	0.0038	1.53×10^{-6}	-0.0183	0.0035	1.59×10^{-7}	-0.0178	0.0080	2.58×10^{-2}	-0.0216	0.0082	8.84×10^{-3}
FVC	rs425102(5:77396400)	T/G	0.0203	0.0044	3.46×10^{-6}	0.0226	0.0040	2.30×10^{-8}	0.0096	0.0094	3.09×10^{-1}	0.0199	0.0098	4.16×10^{-2}
FEV_1/FVC	rs987068(5:95025146)	G/C	0.0325	0.0040	4.58×10^{-16}	0.0306	0.0037	2.55×10^{-16}	0.0188	0.0087	3.17×10^{-2}	0.0264	0.0091	3.52×10^{-3}
FEV_1/FVC	rs10059661(5:121410529)	C/G	-0.0257	0.0049	1.67×10^{-7}	-0.0388	0.0046	2.55×10^{-17}	-0.0335	0.0108	1.86×10^{-3}	-0.0234	0.0113	3.73×10^{-2}
FEV_1/FVC	rs17163397(5:128767384)	A/G	-0.0336	0.0056	2.54×10^{-9}	-0.0297	0.0053	1.64×10^{-8}	-0.0155	0.0124	2.12×10^{-1}	-0.0296	0.0130	2.29×10^{-2}
FVC	rs3843503(5:131466629)	T/A	0.0199	0.0038	1.60×10^{-7}	0.0182	0.0035	2.48×10^{-7}	0.0299	0.0083	3.21×10^{-4}	0.0087	0.0087	3.18×10^{-1}
FEV_1/FVC	rs7733410(5:147856522)	G/A	-0.0579	0.0037	5.18×10^{-54}	-0.0502	0.0035	8.44×10^{-47}	-0.0422	0.0084	5.79×10^{-7}	-0.0250	0.0088	4.40×10^{-3}
FEV_1	rs1800888(5:148206885)	C/T	0.0952	0.0152	3.41×10^{-10}	0.0753	0.0144	1.62×10^{-7}	0.0822	0.0358	2.18×10^{-2}	0.1074	0.0383	5.01×10^{-3}
FEV_1	rs11952673(5:148652302)	T/G	-0.0246	0.0038	1.06×10^{-10}	-0.0189	0.0036	1.02×10^{-7}	-0.0122	0.0082	1.38×10^{-1}	0.0026	0.0086	7.61×10^{-1}
FEV_1/FVC	rs11134766(5:156908317)	T/C	-0.0677	0.0076	8.36×10^{-19}	-0.0695	0.0071	1.51×10^{-22}	-0.0533	0.0159	7.91×10^{-4}	-0.0468	0.0168	5.35×10^{-3}
FEV_1/FVC	rs11134789(5:156944199)	C/A	0.0352	0.0039	3.34×10^{-19}	0.0516	0.0037	2.94×10-45	0.0311	0.0084	2.02×10^{-4}	0.0326	0.0088	1.93×10^{-4}
FEV_1/FVC	rs10059996(5:170901463)	T/G	-0.0291	0.0040	2.43×10^{-13}	-0.0409	0.0037	3.73×10^{-28}	-0.0297	0.0090	9.46×10^{-4}	-0.0366	0.0093	8.27×10^{-5}
FEV_1/FVC	rs79898473(5:179598771)	T/C	-0.0332	0.0040	5.12×10^{-17}	-0.0326	0.0037	1.77×10^{-18}	-0.0165	0.0087	5.87×10^{-2}	-0.0315	0.0092	6.44×10^{-4}
FEV_1/FVC	rs1294417(6:6741932)	T/C	-0.0275	0.0037	1.75×10^{-13}	-0.0377	0.0035	3.77×10^{-27}	-0.0169	0.0075	2.43×10^{-2}	-0.0362	0.0079	4.61×10^{-6}
FEV_1/FVC	rs2076295(6:7563232)	T/G	-0.0212	0.0037	1.60×10^{-8}	-0.0251	0.0035	7.83×10^{-13}	-0.0206	0.0075	6.17×10^{-3}	-0.0276	0.0079	5.07×10^{-4}
FVC	rs12198986(6:7720059)	G/A	0.0247	0.0037	3.65×10^{-11}	0.0238	0.0035	7.73×10^{-12}	0.0074	0.0075	3.19×10^{-1}	0.0321	0.0078	4.30×10^{-5}
FVC	rs10498672(6:7797840)	C/G	0.0319	0.0049	7.02×10^{-11}	0.0367	0.0046	7.66×10^{-16}	0.0307	0.0098	1.66×10^{-3}	0.0454	0.0102	8.34×10^{-6}
FEV_1/FVC	rs13198081(6:22017543)	G/C	-0.0275	0.0039	1.71×10^{-12}	-0.0366	0.0036	7.68×10^{-24}	-0.0193	0.0077	1.29×10^{-2}	-0.0176	0.0082	3.19×10^{-2}
PEF	rs7752448(6:28301099)	A/G	0.0554	0.0056	7.32×10^{-23}	0.0636	0.0052	7.80×10^{-35}	0.0252	0.0195	1.96×10^{-1}	-0.0044	0.0210	8.33×10^{-1}
FEV_1/FVC	rs2070600(6:32151443)	T/C	0.1407	0.0076	2.27×10^{-76}	0.1610	0.0071	8.68×10^{-115}	0.1254	0.0181	4.61×10^{-12}	0.1246	0.0188	3.63×10^{-11}
FEV_1/FVC	rs9274247(6:32631295)	A/G	-0.0442	0.0044	7.06×10^{-24}	-0.0530	0.0041	2.19×10^{-38}	-0.0299	0.0120	1.25×10^{-2}	-0.0275	0.0124	2.66×10^{-2}
FVC	rs9689096(6:34188892)	A/C	-0.0366	0.0077	1.94×10^{-6}	-0.0328	0.0072	6.04×10^{-6}	-0.0024	0.0165	8.83×10^{-1}	-0.0456	0.0181	1.16×10^{-2}
FVC	rs9357446(6:44447598)	G/A	0.0120	0.0037	1.37×10^{-3}	0.0133	0.0035	1.30×10^{-4}	0.0256	0.0075	6.33×10^{-4}	0.0120	0.0079	1.25×10^{-1}
FEV_1/FVC	rs12202314(6:45530471)	T/C	-0.0166	0.0040	3.42×10^{-5}	-0.0242	0.0037	8.99×10^{-11}	-0.0189	0.0080	1.83×10^{-2}	-0.0231	0.0085	6.71×10^{-3}
FVC	rs9472541(6:45622748)	T/A	0.0151	0.0042	2.87×10^{-4}	0.0120	0.0039	1.88×10^{-3}	0.0223	0.0083	7.07×10^{-3}	0.0161	0.0086	6.22×10^{-2}
FEV_1	rs2894837(6:56336406)	A/G	0.0175	0.0039	5.94×10^{-6}	0.0161	0.0036	8.22×10^{-6}	0.0211	0.0084	1.21×10^{-2}	0.0158	0.0087	6.91×10^{-2}
FEV_1/FVC	rs13206405(6:73663814)	C/A	-0.0347	0.0046	6.74×10^{-14}	-0.0411	0.0043	2.12×10^{-21}	-0.0327	0.0097	8.01×10^{-4}	-0.0208	0.0100	3.82×10^{-2}
FEV_1/FVC	rs2798641(6:109268050)	C/T	0.0417	0.0048	4.28×10^{-18}	0.0553	0.0045	2.87×10^{-35}	0.0234	0.0103	2.32×10^{-2}	0.0267	0.0108	1.37×10^{-2}

Trait	SNP (Chr:Pos)	Coded / Non-coded allele	UKB ever smoker			UKB never smoker			SpM ever smoker			SpM never smoker		
			β	SE	P	β	SE	P	β	SE	P	β	SE	P
FVC	rs6918725(6:126990392)	T/G	-0.0186	0.0038	6.83×10^{-7}	-0.0206	0.0035	3.33×10^{-9}	-0.0104	0.0080	1.96×10^{-1}	-0.0284	0.0083	6.41×10^{-4}
FEV_1	rs2627237(6:134339265)	A/G	0.0133	0.0038	3.97×10^{-4}	0.0134	0.0035	1.62×10^{-4}	0.0278	0.0081	6.12×10^{-4}	0.0148	0.0084	7.93×10^{-2}
FEV_1	rs1102077(6:140271357)	A/C	0.0109	0.0043	1.22×10^{-2}	0.0302	0.0041	8.93×10^{-14}	0.0093	0.0095	3.24×10^{-1}	0.0278	0.0097	4.25×10^{-3}
FEV_1	rs9385988(6:142560957)	A/G	-0.0313	0.0041	4.21×10^{-14}	-0.0258	0.0039	2.84×10^{-11}	-0.0305	0.0087	4.72×10^{-4}	-0.0252	0.0090	5.33×10^{-3}
FEV_1/FVC	rs17280293(6:142688969)	A/G	-0.1754	0.0114	3.23×10^{-53}	-0.1970	0.0105	1.47×10^{-78}	-0.1743	0.0267	6.91×10^{-11}	-0.1168	0.0273	1.84×10^{-5}
FEV_1/FVC	rs7753012(6:142745883)	T/G	-0.0742	0.0040	4.44×10^{-75}	-0.0818	0.0038	2.38×10^{-104}	-0.0449	0.0087	2.12×10^{-7}	-0.0395	0.0091	1.28×10^{-5}
FEV_1/FVC	rs4318980(7:7256490)	A/G	-0.0190	0.0038	5.54×10^{-7}	-0.0188	0.0035	1.12×10^{-7}	-0.0172	0.0076	2.30×10^{-2}	-0.0127	0.0080	1.11×10^{-1}
FVC	rs4721442(7:15506007)	T/G	0.0202	0.0051	7.28×10^{-5}	0.0219	0.0047	3.62×10^{-6}	0.0261	0.0098	7.41×10^{-3}	0.0110	0.0100	2.69×10^{-1}
FEV_1/FVC	rs4721457(7:15872324)	T/C	0.0180	0.0053	6.28×10^{-4}	0.0223	0.0049	4.87×10^{-6}	0.0172	0.0102	9.30×10^{-2}	0.0353	0.0106	8.55×10^{-4}
FEV_1	rs559233(7:26848830)	T/C	0.0146	0.0037	9.56×10^{-5}	0.0145	0.0035	3.24×10^{-5}	0.0237	0.0076	1.96×10^{-3}	0.0287	0.0081	3.95×10^{-4}
FVC	rs62454414(7:27182329)	T/G	0.0161	0.0055	3.74×10^{-3}	0.0193	0.0052	1.75×10^{-4}	0.0324	0.0109	2.95×10^{-3}	0.0094	0.0110	3.96×10^{-1}
FEV_1	rs1513272(7:28200097)	C/T	-0.0177	0.0037	1.86×10^{-6}	-0.0235	0.0035	1.21×10^{-11}	-0.0299	0.0074	5.74×10^{-5}	-0.0174	0.0078	2.62×10^{-2}
FVC	rs17232687(7:46448518)	T/C	-0.0199	0.0037	9.89×10^{-8}	-0.0173	0.0035	6.64×10^{-7}	-0.0260	0.0080	1.10×10^{-3}	-0.0094	0.0082	2.52×10^{-1}
FEV_1	rs12707691(7:84569510)	C/G	-0.0158	0.0039	5.60×10^{-5}	-0.0263	0.0037	6.94×10^{-13}	-0.0242	0.0085	4.32×10^{-3}	-0.0183	0.0089	3.83×10^{-2}
FEV_1/FVC	rs2261360(7:99692993)	T/G	0.0227	0.0044	2.54×10^{-7}	0.0269	0.0041	5.89×10^{-11}	0.0258	0.0094	6.15×10^{-3}	0.0157	0.0097	1.04×10^{-1}
FEV_1/FVC	rs193686(7:116431427)	C/T	0.0115	0.0040	3.86×10^{-3}	0.0232	0.0037	5.28×10^{-10}	0.0172	0.0086	4.51×10^{-2}	0.0286	0.0089	1.35×10^{-3}
FEV_1	rs12698403(7:156127246)	G/A	0.0235	0.0037	3.67×10^{-10}	0.0295	0.0035	3.14×10^{-17}	0.0218	0.0081	6.96×10^{-3}	0.0222	0.0083	7.47×10^{-3}
FEV_1/FVC	rs330939(8:9018590)	T/G	0.0227	0.0039	4.05×10^{-9}	0.0279	0.0036	1.44×10^{-14}	0.0202	0.0078	9.59×10^{-3}	0.0203	0.0083	1.46×10^{-2}
FEV_1	rs4128298(8:11823332)	T/C	-0.0182	0.0041	1.03×10^{-5}	-0.0159	0.0039	3.72×10^{-5}	-0.0352	0.0085	3.41×10^{-5}	0.0000	0.0092	9.98×10^{-1}
FEV_1	rs7465401(8:70367248)	T/C	-0.0192	0.0042	4.45×10^{-6}	-0.0202	0.0039	2.34×10^{-7}	-0.0318	0.0088	3.06×10^{-4}	-0.0209	0.0091	2.13×10^{-2}
FVC	rs7838717(8:145504343)	T/C	-0.0224	0.0040	1.34×10^{-8}	-0.0198	0.0037	6.87×10^{-8}	-0.0328	0.0085	1.07×10^{-4}	-0.0237	0.0087	6.60×10^{-3}
FVC	rs771662(9:1568941)	T/C	-0.0194	0.0039	7.91×10^{-7}	-0.0154	0.0036	2.49×10^{-5}	-0.0119	0.0079	1.32×10^{-1}	-0.0027	0.0084	7.44×10^{-1}
FEV_1/FVC	rs1570203(9:4120648)	G/A	-0.0262	0.0038	3.06×10^{-12}	-0.0261	0.0035	1.08×10^{-13}	-0.0109	0.0075	1.47×10^{-1}	-0.0194	0.0079	1.49×10^{-2}
FEV_1	rs7041139(9:18013733)	C/T	0.0166	0.0040	3.10×10^{-5}	0.0187	0.0037	4.48×10^{-7}	0.0243	0.0080	2.48×10^{-3}	0.0178	0.0085	3.55×10^{-2}
FEV_1/FVC	rs1107677(9:23587027)	T/C	0.0260	0.0037	2.86×10^{-12}	0.0242	0.0035	3.25×10^{-12}	0.0092	0.0077	2.27×10^{-1}	0.0133	0.0081	1.01×10^{-1}
FEV_1/FVC	rs28446321(9:98266855)	T/A	0.0590	0.0065	1.38×10^{-19}	0.0501	0.0061	2.16×10^{-16}	0.0388	0.0135	4.13×10^{-3}	0.0219	0.0141	1.20×10^{-1}
FEV_1/FVC	rs72743974(9:98878881)	A/G	-0.0237	0.0050	2.06×10^{-6}	-0.0186	0.0047	6.59×10^{-5}	-0.0185	0.0104	7.66×10^{-2}	-0.0329	0.0108	2.32×10^{-3}
FEV_1/FVC	rs57649467(9:101632854)	G/A	-0.0188	0.0038	9.22×10^{-7}	-0.0141	0.0036	8.37×10^{-5}	-0.0091	0.0083	2.69×10^{-1}	-0.0247	0.0086	4.22×10^{-3}
FEV_1/FVC	rs1491106(9:109483517)	T/G	0.0252	0.0038	5.38×10^{-11}	0.0302	0.0036	2.73×10^{-17}	0.0147	0.0082	7.23×10^{-2}	0.0193	0.0085	2.23×10^{-2}
FEV_1/FVC	rs10983184(9:119234058)	C/T	-0.0227	0.0039	5.54×10^{-9}	-0.0299	0.0036	1.84×10^{-16}	-0.0247	0.0083	3.09×10^{-3}	-0.0270	0.0087	1.93×10^{-3}
FEV_1	rs967497(9:131943843)	A/G	0.0179	0.0040	7.05×10^{-6}	0.0084	0.0037	2.48×10^{-2}	0.0183	0.0087	3.48×10^{-2}	0.0196	0.0090	3.00×10^{-2}
FVC	rs7024579(9:139100413)	C/T	0.0233	0.0040	7.08×10^{-9}	0.0205	0.0037	4.38×10^{-8}	0.0158	0.0087	6.73×10^{-2}	0.0271	0.0089	2.22×10^{-3}
FVC	rs4073153(9:139259349)	A/G	0.0162	0.0038	1.84×10^{-5}	0.0137	0.0035	1.03×10^{-4}	0.0001	0.0084	9.91×10^{-1}	0.0128	0.0087	1.38×10^{-1}

Trait	SNP (Chr:Pos)	Coded / Non-coded allele	UKB ever smoker			UKB never smoker			SpM ever smoker			SpM never smoker		
			β	SE	P	β	SE	P	β	SE	P	β	SE	P
FEV_1/FVC	rs7090277(10:12278021)	T/A	-0.0364	0.0037	1.15×10^{-22}	-0.0470	0.0035	9.05×10^{-42}	-0.0334	0.0074	6.83×10^{-6}	-0.0313	0.0078	6.66×10^{-5}
PEF	rs7914842(10:30268770)	A/G	0.0157	0.0038	3.28×10^{-5}	0.0194	0.0035	4.07×10^{-8}	0.0143	0.0128	2.65×10^{-1}	0.0092	0.0137	5.02×10^{-1}
FEV_1/FVC	rs1274475(10:34480582)	G/A	-0.0203	0.0039	1.56×10^{-7}	-0.0144	0.0036	6.77×10^{-5}	-0.0124	0.0079	1.13×10^{-1}	-0.0231	0.0084	5.96×10^{-3}
FEV_1	rs7082066(10:64998971)	A/G	0.0233	0.0048	9.83×10^{-7}	0.0212	0.0045	2.04×10^{-6}	0.0231	0.0102	2.29×10^{-2}	0.0077	0.0107	4.71×10^{-1}
FVC	rs10998018(10:69962954)	A/G	-0.0183	0.0037	1.01×10^{-6}	-0.0270	0.0035	8.99×10^{-15}	-0.0168	0.0080	3.68×10^{-2}	-0.0165	0.0083	4.63×10^{-2}
FEV_1	rs7098573(10:75580014)	G/A	0.0234	0.0041	1.29×10^{-8}	0.0245	0.0039	1.94×10^{-10}	0.0303	0.0089	6.34×10^{-4}	0.0263	0.0090	3.65×10^{-3}
PEF	rs60820984(10:75639578)	C/T	0.0226	0.0048	2.54×10^{-6}	0.0176	0.0045	8.67×10^{-5}	0.0377	0.0164	2.19×10^{-2}	0.0434	0.0175	1.30×10^{-2}
FVC	rs1259605(10:77119039)	T/C	-0.0119	0.0043	6.10×10^{-3}	-0.0148	0.0040	2.36×10^{-4}	0.0015	0.0093	8.73×10^{-1}	-0.0042	0.0095	6.55×10^{-1}
FEV_1	rs2637254(10:78312002)	G/A	0.0312	0.0037	3.36×10^{-17}	0.0269	0.0035	8.49×10^{-15}	0.0176	0.0080	2.79×10^{-2}	0.0323	0.0082	8.65×10^{-5}
FEV_1/FVC	rs721917(10:81706324)	A/G	0.0210	0.0038	2.48×10^{-8}	0.0203	0.0035	7.59×10^{-9}	0.0109	0.0081	1.78×10^{-1}	0.0185	0.0084	2.72×10^{-2}
FEV_1	rs11191841(10:105639611)	T/C	-0.0089	0.0037	1.64×10^{-2}	-0.0196	0.0035	1.47×10^{-8}	-0.0274	0.0080	5.83×10^{-4}	-0.0226	0.0082	5.91×10^{-3}
FEV_1/FVC	rs4279944(10:124297637)	T/C	0.0269	0.0053	4.22×10^{-7}	0.0179	0.0050	2.98×10^{-4}	-0.0069	0.0120	5.64×10^{-1}	0.0281	0.0125	2.42×10^{-2}
FEV_1/FVC	rs10836366(11:35308988)	T/C	0.0189	0.0043	1.05×10^{-5}	0.0153	0.0040	1.42×10^{-4}	0.0172	0.0085	4.40×10^{-2}	0.0216	0.0089	1.55×10^{-2}
FVC	rs17596617(11:43690717)	C/T	0.0211	0.0040	1.74×10^{-7}	0.0125	0.0037	8.52×10^{-4}	0.0405	0.0081	5.69×10^{-7}	0.0275	0.0084	1.11×10^{-3}
FEV_1	rs10838435(11:45244903)	C/G	0.0173	0.0053	1.02×10^{-3}	0.0220	0.0050	8.96×10^{-6}	0.0293	0.0104	4.97×10^{-3}	0.0299	0.0110	6.66×10^{-3}
FEV_1	rs71490394(11:62370155)	G/A	-0.0299	0.0038	6.94×10^{-15}	-0.0260	0.0036	4.08×10^{-13}	-0.0332	0.0084	7.39×10^{-5}	-0.0100	0.0087	2.52×10^{-1}
FEV_1/FVC	rs2027761(11:73036179)	C/T	-0.0384	0.0059	6.22×10^{-11}	-0.0362	0.0055	4.67×10^{-11}	-0.0469	0.0124	1.51×10^{-4}	-0.0156	0.0129	2.27×10^{-1}
FEV_1/FVC	rs11234768(11:86448839)	T/C	0.0275	0.0051	8.76×10^{-8}	0.0330	0.0048	5.63×10^{-12}	0.0217	0.0111	4.97×10^{-2}	0.0279	0.0114	1.45×10^{-2}
FEV_1/FVC	rs541601(11:126009500)	T/C	-0.0201	0.0048	2.78×10^{-5}	-0.0284	0.0045	2.62×10^{-10}	-0.0312	0.0100	1.90×10^{-3}	-0.0264	0.0106	1.23×10^{-2}
FVC	rs56196860(12:2908330)	C/A	0.0377	0.0107	4.38×10^{-4}	0.0625	0.0099	3.10×10^{-10}	0.0815	0.0271	2.61×10^{-3}	0.0703	0.0276	1.08×10^{-2}
FEV_1	rs12811814(12:4243749)	T/C	0.0098	0.0038	9.36×10^{-3}	0.0183	0.0035	1.92×10^{-7}	0.0216	0.0078	5.48×10^{-3}	0.0199	0.0082	1.48×10^{-2}
FEV_1/FVC	rs10841302(12:19808912)	G/C	-0.0144	0.0037	1.21×10^{-4}	-0.0180	0.0035	2.58×10^{-7}	-0.0322	0.0075	1.52×10^{-5}	-0.0125	0.0078	1.10×10^{-1}
FVC	rs7977418(12:28588242)	T/C	0.0372	0.0038	3.34×10^{-23}	0.0364	0.0035	1.59×10^{-25}	0.0349	0.0076	3.77×10^{-6}	0.0515	0.0080	1.06×10^{-10}
FEV_1	rs1689510(12:56396768)	C/G	-0.0123	0.0039	1.68×10^{-3}	-0.0203	0.0036	2.49×10^{-8}	-0.0024	0.0085	7.81×10^{-1}	-0.0136	0.0087	1.18×10^{-1}
FEV_1/FVC	rs11172113(12:57527283)	T/C	-0.0232	0.0038	8.25×10^{-10}	-0.0233	0.0035	3.51×10^{-11}	-0.0167	0.0081	3.86×10^{-2}	-0.0263	0.0083	1.62×10^{-3}
FEV_1/FVC	rs1244869(12:65075332)	T/G	0.0155	0.0039	5.76×10^{-5}	0.0116	0.0036	1.36×10^{-3}	0.0279	0.0083	7.61×10^{-4}	0.0214	0.0086	1.28×10^{-2}
FEV_1	rs12825748(12:65793153)	G/C	-0.0177	0.0040	1.06×10^{-5}	-0.0201	0.0038	9.11×10^{-8}	-0.0276	0.0085	1.19×10^{-3}	-0.0145	0.0088	9.99×10^{-2}
FEV_1	rs11176001(12:66409367)	C/A	0.0296	0.0055	7.61×10^{-8}	0.0245	0.0051	1.68×10^{-6}	0.0496	0.0117	2.32×10^{-5}	0.0254	0.0123	3.80×10^{-2}
PEF	rs56390486(12:85719906)	A/G	0.0167	0.0041	5.04×10^{-5}	0.0244	0.0038	2.20×10^{-10}	0.0179	0.0139	1.97×10^{-1}	0.0065	0.0148	6.61×10^{-1}
FVC	rs9788269(12:94194890)	A/G	-0.0206	0.0042	1.09×10^{-6}	-0.0109	0.0039	5.65×10^{-3}	-0.0215	0.0090	1.70×10^{-2}	0.0012	0.0093	8.95×10^{-1}
FEV_1/FVC	rs113745635(12:95554771)	T/C	-0.0298	0.0045	4.82×10^{-11}	-0.0301	0.0042	1.34×10^{-12}	-0.0347	0.0095	2.40×10^{-4}	-0.0034	0.0096	7.26×10^{-1}
FEV_1/FVC	rs7970544(12:96242109)	T/G	0.0431	0.0047	8.90×10^{-20}	0.0479	0.0044	2.91×10^{-27}	0.0349	0.0103	7.31×10^{-4}	0.0411	0.0110	1.79×10^{-4}

Trait	SNP (Chr:Pos)	Coded / Non-coded allele	UKB ever smoker			UKB never smoker			SpM ever smoker			SpM never smoker		
			β	SE	P	β	SE	P	β	SE	P	β	SE	P
PEF	rs972936(12:102824921)	T/C	0.0247	0.0042	6.28×10^{-9}	0.0331	0.0040	5.64×10^{-17}	0.0412	0.0145	4.44×10^{-3}	0.0344	0.0156	2.75×10^{-2}
FEV_1	rs2701110(12:114669870)	C/A	-0.0254	0.0049	2.72×10^{-7}	-0.0258	0.0046	2.61×10^{-8}	-0.0231	0.0111	3.64×10^{-2}	-0.0382	0.0117	1.07×10^{-3}
FEV_1	rs10850377(12:115201436)	G/A	-0.0195	0.0039	6.52×10^{-7}	-0.0122	0.0037	8.61×10^{-4}	-0.0417	0.0085	9.02×10^{-7}	-0.0277	0.0088	1.67×10^{-3}
FVC	rs35505(12:115501127)	A/G	0.0173	0.0040	1.72×10^{-5}	0.0266	0.0037	1.11×10^{-12}	0.0274	0.0087	1.75×10^{-3}	0.0126	0.0090	1.60×10^{-1}
FEV_1/FVC	rs9533803(13:44820608)	C/T	0.0247	0.0046	6.68×10^{-8}	0.0275	0.0043	1.11×10^{-10}	0.0257	0.0091	4.86×10^{-3}	0.0319	0.0097	9.60×10^{-4}
FEV_1	rs2812208(13:50707087)	G/C	-0.0645	0.0130	6.94×10^{-7}	-0.0561	0.0121	3.73×10^{-6}	-0.0463	0.0253	6.71×10^{-2}	-0.0792	0.0269	3.22×10^{-3}
FVC	rs803765(13:71647588)	C/A	0.0241	0.0039	8.02×10^{-10}	0.0284	0.0037	8.33×10^{-15}	0.0131	0.0079	9.63×10^{-2}	0.0321	0.0083	1.05×10^{-4}
FEV_1	rs4885681(13:80467235)	C/T	-0.0210	0.0042	4.15×10^{-7}	-0.0156	0.0039	6.28×10^{-5}	-0.0198	0.0090	2.81×10^{-2}	-0.0252	0.0094	7.60×10^{-3}
FEV_1/FVC	rs11620380(13:99665512)	C/A	0.0336	0.0061	3.68×10^{-8}	0.0215	0.0056	1.37×10^{-4}	0.0451	0.0130	5.17×10^{-4}	0.0181	0.0137	1.87×10^{-1}
FEV_1/FVC	rs9634470(13:109918493)	T/C	-0.0251	0.0042	2.94×10^{-9}	-0.0173	0.0039	1.10×10^{-5}	-0.0208	0.0095	2.86×10^{-2}	-0.0342	0.0099	5.45×10^{-4}
FEV_1/FVC	rs1951121(14:23429729)	T/G	0.0213	0.0038	2.42×10^{-8}	0.0162	0.0036	4.97×10^{-6}	0.0228	0.0076	2.82×10^{-3}	0.0135	0.0082	9.84×10^{-2}
FEV_1/FVC	rs74053129(14:54346010)	G/A	-0.0425	0.0063	1.85×10^{-11}	-0.0356	0.0059	1.64×10^{-9}	-0.0402	0.0127	1.50×10^{-3}	-0.0395	0.0137	4.01×10^{-3}
FEV_1/FVC	rs35107139(14:54419106)	A/C	0.0297	0.0039	3.38×10^{-14}	0.0347	0.0037	2.43×10^{-21}	0.0124	0.0082	1.30×10^{-1}	0.0366	0.0086	2.33×10^{-5}
FVC	rs10141786(14:74817418)	A/G	0.0247	0.0038	1.01×10^{-10}	0.0216	0.0036	1.29×10^{-9}	0.0101	0.0082	2.15×10^{-1}	0.0265	0.0085	1.76×10^{-3}
FEV_1/FVC	rs1756281(14:84338431)	A/G	0.0300	0.0041	1.51×10^{-13}	0.0256	0.0038	1.47×10^{-11}	0.0025	0.0087	7.75×10^{-1}	0.0131	0.0090	1.46×10^{-1}
FEV_1	rs11160037(14:92512143)	A/G	-0.0150	0.0038	8.24×10^{-5}	-0.0169	0.0036	2.34×10^{-6}	-0.0354	0.0082	1.49×10^{-5}	-0.0272	0.0085	1.28×10^{-3}
FVC	rs11621587(14:93098339)	G/C	-0.0300	0.0048	4.67×10^{-10}	-0.0383	0.0045	1.05×10^{-17}	-0.0493	0.0105	2.87×10^{-6}	-0.0486	0.0109	9.03×10^{-6}
FVC	rs34245505(15:40397191)	C/G	0.0153	0.0048	1.29×10^{-3}	0.0257	0.0044	5.68×10^{-9}	0.0135	0.0100	1.77×10^{-1}	0.0367	0.0105	4.80×10^{-4}
FEV_1	rs2304645(15:40716253)	G/C	0.0149	0.0037	5.99×10^{-5}	0.0145	0.0035	2.95×10^{-5}	0.0254	0.0074	6.19×10^{-4}	0.0121	0.0079	1.25×10^{-1}
FVC	rs4924525(15:41255396)	C/A	0.0180	0.0038	1.75×10^{-6}	0.0176	0.0035	4.43×10^{-7}	0.0207	0.0076	6.60×10^{-3}	0.0237	0.0081	3.48×10^{-3}
FEV_1/FVC	rs2012453(15:41840238)	A/G	0.0251	0.0038	4.21×10^{-11}	0.0240	0.0035	1.28×10^{-11}	0.0105	0.0075	1.66×10^{-1}	0.0309	0.0079	9.71×10^{-5}
FEV_1/FVC	rs56383987(15:41953211)	T/C	-0.0424	0.0082	2.07×10^{-7}	-0.0370	0.0077	1.40×10^{-6}	-0.0279	0.0183	1.28×10^{-1}	-0.0174	0.0197	3.76×10^{-1}
FEV_1/FVC	rs79234094(15:49409527)	G/A	-0.0203	0.0043	1.90×10^{-6}	-0.0307	0.0040	9.88×10^{-15}	-0.0208	0.0085	1.37×10^{-2}	-0.0468	0.0090	1.81×10^{-7}
FEV_1/FVC	rs35251997(15:49706145)	A/T	-0.0405	0.0073	3.47×10^{-8}	-0.0661	0.0069	4.75×10^{-22}	-0.0438	0.0147	2.81×10^{-3}	-0.0343	0.0157	2.90×10^{-2}
FEV_1/FVC	rs62012772(15:63866877)	T/C	-0.0151	0.0049	1.90×10^{-3}	-0.0340	0.0046	1.12×10^{-13}	-0.0338	0.0096	4.53×10^{-4}	-0.0281	0.0104	6.75×10^{-3}
FVC	rs12917612(15:67491274)	C/A	0.0234	0.0045	1.44×10^{-7}	0.0217	0.0041	1.30×10^{-7}	0.0160	0.0090	7.54×10^{-2}	0.0236	0.0095	1.28×10^{-2}
FEV_1/FVC	rs1441358(15:71612514)	T/G	0.0675	0.0039	3.98×10^{-66}	0.0721	0.0037	1.24×10^{-85}	0.0411	0.0078	1.62×10^{-7}	0.0525	0.0084	3.21×10^{-10}
FEV_1	rs62015883(15:71803450)	C/T	0.0195	0.0048	5.44×10^{-5}	0.0172	0.0045	1.40×10^{-4}	0.0170	0.0098	8.31×10^{-2}	0.0318	0.0104	2.28×10^{-3}
FEV_1/FVC	rs7176074(15:73833600)	G/T	-0.0249	0.0086	4.04×10^{-3}	-0.0387	0.0081	1.97×10^{-6}	-0.0302	0.0169	7.36×10^{-2}	-0.0501	0.0178	4.77×10^{-3}
FEV_1/FVC	rs1896797(15:84274591)	G/A	-0.0287	0.0037	1.38×10^{-14}	-0.0305	0.0035	1.66×10^{-18}	-0.0250	0.0079	1.64×10^{-3}	-0.0215	0.0082	9.02×10^{-3}
FVC	rs3751837(16:3583173)	C/T	0.0276	0.0045	1.18×10^{-9}	0.0374	0.0042	6.29×10^{-19}	0.0223	0.0091	1.37×10^{-2}	0.0389	0.0094	3.76×10^{-5}
FEV_1/FVC	rs56104880(16:4361138)	T/C	0.0250	0.0041	9.70×10^{-10}	0.0187	0.0038	9.23×10^{-7}	0.0236	0.0085	5.72×10^{-3}	0.0211	0.0091	2.00×10^{-2}

Trait	SNP (Chr:Pos)	Coded / Non-coded allele	UKB ever smoker			UKB never smoker			SpM ever smoker			SpM never smoker		
			β	SE	P	β	SE	P	β	SE	P	β	SE	P
FVC	rs11074547(16:10136889)	T/G	-0.0191	0.0043	6.80×10^{-6}	-0.0092	0.0040	1.99×10^{-2}	-0.0272	0.0085	1.39×10^{-3}	-0.0219	0.0089	1.43×10^{-2}
FEV_1/FVC	rs78442819(16:10740982)	G/C	0.0303	0.0048	2.18×10^{-10}	0.0360	0.0045	6.32×10^{-16}	0.0349	0.0099	4.09×10^{-4}	0.0430	0.0104	3.34×10^{-5}
FEV_1	rs12446589(16:28870962)	A/G	-0.0148	0.0038	9.00×10^{-5}	-0.0123	0.0035	5.06×10^{-4}	-0.0063	0.0076	4.07×10^{-1}	-0.0017	0.0080	8.33×10^{-1}
FVC	rs76219171(16:50188929)	G/A	0.0380	0.0080	2.14×10^{-6}	0.0320	0.0073	1.25×10^{-5}	0.0546	0.0170	1.37×10^{-3}	0.0310	0.0176	7.89×10^{-2}
FEV_1/FVC	rs35420030(16:53935407)	T/C	-0.0412	0.0083	7.28×10^{-7}	-0.0364	0.0079	3.69×10^{-6}	-0.0797	0.0173	4.19×10^{-6}	-0.0852	0.0175	1.16×10^{-6}
FEV_1/FVC	rs11648508(16:58063513)	G/T	-0.0414	0.0040	5.91×10^{-25}	-0.0331	0.0037	9.13×10^{-19}	-0.0215	0.0080	7.46×10^{-3}	-0.0153	0.0085	7.15×10^{-2}
FEV_1	rs8047194(16:69891510)	G/T	0.0217	0.0037	4.87×10^{-9}	0.0197	0.0035	1.23×10^{-8}	0.0200	0.0079	1.14×10^{-2}	0.0322	0.0082	8.72×10^{-5}
FEV_1/FVC	rs11858992(16:75411445)	A/C	0.0374	0.0038	5.81×10^{-23}	0.0443	0.0035	4.02×10^{-36}	0.0349	0.0081	1.51×10^{-5}	0.0265	0.0083	1.50×10^{-3}
FEV_1	rs2345443(16:78225633)	A/G	0.0225	0.0040	2.08×10^{-8}	0.0208	0.0037	2.79×10^{-8}	0.0178	0.0086	3.80×10^{-2}	0.0234	0.0089	8.65×10^{-3}
FEV_1/FVC	rs12918140(16:86403821)	G/C	0.0265	0.0058	5.59×10^{-6}	0.0193	0.0054	3.99×10^{-4}	0.0221	0.0128	8.25×10^{-2}	0.0235	0.0133	7.68×10^{-2}
FEV_1	rs6539952(16:86579223)	C/A	0.0153	0.0043	3.34×10^{-4}	0.0168	0.0040	2.32×10^{-5}	0.0290	0.0090	1.30×10^{-3}	0.0228	0.0096	1.73×10^{-2}
FEV_1/FVC	rs8082036(17:3882613)	G/C	-0.0267	0.0037	6.97×10^{-13}	-0.0228	0.0035	5.65×10^{-11}	-0.0186	0.0074	1.22×10^{-2}	-0.0159	0.0079	4.29×10^{-2}
FEV_1	rs4796334(17:6469793)	G/A	0.0180	0.0037	1.37×10^{-6}	0.0083	0.0035	1.77×10^{-2}	0.0311	0.0075	3.59×10^{-5}	0.0169	0.0079	3.33×10^{-2}
FVC	rs1215(17:7163350)	A/G	0.0142	0.0054	8.10×10^{-3}	0.0221	0.0049	8.16×10^{-6}	0.0280	0.0108	9.74×10^{-3}	0.0414	0.0117	3.92×10^{-4}
FEV_1	rs4968200(17:7448457)	C/G	-0.0234	0.0053	1.02×10^{-5}	-0.0142	0.0049	4.18×10^{-3}	-0.0283	0.0109	9.63×10^{-3}	-0.0412	0.0115	3.44×10^{-4}
FVC	rs34351630(17:16030520)	T/C	0.0068	0.0038	6.81×10^{-2}	0.0170	0.0035	1.07×10^{-6}	0.0305	0.0076	6.35×10^{-5}	0.0217	0.0080	6.84×10^{-3}
FEV_1/FVC	rs2244592(17:28072327)	A/G	-0.0327	0.0037	1.59×10^{-18}	-0.0352	0.0035	4.58×10^{-24}	-0.0232	0.0076	2.15×10^{-3}	-0.0349	0.0080	1.46×10^{-5}
FVC	rs62070648(17:29210595)	A/G	0.0292	0.0042	3.70×10^{-12}	0.0218	0.0039	2.50×10^{-8}	0.0090	0.0086	2.91×10^{-1}	0.0061	0.0089	4.95×10^{-1}
FEV_1/FVC	rs35246838(17:36915540)	T/C	0.0428	0.0056	1.70×10^{-14}	0.0366	0.0052	1.77×10^{-12}	0.0300	0.0114	8.45×10^{-3}	0.0228	0.0120	5.73×10^{-2}
FVC	rs8069451(17:37504933)	T/C	0.0187	0.0043	1.44×10^{-5}	0.0186	0.0040	3.54×10^{-6}	0.0167	0.0088	5.77×10^{-2}	0.0278	0.0093	2.88×10^{-3}
FEV_1	rs79412431(17:43940021)	G/A	0.0434	0.0045	9.21×10^{-22}	0.0454	0.0042	2.32×10^{-27}	0.0166	0.0106	1.18×10^{-1}	0.0345	0.0112	2.02×10^{-3}
FVC	rs12945803(17:46552229)	T/C	0.0183	0.0045	5.64×10^{-5}	0.0189	0.0042	8.27×10^{-6}	0.0145	0.0091	1.12×10^{-1}	0.0415	0.0096	1.43×10^{-5}
FVC	rs28519449(17:54195453)	C/T	-0.0144	0.0038	1.55×10^{-4}	-0.0267	0.0035	4.05×10^{-14}	-0.0202	0.0076	8.06×10^{-3}	-0.0248	0.0080	1.97×10^{-3}
FEV_1/FVC	rs8068952(17:59286644)	G/C	0.0355	0.0045	4.20×10^{-15}	0.0254	0.0042	1.71×10^{-9}	0.0152	0.0098	1.20×10^{-1}	0.0236	0.0102	2.02×10^{-2}
FVC	rs77672322(17:62497964)	C/T	0.0374	0.0116	1.22×10^{-3}	0.0393	0.0108	2.85×10^{-4}	0.0515	0.0343	1.34×10^{-1}	0.1050	0.0342	2.12×10^{-3}
FEV_1/FVC	rs11653958(17:62686730)	G/A	-0.0135	0.0043	1.61×10^{-3}	-0.0221	0.0040	2.73×10^{-8}	-0.0200	0.0096	3.66×10^{-2}	-0.0245	0.0099	1.31×10^{-2}
FVC	rs6501431(17:68976415)	T/C	0.0212	0.0046	3.15×10^{-6}	0.0091	0.0042	3.04×10^{-2}	0.0346	0.0097	3.62×10^{-4}	0.0244	0.0099	1.39×10^{-2}
FEV_1	rs6501455(17:69201811)	A/G	0.0241	0.0037	7.78×10^{-11}	0.0360	0.0035	2.13×10^{-25}	0.0266	0.0080	8.39×10^{-4}	0.0195	0.0083	1.86×10^{-2}
FEV_1/FVC	rs996865(17:69371318)	C/T	0.0432	0.0070	6.92×10^{-10}	0.0545	0.0066	1.20×10^{-16}	0.0114	0.0157	4.67×10^{-1}	0.0659	0.0168	9.12×10^{-5}
FEV_1	rs9892893(17:73525670)	T/G	0.0174	0.0042	4.13×10^{-5}	0.0262	0.0040	3.92×10^{-11}	0.0057	0.0093	5.41×10^{-1}	0.0212	0.0096	2.78×10^{-2}
FVC	rs59606152(17:79952944)	C/T	-0.0285	0.0061	2.76×10^{-6}	-0.0398	0.0056	1.75×10^{-12}	-0.0287	0.0158	7.03×10^{-2}	-0.0406	0.0156	9.09×10^{-3}
FEV_1	rs513953(18:8801351)	A/G	-0.0211	0.0043	7.41×10^{-7}	-0.0364	0.0040	6.69×10^{-20}	-0.0172	0.0086	4.54×10^{-2}	-0.0127	0.0090	1.58×10^{-1}

Trait	SNP (Chr:Pos)	Coded / Non-coded allele	UKB ever smoker			UKB never smoker			SpM ever smoker			SpM never smoker		
			β	SE	P	β	SE	P	β	SE	P	β	SE	P
FEV_1/FVC	rs8089099(18:10078071)	G/A	-0.0224	0.0042	9.10×10^{-8}	-0.0254	0.0039	8.46×10^{-11}	-0.0275	0.0085	1.22×10^{-3}	-0.0237	0.0090	8.66×10^{-3}
FEV_1/FVC	rs1985511(18:19816712)	A/T	0.0133	0.0038	4.17×10^{-4}	0.0132	0.0035	1.62×10^{-4}	0.0224	0.0076	3.05×10^{-3}	0.0208	0.0081	9.78×10^{-3}
FEV_1	rs11082051(18:20234336)	A/G	0.0104	0.0037	5.25×10^{-3}	0.0181	0.0035	2.01×10^{-7}	0.0078	0.0075	2.98×10^{-1}	0.0319	0.0079	5.20×10^{-5}
FEV_1	rs9947743(18:20708321)	A/G	-0.0219	0.0046	1.50×10^{-6}	-0.0237	0.0043	2.39×10^{-8}	-0.0082	0.0092	3.69×10^{-1}	-0.0334	0.0096	5.21×10^{-4}
FVC	rs303752(18:21074255)	G/A	0.0192	0.0039	6.71×10^{-7}	0.0135	0.0036	1.56×10^{-4}	0.0210	0.0078	7.28×10^{-3}	0.0131	0.0083	1.13×10^{-1}
FVC	rs1668091(18:22290711)	T/C	-0.0164	0.0040	4.91×10^{-5}	-0.0184	0.0037	9.41×10^{-7}	-0.0150	0.0081	6.32×10^{-2}	-0.0310	0.0084	2.30×10^{-4}
FEV_1	rs9807668(18:42827898)	C/T	-0.0330	0.0063	1.72×10^{-7}	-0.0249	0.0059	2.50×10^{-5}	-0.0262	0.0131	4.54×10^{-2}	-0.0417	0.0138	2.58×10^{-3}
FVC	rs12607758(18:51022606)	T/C	0.0122	0.0038	1.39×10^{-3}	0.0189	0.0035	9.17×10^{-8}	0.0157	0.0077	4.20×10^{-2}	0.0078	0.0080	3.31×10^{-1}
FVC	rs2202572(18:53566471)	A/C	0.0151	0.0040	1.48×10^{-4}	0.0118	0.0037	1.40×10^{-3}	0.0230	0.0081	4.37×10^{-3}	0.0238	0.0085	5.01×10^{-3}
FEV_1	rs11085744(19:10819967)	C/T	0.0161	0.0037	1.74×10^{-5}	0.0110	0.0035	1.63×10^{-3}	0.0230	0.0077	2.65×10^{-3}	0.0192	0.0081	1.71×10^{-2}
FEV_1/FVC	rs9636166(19:31829613)	A/C	0.0351	0.0056	4.14×10^{-10}	0.0447	0.0052	1.33×10^{-17}	0.0166	0.0114	1.45×10^{-1}	0.0188	0.0122	1.24×10^{-1}
FVC	rs2967516(19:36881643)	A/G	-0.0108	0.0041	8.86×10^{-3}	-0.0157	0.0038	3.84×10^{-5}	-0.0227	0.0083	6.06×10^{-3}	-0.0165	0.0087	5.68×10^{-2}
FEV_1/FVC	rs34093919(19:41117300)	G/A	-0.1567	0.0168	1.37×10^{-20}	-0.1572	0.0157	1.22×10^{-23}	-0.1690	0.0323	1.68×10^{-7}	-0.1429	0.0326	1.19×10^{-5}
FVC	rs2145272(20:6626218)	A/G	0.0317	0.0039	3.59×10^{-16}	0.0289	0.0036	1.07×10^{-15}	0.0184	0.0079	1.99×10^{-2}	0.0184	0.0083	2.64×10^{-2}
FEV_1	rs6032942(20:10745545)	G/C	-0.0148	0.0044	7.34×10^{-4}	-0.0162	0.0041	8.12×10^{-5}	-0.0238	0.0088	6.74×10^{-3}	-0.0245	0.0093	8.60×10^{-3}
FEV_1	rs2236180(20:25282608)	T/C	0.0288	0.0048	1.42×10^{-9}	0.0114	0.0045	1.11×10^{-2}	0.0308	0.0095	1.23×10^{-3}	0.0096	0.0101	3.40×10^{-1}
FEV_1/FVC	rs4413223(20:30858967)	A/G	-0.0267	0.0049	5.73×10^{-8}	-0.0182	0.0046	6.79×10^{-5}	-0.0194	0.0100	5.12×10^{-2}	-0.0365	0.0106	5.74×10^{-4}
FVC	rs143384(20:34025756)	A/G	0.0225	0.0038	3.51×10^{-9}	0.0291	0.0035	1.57×10^{-16}	0.0248	0.0078	1.38×10^{-3}	0.0290	0.0082	3.94×10^{-4}
FVC	rs12481092(20:45486817)	C/T	-0.0268	0.0042	2.41×10^{-10}	-0.0296	0.0039	5.97×10^{-14}	-0.0183	0.0083	2.84×10^{-2}	-0.0156	0.0087	7.37×10^{-2}
FVC	rs4809221(20:62372706)	G/A	0.0295	0.0040	1.25×10^{-13}	0.0315	0.0037	1.52×10^{-17}	0.0244	0.0087	5.27×10^{-3}	0.0231	0.0089	9.71×10^{-3}
FEV_1/FVC	rs12627254(21:35368402)	G/T	-0.0305	0.0056	4.23×10^{-8}	-0.0415	0.0052	1.44×10^{-15}	-0.0350	0.0111	1.53×10^{-3}	-0.0400	0.0118	6.67×10^{-4}
FEV_1/FVC	rs62213732(21:35675966)	C/T	-0.0173	0.0039	6.64×10^{-6}	-0.0303	0.0036	2.56×10^{-17}	-0.0304	0.0077	8.39×10^{-5}	-0.0191	0.0082	1.98×10^{-2}
FEV_1	rs1978968(22:18448113)	C/T	-0.0237	0.0043	4.72×10^{-8}	-0.0328	0.0041	6.97×10^{-16}	-0.0299	0.0088	7.16×10^{-4}	-0.0250	0.0094	7.63×10^{-3}
FEV_1	rs9610955(22:20790723)	C/G	-0.0122	0.0047	9.18×10^{-3}	-0.0220	0.0044	4.57×10^{-7}	-0.0304	0.0095	1.34×10^{-3}	-0.0180	0.0101	7.47×10^{-2}
FEV_1/FVC	rs2283847(22:28181399)	T/C	-0.0262	0.0038	7.85×10^{-12}	-0.0202	0.0036	1.84×10^{-8}	-0.0142	0.0078	6.76×10^{-2}	-0.0231	0.0081	4.55×10^{-3}
FEV_1	rs113111175(22:50867711)	C/T	-0.0175	0.0057	2.03×10^{-3}	-0.0210	0.0053	7.29×10^{-5}	-0.0342	0.0125	6.40×10^{-3}	-0.0242	0.0133	6.81×10^{-2}

F. Risk score weighting for the 279 lung function and COPD associated SNPs by phenotype (S = SpiroMeta Consortium, UKB = UK Biobank, N = Novel, P= Previously reported, W = weight)

		FEV_1				FEV_1/FVC				FVC				PEF			
Signal	**SNP (Chr:Pos)**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**
N	rs9661802 (1:6678864)	S	C	-0.0025	0.14	S	C	-0.0217	1.10	S	A	-0.0136	0.92	S	C	-0.0090	0.61
N	rs12737805 (1:22612690)	S	G	-0.0220	1.23	S	G	-0.0040	0.20	S	G	-0.0186	1.25	S	G	-0.0099	0.68
N	rs9438626 (1:26775367)	UKB	G	-0.0180	1.00	S	G	-0.0025	0.13	UKB	G	-0.0164	1.10	UKB	G	-0.0125	0.86
N	rs12096239 (1:26796922)	S	C	-0.0209	1.17	S	C	-0.0128	0.65	S	C	-0.0153	1.03	S	C	-0.0277	1.90
N	rs1416685 (1:51243374)	S	G	-0.0183	1.02	S	G	-0.0192	0.98	S	G	-0.0054	0.36	S	G	-0.0292	2.00
N	rs72673461 (1:60966772)	S	G	-0.0164	0.91	S	G	-0.0439	2.23	S	T	-0.0119	0.80	S	T	-0.0171	1.17
N	rs9661687 (1:78387270)	S	T	-0.0206	1.15	S	T	-0.0295	1.49	S	T	-0.0039	0.26	S	T	-0.0157	1.08
N	rs10874851 (1:92106637)	S	A	-0.0083	0.46	S	A	-0.0168	0.85	S	C	-0.0041	0.27	S	A	-0.0114	0.78
N	rs9970286 (1:111737398)	S	G	-0.0171	0.95	S	G	-0.0235	1.19	S	G	-0.0022	0.15	S	G	-0.0197	1.35
N	rs11205354 (1:150249101)	S	C	-0.0122	0.68	S	C	-0.0150	0.76	S	C	-0.0010	0.07	S	C	-0.0314	2.15
N	rs141942982 (1:155137395)	S	T	-0.0257	1.43	S	T	-0.0355	1.80	S	G	-0.0014	0.09	S	T	-0.0116	0.79
N	rs4651005 (1:178719306)	S	C	-0.0195	1.09	S	C	-0.0027	0.14	S	C	-0.0170	1.14	S	C	-0.0028	0.19
N	rs2146098 (1:186090370)	S	A	-0.0081	0.45	S	G	-0.0150	0.76	S	A	-0.0187	1.26	S	G	-0.0071	0.48
N	rs17531405 (1:186113852)	S	G	-0.0104	0.58	S	G	-0.0252	1.28	S	C	-0.0055	0.37	S	G	-0.0061	0.42
N	rs10919604 (1:198898157)	S	G	-0.0196	1.09	S	G	-0.0224	1.14	S	G	-0.0056	0.38	S	G	-0.0118	0.81
N	rs4309038 (1:201884647)	UKB	G	-0.0053	0.29	UKB	G	-0.0144	0.73	S	G	-0.0004	0.03	UKB	G	-0.0063	0.43
N	rs2799098 (1:218521609)	S	A	-0.0020	0.11	S	A	-0.0246	1.25	S	G	-0.0150	1.00	S	A	-0.0029	0.20
N	rs75128958 (1:219483218)	S	A	-0.0173	0.97	S	A	-0.0404	2.05	S	G	-0.0015	0.10	S	A	-0.0249	1.71
N	rs17009288 (1:221204299)	S	A	-0.0161	0.90	S	C	-0.0101	0.51	S	A	-0.0203	1.36	S	C	-0.0102	0.70
N	rs2544536 (2:15906854)	S	T	-0.0261	1.45	S	T	-0.0227	1.15	S	T	-0.0124	0.83	S	T	-0.0237	1.62
N	rs6751968 (2:18570024)	S	C	-0.0180	1.00	S	A	-0.0136	0.69	S	C	-0.0286	1.92	S	C	-0.0078	0.53
N	rs13430465 (2:18702313)	S	C	-0.0304	1.70	S	T	-0.0071	0.36	S	C	-0.0345	2.32	S	C	-0.0245	1.68
N	rs13009582 (2:24018480)	UKB	G	-0.0149	0.83	UKB	G	-0.0021	0.11	UKB	G	-0.0154	1.03	S	G	-0.0050	0.34
N	rs732990 (2:26842146)	UKB	C	-0.0070	0.39	UKB	G	-0.0117	0.60	UKB	C	-0.0148	0.99	S	G	-0.0028	0.19
N	rs4952564 (2:42243850)	S	A	-0.0144	0.80	S	G	-0.0071	0.36	S	A	-0.0193	1.30	S	A	-0.0073	0.50
N	rs12470864 (2:102926362)	S	A	-0.0097	0.54	S	A	-0.0215	1.09	S	G	-0.0054	0.36	S	G	-0.0043	0.29
N	rs1406225 (2:145797829)	S	T	-0.0055	0.31	S	T	-0.0208	1.05	S	G	-0.0083	0.56	S	T	-0.0020	0.14
N	rs7424771 (2:161276378)	S	A	-0.0217	1.21	S	A	-0.0019	0.10	S	A	-0.0163	1.09	S	A	-0.0036	0.25

		FEV_1				FEV_1/FVC				FVC				PEF			
Signal	**SNP (Chr:Pos)**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**
N	rs2304340 (2:179260382)	UKB	A	-0.0126	0.70	UKB	G	-0.0034	0.17	UKB	A	-0.0157	1.06	S	G	-0.0005	0.03
N	rs2084448 (2:187530520)	S	C	-0.0093	0.52	S	C	-0.0205	1.04	S	T	-0.0010	0.07	S	C	-0.0237	1.62
N	rs1249096 (2:199723365)	S	G	-0.0241	1.34	S	G	-0.0026	0.13	S	G	-0.0224	1.50	S	G	-0.0107	0.73
N	rs985256 (2:201208692)	UKB	C	-0.0138	0.77	UKB	C	-0.0168	0.85	S	C	-0.0052	0.35	UKB	C	-0.0136	0.93
N	rs12997625 (2:202970250)	UKB	T	-0.0115	0.64	UKB	C	-0.0038	0.19	UKB	T	-0.0146	0.98	S	C	-0.0016	0.11
N	rs6435952 (2:217614730)	S	T	-0.0309	1.72	S	T	-0.0335	1.70	S	T	-0.0129	0.87	S	T	-0.0181	1.24
N	rs4294980 (2:218604356)	UKB	G	-0.0165	0.92	S	G	-0.0026	0.13	UKB	G	-0.0175	1.18	UKB	G	-0.0028	0.19
N	rs4674407 (2:220382700)	UKB	T	-0.0079	0.44	UKB	C	-0.0092	0.47	UKB	T	-0.0140	0.94	UKB	T	-0.0046	0.31
N	rs6431620 (2:239604970)	S	G	-0.0119	0.66	S	T	-0.0104	0.53	UKB	G	-0.0176	1.18	UKB	G	-0.0091	0.63
N	rs6437219 (2:241844033)	S	C	-0.0153	0.85	S	T	-0.0090	0.46	S	C	-0.0214	1.44	S	C	-0.0209	1.43
N	rs6733504 (2:242495953)	S	G	-0.0209	1.16	S	G	-0.0014	0.07	S	G	-0.0189	1.27	S	A	-0.0001	0.01
N	rs2974389 (3:13787641)	UKB	G	-0.0146	0.81	UKB	G	-0.0022	0.11	UKB	G	-0.0144	0.97	UKB	G	-0.0115	0.79
N	rs73048404 (3:25179533)	UKB	G	-0.0151	0.84	UKB	T	-0.0038	0.19	UKB	G	-0.0170	1.14	UKB	G	-0.0120	0.82
N	rs35480566 (3:71583177)	S	A	-0.0225	1.25	S	A	-0.0058	0.29	S	A	-0.0216	1.45	S	A	-0.0098	0.67
N	rs586936 (3:73862616)	S	A	-0.0147	0.82	S	A	-0.0191	0.97	S	A	-0.0033	0.22	S	A	-0.0062	0.43
N	rs1610265 (3:99420192)	S	T	-0.0255	1.42	S	C	-0.0149	0.76	S	T	-0.0356	2.39	S	T	-0.0111	0.76
N	rs1799807 (3:165548529)	S	C	-0.0507	2.83	S	C	-0.0871	4.42	S	T	-0.0082	0.55	S	C	-0.0344	2.36
N	rs6780171 (3:185503456)	S	A	-0.0221	1.23	S	A	-0.0076	0.39	S	A	-0.0193	1.30	S	A	-0.0152	1.04
N	rs12331869 (4:56012149)	UKB	A	-0.0161	0.90	S	A	-0.0074	0.38	UKB	A	-0.0137	0.92	S	A	-0.0063	0.43
N	rs62316310 (4:75676529)	S	G	-0.0235	1.31	S	G	-0.0209	1.06	S	G	-0.0066	0.45	S	G	-0.0299	2.05
N	rs11098196 (4:79403952)	S	T	-0.0079	0.44	S	T	-0.0181	0.92	S	G	-0.0011	0.07	S	T	-0.0160	1.09
N	rs13109426 (4:145330628)	S	A	-0.0139	0.78	S	G	-0.0114	0.58	S	A	-0.0215	1.45	S	G	-0.0051	0.35
N	rs13116999 (4:145442364)	S	G	-0.0350	1.95	S	G	-0.0409	2.08	S	G	-0.0098	0.66	S	G	-0.0374	2.56
N	rs11739847 (5:609661)	S	A	-0.0232	1.29	S	A	-0.0083	0.42	S	A	-0.0205	1.38	S	G	-0.0027	0.19
N	rs4866846 (5:43976162)	S	G	-0.0305	1.70	S	G	-0.0183	0.93	S	G	-0.0191	1.28	S	G	-0.0004	0.03
N	rs10059661 (5:121410529)	S	G	-0.0004	0.02	S	C	-0.0298	1.51	S	G	-0.0180	1.21	S	G	-0.0108	0.74
N	rs17163397 (5:128767384)	S	A	-0.0342	1.91	S	A	-0.0314	1.60	S	A	-0.0202	1.35	S	G	-0.0018	0.12
N	rs1800888 (5:148206885)	S	T	-0.0843	4.70	S	T	-0.0679	3.45	S	T	-0.0589	3.95	S	T	-0.0466	3.19
N	rs10059996 (5:170901463)	S	T	-0.0166	0.93	S	T	-0.0324	1.64	S	G	-0.0035	0.24	S	T	-0.0150	1.03
N	rs79898473 (5:179598771)	S	T	-0.0088	0.49	S	T	-0.0203	1.03	S	C	-0.0038	0.25	S	T	-0.0158	1.08
N	rs12198986 (6:7720059)	S	A	-0.0165	0.92	S	A	-0.0006	0.03	S	A	-0.0175	1.17	S	G	-0.0085	0.58
N	rs9689096 (6:34188892)	S	A	-0.0417	2.32	S	A	-0.0113	0.57	S	A	-0.0396	2.66	S	C	-0.0177	1.21
N	rs9357446 (6:44447598)	UKB	A	-0.0121	0.68	UKB	G	-0.0004	0.02	UKB	A	-0.0134	0.90	S	A	-0.0005	0.04
N	rs12202314 (6:45530471)	S	T	-0.0082	0.46	S	T	-0.0214	1.09	S	C	-0.0051	0.34	S	C	-0.0137	0.94

		FEV_1				FEV_1/FVC				FVC				PEF			
Signal	**SNP (Chr:Pos)**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**
N	rs9472541 (6:45622748)	UKB	A	-0.0098	0.54	UKB	T	-0.0067	0.34	UKB	A	-0.0145	0.98	S	A	-0.0026	0.18
N	rs2894837 (6:56336406)	S	G	-0.0205	1.14	S	G	-0.0065	0.33	S	G	-0.0137	0.92	S	G	-0.0167	1.14
N	rs2627237 (6:134339265)	UKB	G	-0.0130	0.73	S	G	-0.0001	0.00	UKB	G	-0.0115	0.77	S	G	-0.0013	0.09
N	rs1102077 (6:140271357)	S	C	-0.0226	1.26	S	A	-0.0025	0.13	S	C	-0.0208	1.40	S	C	-0.0298	2.04
N	rs9385988 (6:142560957)	S	A	-0.0268	1.50	S	A	-0.0165	0.84	S	A	-0.0198	1.33	S	A	-0.0339	2.32
N	rs4721457 (7:15872324)	S	C	-0.0174	0.97	S	C	-0.0290	1.47	S	T	-0.0044	0.29	S	C	-0.0208	1.43
N	rs559233 (7:26848830)	S	C	-0.0216	1.21	S	C	-0.0157	0.80	S	C	-0.0152	1.02	S	C	-0.0128	0.88
N	rs62454414 (7:27182329)	UKB	G	-0.0156	0.87	UKB	T	-0.0044	0.22	UKB	G	-0.0196	1.32	UKB	G	-0.0044	0.30
N	rs1513272 (7:28200097)	S	C	-0.0179	1.00	S	C	-0.0023	0.11	S	C	-0.0162	1.09	S	C	-0.0107	0.73
N	rs17232687 (7:46448518)	S	T	-0.0168	0.94	S	C	-0.0020	0.10	S	T	-0.0184	1.24	S	T	-0.0003	0.02
N	rs12707691 (7:84569510)	S	C	-0.0254	1.42	S	C	-0.0062	0.32	S	C	-0.0216	1.45	S	C	-0.0055	0.38
N	rs193686 (7:116431427)	UKB	T	-0.0029	0.16	UKB	T	-0.0164	0.83	S	C	-0.0013	0.09	S	T	-0.0042	0.29
N	rs330939 (8:9018590)	S	G	-0.0096	0.53	S	G	-0.0179	0.91	S	T	-0.0036	0.24	S	G	-0.0102	0.70
N	rs4128298 (8:11823332)	UKB	T	-0.0166	0.92	S	T	-0.0170	0.86	UKB	T	-0.0080	0.54	UKB	T	-0.0130	0.89
N	rs7465401 (8:70367248)	S	T	-0.0246	1.37	S	T	-0.0070	0.35	S	T	-0.0199	1.33	S	T	-0.0140	0.96
N	rs7838717 (8:145504343)	S	T	-0.0268	1.49	S	C	-0.0007	0.03	S	T	-0.0269	1.81	S	T	-0.0088	0.60
N	rs7041139 (9:18013733)	S	T	-0.0225	1.25	S	T	-0.0064	0.33	S	T	-0.0190	1.28	S	T	-0.0087	0.59
N	rs72743974 (9:98878881)	S	A	-0.0269	1.50	S	A	-0.0287	1.46	S	A	-0.0086	0.58	S	A	-0.0201	1.37
N	rs57649467 (9:101632854)	S	G	-0.0118	0.66	S	G	-0.0193	0.98	S	A	-0.0036	0.24	S	G	-0.0191	1.31
N	rs967497 (9:131943843)	UKB	G	-0.0138	0.77	UKB	G	-0.0097	0.49	UKB	G	-0.0105	0.70	UKB	G	-0.0077	0.53
N	rs1274475 (10:34480582)	S	G	-0.0039	0.22	S	G	-0.0191	0.97	S	A	-0.0074	0.50	S	G	-0.0181	1.24
N	rs60820984 (10:75639578)	UKB	T	-0.0163	0.91	S	T	-0.0148	0.75	UKB	T	-0.0052	0.35	UKB	T	-0.0184	1.26
N	rs11191841 (10:105639611)	S	T	-0.0228	1.27	S	T	-0.0052	0.27	S	T	-0.0180	1.21	S	T	-0.0160	1.10
N	rs10836366 (11:35308988)	UKB	C	-0.0102	0.57	UKB	C	-0.0181	0.92	UKB	C	-0.0021	0.14	UKB	C	-0.0091	0.63
N	rs56196860 (12:2908330)	S	A	-0.0555	3.09	S	C	-0.0190	0.96	S	A	-0.0666	4.47	S	C	-0.0097	0.67
N	rs12811814 (12:4243749)	UKB	C	-0.0137	0.76	S	T	-0.0031	0.16	UKB	C	-0.0107	0.72	UKB	C	-0.0124	0.85
N	rs10841302 (12:19808912)	S	G	-0.0140	0.78	S	G	-0.0226	1.15	S	C	-0.0026	0.17	S	G	-0.0151	1.03
N	rs1244869 (12:65075332)	UKB	G	-0.0076	0.42	UKB	G	-0.0138	0.70	UKB	G	-0.0019	0.13	UKB	G	-0.0098	0.67
N	rs11176001 (12:66409367)	S	A	-0.0346	1.93	S	A	-0.0207	1.05	S	A	-0.0248	1.67	S	A	-0.0285	1.95
N	rs972936 (12:102824921)	S	C	-0.0286	1.59	S	C	-0.0152	0.77	S	C	-0.0194	1.30	S	C	-0.0378	2.59
N	rs2701110 (12:114669870)	S	C	-0.0258	1.44	S	C	-0.0256	1.30	S	C	-0.0098	0.66	S	C	-0.0124	0.85
N	rs9533803 (13:44820608)	S	T	-0.0135	0.75	S	T	-0.0252	1.28	S	C	-0.0026	0.17	S	T	-0.0205	1.40
N	rs2812208 (13:50707087)	S	G	-0.0730	4.07	S	G	-0.0341	1.73	S	G	-0.0513	3.45	S	G	-0.0171	1.17
N	rs803765 (13:71647588)	S	A	-0.0115	0.64	S	C	-0.0128	0.65	S	A	-0.0182	1.22	S	A	-0.0032	0.22
N	rs4885681 (13:80467235)	S	C	-0.0219	1.22	S	C	-0.0035	0.18	S	C	-0.0243	1.63	S	C	-0.0103	0.71

		FEV_1				FEV_1/FVC				FVC				PEF			
Signal	**SNP (Chr:Pos)**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**
N	rs11620380 (13:99665512)	S	A	-0.0185	1.03	S	A	-0.0363	1.84	S	C	-0.0047	0.32	S	A	-0.0224	1.54
N	rs9634470 (13:109918493)	S	T	-0.0007	0.04	S	T	-0.0215	1.09	S	C	-0.0118	0.79	S	T	-0.0083	0.57
N	rs1951121 (14:23429729)	S	G	-0.0156	0.87	S	G	-0.0185	0.94	S	G	-0.0055	0.37	S	G	-0.0240	1.65
N	rs74053129 (14:54346010)	S	A	-0.0020	0.11	S	G	-0.0401	2.04	S	A	-0.0245	1.64	S	G	-0.0245	1.68
N	rs10141786 (14:74817418)	S	G	-0.0187	1.04	S	A	-0.0035	0.18	S	G	-0.0192	1.29	S	G	-0.0097	0.66
N	rs34245505 (15:40397191)	S	G	-0.0105	0.59	S	C	-0.0155	0.79	S	G	-0.0255	1.72	S	G	-0.0181	1.24
N	rs2304645 (15:40716253)	UKB	C	-0.0143	0.80	UKB	C	0.0000	0.00	UKB	C	-0.0158	1.06	UKB	C	-0.0051	0.35
N	rs4924525 (15:41255396)	S	A	-0.0120	0.67	S	C	-0.0112	0.57	S	A	-0.0201	1.35	S	A	-0.0043	0.29
N	rs79234094 (15:49409527)	S	G	-0.0218	1.22	S	G	-0.0330	1.68	S	G	-0.0008	0.05	S	G	-0.0323	2.21
N	rs35251997 (15:49706145)	S	A	-0.0529	2.95	S	A	-0.0401	2.04	S	A	-0.0283	1.90	S	A	-0.0541	3.70
N	rs62012772 (15:63866877)	S	T	-0.0174	0.97	S	T	-0.0381	1.93	S	C	-0.0030	0.20	S	T	-0.0125	0.86
N	rs7176074 (15:73833600)	S	G	-0.0056	0.31	UKB	G	-0.0320	1.62	UKB	T	-0.0061	0.41	UKB	G	-0.0028	0.19
N	rs3751837 (16:3583173)	S	T	-0.0224	1.25	S	C	-0.0055	0.28	S	T	-0.0286	1.92	S	T	-0.0084	0.57
N	rs56104880 (16:4361138)	S	C	-0.0079	0.44	S	C	-0.0242	1.23	S	T	-0.0064	0.43	S	C	-0.0123	0.84
N	rs11074547 (16:10136889)	UKB	T	-0.0118	0.66	UKB	G	-0.0027	0.14	UKB	T	-0.0147	0.98	UKB	T	-0.0042	0.29
N	rs76219171 (16:50188929)	S	A	-0.0362	2.02	S	G	-0.0058	0.29	S	A	-0.0396	2.66	S	A	-0.0006	0.04
N	rs35420030 (16:53935407)	S	T	-0.0374	2.08	S	T	-0.0732	3.72	S	C	-0.0136	0.91	S	T	-0.0299	2.05
N	rs12918140 (16:86403821)	S	C	-0.0066	0.37	S	C	-0.0329	1.67	S	G	-0.0104	0.70	S	G	-0.0197	1.35
N	rs6539952 (16:86579223)	UKB	A	-0.0159	0.89	UKB	A	-0.0156	0.79	UKB	A	-0.0099	0.66	UKB	A	-0.0020	0.14
N	rs8082036 (17:3882613)	S	G	-0.0048	0.27	S	G	-0.0198	1.00	S	C	-0.0067	0.45	S	G	-0.0123	0.84
N	rs4796334 (17:6469793)	UKB	A	-0.0116	0.64	UKB	A	-0.0067	0.34	UKB	A	-0.0089	0.60	UKB	A	-0.0100	0.68
N	rs1215 (17:7163350)	UKB	G	-0.0132	0.74	UKB	A	-0.0059	0.30	UKB	G	-0.0176	1.18	UKB	G	-0.0089	0.61
N	rs4968200 (17:7448457)	UKB	C	-0.0189	1.06	UKB	C	-0.0114	0.58	UKB	C	-0.0152	1.02	UKB	C	-0.0067	0.46
N	rs34351630 (17:16030520)	UKB	C	-0.0107	0.60	UKB	C	-0.0001	0.00	UKB	C	-0.0124	0.83	UKB	C	-0.0019	0.13
N	rs12945803 (17:46552229)	UKB	C	-0.0163	0.91	UKB	T	-0.0003	0.02	UKB	C	-0.0180	1.21	S	T	-0.0124	0.85
N	rs28519449 (17:54195453)	S	C	-0.0178	0.99	S	T	-0.0102	0.52	S	C	-0.0217	1.46	S	T	-0.0023	0.16
N	rs8068952 (17:59286644)	S	C	-0.0218	1.22	S	C	-0.0244	1.24	S	C	-0.0107	0.72	S	C	-0.0247	1.70
N	rs77672322 (17:62497964)	UKB	T	-0.0248	1.38	UKB	C	-0.0270	1.37	UKB	T	-0.0398	2.68	UKB	C	-0.0129	0.88
N	rs11653958 (17:62686730)	S	A	-0.0034	0.19	S	G	-0.0264	1.34	S	A	-0.0202	1.35	S	G	-0.0379	2.60
N	rs996865 (17:69371318)	S	T	-0.0100	0.56	S	T	-0.0418	2.12	S	C	-0.0137	0.92	S	T	-0.0435	2.98
N	rs59606152 (17:79952944)	S	C	-0.0288	1.61	S	T	-0.0244	1.24	S	C	-0.0405	2.72	S	C	-0.0162	1.11
N	rs8089099 (18:10078071)	S	G	-0.0233	1.30	S	G	-0.0211	1.07	S	G	-0.0088	0.59	S	G	-0.0156	1.07
N	rs1985511 (18:19816712)	S	T	-0.0006	0.03	UKB	T	-0.0145	0.73	UKB	A	-0.0028	0.19	S	A	-0.0012	0.08
N	rs303752 (18:21074255)	S	A	-0.0148	0.82	S	G	-0.0009	0.05	S	A	-0.0191	1.28	S	A	-0.0030	0.20

		FEV_1				FEV_1/FVC				FVC				PEF			
Signal	**SNP (Chr:Pos)**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**
N	rs1668091 (18:22290711)	S	T	-0.0170	0.95	S	C	-0.0026	0.13	S	T	-0.0189	1.27	S	T	-0.0072	0.49
N	rs9807668 (18:42827898)	S	C	-0.0359	2.00	S	C	-0.0120	0.61	S	C	-0.0298	2.00	S	T	-0.0096	0.66
N	rs2202572 (18:53566471)	UKB	C	-0.0119	0.66	UKB	C	-0.0002	0.01	UKB	C	-0.0143	0.96	UKB	C	-0.0071	0.49
N	rs11085744 (19:10819967)	UKB	T	-0.0145	0.81	S	T	-0.0090	0.45	UKB	T	-0.0106	0.71	S	C	-0.0089	0.61
N	rs2967516 (19:36881643)	UKB	A	-0.0067	0.37	S	G	-0.0027	0.13	UKB	A	-0.0133	0.90	UKB	A	0.0000	0.00
N	rs6032942 (20:10745545)	UKB	G	-0.0162	0.90	S	G	-0.0173	0.88	UKB	G	-0.0066	0.44	S	G	-0.0160	1.10
N	rs12627254 (21:35368402)	S	G	-0.0129	0.72	S	G	-0.0352	1.79	S	T	-0.0107	0.72	S	G	-0.0456	3.13
N	rs113111175 (22:50867711)	UKB	C	-0.0197	1.10	UKB	C	-0.0080	0.40	UKB	C	-0.0169	1.13	S	T	-0.0195	1.33
P	rs2284746 (1:17306675)	UKB	C	-0.0021	0.12	UKB	G	-0.0382	1.94	UKB	C	-0.0204	1.37	UKB	G	-0.0232	1.59
P	rs17513135 (1:40035686)	S	T	-0.0189	1.05	S	T	-0.0212	1.08	S	T	-0.0050	0.33	S	T	-0.0344	2.35
P	rs1192404 (1:92068967)	S	A	-0.0114	0.64	S	G	-0.0365	1.85	S	A	-0.0330	2.22	S	G	-0.0080	0.55
P	rs12140637 (1:92374517)	S	T	-0.0028	0.16	S	T	-0.0144	0.73	S	C	-0.0070	0.47	S	T	-0.0041	0.28
P	rs60804050 (1:118870373)	S	G	-0.0106	0.59	S	A	-0.0031	0.16	S	G	-0.0138	0.93	S	G	-0.0099	0.68
P	rs6681426 (1:150586971)	UKB	A	-0.0168	0.94	UKB	G	-0.0087	0.44	UKB	A	-0.0227	1.53	UKB	A	-0.0049	0.33
P	rs2821332 (1:200085714)	UKB	T	-0.0123	0.69	UKB	A	-0.0039	0.20	UKB	T	-0.0152	1.02	UKB	T	-0.0054	0.37
P	rs12092943 (1:204434927)	UKB	T	-0.0240	1.34	UKB	T	-0.0136	0.69	UKB	T	-0.0198	1.33	UKB	T	-0.0280	1.92
P	rs512597 (1:215095003)	UKB	T	-0.0113	0.63	UKB	C	-0.0023	0.12	UKB	T	-0.0136	0.91	UKB	T	-0.0036	0.25
P	rs4846480 (1:218598469)	UKB	A	-0.0141	0.79	UKB	A	-0.0139	0.70	UKB	A	-0.0085	0.57	UKB	A	-0.0166	1.13
P	rs993925 (1:218860068)	UKB	C	-0.0097	0.54	UKB	C	-0.0126	0.64	UKB	C	-0.0054	0.36	UKB	C	-0.0042	0.29
P	rs4328080 (1:219963088)	UKB	G	-0.0065	0.36	UKB	G	-0.0223	1.13	UKB	A	-0.0030	0.20	UKB	G	-0.0033	0.23
P	rs6657854 (1:221630555)	UKB	A	-0.0105	0.58	UKB	C	-0.0140	0.71	UKB	A	-0.0185	1.24	UKB	C	-0.0024	0.16
P	rs6688537 (1:239850588)	S	A	-0.0187	1.04	S	A	-0.0137	0.70	S	A	-0.0092	0.62	S	A	-0.0314	2.15
P	rs62126408 (2:18309132)	UKB	T	-0.0337	1.88	UKB	T	-0.0467	2.37	UKB	T	-0.0164	1.10	UKB	T	-0.0234	1.60
P	rs1430193 (2:56120853)	UKB	T	-0.0149	0.83	UKB	A	-0.0156	0.79	UKB	T	-0.0241	1.62	UKB	T	-0.0006	0.04
P	rs2322659 (2:136555659)	UKB	C	-0.0198	1.11	UKB	C	-0.0025	0.12	UKB	C	-0.0203	1.36	UKB	C	-0.0119	0.82
P	rs72904209 (2:157046432)	UKB	C	-0.0334	1.86	UKB	C	-0.0328	1.67	UKB	C	-0.0200	1.35	UKB	C	-0.0186	1.28
P	rs2571445 (2:218683154)	UKB	A	-0.0284	1.58	UKB	A	-0.0221	1.12	UKB	A	-0.0215	1.45	UKB	A	-0.0111	0.76
P	rs10498230 (2:229502503)	UKB	C	-0.0245	1.37	UKB	C	-0.0821	4.17	UKB	T	-0.0137	0.92	UKB	C	-0.0341	2.34
P	rs61332075 (2:239316560)	S	G	-0.0128	0.71	S	G	-0.0175	0.89	S	C	-0.0012	0.08	S	G	-0.0198	1.36
P	rs12477314 (2:239877148)	UKB	C	-0.0287	1.60	UKB	C	-0.0492	2.50	UKB	C	-0.0068	0.46	UKB	C	-0.0204	1.40
P	rs1286664 (3:25529280)	UKB	C	-0.0153	0.85	UKB	C	-0.0432	2.19	UKB	T	-0.0047	0.32	UKB	C	-0.0201	1.38
P	rs17666332 (3:29469675)	UKB	G	-0.0098	0.55	UKB	G	-0.0297	1.51	UKB	T	-0.0040	0.27	UKB	G	-0.0079	0.54
P	rs1458979 (3:55150677)	S	G	-0.0092	0.51	S	G	-0.0146	0.74	S	G	-0.0016	0.11	S	A	-0.0021	0.15
P	rs79294353 (3:57494433)	UKB	A	-0.0165	0.92	UKB	A	-0.0178	0.90	UKB	A	-0.0099	0.67	UKB	A	-0.0114	0.78

		FEV_1				FEV_1/FVC				FVC				PEF			
Signal	**SNP (Chr:Pos)**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**
P	rs1490265 (3:67452043)	S	C	-0.0202	1.12	S	C	-0.0046	0.23	S	C	-0.0170	1.14	S	C	-0.0216	1.48
P	rs6778584 (3:98815640)	UKB	C	-0.0288	1.60	UKB	C	-0.0056	0.29	UKB	C	-0.0282	1.89	UKB	C	-0.0136	0.93
P	rs2811415 (3:127991527)	S	G	-0.0164	0.92	S	G	-0.0150	0.76	S	G	-0.0077	0.52	S	G	-0.0037	0.26
P	rs1595029 (3:158241767)	UKB	C	-0.0192	1.07	UKB	A	-0.0057	0.29	UKB	C	-0.0233	1.57	UKB	A	-0.0009	0.06
P	rs56341938 (3:168715808)	S	A	-0.0228	1.27	S	A	-0.0127	0.64	S	A	-0.0153	1.03	S	A	-0.0040	0.27
P	rs1344555 (3:169300219)	UKB	T	-0.0237	1.32	UKB	T	-0.0116	0.59	UKB	T	-0.0205	1.38	UKB	T	-0.0028	0.19
P	rs28520091 (4:7846240)	UKB	C	-0.0035	0.19	UKB	C	-0.0154	0.78	UKB	T	-0.0044	0.29	UKB	C	-0.0020	0.14
P	rs13110699 (4:89815695)	S	T	-0.0087	0.49	S	G	-0.0463	2.35	S	T	-0.0360	2.42	S	G	-0.0169	1.16
P	rs2045517 (4:89870964)	UKB	T	-0.0040	0.22	UKB	T	-0.0448	2.27	UKB	C	-0.0174	1.17	UKB	T	-0.0172	1.18
P	rs2007403 (4:106131210)	S	T	-0.0166	0.93	S	T	-0.0157	0.80	S	T	-0.0079	0.53	S	T	-0.0107	0.73
P	rs10516526 (4:106688904)	UKB	A	-0.0706	3.94	UKB	A	-0.0466	2.37	UKB	A	-0.0548	3.68	UKB	A	-0.0304	2.08
P	rs34712979 (4:106819053)	S	A	-0.0664	3.70	S	A	-0.0520	2.64	S	A	-0.0367	2.46	S	A	-0.0544	3.73
P	rs12504628 (4:145436324)	UKB	T	-0.0375	2.09	UKB	T	-0.0737	3.74	UKB	T	-0.0026	0.18	UKB	T	-0.0647	4.43
P	rs111898810 (4:146174040)	UKB	A	-0.0257	1.44	UKB	A	-0.0271	1.38	UKB	A	-0.0149	1.00	UKB	A	-0.0407	2.79
P	rs91731 (5:33334312)	S	A	-0.0348	1.94	S	A	-0.0071	0.36	S	A	-0.0368	2.47	S	A	-0.0263	1.81
P	rs1448044 (5:44296986)	UKB	G	-0.0170	0.95	UKB	A	-0.0053	0.27	UKB	G	-0.0206	1.39	UKB	G	-0.0077	0.53
P	rs1551943 (5:52195033)	S	A	-0.0192	1.07	S	A	-0.0201	1.02	S	A	-0.0045	0.30	S	A	-0.0244	1.67
P	rs2441026 (5:53444498)	S	C	-0.0180	1.00	S	C	-0.0013	0.07	S	C	-0.0189	1.27	S	C	-0.0054	0.37
P	rs72776440 (5:77440196)	UKB	C	-0.0147	0.82	UKB	G	-0.0113	0.57	UKB	C	-0.0212	1.43	UKB	C	-0.0029	0.20
P	rs153916 (5:95036700)	UKB	T	-0.0050	0.28	UKB	T	-0.0280	1.42	UKB	C	-0.0088	0.59	UKB	T	-0.0095	0.65
P	rs7713065 (5:131788334)	S	C	-0.0058	0.32	S	A	-0.0089	0.45	S	C	-0.0154	1.04	S	C	-0.0002	0.01
P	rs7715901 (5:147856392)	UKB	A	-0.0330	1.84	UKB	A	-0.0535	2.72	UKB	A	-0.0093	0.62	UKB	A	-0.0298	2.04
P	rs2014787 (5:148611675)	S	C	-0.0106	0.59	S	C	-0.0109	0.55	S	C	-0.0049	0.33	S	G	-0.0076	0.52
P	rs10515750 (5:156810072)	S	T	-0.0274	1.53	S	T	-0.0406	2.06	S	T	-0.0039	0.26	S	T	-0.0257	1.76
P	rs1990950 (5:156920756)	UKB	G	-0.0227	1.27	UKB	G	-0.0374	1.90	UKB	G	-0.0067	0.45	UKB	G	-0.0237	1.63
P	rs1294421 (6:6743149)	UKB	T	-0.0064	0.36	UKB	T	-0.0324	1.64	UKB	G	-0.0083	0.56	UKB	T	-0.0122	0.84
P	rs55938083 (6:7565376)	UKB	C	-0.0022	0.12	UKB	C	-0.0163	0.83	UKB	T	-0.0058	0.39	UKB	C	-0.0049	0.34
P	rs6924424 (6:7801611)	UKB	G	-0.0298	1.66	UKB	T	-0.0056	0.28	UKB	G	-0.0354	2.38	UKB	G	-0.0140	0.96
P	rs1928168 (6:22017738)	UKB	C	-0.0047	0.26	UKB	C	-0.0306	1.55	UKB	T	-0.0102	0.68	UKB	C	-0.0127	0.87
P	rs34864796 (6:27459923)	UKB	A	-0.0489	2.72	UKB	A	-0.0337	1.71	UKB	A	-0.0381	2.56	UKB	A	-0.0558	3.82
P	rs2070600 (6:32151443)	UKB	C	-0.0314	1.75	UKB	C	-0.1498	7.60	UKB	T	-0.0373	2.51	UKB	C	-0.0739	5.07
P	rs114544105 (6:32635629)	S	A	-0.0137	0.77	S	A	-0.0251	1.27	S	A	-0.0032	0.22	S	A	-0.0033	0.23
P	rs16883089 (6:73658053)	S	T	-0.0034	0.19	S	T	-0.0232	1.18	S	C	-0.0131	0.88	S	T	-0.0195	1.34
P	rs2768551 (6:109270656)	UKB	A	-0.0310	1.73	UKB	A	-0.0484	2.45	UKB	A	-0.0095	0.64	UKB	A	-0.0238	1.63

		FEV_1				FEV_1/FVC				FVC				PEF			
Signal	**SNP (Chr:Pos)**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**
P	rs11759026 (6:126792095)	UKB	A	-0.0099	0.55	UKB	G	-0.0036	0.18	UKB	A	-0.0127	0.85	UKB	A	-0.0020	0.14
P	rs7753012 (6:142745883)	UKB	T	-0.0271	1.51	UKB	T	-0.0765	3.88	UKB	G	-0.0080	0.54	UKB	T	-0.0375	2.57
P	rs148274477 (6:142838173)	UKB	C	-0.0221	1.23	UKB	C	-0.1823	9.25	UKB	T	-0.0652	4.38	UKB	C	-0.0969	6.64
P	rs10246303 (7:7286445)	S	T	-0.0094	0.52	S	T	-0.0097	0.49	S	T	-0.0020	0.14	S	T	-0.0036	0.25
P	rs55905169 (7:15506529)	UKB	C	-0.0184	1.03	UKB	G	-0.0055	0.28	UKB	C	-0.0222	1.49	UKB	C	-0.0065	0.44
P	rs72615157 (7:99635967)	S	A	-0.0084	0.47	S	G	-0.0125	0.63	S	A	-0.0166	1.12	S	A	-0.0048	0.33
P	rs12698403 (7:156127246)	S	A	-0.0260	1.45	S	A	-0.0210	1.07	S	A	-0.0127	0.85	S	A	-0.0107	0.74
P	rs771924 (9:1555835)	UKB	A	-0.0159	0.89	UKB	A	-0.0026	0.13	UKB	A	-0.0167	1.12	UKB	A	-0.0078	0.53
P	rs7872188 (9:4124377)	S	T	-0.0207	1.15	S	T	-0.0195	0.99	S	T	-0.0094	0.63	S	T	-0.0267	1.83
P	rs10965947 (9:23588583)	UKB	C	-0.0173	0.96	UKB	C	-0.0243	1.23	UKB	C	-0.0074	0.49	UKB	C	-0.0192	1.31
P	rs16909859 (9:98204792)	UKB	G	-0.0067	0.38	UKB	A	-0.0441	2.24	UKB	G	-0.0273	1.83	UKB	A	-0.0100	0.68
P	rs2451951 (9:109496630)	UKB	C	-0.0071	0.39	UKB	C	-0.0243	1.23	UKB	T	-0.0040	0.27	UKB	C	-0.0098	0.67
P	rs803923 (9:119401650)	UKB	A	-0.0204	1.14	UKB	A	-0.0234	1.19	UKB	A	-0.0106	0.71	UKB	A	-0.0275	1.88
P	rs10858246 (9:139102831)	UKB	C	-0.0172	0.96	UKB	G	-0.0073	0.37	UKB	C	-0.0229	1.54	UKB	C	-0.0027	0.18
P	rs10870202 (9:139257411)	S	C	-0.0050	0.28	S	C	-0.0062	0.31	S	C	-0.0004	0.03	S	C	-0.0124	0.85
P	rs7090277 (10:12278021)	UKB	T	-0.0338	1.89	UKB	T	-0.0438	2.22	UKB	T	-0.0152	1.02	UKB	T	-0.0320	2.19
P	rs3847402 (10:30267810)	S	A	-0.0070	0.39	S	A	-0.0154	0.78	S	G	-0.0053	0.35	S	A	-0.0119	0.82
P	rs7899503 (10:65087468)	UKB	G	-0.0171	0.95	UKB	G	-0.0095	0.48	UKB	G	-0.0137	0.92	UKB	G	-0.0126	0.87
P	rs7095607 (10:69957350)	S	A	-0.0146	0.82	S	A	-0.0047	0.24	S	A	-0.0142	0.95	S	A	-0.0193	1.32
P	rs3849969 (10:75525999)	UKB	G	-0.0233	1.30	UKB	G	-0.0142	0.72	UKB	G	-0.0186	1.25	UKB	G	-0.0142	0.97
P	rs1259524 (10:77004644)	UKB	G	-0.0128	0.71	UKB	G	-0.0047	0.24	UKB	G	-0.0116	0.78	UKB	G	-0.0100	0.68
P	rs2637254 (10:78312002)	UKB	A	-0.0298	1.66	UKB	A	-0.0208	1.06	UKB	A	-0.0237	1.59	UKB	A	-0.0167	1.15
P	rs2256462 (10:81685593)	UKB	T	-0.0089	0.50	UKB	T	-0.0191	0.97	UKB	T	-0.0009	0.06	UKB	T	-0.0082	0.56
P	rs2293871 (10:124273671)	UKB	C	-0.0010	0.05	UKB	C	-0.0208	1.05	UKB	T	-0.0093	0.63	UKB	C	-0.0102	0.70
P	rs4237643 (11:43648368)	UKB	T	-0.0164	0.91	UKB	T	-0.0008	0.04	UKB	T	-0.0169	1.13	UKB	T	-0.0081	0.55
P	rs2863171 (11:45250732)	UKB	A	-0.0154	0.86	UKB	A	-0.0110	0.56	UKB	A	-0.0130	0.87	UKB	A	-0.0018	0.12
P	rs2509961 (11:62310909)	S	T	-0.0181	1.01	S	T	-0.0040	0.20	S	T	-0.0153	1.03	S	T	-0.0066	0.45
P	rs11235809 (11:73290163)	UKB	A	-0.0243	1.35	UKB	A	-0.0328	1.67	UKB	A	-0.0109	0.74	UKB	A	-0.0162	1.11
P	rs7108254 (11:86436086)	S	G	-0.0233	1.30	S	G	-0.0261	1.32	S	G	-0.0067	0.45	S	G	-0.0246	1.69
P	rs567508 (11:126008910)	S	G	-0.0124	0.69	S	G	-0.0284	1.44	S	A	-0.0033	0.22	S	G	-0.0155	1.06
P	rs2348418 (12:28689514)	UKB	C	-0.0319	1.78	UKB	T	-0.0028	0.14	UKB	C	-0.0369	2.48	UKB	C	-0.0155	1.07
P	rs772920 (12:56390364)	UKB	G	-0.0167	0.93	UKB	G	-0.0158	0.80	UKB	G	-0.0105	0.71	UKB	G	-0.0076	0.52
P	rs11172113 (12:57527283)	UKB	T	-0.0082	0.46	UKB	T	-0.0232	1.18	UKB	C	-0.0024	0.16	UKB	T	-0.0166	1.14
P	rs1494502 (12:65824670)	S	A	-0.0158	0.88	S	G	-0.0060	0.30	S	A	-0.0178	1.20	S	A	-0.0118	0.81

		FEV$_1$				FEV$_1$/FVC				FVC				PEF			
Signal	**SNP (Chr:Pos)**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**
P	rs7971039 (12:85724305)	UKB	G	-0.0134	0.75	UKB	G	-0.0036	0.18	UKB	G	-0.0118	0.80	UKB	G	-0.0202	1.38
P	rs11107184 (12:94184082)	UKB	C	-0.0104	0.58	UKB	C	-0.0015	0.07	UKB	C	-0.0109	0.73	UKB	C	-0.0001	0.01
P	rs113745635 (12:95554771)	S	T	-0.0061	0.34	S	T	-0.0169	0.86	S	C	-0.0072	0.48	S	C	-0.0084	0.57
P	rs12820313 (12:96255704)	UKB	C	-0.0099	0.55	UKB	C	-0.0427	2.17	UKB	T	-0.0097	0.65	UKB	C	-0.0237	1.62
P	rs10850377 (12:115201436)	UKB	G	-0.0174	0.97	UKB	G	-0.0100	0.51	UKB	G	-0.0143	0.96	UKB	G	-0.0032	0.22
P	rs35506 (12:115500691)	S	T	-0.0179	1.00	S	A	-0.0020	0.10	S	T	-0.0176	1.18	S	T	-0.0093	0.64
P	rs4444235 (14:54410919)	UKB	C	-0.0041	0.23	UKB	C	-0.0258	1.31	UKB	T	-0.0090	0.60	UKB	C	-0.0195	1.34
P	rs1698268 (14:84309664)	S	T	-0.0021	0.12	S	T	-0.0016	0.08	S	A	-0.0004	0.03	S	T	-0.0152	1.04
P	rs7155279 (14:92485881)	UKB	G	-0.0155	0.87	UKB	G	-0.0073	0.37	UKB	G	-0.0129	0.87	UKB	G	-0.0066	0.45
P	rs72699866 (14:93114787)	UKB	G	-0.0329	1.84	UKB	G	-0.0069	0.35	UKB	G	-0.0326	2.19	UKB	G	-0.0138	0.95
P	rs1200345 (15:41819716)	UKB	C	-0.0101	0.57	UKB	C	-0.0226	1.15	UKB	T	-0.0001	0.01	UKB	C	-0.0099	0.68
P	rs72724130 (15:41977690)	S	A	-0.0126	0.70	S	T	-0.0187	0.95	S	A	-0.0279	1.88	S	T	-0.0128	0.88
P	rs8025774 (15:67483276)	UKB	T	-0.0132	0.73	UKB	C	-0.0143	0.73	UKB	T	-0.0220	1.48	UKB	C	-0.0053	0.36
P	rs10851839 (15:71628370)	UKB	T	-0.0240	1.34	UKB	T	-0.0686	3.48	UKB	A	-0.0089	0.60	UKB	T	-0.0424	2.90
P	rs12591467 (15:71788387)	S	C	-0.0189	1.05	S	C	-0.0289	1.47	S	C	-0.0034	0.23	S	T	-0.0075	0.52
P	rs12438269 (15:84502549)	S	C	-0.0248	1.38	S	C	-0.0163	0.83	S	C	-0.0161	1.08	S	C	-0.0028	0.19
P	rs12149593 (16:10704535)	UKB	A	-0.0057	0.32	UKB	A	-0.0202	1.02	UKB	C	-0.0046	0.31	UKB	A	-0.0099	0.68
P	rs181206 (16:28513403)	UKB	G	-0.0122	0.68	UKB	G	-0.0043	0.22	UKB	G	-0.0110	0.74	UKB	G	-0.0037	0.25
P	rs12447804 (16:58075282)	UKB	T	-0.0085	0.47	UKB	T	-0.0374	1.90	UKB	C	-0.0090	0.61	UKB	T	-0.0107	0.73
P	rs3973397 (16:70040398)	UKB	A	-0.0142	0.79	UKB	A	-0.0056	0.28	UKB	A	-0.0129	0.87	UKB	A	-0.0063	0.43
P	rs3743609 (16:75467021)	UKB	C	-0.0244	1.36	UKB	C	-0.0396	2.01	UKB	C	-0.0064	0.43	UKB	C	-0.0323	2.21
P	rs1079572 (16:78187138)	UKB	A	-0.0094	0.53	UKB	G	-0.0023	0.12	UKB	A	-0.0110	0.74	UKB	A	-0.0037	0.25
P	rs62070270 (17:28263980)	S	G	-0.0152	0.85	S	G	-0.0198	1.00	S	G	-0.0037	0.25	S	G	-0.0146	1.00
P	rs62070631 (17:29087285)	UKB	G	-0.0233	1.30	UKB	G	-0.0044	0.22	UKB	G	-0.0253	1.70	UKB	G	-0.0163	1.12
P	rs11658500 (17:36886828)	S	A	-0.0188	1.05	S	A	-0.0178	0.90	S	A	-0.0080	0.54	S	A	-0.0099	0.68
P	rs8067511 (17:37611352)	UKB	C	-0.0151	0.84	UKB	T	-0.0010	0.05	UKB	C	-0.0166	1.12	UKB	C	-0.0046	0.32
P	rs17577877 (17:44208218)	S	G	-0.0268	1.50	S	A	-0.0004	0.02	S	G	-0.0296	1.99	S	G	-0.0113	0.78
P	rs6501431 (17:68976415)	UKB	C	-0.0089	0.50	UKB	T	-0.0114	0.58	UKB	C	-0.0159	1.07	UKB	T	-0.0045	0.31
P	rs1859962 (17:69108753)	UKB	T	-0.0242	1.35	UKB	T	-0.0128	0.65	UKB	T	-0.0200	1.34	UKB	T	-0.0247	1.69
P	rs7218675 (17:73513185)	S	A	-0.0085	0.47	S	A	-0.0106	0.54	S	A	-0.0010	0.07	S	C	-0.0028	0.19
P	rs633286 (18:8809273)	UKB	T	-0.0285	1.59	UKB	T	-0.0261	1.33	UKB	T	-0.0176	1.18	UKB	T	-0.0224	1.53
P	rs7243351 (18:20148531)	UKB	C	-0.0110	0.61	UKB	C	-0.0029	0.15	UKB	C	-0.0101	0.68	UKB	C	-0.0022	0.15
P	rs7238093 (18:20728158)	UKB	T	-0.0209	1.16	UKB	T	-0.0060	0.30	UKB	T	-0.0201	1.35	UKB	T	-0.0111	0.76
P	rs8089865 (18:50957922)	UKB	G	-0.0070	0.39	UKB	A	-0.0090	0.46	UKB	G	-0.0122	0.82	UKB	A	-0.0035	0.24

		FEV_1				FEV_1/FVC				FVC				PEF			
Signal	**SNP (Chr:Pos)**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**	**Source**	**Eff All**	**β**	**W**
P	rs9636166 (19:31829613)	UKB	C	-0.0242	1.35	UKB	C	-0.0399	2.03	UKB	C	-0.0064	0.43	UKB	C	-0.0188	1.29
P	rs113473882 (19:41124155)	UKB	T	-0.0319	1.78	UKB	T	-0.1569	7.96	UKB	C	-0.0412	2.77	UKB	T	-0.0735	5.04
P	rs6140050 (20:6632901)	S	C	-0.0099	0.55	S	A	-0.0048	0.24	S	C	-0.0125	0.84	S	C	-0.0139	0.95
P	rs6138639 (20:25669052)	UKB	G	-0.0215	1.20	UKB	G	-0.0163	0.83	UKB	G	-0.0160	1.07	UKB	G	-0.0179	1.23
P	rs1737889 (20:31042176)	UKB	T	-0.0173	0.96	UKB	T	-0.0190	0.96	UKB	T	-0.0097	0.65	UKB	T	-0.0133	0.91
P	rs6088813 (20:33975181)	UKB	C	-0.0195	1.09	UKB	A	-0.0057	0.29	UKB	C	-0.0250	1.68	UKB	A	-0.0146	1.00
P	rs2236519 (20:45529571)	UKB	A	-0.0157	0.88	UKB	G	-0.0057	0.29	UKB	A	-0.0201	1.35	UKB	A	-0.0102	0.70
P	rs6062304 (20:62351539)	S	T	-0.0249	1.39	S	T	-0.0026	0.13	S	T	-0.0254	1.71	S	T	-0.0039	0.27
P	rs2834440 (21:35690499)	UKB	A	-0.0007	0.04	UKB	G	-0.0243	1.23	UKB	A	-0.0127	0.86	UKB	G	-0.0098	0.67
P	rs11704827 (22:18450287)	S	A	-0.0270	1.51	S	A	-0.0192	0.97	S	A	-0.0147	0.99	S	A	-0.0147	1.01
P	rs4820216 (22:20854161)	UKB	T	-0.0148	0.83	UKB	T	-0.0194	0.98	UKB	T	-0.0057	0.38	UKB	T	-0.0139	0.95
P	rs2283847 (22:28181399)	S	T	-0.0172	0.96	S	T	-0.0171	0.87	S	T	-0.0082	0.55	S	T	-0.0235	1.61

G. The 14 sentinels identified as part of the genome-wide interaction analysis which were not independent across traits. Each row corresponds to the SNPs which are either the same or part of the same signal. The sentinel chosen to represent the signal was that which had the strongest interaction effect (smallest interaction effect p-value) and is highlighted in bold. Betas (β) and standard errors (SE) are on the inverse normalised scale. CAF = coded allele frequency, MAF = minor allele frequency, and INFO = Imputation score.

Trait	RSID (chr:pos)	Coded / Non-coded allele	CAF	MAF	INFO	β	SE	P value	Gene/closest gene
FEV_1	rs17704183 (4:168390470)	A/G	0.362	0.362	0.989	0.025	0.005	2.97×10^{-6}	*SPOCK3* (upstream)
FVC	**rs17704183 (4:168390470)**	**A/G**	**0.362**	**0.362**	**0.989**	**0.026**	**0.005**	$\mathbf{7.48 \times 10^{-7}}$	***SPOCK3* (upstream)**
FEV_1	**rs117585696 (14:58706512)**	**C/T**	**0.018**	**0.018**	**1.000**	**0.093**	**0.019**	$\mathbf{1.18 \times 10^{-6}}$	***ACTR10***
FVC	rs143449627 (14:58708312)	C/G	0.016	0.016	0.988	0.098	0.020	1.34×10^{-6}	*PSMA3* (upstream)
FEV_1	rs2009746 (15:78754102)	G/A	0.329	0.329	0.998	-0.044	0.005	2.61×10^{-16}	*IREB2*
FVC	rs2009746 (15:78754102)	G/A	0.329	0.329	0.998	-0.028	0.005	3.06×10^{-7}	*IREB2*
FEV_1/FVC	**rs8042849 (15:78817929)**	**T/C**	**0.659**	**0.341**	**0.999**	**0.049**	**0.005**	$\mathbf{7.25 \times 10^{-20}}$	***HYKK***

Trait	RSID (chr:pos)	Coded / Non-coded allele	CAF	MAF	INFO	β	SE	P value	Gene/closest gene
FEV_1/FVC	**rs7173514 (15:78849918)**	**T/C**	0.229	**0.229**	**0.991**	**0.054**	**0.006**	**4.62×10^{-19}**	***CHRNA3*** **(downstream)**
PEF	rs7173514 (15:78849918)	T/C	0.229	0.229	0.991	0.037	0.006	7.77×10^{-10}	*CHRNA3* (downstream)
FEV_1	rs8042059 (15:78907859)	C/A	0.229	0.229	1.000	0.045	0.006	9.26×10^{-14}	*CHRNA3*
FEV_1	rs9930741 (16:20695486)	T/C	0.410	0.410	0.994	-0.024	0.005	4.80×10^{-6}	*ACSM1*
FVC	**rs4783512 (16:20937848)**	**C/A**	0.625	**0.375**	**0.982**	**-0.025**	**0.005**	**2.78×10^{-6}**	***LYRM1*** **(downstream)**
FEV_1	**rs35524777 (18:11424551)**	**T/C**	0.062	**0.062**	**0.953**	**-0.056**	**0.011**	**1.44×10^{-7}**	***PIEZO2*** **(upstream)**
FVC	rs35524777 (18:11424551)	T/C	0.062	0.062	0.953	-0.051	0.011	2.75×10^{-6}	*PIEZO2* (upstream)

H. Ever/never smoker stratified results for the 58 signals independent both within and across the four analysed traits: FEV_1, FEV_1/FVC, FVC and PEF. Betas (β) and standard errors (SE) are on the inverse normalised scale. CAF – coded allele frequency and MAF – minor allele frequency.

			Ever smoker					Never smoker				
Trait	**RSID (chr:pos)**	**Eff allele / Ref allele**	**CAF**	**MAF**	**β**	**SE**	**P value**	**CAF**	**MAF**	**β**	**SE**	**P value**
FEV_1/FVC	rs61787074 (1:27851315)	T/C	0.025	0.025	0.045	0.013	4.20×10^{-4}	0.026	0.026	-0.041	0.012	5.15×10^{-4}
PEF	rs12077425 (1:29744567)	T/C	0.006	0.006	-0.088	0.025	3.40×10^{-4}	0.006	0.006	0.075	0.023	1.05×10^{-3}
FEV_1/FVC	rs116757305 (1:43876305)	T/C	0.023	0.023	-0.040	0.013	1.73×10^{-3}	0.023	0.023	0.043	0.012	3.37×10^{-4}
PEF	rs72772245 (2:1359174)	C/T	0.008	0.008	-0.071	0.021	9.53×10^{-4}	0.009	0.009	0.068	0.020	5.64×10^{-4}
FVC	rs146541032 (2:62358665)	C/T	0.006	0.006	-0.072	0.025	3.98×10^{-3}	0.006	0.006	0.096	0.024	5.74×10^{-5}
FEV_1	rs142312019(2:153914485)	T/C	0.006	0.006	-0.091	0.025	2.92×10^{-4}	0.006	0.006	0.062	0.023	7.38×10^{-3}
FEV_1	rs10497204(2:159950898)	T/C	0.485	0.485	0.011	0.004	4.84×10^{-3}	0.485	0.485	-0.012	0.003	3.70×10^{-4}
FVC	rs6773439(3:15530595)	A/G	0.173	0.173	-0.019	0.005	1.70×10^{-4}	0.172	0.172	0.012	0.005	8.47×10^{-3}
PEF	rs113246660(3:30189600)	C/T	0.009	0.009	-0.074	0.022	7.48×10^{-4}	0.008	0.008	0.069	0.021	8.65×10^{-4}
FEV_1	rs139977403(3:76755996)	A/C	0.009	0.009	0.083	0.021	5.55×10^{-5}	0.009	0.009	-0.055	0.020	4.72×10^{-3}
FEV_1/FVC	rs116799787(3:180022491)	T/C	0.009	0.009	0.067	0.021	1.15×10^{-3}	0.009	0.009	-0.064	0.019	8.97×10^{-4}
FEV_1/FVC	rs74376726(3:182694981)	A/G	0.059	0.059	0.030	0.008	1.79×10^{-4}	0.060	0.060	-0.019	0.007	9.28×10^{-3}
FEV_1	rs74823357(3:193713123)	T/C	0.041	0.041	0.029	0.009	1.91×10^{-3}	0.041	0.041	-0.028	0.009	1.40×10^{-3}
FVC	rs17704183(4:168390470)	A/G	0.363	0.363	0.017	0.004	2.30×10^{-5}	0.361	0.361	-0.010	0.004	7.61×10^{-3}
FEV_1/FVC	rs115906789(5:5809051)	A/G	0.029	0.029	0.039	0.012	8.19×10^{-4}	0.030	0.030	-0.034	0.011	1.25×10^{-3}
PEF	rs55730263(5:83166632)	G/A	0.154	0.154	0.016	0.005	1.91×10^{-3}	0.154	0.154	-0.018	0.005	3.07×10^{-4}
FEV_1	rs146549495(5:93102443)	T/C	0.016	0.016	0.057	0.015	1.36×10^{-4}	0.017	0.017	-0.034	0.014	1.38×10^{-2}
FVC	rs1593464(5:94360291)	A/G	0.879	0.121	-0.019	0.006	1.05×10^{-3}	0.879	0.121	0.016	0.005	3.66×10^{-3}
PEF	rs76412370(5:118115062)	T/G	0.118	0.118	0.015	0.006	1.07×10^{-2}	0.119	0.119	-0.020	0.005	1.79×10^{-4}
FVC	rs7728169(5:155340750)	T/A	0.036	0.036	-0.035	0.010	4.85×10^{-4}	0.036	0.036	0.029	0.009	2.23×10^{-3}
FEV_1/FVC	rs11969624(6:5460085)	C/G	0.117	0.117	0.020	0.006	8.57×10^{-4}	0.117	0.117	-0.019	0.006	8.41×10^{-4}
FEV_1/FVC	rs76004091(6:111716001)	C/G	0.014	0.014	0.056	0.016	6.32×10^{-4}	0.014	0.014	-0.039	0.015	1.13×10^{-2}

			Ever smoker					Never smoker				
Trait	**RSID (chr:pos)**	**Eff allele / Ref allele**	**CAF**	**MAF**	**β**	**SE**	**P value**	**CAF**	**MAF**	**β**	**SE**	**P value**
FEV_1	rs147414811(6:141243046)	C/T	0.025	0.025	0.037	0.012	2.43×10^{-3}	0.024	0.024	-0.042	0.011	2.23×10^{-4}
FEV_1/FVC	rs35535406(7:31427263)	C/T	0.237	0.237	0.016	0.004	1.83×10^{-4}	0.239	0.239	-0.011	0.004	8.90×10^{-3}
FEV_1	rs34154123(7:37786836)	A/T	0.104	0.104	-0.020	0.007	3.67×10^{-3}	0.103	0.103	0.024	0.006	1.15×10^{-4}
PEF	rs73139542(7:63325787)	A/G	0.046	0.046	-0.050	0.009	2.46×10^{-8}	0.045	0.045	0.008	0.008	3.27×10^{-1}
FVC	rs61244245(7:68595239)	T/C	0.406	0.406	-0.011	0.004	2.94×10^{-3}	0.405	0.405	0.013	0.004	2.93×10^{-4}
FVC	rs186074884(7:125325472)	A/G	0.006	0.006	0.111	0.026	1.93×10^{-5}	0.007	0.007	-0.059	0.024	1.21×10^{-2}
FEV_1/FVC	rs7817569(8:25461315)	T/C	0.713	0.287	-0.022	0.004	1.30×10^{-7}	0.715	0.285	0.005	0.004	2.33×10^{-1}
FEV_1/FVC	rs11780592(8:27418747)	G/A	0.176	0.176	0.021	0.005	2.09×10^{-5}	0.184	0.184	-0.010	0.004	2.16×10^{-2}
FVC	rs77608508(9:120844163)	G/A	0.076	0.076	-0.032	0.007	8.59×10^{-6}	0.075	0.075	0.016	0.007	1.81×10^{-2}
FEV_1/FVC	rs73555789(9:136456673)	A/G	0.040	0.040	0.043	0.010	5.66×10^{-6}	0.040	0.040	-0.014	0.009	1.18×10^{-1}
PEF	rs61853871(10:73325318)	T/G	0.075	0.075	0.036	0.007	5.47×10^{-7}	0.075	0.075	-0.008	0.007	2.07×10^{-1}
FVC	rs4962379(10:126228742)	C/T	0.874	0.126	-0.023	0.006	3.36×10^{-5}	0.873	0.127	0.011	0.005	3.26×10^{-2}
FEV_1/FVC	rs142254414(11:58976080)	A/G	0.011	0.011	-0.052	0.019	5.60×10^{-3}	0.010	0.010	0.064	0.018	3.53×10^{-4}
PEF	rs553187851(11:117448580)	T/C	0.007	0.007	0.088	0.025	3.37×10^{-4}	0.006	0.006	-0.071	0.023	2.30×10^{-3}
FVC	rs117575177(12:65696432)	G/A	0.022	0.022	0.049	0.013	1.58×10^{-4}	0.022	0.022	-0.034	0.012	4.90×10^{-3}
FEV_1/FVC	rs117367754(12:67751257)	A/G	0.017	0.017	-0.062	0.014	1.57×10^{-5}	0.018	0.018	0.031	0.013	1.89×10^{-2}
FEV_1/FVC	rs10862408(12:82359619)	C/G	0.275	0.275	0.013	0.004	1.49×10^{-3}	0.274	0.274	-0.016	0.004	3.20×10^{-5}
FEV_1/FVC	rs142704172(12:122496894)	T/C	0.008	0.008	-0.076	0.022	4.36×10^{-4}	0.009	0.009	0.068	0.019	4.48×10^{-4}
FEV_1	rs17120700(14:49349224)	A/G	0.645	0.355	-0.010	0.004	1.18×10^{-2}	0.647	0.352	0.015	0.004	6.73×10^{-5}
FEV_1	rs117585696(14:58706512)	C/T	0.018	0.018	0.039	0.014	5.57×10^{-3}	0.018	0.018	-0.052	0.013	7.89×10^{-5}
PEF	rs61183515(15:55175622)	C/A	0.197	0.197	-0.016	0.005	4.71×10^{-4}	0.197	0.197	0.012	0.004	5.34×10^{-3}
FEV_1/FVC	rs8042849(15:78817929)	T/C	0.662	0.338	0.040	0.004	8.06×10^{-24}	0.657	0.343	-0.006	0.004	1.29×10^{-1}
FEV_1/FVC	rs7173514(15:78849918)	T/C	0.228	0.228	0.040	0.004	3.46×10^{-19}	0.229	0.229	-0.011	0.004	1.09×10^{-2}
FEV_1/FVC	rs62023825(16:11506666)	T/C	0.162	0.162	-0.020	0.005	1.42×10^{-4}	0.161	0.161	0.012	0.005	1.26×10^{-2}
FVC	rs4783512(16:20937848)	C/A	0.626	0.374	-0.015	0.004	8.05×10^{-5}	0.625	0.375	0.009	0.004	1.24×10^{-2}
PEF	rs539865765(16:51905937)	G/C	0.005	0.005	0.097	0.027	3.76×10^{-4}	0.005	0.005	-0.078	0.025	1.81×10^{-3}
FEV_1/FVC	rs80277243(16:52793919)	A/G	0.011	0.011	0.048	0.018	8.95×10^{-3}	0.011	0.011	-0.076	0.017	1.20×10^{-5}
FVC	rs148424048(17:5086198)	A/G	0.013	0.013	-0.080	0.017	2.02×10^{-6}	0.012	0.012	0.022	0.016	1.63×10^{-1}
PEF	rs4796410(17:7273247)	G/A	0.119	0.119	0.023	0.006	1.03×10^{-4}	0.118	0.118	-0.014	0.005	1.22×10^{-2}

			Ever smoker					Never smoker				
Trait	**RSID (chr:pos)**	**Eff allele / Ref allele**	**CAF**	**MAF**	**β**	**SE**	**P value**	**CAF**	**MAF**	**β**	**SE**	**P value**
FVC	rs12452505(17:63556402)	G/C	0.143	0.143	0.017	0.005	1.32E-03	0.141	0.141	-0.021	0.005	2.23×10^{-5}
FEV_1	rs35524777(18:11424551)	T/C	0.061	0.061	-0.034	0.008	1.86E-05	0.063	0.063	0.021	0.007	4.83×10^{-3}
PEF	rs151310656(19:16777665)	A/G	0.029	0.029	0.025	0.012	3.07E-02	0.029	0.029	-0.053	0.011	1.52×10^{-6}
FEV_1/FVC	rs67134151(19:29510958)	G/T	0.071	0.071	0.028	0.007	1.37E-04	0.070	0.070	-0.017	0.007	1.24×10^{-2}
FEV_1	rs2604894(19:41292404)	G/A	0.550	0.450	-0.021	0.004	3.78E-08	0.452	0.452	0.004	0.004	2.94×10^{-1}
FVC	rs28650353(21:35166202)	A/G	0.014	0.014	0.038	0.016	2.07E-02	0.014	0.014	-0.070	0.015	3.38×10^{-6}
FEV_1	rs2225434(21:46634146)	C/T	0.569	0.431	-0.011	0.004	2.18E-03	0.567	0.433	0.013	0.004	1.91×10^{-4}

I. Smoking look up for the 58 interaction signals independent both within and across the four analysed traits: FEV_1, FEV_1/FVC, FVC and PEF. Betas (β) and standard errors (SE) are on the inverse normalised scale. CAF – coded allele frequency, MAF – minor allele frequency, and INFO = imputation score.

RSID (Chr:Pos)	Trait	Coded / Non-coded allele	CAF	MAF	INFO	β	SE	P value
rs61787074(1:27851315)	SI	C/T	0.974	0.026	0.848	0.008	0.009	3.70×10^{-1}
	SC	C/T	0.974	0.026	0.848	0.002	0.014	8.90×10^{-1}
	CPD	C/T	0.974	0.026	0.848	-0.043	0.017	1.20×10^{-2}
rs12077425(1:29744567)	SI	C/T	0.994	0.006	0.971	-0.021	0.018	2.50×10^{-1}
	SC	C/T	0.994	0.006	0.971	0.026	0.026	3.20×10^{-1}
	CPD	C/T	0.994	0.006	0.971	-0.006	0.034	8.50×10^{-1}
rs116757305(1:43876305)	SI	C/T	0.977	0.023	0.954	-0.022	0.009	1.50×10^{-2}
	SC	C/T	0.977	0.023	0.954	0.010	0.014	4.80×10^{-1}
	CPD	C/T	0.977	0.023	0.954	-0.019	0.017	2.70×10^{-1}
rs72772245(2:1359174)	SI	T/C	0.992	0.008	0.911	0.011	0.015	4.80×10^{-1}
	SC	T/C	0.992	0.008	0.911	0.033	0.023	1.50×10^{-1}
	CPD	T/C	0.992	0.008	0.911	0.035	0.029	2.20×10^{-1}
rs146541032(2:62358665)	SI	T/C	0.994	0.006	0.885	-0.044	0.018	1.50×10^{-2}
	SC	T/C	0.994	0.006	0.885	-0.003	0.027	9.00×10^{-1}
	CPD	T/C	0.994	0.006	0.885	-0.024	0.034	4.70×10^{-1}
rs142312019(2:153914485)	SI	C/T	0.994	0.006	0.895	0.008	0.018	6.50×10^{-1}
	SC	C/T	0.994	0.006	0.895	0.017	0.027	5.10×10^{-1}
	CPD	C/T	0.994	0.006	0.895	-0.024	0.033	4.70×10^{-1}
rs10497204(2:159950898)	SI	C/T	0.514	0.486	0.987	0.002	0.003	4.40×10^{-1}
	SC	C/T	0.514	0.486	0.987	0.001	0.004	7.90×10^{-1}
	CPD	C/T	0.516	0.484	0.987	-0.002	0.005	7.20×10^{-1}
rs6773439(3:15530595)	SI	G/A	0.828	0.172	0.975	0.004	0.004	2.80×10^{-1}
	SC	G/A	0.827	0.173	0.975	-0.003	0.005	5.50×10^{-1}
	CPD	G/A	0.827	0.173	0.975	0.000	0.007	9.60×10^{-1}
rs113246660(3:30189600)	SI	T/C	0.991	0.009	0.837	-0.014	0.015	3.70×10^{-1}

RSID (Chr:Pos)	Trait	Coded / Non-coded allele	CAF	MAF	INFO	β	SE	P value
	SC	T/C	0.991	0.009	0.837	-0.016	0.023	4.80×10^{-1}
	CPD	T/C	0.991	0.009	0.837	-0.014	0.028	6.20×10^{-1}
rs139977403(3:76755996)	SI	C/A	0.991	0.009	0.874	0.007	0.015	6.40×10^{-1}
	SC	C/A	0.991	0.009	0.874	0.018	0.022	4.30×10^{-1}
	CPD	C/A	0.991	0.009	0.874	0.018	0.028	5.20×10^{-1}
rs116799787(3:180022491)	SI	C/T	0.991	0.009	0.941	-0.032	0.015	2.80×10^{-1}
	SC	C/T	0.991	0.009	0.941	0.008	0.022	7.10×10^{-1}
	CPD	C/T	0.991	0.009	0.941	0.058	0.027	3.40×10^{-2}
rs74376726(3:182694981)	SI	G/A	0.941	0.059	0.969	0.008	0.006	1.60×10^{-1}
	SC	G/A	0.942	0.058	0.969	0.007	0.009	3.90×10^{-1}
	CPD	G/A	0.942	0.058	0.969	-0.006	0.011	6.00×10^{-1}
rs74823357(3:193713123)	SI	C/T	0.959	0.041	1.000	0.003	0.007	7.10×10^{-1}
	SC	C/T	0.959	0.041	1.000	-0.002	0.010	8.50×10^{-1}
	CPD	C/T	0.959	0.041	1.000	0.000	0.013	9.90×10^{-1}
rs17704183(4:168390470)	SI	G/A	0.638	0.362	0.989	-0.006	0.003	4.50×10^{-2}
	SC	G/A	0.637	0.363	0.989	0.004	0.004	3.60×10^{-1}
	CPD	G/A	0.638	0.362	0.989	0.000	0.005	9.30×10^{-1}
rs115906789(5:5809051)	SI	G/A	0.971	0.029	0.911	0.007	0.008	4.30×10^{-1}
	SC	G/A	0.971	0.029	0.911	0.006	0.012	6.30×10^{-1}
	CPD	G/A	0.971	0.029	0.911	-0.005	0.016	7.70×10^{-1}
rs55730263(5:83166632)	SI	A/G	0.847	0.153	0.965	0.001	0.004	7.40×10^{-1}
	SC	A/G	0.848	0.152	0.965	0.007	0.006	2.20×10^{-1}
	CPD	A/G	0.847	0.153	0.965	0.007	0.007	3.40×10^{-1}
rs146549495(5:93102443)	SI	C/T	0.984	0.016	0.967	0.007	0.011	5.20×10^{-1}
	SC	C/T	0.984	0.016	0.967	0.015	0.016	3.50×10^{-1}
	CPD	C/T	0.984	0.016	0.967	0.014	0.020	4.70×10^{-1}
rs1593464(5:94360291)	SI	G/A	0.120	0.120	0.995	0.002	0.004	6.60×10^{-1}
	SC	G/A	0.120	0.120	0.995	-0.003	0.006	5.70×10^{-1}
	CPD	G/A	0.119	0.119	0.995	0.000	0.008	9.50×10^{-1}
rs76412370(5:118115062)	SI	G/T	0.881	0.119	0.963	-0.002	0.004	7.00×10^{-1}
	SC	G/T	0.881	0.119	0.963	-0.011	0.006	6.90×10^{-2}
	CPD	G/T	0.882	0.118	0.963	-0.022	0.008	5.50×10^{-3}

RSID (Chr:Pos)	Trait	Coded / Non-coded allele	CAF	MAF	INFO	β	SE	P value
rs7728169(5:155340750)	SI	A/T	0.964	0.036	0.971	-0.003	0.007	6.40×10^{-1}
	SC	A/T	0.964	0.036	0.971	-0.004	0.011	7.00×10^{-1}
	CPD	A/T	0.964	0.036	0.971	-0.015	0.014	2.50×10^{-1}
rs11969624(6:5460085)	SI	G/C	0.884	0.116	0.918	0.005	0.004	2.70×10^{-1}
	SC	G/C	0.884	0.116	0.918	0.011	0.006	9.40×10^{-2}
	CPD	G/C	0.884	0.116	0.918	-0.011	0.008	1.80×10^{-1}
rs76004091(6:111716001)	SI	G/C	0.986	0.014	0.928	-0.023	0.012	4.70×10^{-2}
	SC	G/C	0.986	0.014	0.928	0.006	0.017	7.10×10^{-1}
	CPD	G/C	0.986	0.014	0.928	0.005	0.022	8.30×10^{-1}
rs147414811(6:141243046)	SI	T/C	0.976	0.024	0.979	-0.006	0.009	5.00×10^{-1}
	SC	T/C	0.976	0.024	0.979	0.031	0.013	1.80×10^{-2}
	CPD	T/C	0.976	0.024	0.979	0.017	0.017	3.00×10^{-1}
rs35535406(7:31427263)	SI	T/C	0.762	0.238	0.997	0.002	0.003	5.30×10^{-1}
	SC	T/C	0.762	0.238	0.997	-0.002	0.005	7.40×10^{-1}
	CPD	T/C	0.762	0.238	0.997	0.000	0.006	9.80×10^{-1}
rs34154123(7:37786836)	SI	T/A	0.897	0.103	0.809	-0.010	0.005	4.40×10^{-2}
	SC	T/A	0.896	0.104	0.809	0.018	0.007	1.30×10^{-2}
	CPD	T/A	0.897	0.103	0.809	0.006	0.009	5.30×10^{-1}
rs73139542(7:63325787)	SI	G/A	0.954	0.046	0.972	-0.006	0.006	3.40×10^{-1}
	SC	G/A	0.954	0.046	0.972	0.002	0.010	8.60×10^{-1}
	CPD	G/A	0.954	0.046	0.972	0.007	0.012	5.40×10^{-1}
rs61244245(7:68595239)	SI	C/T	0.596	0.404	0.991	-0.002	0.003	5.00×10^{-1}
	SC	C/T	0.595	0.405	0.991	-0.002	0.004	5.50×10^{-1}
	CPD	C/T	0.594	0.406	0.991	0.008	0.005	9.70×10^{-2}
rs186074884(7:125325472)	SI	G/A	0.994	0.006	0.840	-0.005	0.018	7.80×10^{-1}
	SC	G/A	0.994	0.006	0.840	0.024	0.027	3.70×10^{-1}
	CPD	G/A	0.994	0.006	0.840	-0.041	0.034	2.30×10^{-1}
rs7817569(8:25461315)	SI	C/T	0.286	0.286	0.972	0.002	0.003	6.00×10^{-1}
	SC	C/T	0.287	0.287	0.972	-0.001	0.004	8.60×10^{-1}
	CPD	C/T	0.288	0.288	0.972	-0.006	0.006	2.90×10^{-1}
rs11780592(8:27418747)	SI	A/G	0.819	0.181	0.997	0.020	0.003	7.80×10^{-9}
	SC	A/G	0.823	0.177	0.997	0.013	0.005	1.60×10^{-1}

RSID (Chr:Pos)	Trait	Coded / Non-coded allele	CAF	MAF	INFO	β	SE	P value
	CPD	A/G	0.824	0.176	0.997	-0.017	0.007	8.40×10^{-3}
rs77608508(9:120844163)	SI	A/G	0.924	0.076	0.983	0.009	0.005	7.00×10^{-2}
	SC	A/G	0.924	0.076	0.983	-0.001	0.008	8.40×10^{-1}
	CPD	A/G	0.923	0.077	0.983	0.000	0.009	9.80×10^{-1}
rs73555789(9:136456673)	SI	G/A	0.960	0.040	0.983	0.004	0.007	5.90×10^{-1}
	SC	G/A	0.960	0.040	0.983	0.034	0.010	7.90×10^{-4}
	CPD	G/A	0.961	0.039	0.983	0.037	0.013	4.50×10^{-3}
rs61853871(10:73325318)	SI	G/T	0.925	0.075	0.984	-0.003	0.005	5.60×10^{-1}
	SC	G/T	0.925	0.075	0.984	0.006	0.008	4.60×10^{-1}
	CPD	G/T	0.925	0.075	0.984	0.007	0.010	4.70×10^{-1}
rs4962379(10:126228742)	SI	T/C	0.127	0.127	0.989	-0.001	0.004	8.00×10^{-1}
	SC	T/C	0.126	0.126	0.989	0.004	0.006	5.10×10^{-1}
	CPD	T/C	0.125	0.125	0.989	0.007	0.008	$3. \times 10^{-1}$
rs142254414(11:58976080)	SI	G/A	0.989	0.011	0.910	0.018	0.014	1.70×10^{-1}
	SC	G/A	0.989	0.011	0.910	-0.023	0.020	2.60×10^{-1}
	CPD	G/A	0.989	0.011	0.910	0.000	0.025	9.90×10^{-1}
rs553187851(11:117448580)	SI	C/T	0.993	0.007	0.875	-0.024	0.018	1.80×10^{-1}
	SC	C/T	0.993	0.007	0.875	0.025	0.026	3.30×10^{-1}
	CPD	C/T	0.994	0.006	0.875	-0.033	0.033	3.30×10^{-1}
rs117575177(12:65696432)	SI	A/G	0.978	0.022	0.966	0.008	0.009	3.90×10^{-1}
	SC	A/G	0.978	0.022	0.966	0.025	0.014	7.10×10^{-2}
	CPD	A/G	0.978	0.022	0.966	-0.007	0.017	6.90×10^{-1}
rs117367754(12:67751257)	SI	G/A	0.983	0.017	0.993	0.005	0.010	6.40×10^{-1}
	SC	G/A	0.983	0.017	0.993	-0.009	0.015	5.80×10^{-1}
	CPD	G/A	0.982	0.018	0.993	-0.003	0.019	8.60×10^{-1}
rs10862408(12:82359619)	SI	G/C	0.725	0.275	0.992	-0.001	0.003	7.10×10^{-1}
	SC	G/C	0.725	0.275	0.992	0.008	0.004	6.90×10^{-2}
	CPD	G/C	0.725	0.275	0.992	-0.001	0.006	9.20×10^{-1}
rs142704172(12:122496894)	SI	C/T	0.992	0.008	0.919	0.029	0.015	5.90×10^{-2}
	SC	C/T	0.992	0.008	0.919	-0.025	0.023	2.80×10^{-1}
	CPD	C/T	0.992	0.008	0.919	-0.026	0.029	3.80×10^{-1}
rs17120700(14:49349224)	SI	G/A	0.351	0.351	0.987	0.005	0.003	1.00×10^{-1}

RSID (Chr:Pos)	Trait	Coded / Non-coded allele	CAF	MAF	INFO	β	SE	P value
	SC	G/A	0.352	0.352	0.987	-0.001	0.004	7.70×10^{-1}
	CPD	G/A	0.351	0.351	0.987	0.002	0.005	6.40×10^{-1}
rs117585696(14:58706512)	SI	T/C	0.982	0.018	1.000	-0.005	0.010	6.10×10^{-1}
	SC	T/C	0.981	0.019	1.000	0.012	0.015	4.00×10^{-1}
	CPD	T/C	0.982	0.018	1.000	0.024	0.019	2.10×10^{-1}
rs61183515(15:55175622)	SI	A/C	0.803	0.197	0.995	0.004	0.003	2.60×10^{-1}
	SC	A/C	0.802	0.198	0.995	-0.005	0.005	3.40×10^{-1}
	CPD	A/C	0.801	0.199	0.995	-0.002	0.006	7.60×10^{-1}
rs8042849(15:78817929)	SI	C/T	0.339	0.339	0.999	-0.011	0.003	9.60×10^{-5}
	SC	C/T	0.335	0.335	0.999	-0.001	0.004	8.70×10^{-1}
	CPD	C/T	0.343	0.343	0.999	0.098	0.005	1.10×10^{-78}
rs7173514(15:78849918)	SI	C/T	0.771	0.229	0.991	-0.001	0.003	8.10×10^{-1}
	SC	C/T	0.770	0.230	0.991	0.007	0.005	1.40×10^{-1}
	CPD	C/T	0.778	0.222	0.991	0.076	0.006	1.50×10^{-36}
rs62023825(16:11506666)	SI	C/T	0.839	0.161	0.959	-0.007	0.004	5.70×10^{-2}
	SC	C/T	0.837	0.163	0.959	0.007	0.005	1.70×10^{-1}
	CPD	C/T	0.837	0.163	0.959	-0.011	0.007	1.30×10^{-1}
rs4783512(16:20937848)	SI	A/C	0.375	0.375	0.982	0.003	0.003	2.70×10^{-1}
	SC	A/C	0.376	0.376	0.982	-0.005	0.004	2.10×10^{-1}
	CPD	A/C	0.375	0.375	0.982	-0.002	0.005	6.90×10^{-1}
rs539865765(16:51905937)	SI	C/G	0.995	0.005	0.896	-0.002	0.019	9.00×10^{-1}
	SC	C/G	0.995	0.005	0.896	0.037	0.029	2.00×10^{-1}
	CPD	C/G	0.995	0.005	0.896	0.029	0.036	4.30×10^{-1}
rs80277243(16:52793919)	SI	G/A	0.989	0.011	0.902	0.006	0.013	6.50×10^{-1}
	SC	G/A	0.989	0.011	0.902	0.018	0.020	3.70×10^{-1}
	CPD	G/A	0.989	0.011	0.902	-0.044	0.025	7.30×10^{-2}
rs148424048(17:5086198)	SI	G/A	0.987	0.013	1.000	0.000	0.012	9.80×10^{-1}
	SC	G/A	0.987	0.013	1.000	-0.009	0.018	5.90×10^{-1}
	CPD	G/A	0.987	0.013	1.000	-0.013	0.022	5.60×10^{-1}
rs4796410(17:7273247)	SI	A/G	0.883	0.117	0.988	0.003	0.004	5.40×10^{-1}
	SC	A/G	0.883	0.117	0.988	0.005	0.006	4.20×10^{-1}
	CPD	A/G	0.882	0.118	0.988	-0.009	0.008	2.40×10^{-1}

RSID (Chr:Pos)	Trait	Coded / Non-coded allele	CAF	MAF	INFO	β	SE	P value
rs12452505(17:63556402)	SI	C/G	0.859	0.141	0.992	-0.009	0.004	2.10×10^{-2}
	SC	C/G	0.857	0.143	0.992	-0.003	0.006	5.40×10^{-1}
	CPD	C/G	0.857	0.143	0.992	0.005	0.007	4.90×10^{-1}
rs35524777(18:11424551)	SI	C/T	0.938	0.062	0.953	0.012	0.006	3.20×10^{-2}
	SC	C/T	0.939	0.061	0.953	0.000	0.009	9.70×10^{-1}
	CPD	C/T	0.939	0.061	0.953	0.003	0.011	7.70×10^{-1}
rs151310656(19:16777665)	SI	G/A	0.971	0.029	0.878	0.011	0.008	1.80×10^{-1}
	SC	G/A	0.971	0.029	0.878	0.007	0.012	5.80×10^{-1}
	CPD	G/A	0.971	0.029	0.878	0.027	0.016	8.90×10^{-2}
rs67134151(19:29510958)	SI	T/G	0.930	0.070	0.991	0.004	0.005	4.90×10^{-1}
	SC	T/G	0.930	0.070	0.991	0.010	0.008	2.00×10^{-1}
	CPD	T/G	0.930	0.070	0.991	-0.013	0.010	2.00×10^{-1}
rs2604894(19:41292404)	SI	A/G	0.452	0.452	0.983	-0.008	0.003	2.20×10^{-3}
	SC	A/G	0.451	0.451	0.983	0.011	0.004	4.40×10^{-3}
	CPD	A/G	0.453	0.453	0.983	-0.035	0.005	3.00×10^{-12}
rs28650353(21:35166202)	SI	G/A	0.986	0.014	0.958	0.004	0.012	7.00×10^{-1}
	SC	G/A	0.986	0.014	0.958	0.008	0.017	6.40×10^{-1}
	CPD	G/A	0.986	0.014	0.958	0.007	0.022	7.60×10^{-1}
rs2225434(21:46634146)	SI	T/C	0.432	0.432	0.995	-0.005	0.003	7.00×10^{-2}
	SC	T/C	0.431	0.431	0.995	0.002	0.004	6.90×10^{-1}
	CPD	T/C	0.428	0.428	0.995	-0.011	0.005	3.30×10^{-2}

New genetic signals for lung function highlight pathways and pleiotropy, and chronic obstructive pulmonary disease associations across multiple ancestries.

Authors:

Nick Shrine*[1]; Anna L Guyatt*[1]; A Mesut Erzurumluoglu*[1]; Victoria E Jackson[1,2,3]; Brian D Hobbs[4,5]; Carl Melbourne[1]; Chiara Batini[1]; Katherine A Fawcett[1]; Kijoung Song[6]; Phuwanat Sakornsakolpat[4,7]; Xingnan Li[8]; Ruth Boxall[9,10]; Nicola F Reeve[1]; Ma'en Obeidat[11]; Jing Hua Zhao[12]; Matthias Wielscher[13]; Understanding Society Scientific Group[14]; Stefan Weiss[15]; Katherine A Kentistou[16,17]; James P Cook[18]; Benjamin B Sun[19]; Jian Zhou[20]; Jennie Hui[21,22,23,24]; Stefan Karrasch[25,26,27]; Medea Imboden[28,29]; Sarah E Harris[30,61]; Jonathan Marten[32]; Stefan Enroth[33]; Shona M Kerr[32]; Ida Surakka[34,35]; Veronique Vitart[32]; Terho Lehtimäki[36]; Richard J Allen[1]; Per S Bakke[37]; Terri H Beaty[38]; Eugene R Bleecker[8]; Yohan Bossé[39,40]; Corry-Anke Brandsma[41]; Zhengming Chen[9]; James D Crapo[42,43]; John Danesh[19,44,45,46]; Dawn L DeMeo[4]; Frank Dudbridge[1]; Ralf Ewert[47]; Christian Gieger[48]; Amund Gulsvik[37]; Anna L Hansell[49,50]; Ke Hao[51]; Josh D Hoffman[6]; John Hokanson[52]; Georg Homuth[15]; Peter K Joshi[16]; Philippe Joubert[40,53]; Claudia Langenberg[12]; Xuan Li[11]; Liming Li[54]; Kuang Lin[9]; Lars Lind[55]; Nick Locantore[56]; Jian'an Luan[12]; Anubha Mahajan[57]; Joseph C Maranville[58]; Alison Murray[59]; David C Nickle[60]; Richard Packer[1]; Margaret M Parker[4]; Megan L Paynton[1]; David Porteous[61,30]; Dmitry Prokopenko[4]; Dandi Qiao[4]; Rajesh Rawal[48]; Heiko Runz[58]; Ian Sayers[63]; Don D Sin[11,64]; Blair H Smith[65]; María Soler Artigas[66,67,68]; David Sparrow[69,70]; Ruth Tal-Singer[56]; Paul RHJ Timmers[16]; Maarten Van den Berge[71]; John C Whittaker[72]; Prescott Woodruff[73]; Laura M Yerges Armstrong[6]; Olga G Troyanskaya[74,75]; Olli T Raitakari[76,77]; Mika Kähönen[78]; Ozren Polasek[79,16]; Ulf Gyllensten[33]; Igor Rudan[16]; Ian J Deary[30,81]; Nicole M Probst-Hensch[28,29]; Holger Schulz[25,27]; Alan L James[21,82,83]; James F Wilson[16,32]; Beate Stubbe[47]; Eleftheria Zeggini[85]; Marjo-Riitta Jarvelin[13,86,87,88]; Nick Wareham[12]; Edwin K Silverman[4,5]; Caroline Hayward[32]; Andrew P Morris[18,57]; Adam S Butterworth[19,46]; Robert A Scott[72]; Robin G Walters[9]; Deborah A Meyers[8]; Michael H Cho[4,5]; David P Strachan[91]; Ian P Hall*[63]; Martin D Tobin*†[1,92]; Louise V Wain*†[1,92];

* = Contributed equally to this work.

† = Corresponding Authors.

Abstract

Reduced lung function predicts mortality and is key to the diagnosis of COPD. In a genome-wide association study in 400,102 individuals of European ancestry, we define 279 lung function signals, one-half of which are new. In combination these variants strongly predict COPD in deeply-phenotyped patient populations. Furthermore, the combined effect of these variants showed generalisability across smokers and never-smokers, and across ancestral groups. We highlight biological pathways, known and potential drug targets for COPD and, in phenome-wide association studies, autoimmune-related and other pleiotropic effects of lung function associated variants. This new genetic evidence has potential to improve future preventive and therapeutic strategies for COPD.

Introduction:

Impaired lung function is predictive of mortality[1] and is the key diagnostic criterion for chronic obstructive pulmonary disease (COPD). Globally, COPD accounted for 2.9 million deaths in 2016[2], being one of the key causes of both Years of Life Lost and Years Lived with Disability worldwide[3]. Determinants of maximally attained lung function and of lung function decline can influence the risk

of developing COPD. Tobacco smoking is the single largest risk factor for COPD, although other environmental exposures and genetic makeup are important[4,5]. Genetic variants associated with lung function and COPD susceptibility can be causally informative, assisting with risk prediction, as well as drug target identification and validation[6]. Whilst there has been considerable progress in identifying genetic markers associated with lung function and risk of COPD[4,7-19] seeking a high yield of associated genetic variants is key to progressing knowledge because: (i) implication of multiple molecules in each pathway will be needed to build an accurate picture of the pathways underpinning development of COPD; (ii) not all proteins identified will be druggable and; (iii) combining information across multiple variants can improve prediction of disease susceptibility.

Through new detailed quality control and analyses of spirometric measures of lung function in UK Biobank, completion of genome-wide genotyping in UK Biobank, and expansion of the SpiroMeta Consortium, we undertook the largest genome-wide association study of lung function performed to date. Comprising a total of 400,102 individuals of European ancestry, our study entailed a near seven-fold increase in sample size over previous studies of similar ancestry to address the following aims: (i) to generate a high yield of genetic markers associated with lung function; (ii) to confirm and fine-map previously reported lung function signals; (iii) to investigate the putative causal genes and biological pathways through which lung function associated variants act, and their wider pleiotropic effects on other traits; and (iv) to generate a weighted genetic risk score for lung function and test its association with COPD susceptibility in individuals of European and other ancestries.

Results:

139 new signals for lung function

Here we present a total of 279 distinct association signals for lung function, of which a half (139 variants) are new having reached genome-wide significance ($P<5x10^{-9}$) in this study. We increased the sample size available for the study of quantitative measures of lung function in UK Biobank by refining the quality control of spirometry based on recommendations of the UK Biobank Outcomes Adjudication Working Group, utilising additional metrics derived from the blow curve time series measurements, and relaxing the reproducibility threshold for repeat measures (**Supplementary Note**). Genome-wide association analyses of forced expired volume in 1 second (FEV_1), forced vital capacity (FVC) and FEV_1/FVC were then undertaken in 321,047 individuals in UK Biobank (**Supplementary Table 1**) and in 79,055 individuals from the SpiroMeta Consortium (**Supplementary Tables 2 and 3**). A linear mixed model approach implemented in BOLT-LMM[20] was used for UK Biobank to account for relatedness and fine-scale population structure (**Online Methods**). A total of 19,871,028 variants imputed in both UK Biobank and SpiroMeta were analysed. Peak expiratory flow (PEF) was also analysed genome-wide in UK Biobank and up to 24,218 samples from SpiroMeta. All individuals included in the genome-wide analyses were of European ancestry (**Supplementary Figure 1** and **Supplementary Table 2**).

To maximise statistical power for discovery of new signals, whilst maintaining stringent significance thresholds to minimise reporting of false positives, we adopted a study design incorporating both two-stage and one-stage approaches (**Figure 1**). In the two-stage analysis, 99 new distinct signals, defined using conditional analyses, were associated with one or more traits at $P<5x10^{-9}$ in UK Biobank and showed association ($P<10^{-3}$) with a consistent direction of effect in SpiroMeta ("Tier 1" signals, **Supplementary Figure 2; Supplementary Table 4**). In the one-stage analysis, we meta-analysed UK Biobank and SpiroMeta (up to 400,102 individuals) and 40 additional new distinct

signals associated with one or more lung function traits reaching $P<5x10^{-9}$ were identified (**Supplementary Figure 2, Supplementary Table 4**) that were also associated with $P<10^{-3}$ separately in UK Biobank and in SpiroMeta, with consistent direction of effect ("Tier 2" signals). An additional 323 signals were significantly associated with one or more lung function traits in the meta-analysis of UK Biobank and SpiroMeta ($P<5x10^{-9}$) and reached $P<10^{-3}$ for association in only one of UK Biobank or SpiroMeta ("Tier 3" signals, **Supplementary Table 5**). Only the 139 signals meeting Tier 1 and Tier 2 criteria were followed up further. The strength and direction of association of the sentinel variant (the variant in each signal with the lowest P value) for these 139 new signals across all 4 lung function traits are shown in **Figure 2**.

To assess whether any of these 139 signals associated with lung function could be driven via an underlying association with smoking, we examined association of the sentinel variants with smoking behaviour in UK Biobank (**Online Methods**). The only new sentinel associated with smoking behaviour was rs193686 (in an intron of *MET*, **Supplementary Table 6**). Therefore, we tested for association between this variant and lung function in never smokers (n=173,658). Whilst rs193686 was associated with smoking initiation ($P=9.18x10^{-6}$), the allele associated with smoking initiation was associated with increased lung function in never smokers (FEV_1/FVC $P=5.28x10^{-10}$, **Supplementary Table 7**). Therefore, this signal was retained for further analysis.

A total of 279 signals of association for lung function

Of 157 previously published signals of association with lung function and COPD[3,6-18], 142 were associated at $P<10^{-5}$ in UK Biobank (**Online Methods, Supplementary Figure 3**, **Supplementary Table 8**). Two sentinel variants (rs1689510 near *RAB5B* and rs11134789 in an intron of *ADAM19*) were associated with smoking initiation ($P=9.72x10^{-6}$ and $P=2.13x10^{-5}$, respectively) (**Supplementary Table 6**), but were also associated with lung function in never smokers ($P=2.49x10^{-8}$ for FEV_1 and $P=2.94x10^{-45}$ for FEV_1/FVC, respectively, **Supplementary Table 7**). SNP rs17486278 at *CHRNA5* and rs11667314 near *CYP2A6* were each associated with cigarettes per day ($P=1.35x10^{-79}$ and $P=6.47\times10^{-24}$, respectively; **Supplementary Table 6**); neither were significantly associated with lung function among never smokers, hence these latter two signals were excluded from further analysis. This brings the total number of distinct signals of association with lung function to 279 (**Supplementary Table 9**). None of these variants showed interaction with ever-smoking status ($P>1.8x10^{-4}$, Online Methods**, Supplementary Table 7**). The 140 previously reported lung function signals showing association in this study (UK Biobank $P<10^{-5}$) explained 5.0%, 3.4%, 9.2% and 4.5% of the estimated heritability of FEV_1, FVC, FEV_1/FVC and PEF, respectively (**Online Methods**). The 139 new signals reported here, explain an additional 4.3%, 3.3%, 3.9% and 3.3% of the estimated heritability, respectively.

Identification of putative causal genes

Bayesian refinement was undertaken for each signal to identify the set of variants that were 99% likely to contain the underlying causal variant (assuming the causal variant has been analysed). The signals in the HLA region were excluded due to extended linkage disequilibrium. The results from the meta-analysis of UK Biobank and SpiroMeta were used to define the 99% credible sets (**Online Methods**, **Supplementary Table 10, Supplementary File–Region Plots**).

To identify putative causal genes for each signal, we identified deleterious variants and variants associated with gene expression (eQTLs) or protein levels (pQTLs) within each 99% credible set for all new and previously reported signals outside the HLA region (**Online Methods**).

There were 25 SNPs, located in 22 unique genes, which were exonic, at a splice site or in the untranslated regions and additionally annotated as potentially deleterious (**Online Methods**, **Supplementary Table 11**). Amongst our new signals, there were 10 variants annotated as deleterious in 9 different genes: *DOCK9* (rs117633128, MAF=10.6%), *CEP72* (rs12522955, MAF=20.2%), *BCHE* (rs1799807, MAF=1.95%), *DST* (rs11756977, MAF=28.9%), *KIAA0753* (rs2304977, MAF=37.7%; rs9889363, MAF=37.7%), *LRRC45* (rs72861736, MAF=10.9%), *BTC* (rs11938093, MAF=26.6%), *C2orf54* (rs6709469, MAF=49.9%) and *IER5L* (rs184457, MAF=31.5%). Of these, the missense variant in *BCHE* (rs1799807) had the highest posterior probability (0.996) in its respective credible set, was low frequency (MAF=1.95%) and resulted in an amino acid change from aspartic acid (D) to glycine (G), known to affect the function of the encoded butyrylcholinesterase enzyme by altering substrate binding[21]. The two common missense variants in *KIAA0753* were within the credible set of new signal rs4796334. *KIAA0753*, *CEP72* and *LRRC45* all encode proteins with a role in ciliogenesis or cilia maintenance[22-26], and all are highly expressed in the airway epithelium[27].

Variants in the 99% credible sets (n=9,698) were queried in three eQTL resources to identify associations with gene expression in lung[28-30] (sample size n=1,111; **Supplementary Table 12**), blood[31] (n=4,896) and a subset of GTEx[32] tissues (max n=388, **Online Methods**). The tissues included from GTEx were lung and blood, plus nine tissues known to contain smooth muscle (**Online Methods**). The latter were chosen based on previous reports of enrichment of lung function GWAS signals in smooth muscle-containing tissues[18,33]. We identified 88 genes for which the most significant SNP associated with expression of that gene in the respective eQTL resource was within one of the 99% credible sets. These 88 genes were implicated by 58 of the 279 signals (**Supplementary Table 13**).

We checked credible set variants for association with protein levels in a pQTL study[34] comprising SNP associations for 3,600 plasma proteins. Using a Bonferroni-corrected 5% significance threshold for 276 tests for these 3,600 proteins ($P < 5.03\times10^{-8}$), we found 1,076 pQTLs in our credible sets covering 26 lung function sentinels implicating 34 proteins. For 5 of these proteins the pQTL sentinel was contained within our lung function credible set: ECM1, THBS4, NPNT, C1QTNF5 and SCARF2 (**Supplementary Table 14**).

In total, 107 putative causal genes were identified (**Table 1**), 8 by both a deleterious variant and an eQTL signal (including *KIAA0753* implicated by two deleterious variants), 1 (*NPNT*) by both an eQTL and a pQTL signal, 1 (*SCARF2*) by both a deleterious variant and a pQTL signal, 13 by a deleterious variant only, 81 by an eQTL signal only and 3 by a pQTL signal only. Among these 107 genes, we highlight 75 for the first time as putative causal genes for lung function (43 implicated by a new signal and 32 newly implicated by a previous signal[18]).

Pathway analysis

We tested whether these 107 putative causal genes were enriched in gene sets and biological pathways (**Online Methods**), finding an enrichment of genes in elastic fibre and extracellular matrix organisation pathways, and a number of gene ontologies including gene sets relating to the

cytoskeleton and processes involved in ciliogenesis (for example, cytoskeleton organisation, organelle organisation, centriole replication and microtubule-based processes) (**Supplementary Table 15**). Whilst the enrichment in elastic fibre-related pathways is consistent with our previous study[18], enrichment in these pathways was further supported in this analysis by two new genes, *ITGAV* (at a new signal) and *GDF5* (a newly implicated gene for a previously reported signal), and by strengthened eQTL evidence for *TGFB2* and *MFAP2* as the putative causal genes at two previously reported signals. The presence of *TGFB2*, *GDF5* and *SMAD3* in our list of 107 genes resulted in enrichment of a TGF-β superfamily signalling pathway (TGF-Core) and multiple related gene ontology terms (**Supplementary Table 15**).

Functional enrichment analyses

We tested for enrichment of the 279 lung function signals in DNase I hypersensitivity sites in 125 cell lines from ENCODE and 299 cell lines and tissues from RoadMap Epigenome Project using FORGE v1.1[35]. There was significant tissue specific overlap (**Online Methods**) of the 279 signals with DNAse1 hotspots in adult and foetal lung, foetal muscle (skeletal), foetal stomach, foetal heart, and fibroblasts (**Supplementary Figure 4**).

We used DeepSEA[36], a variant effect predictor which utilises a deep-learning algorithm, to identify whether our signals were predicted to have a chromatin effect in lung-related cell lines. We identified 10 signals (including 5 new signals) for which the SNP with the largest posterior probability of being causal also had a significant predicted effect on a DNase I hypersensitivity site in lung-related cells (**Supplementary Table 16**). This included a new signal near *SMURF2* (17q24.1, rs11653958) that also had a predicted functional effect on histone marks (DNase I hypersensitivity sites, H3K9ac, H3K27ac, H3K4me1, H3K4me2, H3K4me3) and on *CEBPB*, *FOSL2*, *SIN2AK-20* and *TCF12* transcription factor binding sites, and a new signal near *PDZRN3-AS1* (rs586936) had a large predicted effect on a CEBPB transcription factor binding site.

Drug targets

All 107 putative causal genes were interrogated against the gene-drug interactions table of the Drug-Gene Interactions Database (DGIDB)[37] (**Supplementary Table 17**). We highlight two examples of new genetic signals implicating targets for drugs in development for indications other than COPD. One of our new signals is an eQTL for *ITGAV*. *ITGAV* encodes a component of the avβ6 integrin heterodimer, which is inhibited by a monoclonal antibody, STX-100, in development for pulmonary fibrosis (ClinicalTrials.gov Identifier: NCT01371305) and for which the small molecule GSK3008348 (ClinicalTrials.gov Identifier: NCT03069989) is an antagonist[38]. Integrins have an emerging role as local activators of TGFβ and specifically the avb6 integrin heterodimer can activate latent-TGFβ[39]. In our study, the allele associated with reduced expression of *ITGAV* (**Supplementary Table 13**) was associated with reduced risk of COPD (**Supplementary Table 9**) suggesting that inhibitors of αvβ6 integrin might also have a beneficial effect in COPD. Another of our new signals is associated with expression of *TNFSF13* (synonym *APRIL*), a cytokine which is a member of the TNF ligand family. Atacicept blocks B cell stimulation by *TNFSF13* (as well as by BLyS) and reduced systemic lupus erythematosus disease activity in a recent Phase IIb trial[40]. In our study, the allele associated with decreased expression of *TNFSF13* was associated with reduced FEV_1, indicating that vigilance for pulmonary consequences of atacicept may be warranted.

Genetic Risk Score: association with FEV_1/FVC and COPD in multiple ancestries

We constructed a genetic risk score (GRS) weighted by FEV_1/FVC effect sizes comprising all 279 new or previously reported sentinel variants, and tested the association of the GRS with FEV_1/FVC and GOLD Stage 2-4 COPD (FEV_1/FVC<0.7 and FEV_1<80% predicted) in different ancestry groups in UK Biobank, and China Kadoorie Biobank (**Online Methods**, **Supplementary Table 18**). The GRS was associated with FEV_1/FVC and COPD in each of the ancestry groups (**Figure 3A**).

We tested for a GRS interaction with smoking in European ancestry individuals in UK Biobank[41]. No statistical interaction was seen for FEV_1/FVC (interaction term -0.002 per SD change in GRS, 95% CI: [0.009, 0.005], P=0.532), whilst the findings for COPD were consistent with a slightly smaller effect of the GRS in ever-smokers (OR for ever-smoking-GRS interaction term per SD change in GRS 0.96, 95% CI: [0.92, 0.99], P=0.015).

The association of the GRS with COPD susceptibility was additionally tested in deeply-phenotyped case-control studies (**Supplementary Table 19**). Similar effect size estimates were seen across each of the 5 European ancestry studies (**Figure 3B**); in the meta-analysis of these studies (n=6,979 cases and 3,915 controls), the odds ratio for COPD per standard deviation of the weighted GRS was 1.55 (95% CI: [1.48, 1.62]), $P=2.87\times10^{-75}$ (**Supplementary Table 20**). The GRS was also associated with COPD in individuals of African-American ancestry in COPDGene ($P=8.36\times10^{-7}$), albeit with a smaller effect size estimate, odds ratio=1.26 (95% CI: [1.15, 1.37]).

To aid clinical interpretation, we divided individuals in each of the European ancestry deeply-phenotyped COPD case-control studies into deciles, according to their value of the weighted GRS. The odds ratio for COPD in members of the highest GRS decile compared to the lowest GRS decile was 4.73 (95% CI: [3.79, 5.90]), $P=3.00\times10^{-43}$ (**Figure 3C**, **Supplementary Table 21**). We calculated the population attributable risk fraction and estimated that the proportion of COPD cases attributable to risk scores above the first GRS decile was 54.6% (95% CI: [50.6%, 58.4%]).

Pleiotropy and phenome-wide association studies

As phenome-wide association studies (PheWAS) can provide evidence mimicking pharmacological interventions of drug targets in humans and informing drug development[42], we undertook a PheWAS of 2,411 phenotypes in UK Biobank (**Online Methods, Figure 4**); 226 of the 279 sentinel variants were associated (FDR <1%) with one or more traits and diseases (excluding quantitative lung function traits). Eighty-five of the lung function signals were associated with standing height. In order to investigate whether the genetic association signals for lung function were driven by incomplete adjustment for height, we tested for correlation of effects on lung function in UK Biobank and height in the GIANT consortium for 247 of the 279 signals that had a proxy variant in GIANT[43]; there was no significant correlation (r=-0.096, **Supplementary Figure 5**). Additionally, the PheWAS revealed associations with body composition measures such as fat free mass (54 SNPs) and hip circumference (40 SNPs), as well as muscle strength (32 SNPs, grip strength). One hundred and fourteen of the 279 SNPs were associated with several quantitative measures of blood count, including eosinophil counts and percentages (25 SNPs). Twenty-five of our SNPs were also associated with asthma including 12 SNPs associated both with asthma and eosinophil measures. Five of these SNPs were in LD (r^2>0.1) with a SNP reported for association both with asthma and eosinophil measures in previously published genome-wide association studies. To assess whether any of the lung function associations could be driven by an association with asthma, we compared the effect

size estimated before and after exclusion of all self-reported asthma cases, observing remarkably similar estimates (**Supplementary Figure 6**) suggesting that the lung function associations we report are not primarily driven via known asthma signals.

We examined the specificity of genetic associations, given the potential for this to predict specificity of drug target modification, and found that 53 of the 279 signals were associated only with lung function and COPD-related traits. In contrast, three of our 279 signals were associated with over 100 traits across multiple categories – among these rs3844313, a known intergenic signal near *HLA-DQB1* was associated with 163 traits, and also had the strongest signal in the PheWAS, which was for association with intestinal malabsorption and coeliac disease.

In our 279-variant weighted GRS PheWAS analysis (**Supplementary Table 22**), we found association with respiratory traits including COPD, chronic bronchitis, emphysema, respiratory failure, corticosteroid use and both paediatric and adult-onset asthma (**Figure 5a**). The GRS was also associated with non-respiratory traits including coeliac disease, an intestinal autoimmune disorder (**Figure 5b**). These pleiotropic effects on risk of autoimmune diseases was further confirmed by analysis of previously reported GWAS (**Online Methods, Supplementary Table 23**) which showed overlapping single variant associations with Crohn's disease, ulcerative colitis, psoriasis, systemic lupus erythematosus, IgA nephropathy, pediatric autoimmune disease and type 1 diabetes.

Discussion:

The large sample size of our study, achieved by our refinement of the spirometry in UK Biobank and inclusion of the substantially expanded SpiroMeta consortium data set, has doubled the yield of lung function signals to 279. Fine-mapping of all new and previously reported signals, together with gene and protein expression analyses with improved tissue specificity and stringency, has implicated new genes and pathways, highlighting the importance of cilia development, TGFB-signalling via SMAD3, and elastic fibres in the aetiology of airflow obstruction. Many of the genes and pathways reported here contain druggable targets; we highlight examples where the genetic variants mimicking therapeutic modulation of targets may have opposing effects on lung function. We have developed and applied the first weighted GRS for lung function and tested it in deeply-phenotyped COPD case-control studies. Our GRS shows stronger association and larger effect size estimates (4.73 fold change in COPD risk between highest and lowest risk deciles) than a previous GRS in European ancestry populations[18], as well as generalisability to African, South Asian and Chinese ancestry groups. We undertook the first comprehensive PheWAS for lung function signals, and report genetic variants with apparent specificity of effects and others with pleiotropic effects that might indicate shared biological pathways between different diseases.

For the first time in a GWAS of lung function, we report an enrichment of genes involved in ciliogenesis (including *KIAA0753, CDK2* and *CEP72*). Defects in primary cilia as a result of highly deleterious mutations in essential genes result in ciliopathies known to affect multiple organ systems. We found an enrichment of genes with a role in centriolar replication and duplication, core processes in primary and motile cilia formation. Mutations in *KIAA0753* cause the ciliopathies Joubert Syndrome and Orofaciodigital Syndrome[23]. Reduced airway motile cilia function impacting mucus clearance is a feature of COPD, but it has not been clear whether this is causal or the consequence of damage by external factors such as smoking or infection. Our findings suggest that impaired ciliary function might be a driver of the disease process. We have previously shown,

294 through whole exome re-sequencing, an enrichment of rare variants in cilia-related genes in heavy
295 smokers without airflow obstruction[44].

296 New signals, implicating *ITGAV* and *GDF5*, as well as stronger support for *TGFB2* and *MFAP2* as likely
297 causal genes, provide new genetic support for the importance of elastic fibre pathways in lung
298 function and COPD[18]. The elastic fibres of the extracellular matrix are known to be disrupted in
299 COPD[45]. As the breakdown of elastic fibres by neutrophil elastase leads to emphysema in individuals
300 with alpha$_1$-antitrypsin deficiency, we also assessed the association with the *SERPINA1* Z allele,
301 which was not associated with lung function in our study (rs28929474, P=0.109 for FEV$_1$/FVC in UK
302 Biobank).

303 Smoking and genetic risk both have important effects on lung function and COPD. We found no
304 interaction of smoking with individual lung function associated variants. Our weighted 279-SNP GRS
305 showed no interaction with smoking status for FEV$_1$/FVC, whilst a weak smoking-GRS interaction was
306 observed for COPD susceptibility. Thus our findings are consistent with the effects of smoking and
307 genetic risk being approximately additive on lung function (and multiplicative on COPD risk). Whilst
308 the weighted 279-SNP GRS showed a strong association with COPD susceptibility, and a high
309 attributable risk, we do not claim that this would represent an appropriate method of screening for
310 COPD risk. Incorporation of the GRS into a risk model already comprising available clinical
311 information (including age, sex, height and pack-years of smoking in COPDGene non-Hispanic
312 Whites) leads to an increase in the area under the curve from 0.751 to 0.771, which although
313 statistically significant ($p=3.33x10^{-10}$) is of modest magnitude. Importantly, our findings demonstrate
314 the high absolute risk among genetically susceptible smokers. Based on our estimated GRS relative
315 risk and absolute risk estimates of COPD shown by Lokke *et al.*[46], one would expect the highest GRS
316 risk decile group of smokers to have an absolute risk of developing COPD by approximately 70 years
317 of age of 82.4%, versus 17.4% for the lowest GRS decile.

318 The unprecedented sample size of UK Biobank as a single cohort has revolutionised genetic studies.
319 We used two complementary study designs to maximise sample size for discovery and ensure
320 robustness of findings by requiring independent support for association. Furthermore, through
321 additional analysis of the spirometry data in UK Biobank and substantial expansion of the SpiroMeta
322 consortium, we have markedly increased samples sizes to almost seven times those included in
323 previous studies. As no lower MAF threshold was applied in our analyses, an overall threshold of
324 $P<5x10^{-9}$, as recommended for re-sequencing analyses of European ancestry individuals[47], was
325 applied. We identified the largest number of new signals in our more stringent two-stage design
326 ("Tier 1", 99 new signals). Amongst the signals that we report as "Tier 3" (and did not include in
327 further analyses), all reached $P<10^{-3}$ in UK Biobank and 183 met a less stringent threshold of P<0.05
328 in SpiroMeta.

329 Our study is the first to investigate genome-wide associations with PEF. PEF is determined by various
330 physiological factors including lung volume, large airway calibre, elasticity of the lung and expiratory
331 muscle strength, is used for monitoring asthma, and was incorporated in a recently evaluated clinical
332 score for diagnosing COPD and predicting acute exacerbations of COPD[48]. Overall, 133 of the 279
333 signals were also associated with PEF ($P<10^{-5}$) and for 15 signals (including 4 new signals), PEF was
334 the most significantly associated trait. Of note, a signal near *SLC26A9*, a known cystic fibrosis
335 modifier gene[49], was highly significantly associated with PEF in UK Biobank ($P=3.97x10^{-66}$) and was
336 nominally significant in SpiroMeta ($P=6.93x10^{-3}$), with consistent direction of effect, but did not meet

the Tier 2 criteria ($P<10^{-3}$ in each of SpiroMeta and UK Biobank). This could reflect the limited power for PEF in SpiroMeta (up to 24,218 for PEF compared to 79,055 for the other three traits).

Examining associations of a given genetic variant with a wide range of human phenotypes is a valuable tool in therapeutic target validation. As in our PheWAS, it can highlight variants which show associations with one or more respiratory traits that might be expected to demonstrate greater target specificity than variants associated with many traits. Additionally, in some instances, association with multiple traits may indicate the relevance of drug repurposing. Association of a given SNP with multiple traits does not necessarily imply shared aetiology, and further investigation is warranted. Our GRS PheWAS assesses broader genetic overlap between lung function and other traits and supports the evidence for some shared genetic determinants with autoimmune diseases.

In summary, our study has doubled the number of signals for lung function and, based on relating fine-mapped, annotated variants to gene and protein expression, epigenetic marks, gene sets, biological pathways and druggable proteins, it provides new understanding and resources of utility for the development of therapeutics. The 279-variant GRS we constructed was associated with a 4.71-fold increased relative risk of moderate-severe COPD between highest and lowest deciles, such that one would expect over 80% of smokers in the highest genetic risk decile to develop COPD. The GRS was also predictive of COPD across multiple ancestral groups. Our PheWAS highlights both expected and unexpected associations relevant to respiratory and other systemic diseases. Investigating the nature of the pleiotropic effects of some of these variants will be of benefit for drug target identification and validation.

Online Methods:

Study Design Overview and rationale

For the two-stage approach, we firstly selected distinct signals of association (defined using conditional analyses) with one or more traits achieving $P<5x10^{-9}$ in UK Biobank only (n up to 321,047). A threshold of $P<5x10^{-9}$ was selected to maximise stringency of findings and to be consistent with currently recommended genome-wide significance thresholds for re-sequencing analyses of European ancestry individuals[50]. We then reported as new those signals which additionally met $P<10^{-3}$ in SpiroMeta (N effective >70% of n up to 79,055; **Supplementary Note**, **Supplementary Figure 7**), with consistent directions of effect and term them "Tier 1" signals as they meet our highest level of stringency.

For the one-stage approach, we selected distinct signals of association (defined using conditional analyses) with one or more traits reaching $P<5x10^{-9}$ in the meta-analysis of UK Biobank and SpiroMeta (n up to 400,102) and reported as new those which additionally met $P<10^{-3}$ in both UK Biobank and SpiroMeta with a consistent direction of effect. We term these signals "Tier 2" as they meet our second-highest level of stringency.

All signals meeting either set of criteria described above, and that had not been previously published, were reported as new signals of association with lung function. Signals that reached $P<5x10^{-9}$ in the meta-analysis of UK Biobank and SpiroMeta, had a consistent direction of effect in UK Biobank and SpiroMeta, but which did not reach $P<10^{-3}$ in both UK Biobank and SpiroMeta are presented as "Tier 3" and were not included in further analyses.

UK Biobank

The UK Biobank data resource is described elsewhere (see URLs). Individuals were selected for inclusion in this study if they met the following criteria: (i) had complete data for age, sex, height and smoking status; (ii) had spirometry meeting quality control requirements (based on analyses of acceptability, reproducibility and blow curve metrics; **Supplementary Note**); (iii) had genome-wide imputed genetic data and; (iv) were of European ancestry based on genetic data (**Supplementary Note**; **Supplementary Figure 1**). Genotyping was undertaken using the Affymetrix Axiom® UK BiLEVE and UK Biobank arrays[13]. Genotypes were imputed to the Haplotype Reference Consortium panel[51] (**Supplementary Note**), and retained if minor allele count ≥3 and imputation quality (info) > 0.5. A total of 321,047 individuals were included in this analysis (**Supplementary Table 1**).

Residuals from linear regression of each trait (FEV_1, FVC, FEV_1/FVC and PEF) against age, age^2, sex, height, smoking status (ever/never) and genotyping array were ranked and inverse-normal transformed to obtain adjusted, normally distributed Z-scores. These Z-scores were then used for genome-wide association testing under an additive genetic model using BOLT-LMM v2.3[20]. Principal components were not included as BOLT-LMM uses a linear mixed model to account for relatedness and fine-scale population structure.

Linkage disequilibrium (LD) score regression implemented in LDSC[52] was used to estimate inflation of test statistics due to confounding. Genomic control was applied, adjusting all test statistics by LD score regression intercepts: 1.12 for FEV_1, 1.14 for FVC, 1.19 for FEV_1/FVC and 1.13 for PEF (**Supplementary Figure 8**; **Supplementary Table 24**).

SpiroMeta consortium

The SpiroMeta consortium meta-analysis was comprised of a total of 79,055 individuals from 22 studies. Thirteen studies (n=21,436 individuals) were imputed to the 1000 Genomes Project Phase 1 reference panel[53] (B58C [T1DGC and WTCCC], BHS1&2, three Croatian studies [CROATIA-Korcula, CROATIA-Split and CROATIA-Vis], Health 2000, KORA F4, KORA S3, LBC1936, NSPHS, ORCADES, SAPALDIA and YFS and 9 studies (n=61,682 individuals) were imputed to the Haplotype Reference Consortium (HRC) panel[54] (EPIC [obese cases and population-based studies], GS:SFHS, NFBC1966, NFBC1986, PIVUS, SHIP, SHIP-TREND, UKHLS and VIKING). See **Supplementary Tables 2** and **3** for the definitions of all abbreviations, study characteristics, details of genotyping platforms and imputation panels and methods). Measurements of spirometry for each study are described in the **Supplementary Note**.

In each study, linear regression models were fitted for each lung function trait (FEV_1, FEV_1/FVC, FVC and PEF, where available), with adjustment for age, age^2, sex and height. For studies with unrelated individuals, these models were fitted separately in ever smokers and never smokers, with additional adjustment for principal components of ancestry. Studies with related individuals fitted mixed models in all individuals to account for relatedness, with ever smoking status as a covariate.

In all studies, rank-based inverse normal transformations were undertaken on the residuals, with these transformed residuals used as the phenotype for association testing under an additive genetic model (**Supplementary Table 3**).

In the study level results, variants were excluded if they had a very low MAC (**Supplementary Table 3**) or imputation quality (info) <0.3. In studies with unrelated individuals, the ever and never smokers results were combined, using inverse variance weighted meta-analysis, to give an overall study

result. Genomic control was then applied to all study level results, before combining results across all studies using inverse variance weighted meta-analysis. LD score regression intercepts for the meta-analysis were close to 1 (**Supplementary Figure 8**; **Supplementary Table 24**) and so genomic control was not applied.

Meta-analyses

A total of 19,871,028 variants (imputed or genotyped) in both UK Biobank and SpiroMeta were meta-analysed using inverse-variance weighted fixed effect meta-analysis, and no further genomic control was applied as LD score regression intercepts were close to 1 (**Supplementary Table 24**).

Selection of new signals using conditional analyses

All SNPs ±1Mb were extracted around each sentinel variant. GCTA[55] was then used to perform stepwise conditional analysis to select independently associated SNPs within each 2Mb region. Any secondary signals identified within each 2Mb region were required to meet Tier 1 or Tier 2 criteria (described above) after conditioning on the primary sentinel variant. A combined list of distinct lung function signals was then made across the 4 phenotypes, FEV_1, FVC, FEV_1/FVC and PEF as follows: where sentinel variants for 2 signals for different phenotypes were in high LD ($r^2 > 0.5$), we retained the most significant variant; where 2 signals were in moderate LD ($0.1 > r^2 > 0.5$), we retained variants if, after conditional analysis, they still met the Tier 1 or Tier 2 threshold; for signals in low LD ($r^2 < 0.1$) we retained both variants. We then used the same criteria to identify a subset of new signals which were distinct from previously published independent signals (see below).

Assessment of previously reported lung function signals

We identified 184 autosomal signals from previous GWAS analyses of lung function and COPD[1,4-14]. After LD pruning (keeping only those signals with LD of $r^2 < 0.1$), we removed 24 non-independent SNPs, leaving 160 previously reported independent signals. Of 6 previously reported signals in the HLA region, we included only the 3 independent lung function HLA signals reported from conditional analysis using all imputed HLA genotypes[18]: *AGER* (rs2070600), *HLA-DQB1* (rs114544105) and near *ZNF184* (rs34864796) leaving 157 signals.

We confirmed association of previously reported signals in our data if they met any of three criteria: (i) the previously reported sentinel was associated ($P<10^{-5}$) with any lung function trait in UK Biobank; (ii) a proxy for the previously reported sentinel with $r^2>0.5$ was associated ($P<10^{-5}$) with any lung function trait in UK Biobank; (iii) a proxy for the previously reported sentinel with $r^2>0.1$ was associated with any lung function trait meeting tier 1 or tier 2 criteria (**Supplementary Figure 3**).

Effect on COPD susceptibility – genetic risk score in multiple ancestries

To test association of all lung function signals and COPD susceptibility, we constructed a 279-variant weighted GRS comprising the 139 novel and 140 previously reported signals; we used the previously reported sentinel SNP for published signals. Weights were derived using the FEV_1/FVC ratio decreasing (i.e. COPD risk *increasing*) alleles. For previously reported signals (n=140), results from the UK Biobank analysis were used to derive weights for the 94 signals that were not discovered using UK Biobank data and weights were taken from SpiroMeta for 46 signals where UK Biobank was included in the discovery of those signals. For novel signals identified in this study, weights were taken from SpiroMeta for two-stage (tier 1) signals (n=99), and the smallest absolute effect size from either of UK Biobank or SpiroMeta was used for one-stage (tier 2) signals (n=40) (**Supplementary**

460 **Table 25**). For the weighted GRS the number of risk alleles at each variant was multiplied by its
461 weight.

462 The GRS was first calculated in unrelated individuals (KING kinship coefficient of < 0.0884) within 6
463 ancestral groups of UK Biobank: Europeans, South Asians, Africans, Chinese, Mixed African and
464 Europeans, and Mixed Other (total sample of unrelated individuals across six ancestries: 323,001)
465 using PLINK. Weights and alleles were as described above. COPD was defined as $FEV_1/FVC < 0.7$ and
466 $FEV_1 < 0.8$ of the predicted value, i.e. GOLD stage 2-4 categorisation. Associations with the GRS were
467 then tested using COPD (in ancestral groups with at least 100 COPD cases) and FEV_1/FVC as the
468 outcomes.

469 In addition, we calculated the GRS in individuals from the China Kadoorie Biobank (CKB). Four of the
470 279 SNPs were not available in CKB (rs1800888, rs56196860, rs72724130 and rs77672322), and for
471 12 SNPs, proxies were used (minimum r^2=0.3). Analyses were undertaken in all COPD GOLD stage 2-4
472 cases ($FEV_1/FVC < 0.7$ and $FEV_1 < 0.8$ of the predicted value, in 6,013 cases and 69,567 controls),
473 against an unbiased set of population controls. The GRS was also tested for association with
474 FEV_1/FVC in CKB (n=72,796).

475 Logistic regression of COPD case-control status with the GRS in UK Biobank and China Kadoorie
476 Biobank assumed an additive genetic effect and was adjusted for age, age^2, sex, height, and smoking
477 (**Supplementary Table 18**). Ten principal components were also included in UK Biobank analyses. In
478 China Kadoorie Biobank, analyses were stratified by geographical regions and then meta-analysed
479 using an inverse-variance fixed effect model. Linear models assessing the association with FEV_1/FVC
480 were fitted using the same transformed outcome as in the main GWAS analysis.

481 We then tested association in 5 European ancestry COPD case-control studies: COPDGene (Non-
482 Hispanic White Population) (3,068 cases and 2,110 controls), ECLIPSE (1,713 cases and 147 controls),
483 GenKOLS (836 cases and 692 controls), NETT-NAS (374 cases and 429 controls) and SPIROMICS (988
484 cases and 537 controls) (**Supplementary Table 19**). In addition, we tested this GRS in the COPDGene
485 African American population study (910 cases and 1,556 controls). Logistic regression models using
486 COPD as outcome and the GRS as exposure were adjusted for age, age^2, sex, height, and principal
487 components (**Supplementary Table 20**).

488 Next, we divided individuals in the external COPD case-control studies into deciles according to their
489 values of the weighted GRS. This was undertaken separately by study group, and for each decile
490 logistic models were fitted, comparing the risk of COPD for members of each decile group compared
491 to those in the lowest decile (i.e. those with lowest values of the weighted GRS). Covariates were as
492 for the COPD analyses. Results were combined across European-ancestry study groups by fixed
493 effect meta-analysis (**Supplementary Table 21**).

494 We calculated the population attributable risk fraction (PARF) as follows:

495 $$PARF = \frac{P(E)(OR - 1)}{1 + P(E)(OR - 1)}$$

496

497 where *P(E)* is set to 0.9, i.e. the probability of carrying more risk alleles than those in the lowest risk
498 score decile of the risk score (the 'probability of the exposure'). *OR* refers to the odds of having

COPD in individuals across deciles 2 to 10 of the risk score compared to the odds of having COPD for individuals in the lowest decile (decile 1) of the risk score (**Supplementary Note**).

Effects on smoking behaviour

As our discovery GWAS in UK Biobank was adjusted for ever vs. never smoking status, and not for pack years of smoking (pack years information was missing for 32% of smokers), we evaluated whether any signals of association with lung function might be driven by an association with smoking behaviour by testing for association with smoking initiation (123,890 ever smokers vs. 151,706 never smokers) and cigarettes per day (n=80,015) in UK Biobank (full methods in **Supplementary Note**). We also tested for association with lung function in never smokers only (n=173,658). We excluded any signals associated with smoking behaviour (**Supplementary Table 6**), but not with lung function in never smokers.

Smoking interaction

For associated variants (new and previously reported), we repeated association testing for lung function separately in UK Biobank and SpiroMeta (up to 176,701 ever smokers and 197,999 never smokers), and tested for an interaction effect with smoking using the Welch test (**Supplementary Note**). A threshold of $P<1.79\times10^{-4}$ (Bonferroni corrected for 279 tests) indicated significance.

We further tested for interaction between the weighted GRS and smoking, within 303,619 unrelated individuals of European ancestry in UK Biobank, using COPD and FEV_1/FVC as outcomes (the FEV_1/FVC phenotype was pre-adjusted for age, age^2, sex, and height, and the residuals transformed as per the main GWAS analysis). For COPD (defined as $FEV_1/FVC<0.7$, and FEV_1 <80% predicted) the following logistic model was fitted:

COPD ~ genotyping array + 10 principal components + age + age² + sex + height + smoking status + weighted risk score + (smoking status × weighted risk score).

For FEV_1/FVC the following linear model was fitted:

FEV_1/FVC ~ genotyping array + 10 principal components + smoking status + weighted risk score + (smoking status x weighted risk score).

Proportion of variance explained

We calculated the proportion of variance explained by each of the previously reported (n=140) and new variants (n=139) associated with lung function using the formula:

$$\frac{\sum_{i=1}^{n} 2f_i(1-f_i)\beta_i^2}{V}$$

where n is the number of variants f_i and β_i are the frequency and effect estimate of the i'th variant, and V is the phenotypic variance (always 1 as our phenotypes were inverse-normal transformed). We used the same unbiased effect estimates (β) as used to calculate GRS weights at the same set of 279 sentinel variants used for the GRS, which uses either UK Biobank or SpiroMeta effect estimates (described above). Our previously published estimate of proportion of variance explained[18] used effect estimates derived from UK Biobank. We assumed a heritability of 40%[56,57] to estimate the proportion of additive polygenic variance.

Fine-mapping

A Bayesian method[58] was used to fine-map lung-function-associated signals to the set of variants that were 99% likely to contain the underlying causal variant (assuming that the causal variant has been analysed). This was undertaken for new signals and for previously reported signals reaching $P<10^{-5}$ in UK Biobank. For the previously reported signals, the top sentinel variant from the current analysis in UK Biobank was used, instead of the previously reported variant. We used a value of 0.04 for the prior W in the approximate Bayes factor formula[59]. Effect sizes and standard errors for fine-mapping were obtained from an inverse variance weighted meta-analysis of UK Biobank and SpiroMeta (n up to 400,102). Signals in the HLA region were not included.

Implication of potentially causal genes

Annotation of deleterious variants

Variants in the 99% credible sets were checked for predicted functional effect if they were annotated as "exonic", "splicing", "ncRNA_exonic", "5' UTR" or "3' UTR" (untranslated region) by ANNOVAR[60]. We then used SIFT, PolyPhen-2 (implemented using the Ensembl GRCh37 Variant Effect Predictor, see URLs, accessed 1 February 2018) and FATHMM[61] to annotate missense variants, and CADD (also implemented using VEP) to annotate non-coding variation. Variants were annotated as deleterious in our study if they were labelled 'deleterious' by SIFT, 'probably damaging' or 'possibly damaging' by PolyPhen-2, 'damaging' by FATHMM (specifying the 'Inherited disease' option of the coding variants methods, and setting the prediction algorithm to 'Unweighted') or had a CADD scaled score ≥20 [4]. The union of the four methods was taken to establish the number of potentially deleterious variants and their unique genes.

Gene expression and protein levels

At each novel and previously reported signal, the sentinel variant and 99% credible set[58] were used to query three eQTL resources: lung eQTL (n=1,111)[13], blood eQTL (n=4,896)[62] and GTEx (V7; with n up to 388 depending on tissue: Artery Aorta (n=267), Artery Coronary (n=152), Artery Tibial (n=388), Colon Sigmoid (n=203), Colon Transverse (n=246), Esophagus Gastroesophageal Junction (n=213), Esophagus Muscularis (n=335), Lung (n=383), Small Intestine Terminal Ileum (n=122), Stomach (n=237), and Whole Blood (n=369))[63], and one blood pQTL resource (n=3,301)[34].

A gene was classified as a 'putative causal gene' if the sentinel SNP or any SNP in the respective 99% credible set was associated with expression of this gene or its protein levels (FDR<5% for eQTL, $P<5.03\times10^{-8}$ [for 276 tests at 3,600 proteins] for pQTL) and if the GWAS sentinel SNP or any SNP in the respective 99% credible set was also the variant most strongly associated with expression of the respective gene or level of the respective protein (i.e. the sentinel eQTL/pQTL SNP) in one or more of the eQTL and pQTL data sets. The HLA region was excluded from these analyses.

Pathway analysis

We tested for enrichment of genes identified via variant function annotation, gene expression or protein level analyses in pathway and gene set ontology databases using ConsensusPathdb. Pathways or gene sets represented entirely by genes implicated by the same association signal were excluded. Gene sets and pathways with FDR<5% are reported.

Functional enrichment analyses

We tested for cell-specific enrichment of lung function associated variants in regulatory regions using FORGE[35] (v1.1). One thousand background SNP set repetitions were used. Thresholds $P<1.68\times10^{-4}$ (FDR<2%; >99th percentile) and $P<3.37\times10^{-5}$ (FDR<0.5%; >99.9th percentile) were taken as being 'indicative' and 'significant', respectively. FORGE analysis was carried out for the cell lines in the RoadMap Epigenome project[33] (n=299 cell lines) and ENCODE projects[64] (n=125) separately.

Using DeepSEA[36], we analysed all SNPs in the 99% credible set for predicted chromatin effects. We reported effects for any chromatin effect and lung-related cell line that had an E-value<0.05 (i.e. the expected proportion of SNPs with a larger predicted effect based on empirical distributions of predicted effects for 1000 Genomes SNPs) and an absolute difference in probability of >0.1 (threshold for "high confidence") between the reference and alternative allele.

Drug targets

Genes identified as potentially causal using eQTL, pQTL or variant annotation were interrogated against the gene-drug interactions table of the Drug-Gene Interactions Database (DGIDB) (see URLs), accessed 16th October 2017. Drugs were mapped to CHEMBL IDs (see URLs), and indications (as MeSH headings) were added.

Phenome-wide association studies

To identify whether any of the new or previously reported signals overlap with signals of association for other traits and diseases, the 279 variant weighted GRS was calculated in UK Biobank samples (n up to 379,337) and a phenome-wide association study (PheWAS) across all available traits was performed, with the risk score as the exposure. Traits included UK Biobank baseline measures (from both questionnaires and physical measures), self-reported medication usage, and operative procedures, as well as those captured in Office of Population Censuses and Surveys codes from the electronic health record. We also included self-reported disease variables and those from hospital episode statistics (ICD-10 codes truncated to three-character codes and combined in block and chapter groups) as well as combining both self-report and hospital diagnosed diseases where possible to maximise power. The GRS analysis included 2,453 traits, of which 2,411 were also included in the single-variant analysis (traits with >200 cases were included for the individual SNP PheWAS, whereas traits with >50 cases were included in the risk score PheWAS). Analyses were conducted in unrelated European ancestry individuals (KING kinship coefficient of <0.0442), and were adjusted for age, sex, genotyping array, and ten principal components. Logistic models were fitted for binary outcome, and linear models were fitted for quantitative outcomes. False discovery rates were calculated according to the number of the traits in each analysis (2,453 or 2,411, for the risk score and single-variant PheWAS, respectively).

In addition, the sentinel variants and variants within the 99% credible sets were queried against the GWAS catalog[65] (see URLs, accessed 5th February 2018) and GRASP[66] (see URLs, accessed 6th February 2018) for reported associations significant at $P<5\times10^{-8}$. Associations relating to methylation, expression, metabolite or protein levels, as well as lung function and COPD, were not included.

614 **Data availability statement**

615 UK Biobank GWAS summary statistics will be available via UK Biobank
616 (http://www.ukbiobank.ac.uk/). SpiroMeta GWAS summary statistics, and single-variant PheWAS
617 results will be made available by request.

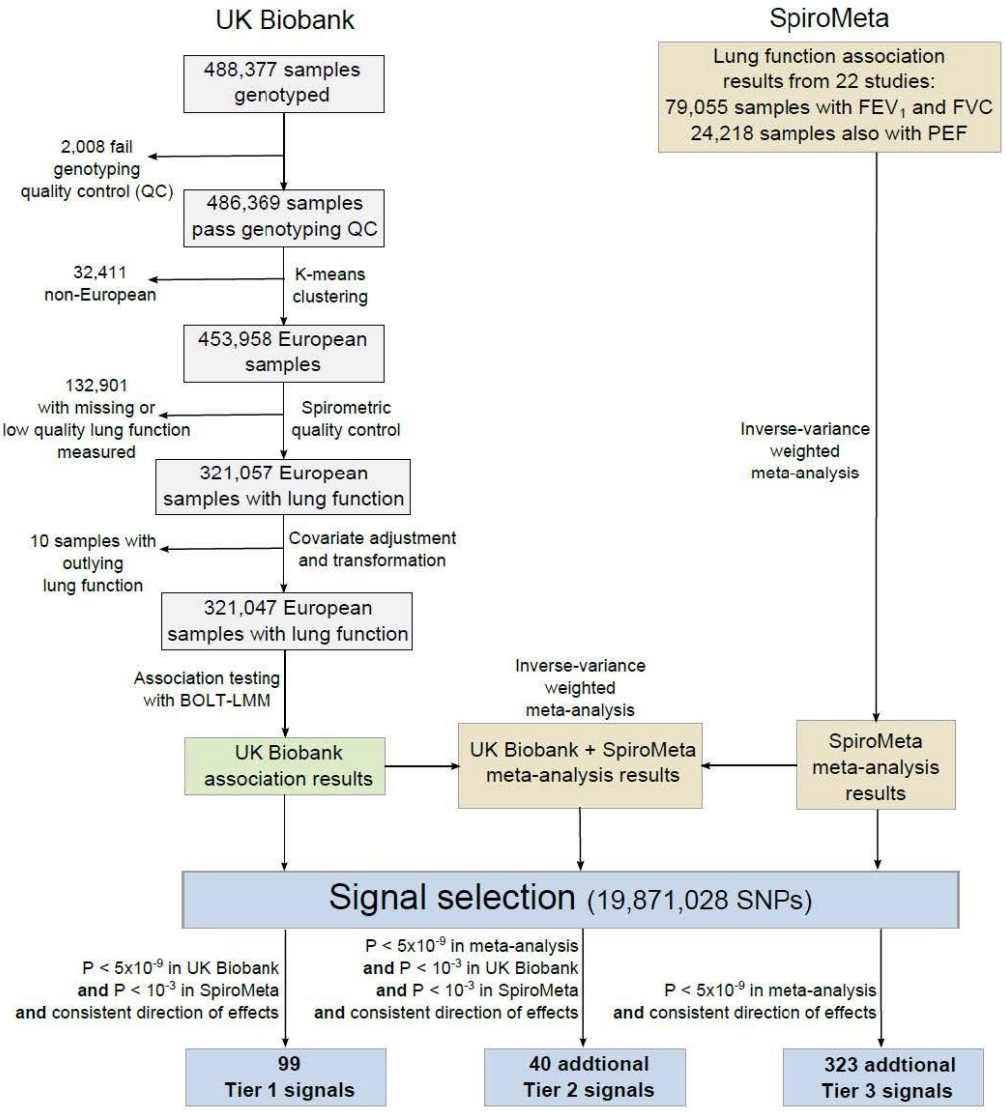


Figure 1: Study design

Tier 1 signals had $P<5\times10^{-9}$ in UK Biobank and $P<10^{-3}$ in SpiroMeta with consistent direction of effect. Tier 2 signals had $P<5\times10^{-9}$ in the meta-analysis of UK Biobank and SpiroMeta with $P<10^{-3}$ in UK Biobank and $P<10^{-3}$ in SpiroMeta with consistent directions of effect. Signals with $P<5\times10^{-9}$ in the meta-analysis of UK Biobank and SpiroMeta, and that had consistent directions of effect but did not meet $P<10^{-3}$ in both cohorts were reported as Tier 3.

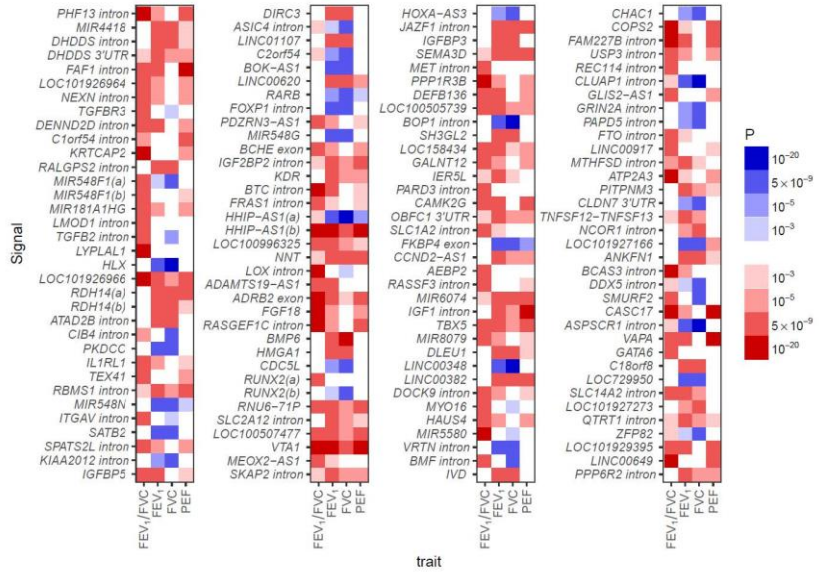
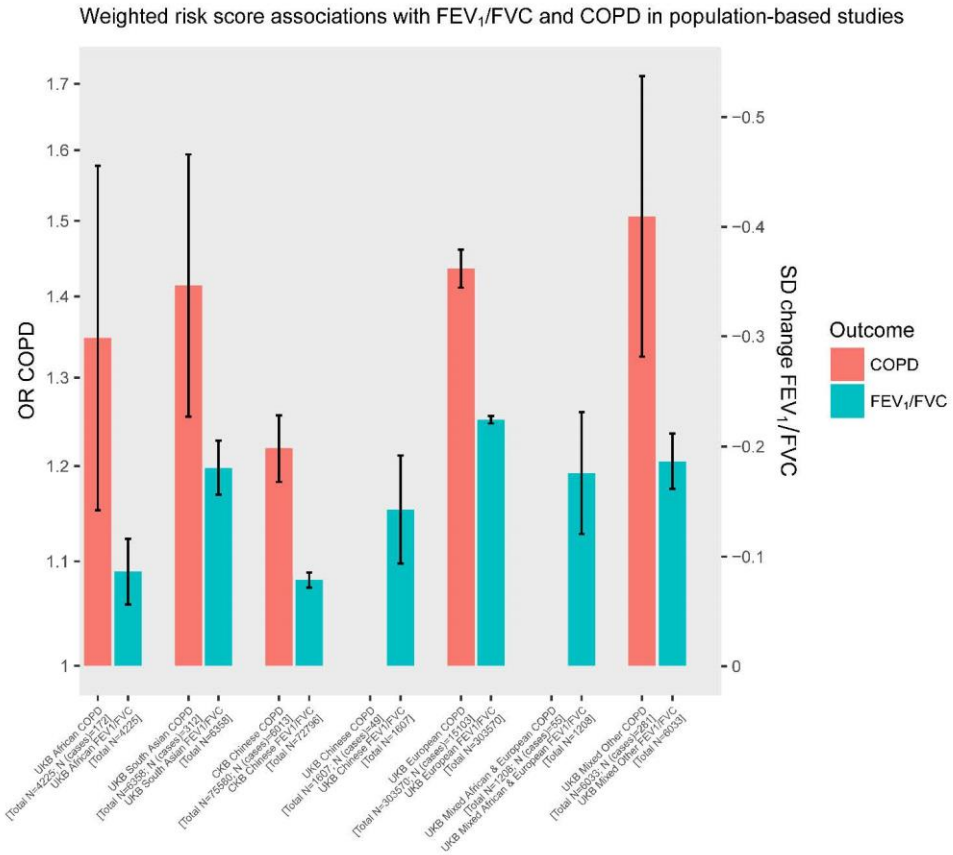


Figure 2: Strength and direction of association across four lung function traits for 139 novel signals: Red indicates decrease in the lung function trait; blue indicates an increase. All effects are aligned to the allele associated with decreased FEV_1/FVC, hence the FEV_1/FVC column is only red or white. P-values are from the meta-analysis of UK Biobank and SpiroMeta (n=400,102). The scale points are thresholds used for (i) confirmation in 2-stage analysis and 1-stage analysis ($P<10^{-3}$); (ii) confirmation of association of previous signals ($P<10^{-5}$); (iii) signal selection in 2-stage and 1-stage analysis ($P<5\times10^{-9}$); capped at ($P<10^{-20}$).



Ancestral group and phenotype studied in UK Biobank or China Kadoorie Biobank

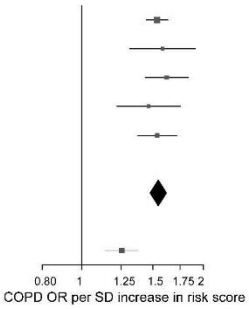
630

19

631 **Figure 3: Association of weighted genetic risk score (GRS) with COPD and FEV_1/FVC.**
632 **A.** Association of weighted genetic risk score (GRS) with COPD and FEV_1/FVC in UK Biobank and China Kadoorie
633 Biobank (CKB). The left axis denotes odds ratios (OR) for COPD per 1 standard deviation (SD) increase in
634 weighted GRS (OR for COPD shown only for ancestries in UK Biobank with > 100 cases of COPD). COPD was
635 defined as FEV_1/FVC < 0.7 and FEV_1 < 0.8 of the predicted value, i.e. GOLD stage 2-4 categorisation. Bars (in
636 red) are labelled with ancestral groups, and the total sample size and number of COPD cases are given. The
637 right-hand axis denotes change in standard deviation (SD) units of FEV_1/FVC per 1 SD increase in weighted
638 GRS in the same individuals (blue bars). For means and standard deviations of the risk scores in each group,
639 see **Supplementary Table 18.** Note some variants featuring in the GRS were discovered in UK Biobank
640 individuals of European ancestry. The height of the bars represents the effect estimate, and the black
641 whiskers represent 95% confidence intervals. There were 13 SNPs with MAF <0.1% in at least one ancestral
642 group: 13/279 in Chinese (of which 4/13 were monomorphic). Two of the 13 SNPs that were monomorphic in
643 Chinese people had MAF<0.1% in Africans.

Weighted risk score associations with COPD susceptibility in COPD case-control studies

Ancestry	Cohort	OR	95%LCI	95%UCI	P	Cases	Controls
European	COPDGene (EUR)	1.54	1.44	1.63	1.97×10^{-41}	3068	2110
	ECLIPSE	1.59	1.31	1.91	1.42×10^{-06}	1713	147
	GenKOLS	1.62	1.44	1.83	8.99×10^{-15}	836	692
	NETT−NAS	1.46	1.22	1.75	3.13×10^{-05}	374	429
	SPIROMICS	1.54	1.38	1.72	4.47×10^{-14}	988	537
	Meta−analysis	**1.55**	**1.48**	**1.62**	$\mathbf{1.48 \times 10^{-75}}$	**6979**	**3915**
African	COPDGene (AFR)	1.26	1.15	1.37	8.36×10^{-07}	910	1556

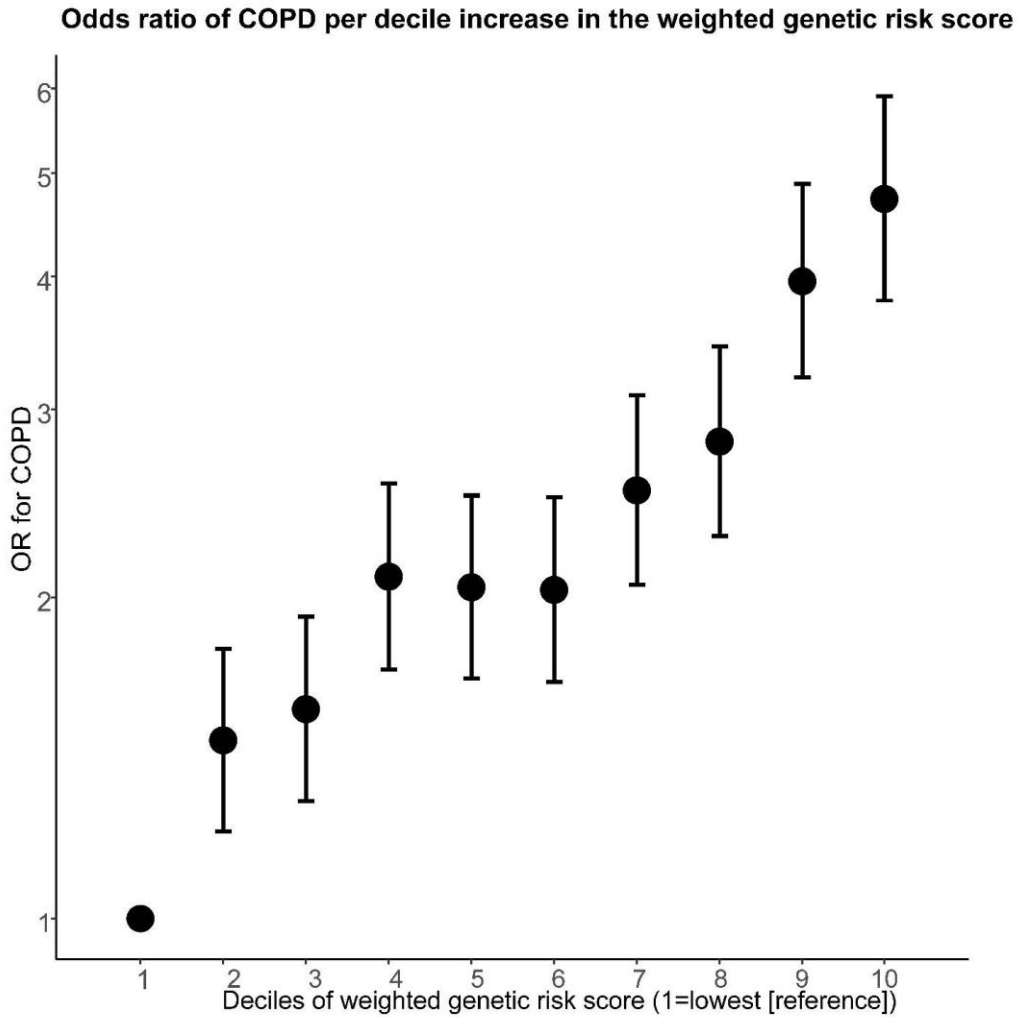


644

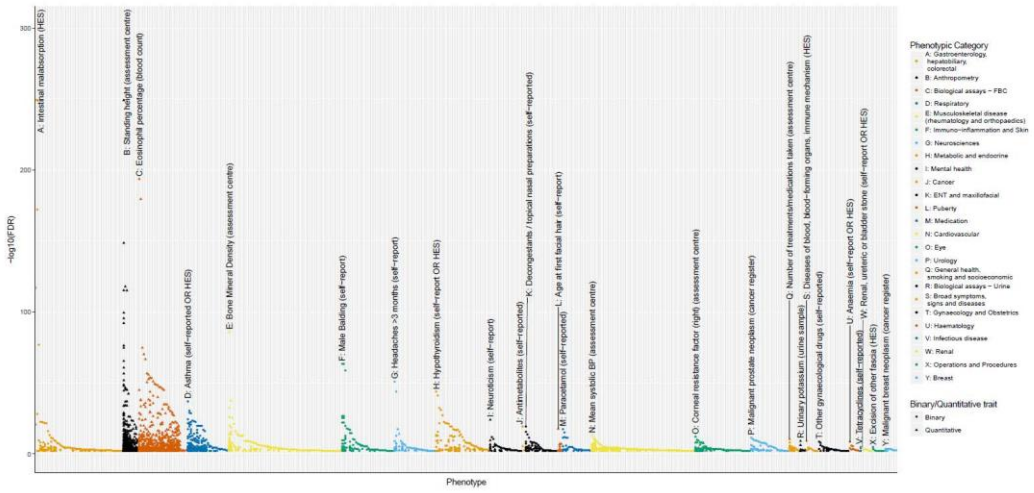
21

645 **B.** Odds ratio (OR) for COPD per 1 standard deviation (SD) increase in weighted genetic risk score in each of six study groups (COPDGene [Non-Hispanic White],
646 COPDGene [African-American], ECLIPSE, GenKOLS, SPIROMICS, NETT-NAS). COPD was defined using GOLD 2-4 criteria. For means and standard deviations of the risk
647 scores in each group see **Supplementary Table 20**. The vertical black line indicates the null effect (an OR of 1). The point estimate of each study is represented by a
648 box proportional to the study's weight, with the lines representing the lower and upper bounds of the 95% confidence interval. A fixed effect meta-analysis of the
649 five European-ancestry groups is denoted with a diamond, the width of which represents the 95% confidence interval for the estimate (I^2 statistic=0).

22



650
651
652 **C.** Odds ratios (OR) for COPD according to membership of deciles 2-10 of the weighted genetic risk score, with
653 decile 1 as the reference group (the 10% of individuals with the lowest genetic risk score). Each point
654 represents a meta-analysis of results for a given comparison (i.e. decile 2 vs reference, decile 3 vs reference ...
655 decile 10 versus reference) in five external European-ancestry study groups (COPDGene, ECLIPSE, GenKOLS,
656 SPIROMICS, NETT-NAS). Deciles were calculated and models were run in each group separately. Points
657 represent odds ratios, and error bars correspond to 95% confidence intervals (**Supplementary Table 21**).



658

659 **Figure 4: Individual PheWAS with 279 variants (traits passing FDR 1% threshold)**
660 Separate association of 279 variants with 2,411 traits (FDR<1%) in UK Biobank (n up to 379,337). In each category, the trait with the strongest association, i.e. highest –
661 $\log_{10}$(FDR), is shown first, followed by other traits in that category in descending order of $-\log_{10}$(FDR). Categories are colour-coded, and outcomes are denoted with a
662 circular or triangular point, according to whether they were coded as binary or quantitative. The top association per-category is labelled with its rsID number, and a plain
663 English label describing the trait. The letter at the beginning of each label allows easy cross-reference with the categories labelled in the legend. Zoomed in versions of each
664 category with visible trait names are available in **Supplementary Figure 9**.

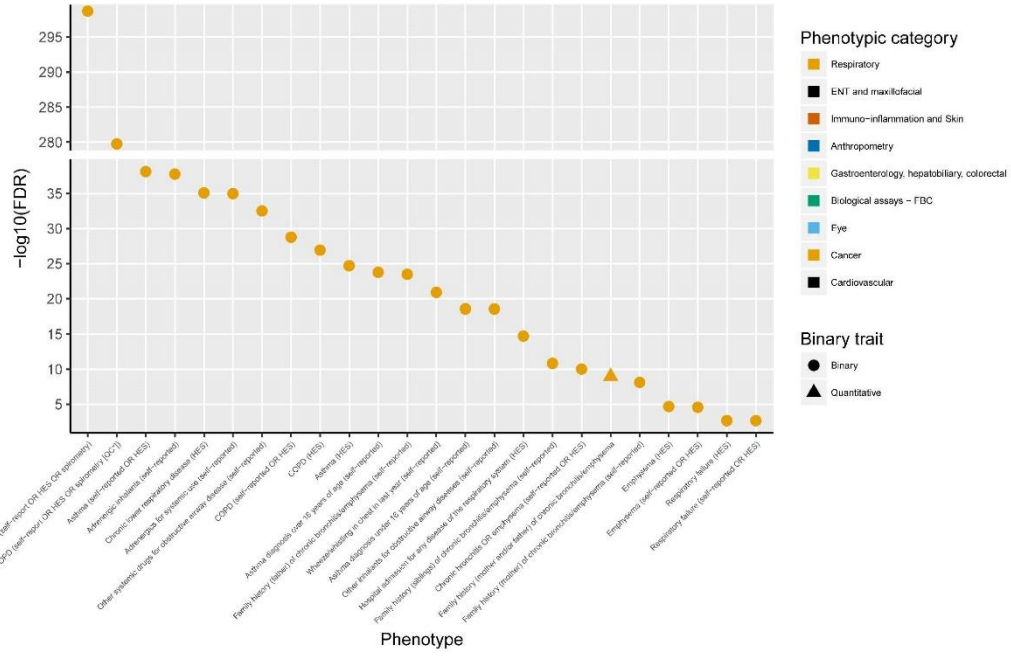
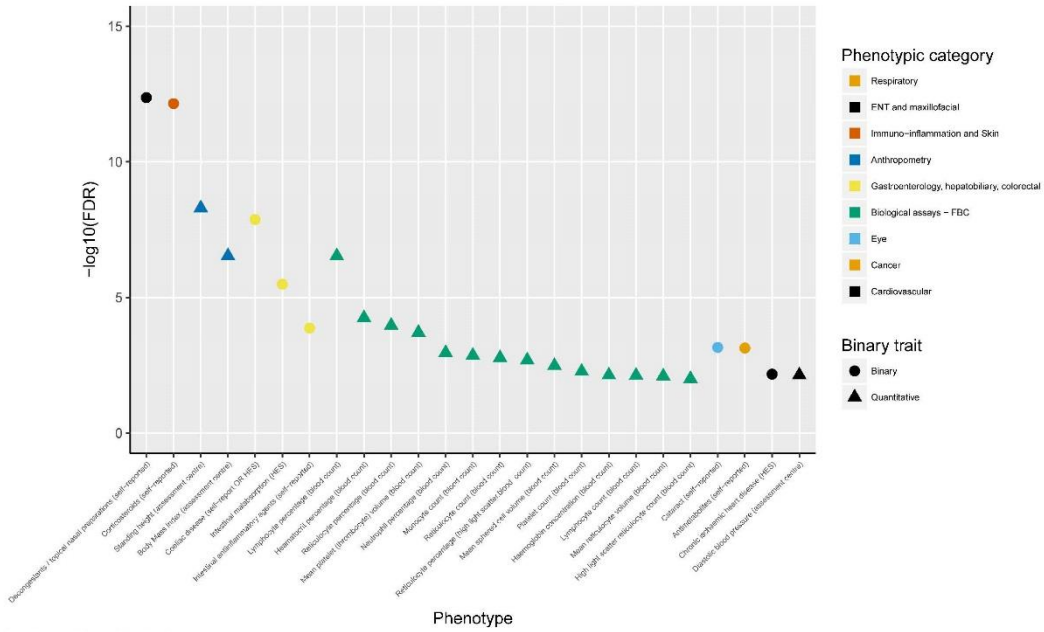


Figure 5: PheWAS with genetic risk score (traits passing FDR 1% threshold)
Association of 279 variant weighted genetic risk score with 2,453 traits (FDR<1%) in UK Biobank (n up to 379,337). In each panel, the category with the strongest association, i.e. highest $-\log_{10}$(FDR), is shown first, followed by all other associations in that category, ordered by descending order of $-\log_{10}$(FDR). Sample sizes varied across traits and are available in **Supplementary Table 22**, along with the full summary statistics for each association, plus details of categorisation and plain English labels for each trait. Trait categories are colour coded, and outcomes are denoted with a circular or triangular point, according to whether they were coded as binary or quantitative. *QC refers to spirometry passing ERS/ATS criteria.

A. Associations with respiratory traits.



673
674 **B.** Associations with all other traits.

28

675 **Table 1: Genes implicated using gene expression data, protein level data and functional annotation**
676 †Genes implicated by eQTL signals: Lung eQTL (n=1,111) and Blood eQTL (n=4,896) datasets and eleven GTEx (V7) tissues were screened: Artery Aorta (n=267), Artery
677 Coronary (n=152), Artery Tibial (n=388), Colon Sigmoid (n=203), Colon Transverse (n=246), Esophagus Gastroesophageal Junction (n=213), Esophagus Muscularis (n=335),
678 Lung (n=383), Small Intestine Terminal Ileum (n=122), Stomach (n=237), and Whole Blood (n=369); see **Supplementary Table 13** for direction of gene expression for the
679 COPD (lung function reducing) risk allele.
680 ‡Genes implicated by pQTL signals: pQLT look up in 3,600 plasma proteins (n up to 3,300).
681 *Genes implicated because they contain a deleterious variant (**Supplementary Table** 11).
682 "Other traits" column lists the other lung function traits for which the sentinel was associated at $P<5\times10^{-9}$ in the meta-analysis of UK Biobank and SpiroMeta.

Gene	Phenotype	Other traits	Novel Tier/ Previous	Sentinel SNP	Position (b37)	COPD risk/alt	Functionally implicated genes
DHDDS (intron)	FVC	FEV_1	Tier 2	rs9438626	1:26,775,367	G/C	*DHDDS†, DRAM2†*
DHDDS (3' UTR)	FEV_1		Tier 1	rs12096239	1:26,796,922	C/G	*HMGN2†, DHDDS†*
NEXN (intron)	FEV_1/FVC	FEV_1	Tier 1	rs9661687	1:78,387,270	T/C	*NEXN†*
DENND2D (intron)	FEV_1/FVC		Tier 1	rs9970286	1:111,737,398	G/A	*CEPT1†, CHI3L2†*
C1orf54 (intron)	PEF	FVC	Tier 1	rs11205354	1:150,249,101	C/A	*MRPS21†, RPRD2†, ECM1‡*
KRTCAP2	FEV_1/FVC		Tier1	rs141942982	1: 155153537	T/C	*THBS4‡*
RALGPS2 (intron)	FEV_1		Tier 1	rs4651005	1:178,719,306	C/T	*ANGPTL1†*
LMOD1 (intron)	FEV_1/FVC	FEV_1	Tier 2	rs4309038	1:201,884,647	G/C	*SHISA4†*
ATAD2B (intron)	FVC	FEV_1	Tier 2	rs13009582	2:24,018,480	G/A	*UBXN2A†*
PKDCC	FVC		Tier 1	rs4952564	2:42,243,850	A/G	*PKDCC†*
ITGAV (intron)	FEV_1/FVC		Tier 1	rs2084448	2:187,530,520	C/T	*ITGAV†*
SPATS2L (intron)	FEV_1/FVC		Tier 2	rs985256	2:201,208,692	C/A	*SPATS2L†*
C2orf54	FVC	FEV_1	Tier 1	rs6437219	2:241,844,033	C/T	*C2orf54†**
MIR548G	FVC		Tier 1	rs1610265	3:99,420,192	T/C	*FILIP1L†*
BCHE (exon)	FEV_1/FVC	FEV_1	Tier 1	rs1799807	3:165,548,529	C/T	*BCHE**
BTC (intron)	FEV_1/FVC	FEV_1/FVC	Tier 1	rs62316310	4:75,676,529	G/A	*BTC**
LOC100996325	FEV_1	FEV_1/FVC, PEF	Tier 1	rs11739847	5:609,661	A/G	*CEP72**
RNU6-71P	FEV_1	FVC, PEF	Tier 1	rs2894837	6:56,336,406	G/A	*DST**
JAZF1 (intron)	FEV_1		Tier 1	rs1513272	7:28,200,097	C/T	*JAZF1†*
MET (intron)	FEV_1/FVC		Tier 2	rs193686	7:116,431,427	T/C	*MET†*
IER5L	FEV_1		Tier 2	rs967497	9:131,943,843	G/A	*CRAT†, PPP2R4†, IER5L**
DOCK9	FEV_1/FVC		Tier 1	rs11620380	13:99,665,512	A/C	*DOCK9**
CHAC1	FVC		Tier 1	rs4924525	15:41,255,396	A/C	*INO80†, CHP1†, RAD51†*
ATP2A3	FEV_1/FVC		Tier 1	rs8082036	17:3,882,613	G/C	*ATP2A3†*
PITPNM3	FEV_1		Tier 2	rs4796334	17:6,469,793	A/G	*KIAA0753†*, TXNDC17†, PITPNM3†*

29

Gene	Phenotype	Other traits	Novel Tier/ Previous	Sentinel SNP	Position (b37)	COPD risk/alt	Functionally implicated genes
TNFSF12-TNFSF13	FEV_1		Tier 2	rs4968200	17:7,448,457	C/G	*TNFSF13*†, *SENP3*†
NCOR1 (intron)	FVC	FEV_1	Tier 2	rs34351630	17:16,030,520	C/T	*ADORA2B*†, *TTC19*†
ASPSCR1 (intron)	FVC	FEV_1	Tier 1	rs59606152	17:79,952,944	C/T	*LRRC45**
C18orf8	FVC		Tier 1	rs303752	18:21,074,255	A/G	*C18orf8*†
ZFP82	FVC	FVC, PEF	Tier 2	rs2967516	19:36,881,643	A/G	*ZFP14*†, *ZFP82*†
MFAP2	FEV_1/FVC	FEV_1, PEF	Previous	rs9435733	1:17,308,254	C/T	*MFAP2*†
LOC101929516	FEV_1/FVC		Previous	rs755249	1:39,995,074	T/C	*PABPC4*†
TGFB2	PEF	FEV_1/FVC	Previous	rs6604614	1:218,631,452	C/G	*TGFB2*†
TRAF3IP1	FEV_1	FVC, FEV_1/FVC, PEF	Previous	rs6710301	2:239,441,308	C/A	*ASB1**
SLMAP (intron)	FEV_1	FEV_1	Previous	rs6445932	3:57,879,611	T/G	*SLMAP*†
RSRC1 (intron)	FVC	FVC, FEV_1/FVC	Previous	rs12634907	3:158,226,886	G/A	*RSRC1*†
GSTCD (intron)	FEV_1	FEV_1, FVC, PEF	Previous	rs11722225	4:106,766,430	T/C	*INTS12*†
NPNT (intron)	FEV_1/FVC		Previous	rs34712979	4:106,819,053	A/G	*NPNT*†‡
AP3B1 (intron)	FVC		Previous	rs425102	5:77,396,400	G/T	*AP3B1*†
SPATA9	FEV_1/FVC		Previous	rs987068	5:95,025,146	C/G	*RHOBTB3*†
P4HA2-AS1	FVC	FEV_1, PEF	Previous	rs3843503	5:131,466,629	A/T	*SLC22A5*†, *P4HA2*†, *C1QTNF5*‡
CYFIP2 (intron)	FEV_1/FVC	FEV_1, PEF	Previous	rs11134766	5:156,908,317	T/C	*ADAM19*†
ADAM19 (intron)	FEV_1/FVC		Previous	rs11134789	5:156,944,199	A/C	*ADAM19*†*
DSP (intron)	FEV_1/FVC	FEV_1	Previous	rs2076295	6:7,563,232	T/G	*DSP*†
MIR588	FVC	FVC, PEF	Previous	rs6918725	6:126,990,392	T/G	*CENPW*†
GPR126 (exon)	FEV_1/FVC		Previous	rs17280293	6:142,688,969	A/G	*GPR126**
C1GALT1 (intron)	FEV_1/FVC	FEV_1	Previous	rs4318980	7:7,256,490	A/G	*C1GALT1*†
QSOX2 (3' UTR)	FVC		Previous	rs7024579	9:139,100,413	T/C	*QSOX2*†
DNLZ (intron)	FVC	FEV_1, FVC, PEF	Previous	rs4073153	9:139,259,349	G/A	*SNAPC4*†, *CARD9*†, *INPP5E*†
CDC123 (intron)	FEV_1/FVC	FEV_1	Previous	rs7090277	10:12,278,021	T/A	*NUDT5*†
MYPN (intron)	FVC	FVC	Previous	rs10998018	10:69,962,954	A/G	*MYPN**
EML3 (intron)	FEV_1	FEV_1	Previous	rs71490394	11:62,370,155	G/A	*EEF1G*†, *ROM1*†*, *EML3*†*
ARHGEF17 (intron)	FEV_1/FVC		Previous	rs2027761	11:73,036,179	C/T	*FAM168A*†, *ARHGEF17*†*
RAB5B (intron)	FEV_1	PEF	Previous	rs1689510	12:56,396,768	C/G	*CDK2*†
LRP1 (intron)	FEV_1/FVC		Previous	rs11172113	12:57,527,283	T/C	*LRP1*†
FGD6 (intron)	FEV_1/FVC		Previous	rs113745635	12:95,554,771	T/C	*FGD6*†
RPAP1	FEV_1/FVC		Previous	rs2012453	15:41,840,238	G/A	*ITPKA*†, *LTK*†, *TYRO3*†, *RPAP1*†

30

Gene	Phenotype	Other traits	Novel Tier/ Previous	Sentinel SNP	Position (b37)	COPD risk/alt	Functionally implicated genes
AAGAB	FVC	FEV_1, PEF	Previous	rs12917612	15:67,491,274	A/C	*AAGAB*†, *SMAD3*†, *IQCH*†
THSD4 (intron)	FEV_1/FVC		Previous	rs1441358	15:71,612,514	G/T	*THSD4*†
IL27	FEV_1		Previous	rs12446589	16:28,870,962	A/G	*SBK1*†, *TUFM*†, *CCDC101*†, *SULT1A1*†, *SULT1A2*†*, *SH2B1*†, *NPIPB7*†, *CLN3*†, *ATXN2L*†, *EIF3C*†
MMP15 (intron)	FEV_1/FVC	PEF	Previous	rs11648508	16:58,063,513	G/T	*MMP15*†
SSH2 (intron)	FEV_1/FVC	FEV_1	Previous	rs2244592	17:28,072,327	A/G	*EFCAB5*†
FBXL20 (intron)	FVC	FVC, PEF	Previous	rs8069451	17:37,504,933	C/T	*CRKRS*†, *FBXL20*†
MAPT-AS1	FEV_1		Previous	rs79412431	17:43,940,021	A/G	*LRRC37A4*†, *MAPT**
TSEN54 (intron)	FEV_1	PEF	Previous	rs9892893	17:73,525,670	G/T	*CASKIN2*†, *TSEN54**
LTBP4 (exon)	FEV_1/FVC		Previous	rs34093919	19:41,117,300	G/A	*LTBP4**
ABHD12 (intron)	FEV_1	FEV_1, PEF	Previous	rs2236180	20:25,282,608	C/T	*PYGB*†*
UQCC1 (5' UTR)	FVC	FEV_1	Previous	rs143384	20:34,025,756	G/A	*UQCC1*†, *GDF5*†
SLC2A4RG (intron)	FVC	FEV_1/FVC	Previous	rs4809221	20:62,372,706	A/G	*LIME1*†
SCARF2 (intron)	FEV_1	FEV_1	Previous	rs9610955	22:20,790,723	C/G	*SCARF2**‡

683

31

References

1. Young, R.P., Hopkins, R. & Eaton, T.E. Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes. *Eur Respir J* **30**, 616-22 (2007).
2. Lefaudeux, D. *et al.* U-BIOPRED clinical adult asthma clusters linked to a subset of sputum omics. *J Allergy Clin Immunol* **139**, 1797-1807 (2017).
3. Hekking, P.P. *et al.* Pathway discovery using transcriptomic profiles in adult-onset severe asthma. *J Allergy Clin Immunol* (2017).
4. Hobbs, B.D. *et al.* Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat Genet* **49**, 426-432 (2017).
5. Salvi, S.S. & Barnes, P.J. Chronic obstructive pulmonary disease in non-smokers. *Lancet* **374**, 733-43 (2009).
6. Nelson, M.R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856-60 (2015).
7. Wilk, J.B. *et al.* A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet* **5**, e1000429 (2009).
8. Repapi, E. *et al.* Genome-wide association study identifies five loci associated with lung function. *Nat Genet* **42**, 36-44 (2010).
9. Hancock, D.B. *et al.* Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet* **42**, 45-52 (2010).
10. Soler Artigas, M. *et al.* Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet* **43**, 1082-90 (2011).
11. Cho, M.H. *et al.* A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet* **21**, 947-57 (2012).
12. Loth, D.W. *et al.* Genome-wide association analysis identifies six new loci associated with forced vital capacity. **46**, 669-77 (2014).
13. Wain, L.V. *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* **3**, 769-81 (2015).
14. Lutz, S.M. *et al.* A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet* **16**, 138 (2015).
15. Soler Artigas, M. *et al.* Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nat Commun* **6**, 8658 (2015).
16. Hobbs, B.D. *et al.* Exome Array Analysis Identifies a Common Variant in IL27 Associated with Chronic Obstructive Pulmonary Disease. **194**, 48-57 (2016).
17. Jackson, V. *et al.* Meta-analysis of exome array data identifies six novel genetic loci for lung function [version 1; referees: 1 approved with reservations]. *Wellcome Open Research* **3**(2018).
18. Wain, L.V. *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat Genet* **49**, 416-425 (2017).
19. Wyss, A.B. *et al.* Multiethnic Meta-analysis Identifies New Loci for Pulmonary Function. *bioRxiv* (2017).
20. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-90 (2015).

21. Benyamin, B. *et al.* GWAS of butyrylcholinesterase activity identifies four novel loci, independent effects within BCHE and secondary associations with metabolic risk factors. *Hum Mol Genet* **20**, 4504-14 (2011).
22. Hammarsjo, A., Wang, Z., Vaz, R. & Taylan, F. Novel KIAA0753 mutations extend the phenotype of skeletal ciliopathies. **7**, 15585 (2017).
23. Stephen, J. *et al.* Mutations in KIAA0753 cause Joubert syndrome associated with growth hormone deficiency. *Hum Genet* **136**, 399-408 (2017).
24. Loukil, A., Tormanen, K. & Sütterlin, C. The daughter centriole controls ciliogenesis by regulating Neurl-4 localization at the centrosome. *The Journal of Cell Biology* (2017).
25. He, R. *et al.* LRRC45 is a centrosome linker component required for centrosome cohesion. *Cell Rep* **4**, 1100-7 (2013).
26. Conkar, D. *et al.* The centriolar satellite protein CCDC66 interacts with CEP290 and functions in cilium formation and trafficking. **130**, 1450-1462 (2017).
27. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**(2015).
28. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet* **8**, e1003029 (2012).
29. Lamontagne, M. *et al.* Refining susceptibility loci of chronic obstructive pulmonary disease with lung eqtls. *PLoS One* **8**, e70220 (2013).
30. Obeidat, M. *et al.* GSTCD and INTS12 regulation and expression in the human lung. *PLoS One* **8**, e74630 (2013).
31. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-1243 (2013).
32. Consortium, G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
33. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
34. Sun, B.B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).
35. Dunham, I., Kulesha, E., Iotchkova, V., Morganella, S. & Birney, E. FORGE: A tool to discover cell specific enrichments of GWAS associated SNPs in regulatory regions. *F1000Research* **4**(2015).
36. Zhou, J. & Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931-4 (2015).
37. Cotto, K.C. *et al.* DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Research*, gkx1143-gkx1143 (2017).
38. Slack, R. *et al.* P112 Discovery of a Novel, High Affinity, Small Molecule αvβ6 Inhibitor for the Treatment of Idiopathic Pulmonary Fibrosis. *QJM: An International Journal of Medicine* **109**, S60-S60 (2016).
39. Raab-Westphal, S., Marshall, J.F. & Goodman, S.L. Integrins as Therapeutic Targets: Successes and Cancers. *Cancers (Basel)* **9**(2017).
40. Merrill, J.T. *et al.* Efficacy and Safety of Atacicept in Patients With Systemic Lupus Erythematosus: Results of a Twenty-Four-Week, Multicenter, Randomized, Double-Blind, Placebo-Controlled, Parallel-Arm, Phase IIb Study. *Arthritis Rheumatol* **70**, 266-276 (2018).
41. Aschard, H. *et al.* Evidence for large-scale gene-by-smoking interaction effects on pulmonary function. *Int J Epidemiol* **46**, 894-904 (2017).
42. Pulley, J.M. *et al.* Accelerating Precision Drug Development and Drug Repurposing by Leveraging Human Genetics. *ASSAY and Drug Development Technologies* **15**, 113-119 (2017).

781 43. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and
782 body mass index in ~700,000 individuals of European ancestry. *bioRxiv* (2018).
783 44. Wain, L.V. *et al.* Whole exome re-sequencing implicates CCDC38 and cilia structure
784 and function in resistance to smoking related airflow obstruction. *PLoS Genet* **10**,
785 e1004314 (2014).
786 45. Black, P.N. *et al.* Changes in elastic fibres in the small airways and alveoli in COPD.
787 *Eur Respir J* **31**, 998-1004 (2008).
788 46. Løkke, A., Lange, P., Scharling, H., Fabricius, P. & Vestbo, J. Developing COPD: a
789 25 year follow up study of the general population. *Thorax* **61**, 935-939 (2006).
790 47. Pulit, S.L., With, S.A.J. & Bakker, P.I.W. Resetting the bar: Statistical significance in
791 whole-genome sequencing-based association studies of global populations. *Genetic*
792 *Epidemiology* **41**, 145-151 (2017).
793 48. Martinez, F.J. *et al.* A New Approach for Identifying Patients with Undiagnosed
794 Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* **195**, 748-756
795 (2017).
796 49. Strug, L.J. *et al.* Cystic fibrosis gene modifier SLC26A9 modulates airway response
797 to CFTR-directed therapeutics. *Hum Mol Genet* **25**, 4590-4600 (2016).
798 50. Pulit, S.L., de With, S.A. & de Bakker, P.I. Resetting the bar: Statistical significance
799 in whole-genome sequencing-based association studies of global populations. *Genet*
800 *Epidemiol* **41**, 145-151 (2017).
801 51. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants.
802 *bioRxiv* (2017).
803 52. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from
804 polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291 (2015).
805 53. The 1000 Genomes Project, C. A map of human genome variation from population-
806 scale sequencing. *Nature* **467**, 1061 (2010).
807 54. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation.
808 *Nature genetics* **48**, 1279-1283 (2016).
809 55. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide
810 complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
811 56. Palmer, L.J. *et al.* Familial aggregation and heritability of adult lung function: results
812 from the Busselton Health Study. *Eur Respir J* **17**, 696-702 (2001).
813 57. Wilk, J.B. *et al.* Evidence for major genes influencing pulmonary function in the
814 NHLBI family heart study. *Genet Epidemiol* **19**, 81-94 (2000).
815 58. Wakefield, J. Reporting and interpretation in genome-wide association studies. *Int J*
816 *Epidemiol* **37**, 641-53 (2008).
817 59. van de Bunt, M., Cortes, A., Brown, M.A., Morris, A.P. & McCarthy, M.I. Evaluating
818 the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. *PLoS*
819 *Genet* **11**, e1005535 (2015).
820 60. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic
821 variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
822 61. Shihab, H.A. *et al.* Predicting the functional, molecular, and phenotypic consequences
823 of amino acid substitutions using hidden Markov models. *Hum Mutat* **34**, 57-65
824 (2013).
825 62. Jansen, R. *et al.* Conditional eQTL analysis reveals allelic heterogeneity of gene
826 expression. *Hum Mol Genet* **26**, 1444-1451 (2017).
827 63. Battle, A., Brown, C.D., Engelhardt, B.E. & Montgomery, S.B. Genetic effects on
828 gene expression across human tissues. *Nature* **550**, 204-213 (2017).
829 64. Consortium, E.P. *et al.* An integrated encyclopedia of DNA elements in the human
830 genome. *Nature* **489**, 57-74 (2012).

831 65. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide
832 association studies (GWAS Catalog). *Nucleic Acids Research* **45**, D896-D901 (2017).
833 66. Leslie, R., O'Donnell, C.J. & Johnson, A.D. GRASP: analysis of genotype-phenotype
834 results from 1390 genome-wide association studies and corresponding open access
835 database. *Bioinformatics* **30**, i185-94 (2014).

836

837 **URLs**

838 UK Biobank: http://www.ukbiobank.ac.uk

839 Variant Effect Predictor: https://www.ensembl.org/vep

840 Drug-Gene Interactions Database (DGIDB): http://www.dgidb.org/data/

841 CHEMBL: https://www.ebi.ac.uk/chembl/drug/indications

842 GWAS catalog: https://www.ebi.ac.uk/gwas/

843 GRASP: https://grasp.nhlbi.nih.gov/Overview.aspx

844

Affiliations:

1. Department of Health Sciences, University of Leicester, Leicester, LE1 7RH, UK
2. Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Pde, 3052 Parkville, Australia
3. Department of Medical Biology, University of Melbourne, 3010 Parkville, Australia
4. Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA
5. Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA
6. Target Sciences, GlaxoSmithKline, Collegeville, PA, US
7. Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand
8. Division of Genetics, Genomics and Precision Medicine, Department of Medicine, University of Arizona, Tucson AZ, USA
9. Nuffield Department of Population Health, University of Oxford, Oxford, OX3 7LF, UK
10. Medical Research Council Population Health Research Unit, University of Oxford, Oxford, OX3 7LF, UK
11. The University of British Columbia Centre for Heart Lung Innovation, St Paul's Hospital, Vancouver, BC, Canada
12. MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, CB2 0QQ, UK
13. Department of Epidemiology and Biostatistics, MRC–PHE Centre for Environment & Health, School of Public Health, Imperial College London, London, W2 1PG, UK
14. A list of contributors can be found in the Supplementary Appendix
15. Interfaculty Institute for Genetics and Functional Genomics, Department of Functional Genomics, University Medicine Greifswald, 17475 Greifswald, Germany
16. Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, EH8 9AG, Scotland
17. Centre for Cardiovascular Sciences, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, EH16 4TJ, Scotland
18. Department of Biostatistics, University of Liverpool, Liverpool L69 3GL, UK
19. MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK
20. Flatiron Institute, Simons Foundation, New York, New York, USA
21. Busselton Population Medical Research Institute, Sir Charles Gairdner Hospital, Nedlands WA 6009, Australia
22. School of Population Health, The University of Western Australia, Crawley WA 6009, Australia
23. PathWest Laboratory Medicine of WA, Sir Charles Gairdner Hospital, Crawley WA 6009, Australia
24. School of Pathology and Laboratory Medicine, The University of Western Australia, Crawley WA 6009, Australia
25. Institute of Epidemiology, Helmholtz Zentrum Muenchen – German Research Center for Environmental Health, Neuherberg, Germany
26. Institute and Outpatient Clinic for Occupational, Social and Environmental Medicine, Ludwig-Maximilians-Universität, Munich, Germany

889 27. Comprehensive Pneumology Center Munich (CPC-M), Member of the German Center for Lung
890 Research (DZL), Neuherberg, Germany
891 28. Swiss Tropical and Public Health Institute, Basel, Switzerland
892 29. University of Basel, Switzerland
893 30. Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh
894 EH8 9JZ, UK
895 31. Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics
896 and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK
897 32. Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine,
898 University of Edinburgh, Edinburgh EH4 2XU, UK
899 33. Department of Immunology, Genetics and Pathology, Uppsala Universitet, Science for Life
900 Laboratory, Husargatan 3, Uppsala, SE-75108, Sweden
901 34. Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
902 35. The National Institute for Health and Welfare (THL), Helsinki, Finland
903 36. Department of Clinical Chemistry, Fimlab Laboratories, and Finnish Cardiovascular Research
904 Center - Tampere, Faculty of Medicine and Life Sciences, University of Tampere, Tampere 33520,
905 Finland
906 37. Department of Clinical Science, University of Bergen, Norway
907 38. Department of Epidemiology, Johns Hopkins University School of Public Health, Baltimore, M.D.,
908 USA 21205
909 39. Department of Molecular Medicine, Laval University, Québec, Canada
910 40. Institut Universitaire de Cardiologie et de Pneumologie de Québec, Laval University, Québec,
911 Canada
912 41. University of Groningen, University Medical Center Groningen, Department of Pathology and
913 Medical Biology, GRIAC Research Institute, University of Groningen, Groningen, The Netherlands
914 42. National Jewish Health, Denver, CO, USA
915 43. Division of Pulmonary, Critical Care and Sleep Medicine, National Jewish Health, Denver, CO,
916 USA
917 44. British Heart Foundation Cambridge Centre of Excellence, Division of Cardiovascular Medicine,
918 Addenbrooke's Hospital, Cambridge CB2 0QQ, UK
919 45. Department of Human Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome
920 Campus, Hinxton, Cambridge CB10 1RQ, UK
921 46. NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Department of Public
922 Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK
923 47. Department of Internal Medicine B - Cardiology, Intensive Care, Pulmonary Medicine and
924 Infectious Diseases, University Medicine Greifswald, 17475 Greifswald, Germany
925 48. Department of Molecular Epidemiology, Institute of Epidemiology II, Helmholtz Zentrum
926 Muenchen – German Research Center for Environmental Health, Neuherberg, Germany
927 49. UK Small Area Health Statistics Unit, MRC-PHE Centre for Environment and Health, School of
928 Public Health, Imperial College London, London, UK
929 50. Imperial College Healthcare NHS Trust, St Mary's Hospital, Paddington, London, UK
930 51. Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai,
931 New York, NY, USA
932 52. Department of Epidemiology, University of Colorado Anschutz Medical Campus, Aurora,
933 Colorado, USA

934 53. Department of Molecular Biology, Medical Biochemistry, and Pathology, Laval University,
935 Québec, Canada
936 54. Department of Epidemiology & Biostatistics, Peking University Health Science Centre, Peking
937 University, Beijing 100191, China
938 55. Department of Medical Sciences, Cardiovascular Epidemiology, Uppsala University, Uppsala 751
939 85, Sweden
940 56. GSK R&D, Collegeville, PA, US
941 57. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK
942 58. MRL, Merck & Co., Inc., Kenilworth, New Jersey, USA
943 59. The Institute of Medical Sciences, Aberdeen Biomedical Imaging Centre, University of Aberdeen,
944 Aberdeen AB25 2ZD, UK
945 60. Merck Research Laboratories, Genetics and Pharmacogenomics, Boston, MA, USA
946 61. Centre for Genomic and Experimental Medicine, Institute of Genetics & Molecular Medicine,
947 University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK
948 62. Centre for Cognitive Ageing and Cognitive Epidemiology, Department of Psychology, The
949 University of Edinburgh, 7 George Square, Edinburgh, EH8 9JZ, UK
950 63. Division of Respiratory Medicine and NIHR-Nottingham Biomedical Research Centre, University
951 of Nottingham
952 64. Respiratory Division, Department of Medicine, University of British Columbia, Vancouver, BC,
953 Canada
954 65. Division of Population Health Sciences, Ninewells Hospital and Medical School, University of
955 Dundee, Dundee DD1 9SY, UK
956 66. Psychiatric Genetics Unit, Group of Psychiatry, Mental Health and Addiction, Vall d'Hebron
957 Research Institute (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain
958 67. Department of Psychiatry, Hospital Universitari Vall d'Hebron, Barcelona, Spain
959 68. Biomedical Network Research Centre on Mental Health (CIBERSAM), Instituto de Salud Carlos
960 III, Barcelona, Spain
961 69. VA Boston Healthcare System, Boston, MA, USA
962 70. Department of Medicine, Boston University School of Medicine, Boston, MA USA
963 71. University of Groningen, University Medical Center Groningen, Department of Pulmonology,
964 GRIAC Research Institute, University of Groningen, Groningen, The Netherlands
965 72. Target Sciences - R&D, GSK Medicines Research Centre, Gunnels Wood Road, Stevenage,
966 Hertfordshire, SG1 2NY, UK
967 73. UCSF Pulmonary, Critical Care, Allergy and Sleep Medicine, University of California San
968 Francisco, USA
969 74. Department of Computer Science, Princeton University, Princeton, New Jersey, USA
970 75. Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey,
971 USA
972 76. Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku
973 20521, Finland
974 77. Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku
975 20520, Finland
976 78. Department of Clinical Physiology, Tampere University Hospital, and Finnish Cardiovascular
977 Research Center - Tampere, Faculty of Medicine and Life Sciences, University of Tampere, Tampere
978 33521, Finland

79. University of Split School of Medicine, Split, Croatia
80. Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, Scotland
81. Department of Psychology, University of Edinburgh, Edinburgh, EH8 9JZ, UK
82. Department of Pulmonary Physiology and Sleep Medicine, Sir Charles Gairdner Hospital, Nedlands WA 6009, Australia
83. School of Medicine and Pharmacology, The University of Western Australia, Crawley 6009, Australia
84. MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, Scotland
85. Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK
86. Center for Life Course Health Research, Faculty of Medicine, University of Oulu, 90014 Oulu, Finland
87. Biocenter Oulu, University of Oulu, Finland
88. Unit of Primary Care, Oulu University Hospital, Kajaanintie 50, P.O. Box 20, FI-90220 Oulu 90029 OYS, Finland
89. MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH42 XU, UK
90. Division of Genetics, Genomics and Precision Medicine, University of Arizona, Tucson AZ, USA
91. Population Health Research Institute, St George's, University of London, London SW17 0RE, UK
92. National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, LE3 9QP, UK

Study contributions:

All authors critically reviewed the manuscript prior to submission.

Contributed to the conception and design of the study: K.S., U.S.S.G., S.K., S.M.K., T.L., P.S.B., T.H.B., E.R.B., Y.B., Z.C., J.D.C., J.D., D.L.D., C.G., A.G., K.H., J.D.H., J.Hokanson, P.J., C.L., L.Li, N.L., J.C.M., H.R., I.Sayers, D.D.S., R. T-S., J.C.W., P.W., L.M.Y., O.T.R., M.K., O.P., U.G., I.R., I.J.D., N.M.P., H.S., A.L.J., J.F.W., E.Z., M.J., N.W., A.S.B., R.A.S., D.A.M., M.H.C., D.P.S., I.P.H., M.D.T., L.V.W.

Undertook data analysis: N.S., A.L.G., A.M.E., V.E.J., B.D.H., C.M., C.Batini, K.A.F., K.S., P.S., Xingnan Li, R.B., N.F.R., M.O., J.Zhao, M.W., S.W., K.A.K., J.P.C., B.B.S., J.Zhou, J.Hui, M.I., S.E.H., J.M., S.E., I.Surakka, V.V., T.L., R.J.A., C.Brandsma, F.D., J.D.H., P.K.J., Xuan Li, A.Mahajan, J.C.M., D.C.N., M.M.P., D.Prokopenko, D.Q., R.R., H.R., D.S., P.R.H.J.T., M.V., L.M.Y., O.G.T., N.M.P., N.W., E.K.S., C.H., A.P.M., A.S.B., R.A.S., M.H.C., D.P.S., M.D.T., L.V.W.

Contributed to data acquisition and/or interpretation: N.S., A.L.G., A.M.E., V.E.J., C.M., C.Batini, K.A.F., K.S., P.S., Xingnan Li, N.F.R., M.O., M.W., K.A.K., B.B.S., S.K., M.I., R.J.A., J.D., F.D., R.E., C.G., A.G., A.L.H., J.D.H., G.H., P.K.J., C.L., Xuan Li, K.L., L.Lind, J.L., J.C.M., A.Murray, R.P., M.M.P., M.L.P., D.Porteous, D.Prokopenko, D.Q., R.R., H.R., I.Sayers, B.H.S., M.S., L.M.Y., O.G.T., N.M.P., H.S., J.F.W., B.S., M.J., N.W., C.H., A.P.M., A.S.B., R.A.S., R.G.W., M.H.C., D.P.S., I.P.H., M.D.T., L.V.W.

Drafted the manuscript: N.S., A.L.G., A.M.E., I.P.H., M.D.T., L.V.W.

Funding and Acknowledgments:

The following authors report specific personal funding from the following grants:

B.D.H.: NIH K08 HL136928, Parker B. Francis Research Opportunity Award

M.W.: EU Horizon2020 (633212 ALEC)

U.S.S.G.: Economic and Social Research Council (ES/H029745/1)

B.B.S.: Cambridge School of Clinical Medicine MB-PhD programme and MRC/Sackler Prize PhD Studentship (MR/K50127X/1).

J.D.: J.D. is a BHF Professor, European Research Council Senior Investigator, and NIHR Senior Investigator.

I.Sayers: MRC (G1000861)

D.S.: D.S. is supported by a VA Research Career Scientist award.

E.Z.: Wellcome Trust (WT098051)

A.P.M.: Wellcome Trust (WT098017 & WT064890)

M.H.C.: This work was supported by NHLBI grants R01HL113264, R01HL137927 (M.H.C. and E.K.S.), R01HL135142 (M.H.C.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funding body has no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

I.P.H.: The research was partially supported by the NIHR Nottingham Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

M.D.T.: M.D. Tobin is supported by a Wellcome Trust Investigator Award (WT202849/Z/16/Z). M.D. Tobin and L.V. Wain have been supported by the MRC (MR/N011317/1). The research was partially

supported by the NIHR Leicester Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

L.V.W.: L.V. Wain holds a GSK/British Lung Foundation Chair in Respiratory Research.

Funding and acknowledgment statements from individual cohorts:

China Kadoorie Biobank: Kadoorie Charitable Foundation in Hong Kong; the UK Wellcome Trust (grant numbers 202922/Z/16/Z, 088158/Z/09/Z, 104085/Z/14/Z); Chinese Ministry of Science and Technology (grant number 2011BAI09B01); National Natural Science Foundation of China (Grant numbers 81390540, 81390541, 81390544); UK Medical Research Council (grant numbers MC_PC_13049, MC_PC_14135); GlaxoSmithKline; BHF Centre of Research Excellence, Oxford (grant number RE/13/1/30181); British Heart Foundation; and Cancer Research UK. The China Kadoorie Biobank participants; CKB project staff based at Beijing, Oxford and the 10 regional centres; the China National Centre for Disease Control and Prevention (CDC) and its regional offices for assisting with the fieldwork. We acknowledge the assistance of BGI (Shenzhen, China) for conducting DNA extraction and genotyping.

COPDGene: See Supplementary Note.

ECLIPSE: The ECLIPSE study (NCT00292552; GSK code SCO104960) was funded by GSK.

GenKOLS: The Norway GenKOLS study (Genetics of Chronic Obstructive Lung Disease, GSK code RES11080) was funded by GSK.

NETTNAS: The National Emphysema Treatment Trial was supported by the NHLBI N01HR76101, N01HR76102, N01HR76103, N01HR76104, N01HR76105, N01HR76106, N01HR76107, N01HR76108, N01HR76109, N01HR76110, N01HR76111, N01HR76112, N01HR76113, N01HR76114, N01HR76115, N01HR76116, N01HR76118 and N01HR76119, the Centers for Medicare and Medicaid Services and the Agency for Healthcare Research and Quality. The Normative Aging Study is supported by the Cooperative Studies Program/ERIC of the US Department of Veterans Affairs and is a component of the Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC).

SPIROMICS: NIH/NHLBI grants: HHSN268200900013C, HHSN268200900014C, HHSN268200900015C, HHSN268200900016C, HHSN268200900017C, HHSN268200900018C, HHSN268200900019C (Eugene Bleecker), HHSN268200900020C. Supplemented by contributions made through the Foundation for the NIH from AstraZeneca; Bellerophon Pharmaceuticals; Boehringer-Ingelheim Pharmaceuticals, Inc; Chiesi Farmaceutici SpA; Forest Research Institute, Inc; GSK; Grifols Therapeutics, Inc; Ikaria, Inc; Nycomed GmbH; Takeda Pharmaceutical Company; Novartis Pharmaceuticals Corporation; Regeneron Pharmaceuticals, Inc; and Sanofi. The authors thank the SPIROMICS participants and participating physicians, investigators and staff for making this research possible. More information about the study and how to access SPIROMICS data is at www.spiromics.org. We would like to acknowledge the following current and former investigators of the SPIROMICS sites and reading centers: Neil E Alexis, PhD; Wayne H Anderson, PhD; R Graham Barr, MD, DrPH; Eugene R Bleecker, MD; Richard C Boucher, MD; Russell P Bowler, MD, PhD; Elizabeth E Carretta, MPH; Stephanie A Christenson, MD; Alejandro P Comellas, MD; Christopher B Cooper, MD, PhD; David J Couper, PhD; Gerard J Criner, MD; Ronald G Crystal, MD; Jeffrey L Curtis, MD; Claire M Doerschuk, MD; Mark T Dransfield, MD; Christine M Freeman, PhD; MeiLan K Han, MD, MS; Nadia N Hansel, MD, MPH; Annette T Hastie, PhD; Eric A Hoffman, PhD; Robert J Kaner, MD; Richard E Kanner, MD; Eric C

1083 Kleerup, MD; Jerry A Krishnan, MD, PhD; Lisa M LaVange, PhD; Stephen C Lazarus, MD; Fernando J
1084 Martinez, MD, MS; Deborah A Meyers, PhD; John D Newell Jr, MD; Elizabeth C Oelsner, MD, MPH;
1085 Wanda K O'Neal, PhD; Robert Paine, III, MD; Nirupama Putcha, MD, MHS; Stephen I. Rennard, MD;
1086 Donald P Tashkin, MD; Mary Beth Scholand, MD; J Michael Wells, MD; Robert A Wise, MD; and
1087 Prescott G Woodruff, MD, MPH. The project officers from the Lung Division of the National Heart,
1088 Lung, and Blood Institute were Lisa Postow, PhD, and Thomas Croxton, PhD, MD.

1089 DeepSEA: The DeepSEA project was primarily supported by US National Institutes of Health (NIH)
1090 grants R01 GM071966 to O.G.T and supported in part by US NIH grant P50 GM071508. O.G.T. is a
1091 senior fellow of the Genetic Networks program of the Canadian Institute for Advanced Research
1092 (CIFAR). Jian Zhou and Olga G Troyanskaya carried out SNV effect prediction analyses in an updated
1093 version of DeepSEA (deepsea.princeton.edu), which contained more lung-related cell lines than the
1094 original version published in Nature Methods in August 2015
1095 (www.nature.com/articles/nmeth.3547).

1096 GSK: This research has been conducted using the UK Biobank Resource under application number
1097 26041.

1098 B58C: Genotyping for the B58C-WTCCC subset was funded by the Wellcome Trust (076113/B/04/Z).
1099 The B58C-T1DGC genotyping utilized resources provided by the Type 1 Diabetes Genetics
1100 Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and
1101 Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID),
1102 National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human
1103 Development (NICHD), and Juvenile Diabetes Research Foundation International (JDRF) and
1104 supported by U01 DK062418. B58C-T1DGC GWAS data were deposited by the Diabetes and
1105 Inflammation Laboratory, Cambridge Institute for Medical Research (CIMR), University of
1106 Cambridge, which is funded by Juvenile Diabetes Research Foundation International, the Wellcome
1107 Trust and the National Institute for Health Research Cambridge Biomedical Research Centre; the
1108 CIMR is in receipt of a Wellcome Trust Strategic Award (079895). The B58C-GABRIEL genotyping was
1109 supported by a contract from the European Commission Framework Programme 6 (018996) and
1110 grants from the French Ministry of Research.

1111 BHS: Generous support for the 1994/5 follow-up study from Healthway, Western Australia, and
1112 support from The Great Wine Estates of the Margaret River region of Western Australia. The study
1113 also acknowledges the numerous Busselton community volunteers who assisted with data collection
1114 and the study participants from the Shire of Busselton.

1115 CROATIA-Korcula/Split/Vis: MRC, the Ministry of Science, Education and Sport in the Republic of
1116 Croatia (216-1080315-0302) and the Croatian Science Foundation (grant 8875) (funding to I. Rudan,
1117 C. Hayward, S.M. Kerr, O. Polasek, V. Vitart, and J. Marten).

1118 H2000: Medical Research Fund of the Tampere University Hospital.

1119 KORA F4 and KORA S3: The KORA study was initiated and financed by the Helmholtz Zentrum
1120 München – German Research Center for Environmental Health, which is funded by the German
1121 Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA
1122 research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-
1123 Maximilians-Universität, as part of LMUinnovativ and by the Competence Network Asthma and

1124 COPD (ASCONET), network COSYCONET (subproject 2, BMBF FKZ 01GI0882) funded by the German
1125 Federal Ministry of Education and Research (BMBF).

1126 LBC1936: Phenotype collection in the Lothian Birth Cohort 1936 was supported by Age UK (The
1127 Disconnected Mind project). Genotyping was funded by the Biotechnology and Biological Sciences
1128 Research Council (BBSRC). The work was undertaken by The University of Edinburgh Centre for
1129 Cognitive Ageing and Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing
1130 Initiative (MR/K026992/1). Funding from the BBSRC and MRC is gratefully acknowledged.

1131 NSPHS: Swedish Medical Research Council (K2007-66X-20270-01-3, 2011-5252, 2012-2884 and
1132 2011-2354), the Foundation for Strategic Research (SSF). NSPHS as part of European Special
1133 Populations Research Network (EUROSPAN) was also supported by the European Commission FP6
1134 STRP (01947, LSHG-CT-2006-01947).

1135 ORCADES: Supported by the Chief Scientist Office of the Scottish Government (CZB/4/276,
1136 CZB/4/710), the Royal Society, the MRC Human Genetics Unit, Arthritis Research UK and the
1137 European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947).
1138 ORCADES DNA extractions were performed at the Wellcome Trust Clinical Research Facility in
1139 Edinburgh.

1140 SAPALDIA: Swiss National Science Foundation (33CS30-148470/1&2, 33CSCO-134276/1, 33CSCO-
1141 108796, 324730_135673, 3247BO-104283, 3247BO-104288, 3247BO-104284, 3247-065896, 3100-
1142 059302, 3200-052720, 3200-042532, 4026-028099, PMPDP3_129021/1, PMPDP3_141671/1), the
1143 Federal Office for the Environment, the Federal Office of Public Health, the Federal Office of Roads
1144 and Transport, the canton's government of Aargau, Basel-Stadt, Basel-Land, Geneva, Luzern, Ticino,
1145 Valais, and Zürich, the Swiss Lung League, the canton's Lung League of Basel Stadt/ Basel Landschaft,
1146 Geneva, Ticino, Valais, Graubünden and Zurich, Stiftung ehemals Bündner Heilstätten, SUVA,
1147 Freiwillige Akademische Gesellschaft, UBS Wealth Foundation, Talecris Biotherapeutics GmbH,
1148 Abbott Diagnostics, European Commission 018996 (GABRIEL) and the Wellcome Trust (WT
1149 084703MA). The SAPALDIA study could not have been done without the help of the study
1150 participants, technical and administrative support and the medical teams and field workers at the
1151 local study sites. Local fieldworkers : Aarau: S Brun, G Giger, M Sperisen, M Stahel, Basel: C Bürli, C
1152 Dahler, N Oertli, I Harreh, F Karrer, G Novicic, N Wyttenbacher, Davos: A Saner, P Senn, R Winzeler,
1153 Geneva: F Bonfils, B Blicharz, C Landolt, J Rochat, Lugano: S Boccia, E Gehrig, MT Mandia, G Solari, B
1154 Viscardi, Montana: AP Bieri, C Darioly, M Maire, Payerne: F Ding, P Danieli A Vonnez, Wald: D
1155 Bodmer, E Hochstrasser, R Kunz, C Meier, J Rakic, U Schafroth, A Walder.

1156 YFS: The Young Finns Study has been financially supported by the Academy of Finland: grants
1157 286284, 134309 (Eye), 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi), and 41071 (Skidi);
1158 the Social Insurance Institution of Finland; Competitive State Research Financing of the Expert
1159 Responsibility area of Kuopio, Tampere and Turku University Hospitals (grant X51001); Juho Vainio
1160 Foundation; Paavo Nurmi Foundation; Finnish Foundation for Cardiovascular Research ; Finnish
1161 Cultural Foundation; Tampere Tuberculosis Foundation; Emil Aaltonen Foundation; Yrjö Jahnsson
1162 Foundation; Signe and Ane Gyllenberg Foundation; Diabetes Research Foundation of Finnish
1163 Diabetes Association; and EU Horizon 2020 (grant 755320 for TAXINOMISIS).

1164 EPIC: Cancer Research UK and the MRC. The authors thank the participants, General Practitioners
1165 and staff of the EPIC-Norfolk study team, and collaborators for their contribution.

Generation Scotland: Chief Scientist Office of the Scottish Government Health Directorate (CZD/16/6), the Scottish Funding Council (HR03006), Medical Research Council UK, and a Wellcome Trust Strategic Award 'STratifying Resilience and Depression Longitudinally' (104036/Z/14/Z)(STRADL). Genotyping of the GS:SFHS samples was carried out by the Edinburgh Clinical Research Facility, University of Edinburgh. We are grateful to all the families who took part in the Generation Scotland: Scottish Family Health Study, the general practitioners and Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes academic researchers, IT staff, laboratory technicians, statisticians and research managers.

NFBC1966: received financial support from the Academy of Finland (project grants 104781, 120315, 129269, 1114194, 24300796, Center of Excellence in Complex Disease Genetics and SALVE), University Hospital Oulu, Biocenter, University of Oulu, Finland (75617), NHLBI grant 5R01HL087679-02 through the STAMPEED program (1RL1MH083268-01), NIH/NIMH (5R01MH63706:02), ENGAGE project and grant agreement HEALTH-F4-2007-201413, EU FP7 EurHEALTHAgeing -277849, the Medical Research Council, UK (G0500539, G0600705, G1002319, PrevMetSyn/SALVE) and the MRC, Centenary Early Career Award. The program is currently being funded by the H2020-633595 DynaHEALTH action, academy of Finland EGEA-project (285547) and EU H2020 ALEC project (Grant Agreement 633212).

NFBC1986: received financial support from EU QLG1-CT-2000-01643 (EUROBLCS) Grant no. E51560, NorFA Grant no. 731, 20056, 30167, USA / NIHH 2000 G DF682 Grant no. 50945.

PIVUS: Swedish Research Council, Swedish Heart-Lung Foundation, Swedish Diabetes Foundation and Uppsala University. Genotyping and analysis was funded by the Wellcome Trust under awards WT064890 and WT098017. The PIVUS investigators express their deepest gratitude to the study participants.

SHIP / SHIP-Trend: Federal Ministry of Education and Research (01ZZ9603, 01ZZ0103, 01ZZ0403, 03ZIK012), and the German Research Foundation (GR 1912/5-1).

UKHLS: Economic and Social Research Council (ES/H029745/1). These data are from Understanding Society: The UK Household Longitudinal Study, which is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council. The data were collected by NatCen and the genome wide scan data were analysed by the Wellcome Trust Sanger Institute. The Understanding Society DAC have an application system for genetics data and all use of the data should be approved by them. The application form is at: https://www.understandingsociety.ac.uk/about/health/data. A full list of authors is available in the Supplementary Note.

VIKING: Supported by a MRC Human Genetics Unit quinquennial programme grant "QTL in Health and Disease". VIKING DNA extractions and genotyping were performed at the Edinburgh Clinical Research Facility, University of Edinburgh.

pQTL: The INTERVAL study: NHSBT (11-01-GEN) and the NIHR-BTRU in Donor Health and Genomics (NIHR BTRU-2014-10024) at the University of Cambridge in partnership with NHSBT. This study was partially funded by Merck. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health of England, or NHSBT. The Cardiovascular Epidemiology Unit at the University of Cambridge: UK MRC (G0800270), BHF (SP/09/002), UK NIHR

Cambridge Biomedical Research Centre, ERC (268834), and European Commission Framework Programme 7 (HEALTH-F2-2012-279233).

Competing interests:

The following authors report potential competing interests:

L.V.W.: Louise V. Wain has received grant support from GSK.

M.D.T.: Martin D. Tobin has received grant support from GSK.

I.P.H.: Ian P. Hall has received support from GSK and BI.

K.S.: Kijoung Song is an employee of GlaxoSmithKline and may own company stock.

J.D.: John Danesh reports personal fees and non-financial support from Merck Sharp & Dohme (MSD) and Novartis, and grants from British Heart Foundation, European Research Council, MSD, NIHR, NHS Blood and Transplant, Novartis, Pfizer, UK MRC, Wellcome Trust, and AstraZeneca.

J.D.H.: Josh D. Hoffman is an employee of GlaxoSmithKline and may own company stock.

N.L.: Nick Locantore is an employee of GlaxoSmithKline and may own company stock.

J.C.M.: Joseph. C. Maranville was a Merck employee during this study, and is now a Celgene employee.

H.R.: Heiko Runz was a Merck employee during this study, and is now a Biogen employee.

R. T-S.: Ruth Tal-Singer is an employee of GlaxoSmithKline and owns company stock.

J.C.W.: John C. Whittaker is an employee of GlaxoSmithKline and may own company stock.

L.M.Y.: Laura M. Yerges Armstrong is an employee of GlaxoSmithKline and may own company stock.

E.K.S.: In the past three years, Edwin K. Silverman received honoraria from Novartis for Continuing Medical Education Seminars and grant and travel support from GlaxoSmithKline

A.S.B.: Adam S. Butterworth reports grants from Merck, Pfizer, Novartis, Biogen and Bioverativ and personal fees from Novartis

R.A.S.: Robert A. Scott is an employee of GlaxoSmithKline and may own company stock.

M.H.C.: Michael H. Cho has received grant support from GSK.

References

(1) DeWan AT. Five classic articles in genetic epidemiology. - Yale J Biol Med.2010 Jun;83(2):87-90. (1551-4056 (Electronic); 0044-0086 (Linking)).

(2) Seyerle AA, Avery CL. Genetic epidemiology: the potential benefits and challenges of using genetic information to improve human health. - N C Med J.2013 Nov-Dec;74(6):505-8. (0029-2559 (Print); 0029-2559 (Linking)).

(3) Abbott S, Fairbanks DJ, Auid- Orcid: 0000-0001-7422-0549. Experiments on Plant Hybrids by Gregor Mendel. - Genetics.2016 Oct;204(2):407-422.doi: 10.1534/genetics.116.195198. (1943-2631 (Electronic); 0016-6731 (Linking)).

(4) Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. - Cell.2017 Jun 15;169(7):1177-1186.doi: 10.1016/j.cell.2017.05.038. (1097-4172 (Electronic); 0092-8674 (Linking)).

(5) Brown T. Genomes. 2nd ed.: Wiley-Liss; 2002.

(6) Lyle J. Palmer, Paul R. Burton, George Davey Smith. An introduction to genetic epidemiology. : Policy Press; 2011.

(7) Bush WS, Moore JH. Chapter 11: Genome-wide association studies. - PLoS Comput Biol.2012;8(12):e1002822.doi: 10.1371/journal.pcbi.1002822.Epub 2012 Dec 27. (1553-7358 (Electronic); 1553-734X (Linking)).

(8) Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. - Annu Rev Genet.2010;44:293-308.doi: 10.1146/annurev-genet-102209-163421. (1545-2948 (Electronic); 0066-4197 (Linking)).

(9) LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. - Nucleic Acids Res.2009 Jul;37(13):4181-93.doi: 10.1093/nar/gkp552.Epub 2009 Jul 1. (1362-4962 (Electronic); 0305-1048 (Linking)).

(10) Wang Z, Chatterjee N. Increasing mapping precision of genome-wide association studies: to genotype and impute, sequence, or both? - Genome Biol.2017 Jun 19;18(1):118.doi: 10.1186/s13059-017-1255-6. (1474-760X (Electronic); 1474-7596 (Linking)).

(11) Marchini J, Howie B, FAU - Myers S, Myers S, FAU - McVean G, McVean G, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. - Nat Genet.2007 Jul;39(7):906-13.doi: 10.1038/ng2088.Epub 2007 Jun 17. (1061-4036 (Print); 1061-4036 (Linking)).

(12) Li Y, Willer CJ, FAU - Ding J, Ding J, FAU - Scheet P, Scheet P, et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. - Genet Epidemiol.2010 Dec;34(8):816-34.doi: 10.1002/gepi.20533. (1098-2272 (Electronic); 0741-0395 (Linking)).

(13) Walter K, FAU - Min JL, Min JL, FAU - Huang J, Huang J, FAU - Crooks L, et al. The UK10K project identifies rare variants in health and disease. - Nature.2015 Oct 1;526(7571):82-90.doi: 10.1038/nature14962.Epub 2015 Sep 14. (1476-4687 (Electronic); 0028-0836 (Linking)).

(14) Auton A, FAU - Brooks LD, Brooks LD, FAU - Durbin RM, Durbin RM, FAU - Garrison EP, et al. A global reference for human genetic variation. - Nature.2015 Oct 1;526(7571):68-74.doi: 10.1038/nature15393. (1476-4687 (Electronic); 0028-0836 (Linking)).

(15) McCarthy S, Auid- Orcid: 0000-0002-2715-4187, Das S, Auid- Orcid: 0000-0001-6346-1590, Kretzschmar W, Auid- Orcid: 0000-0002-2575-0807, et al. A reference panel of 64,976 haplotypes for genotype imputation. - Nat Genet.2016 Oct;48(10):1279-83.doi: 10.1038/ng.3643.Epub 2016 Aug 22. (1546-1718 (Electronic); 1061-4036 (Linking)).

(16) Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. - Nat Rev Genet.2013 Feb;14(2):139-49.doi: 10.1038/nrg3377. (1471-0064 (Electronic); 1471-0056 (Linking)).

(17) Cichon S, FAU - Craddock N, Craddock N, FAU - Daly M, Daly M, FAU - Faraone SV, et al. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. - Am J Psychiatry.2009 May;166(5):540-56.doi: 10.1176/appi.ajp.2008.08091354.Epub 2009 Apr 1. (1535-7228 (Electronic); 0002-953X (Linking)).

(18) Dawn Teare M, Barrett JH. Genetic linkage studies. - Lancet.2005 Sep 17-23;366(9490):1036-44.doi: 10.1016/S0140-6736(05)67382-5. (1474-547X (Electronic); 0140-6736 (Linking)).

(19) Gusella JF, FAU - Wexler NS, Wexler NS, FAU - Conneally PM, Conneally PM, FAU - Naylor SL, et al. A polymorphic DNA marker genetically linked to Huntington's disease. - Nature.1983 Nov 17-23;306(5940):234-8. (0028-0836 (Print); 0028-0836 (Linking)).

(20) Ozaki K, Ohnishi Y, FAU - Iida A, Iida A, FAU - Sekine A, Sekine A, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. - Nat Genet.2002 Dec;32(4):650-4.doi: 10.1038/ng1047.Epub 2002 Nov 11. (1061-4036 (Print); 1061-4036 (Linking)).

(21) The International HapMap Project. - Nature.2003 Dec 18;426(6968):789-96.doi: 10.1038/nature02168. (1476-4687 (Electronic); 0028-0836 (Linking)).

(22) Anderson CA, Pettersson FH, FAU - Clarke GM, Clarke GM, FAU - Cardon LR, Cardon LR, et al. Data quality control in genetic case-control association studies. - Nat Protoc.2010 Sep;5(9):1564-73.doi: 10.1038/nprot.2010.116.Epub 2010 Aug 26. (1750-2799 (Electronic); 1750-2799 (Linking)).

(23) Stevens EL, Heckenberg G, FAU - Roberson EDO, Roberson ED, FAU - Baugher JD, Baugher JD, et al. Inference of relationships in population data using identity-by-

descent and identity-by-state. - PLoS Genet.2011 Sep;7(9):e1002287.doi: 10.1371/journal.pgen.1002287.Epub 2011 Sep 22. (1553-7404 (Electronic); 1553-7390 (Linking)).

(24) Patterson N, Price AL, FAU - Reich D, Reich D. Population structure and eigenanalysis. - PLoS Genet.2006 Dec;2(12):e190.doi: 10.1371/journal.pgen.0020190. (1553-7404 (Electronic); 1553-7390 (Linking)).

(25) Bouaziz M, Ambroise C, FAU - Guedj M, Guedj M. Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. - PLoS One.2011;6(12):e28845.doi: 10.1371/journal.pone.0028845.Epub 2011 Dec 21. (1932-6203 (Electronic); 1932-6203 (Linking)).

(26) Turner S, Armstrong LL, FAU - Bradford Y, Bradford Y, FAU - Carlson CS, Carlson CS, et al. Quality control procedures for genome-wide association studies. - Curr Protoc Hum Genet.2011 Jan;Chapter 1:Unit1.19.doi: 10.1002/0471142905.hg0119s68. (1934-8258 (Electronic); 1934-8258 (Linking)).

(27) Devlin B, Roeder K. Genomic control for association studies. - Biometrics.1999 Dec;55(4):997-1004. (0006-341X (Print); 0006-341X (Linking)).

(28) Yang J, Weedon MN, FAU - Purcell S, Purcell S, FAU - Lettre G, Lettre G, et al. Genomic inflation factors under polygenic inheritance. - Eur J Hum Genet.2011 Jul;19(7):807-12.doi: 10.1038/ejhg.2011.39.Epub 2011 Mar 16. (1476-5438 (Electronic); 1018-4813 (Linking)).

(29) Pruim RJ, Welch RP, FAU - Sanna S, Sanna S, FAU - Teslovich TM, Teslovich TM, et al. LocusZoom: regional visualization of genome-wide association scan results. - Bioinformatics.2010 Sep 15;26(18):2336-7.doi: 10.1093/bioinformatics/btq419.Epub 2010 Jul 15. (1367-4811 (Electronic); 1367-4803 (Linking)).

(30) Purcell S, Neale B, FAU - Todd-Brown K, Todd-Brown K, FAU - Thomas L, Thomas L, 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.doi: 10.1086/519795.Epub 2007 Jul 25. (0002-9297 (Print); 0002-9297 (Linking)).

(31) Panagiotou OA, Ioannidis JP. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. - Int J Epidemiol.2012 Feb;41(1):273-86.doi: 10.1093/ije/dyr178.Epub 2011 Dec 5. (1464-3685 (Electronic); 0300-5771 (Linking)).

(32) Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne C, et al. New genetic signals for lung function highlight pathways and pleiotropy, and chronic obstructive pulmonary disease associations across multiple ancestries. bioRxiv 2018 Cold Spring Harbor Laboratory:343293.

(33) Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA,et al. New genetic signals for lung function highlight pathways and chronic

obstructive pulmonary disease associations across multiple ancestries. - Nat Genet [In press] doi: 10.1038/s41588-018-0321-7 .

(34) Zhou JJ, Cho MH, FAU - Castaldi PJ, Castaldi PJ, FAU - Hersh CP, Hersh CP, et al. Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. - Am J Respir Crit Care Med.2013 Oct 15;188(8):941-7.doi: 10.1164/rccm.201302-0263OC. (1535-4970 (Electronic); 1073-449X (Linking)).

(35) Winham SJ, Biernacka JM. Gene-environment interactions in genome-wide association studies: current approaches and new directions. - J Child Psychol Psychiatry.2013 Oct;54(10):1120-34.doi: 10.1111/jcpp.12114.Epub 2013 Jun 28. (1469-7610 (Electronic); 0021-9630 (Linking)).

(36) MacNee W, Calverley PM. Chronic obstructive pulmonary disease . 7: Management of COPD. - Thorax.2003 Mar;58(3):261-5. (0040-6376 (Print); 0040-6376 (Linking)).

(37) Miller MR, Hankinson J, FAU - Brusasco V, Brusasco V, FAU - Burgos F, Burgos F, et al. Standardisation of spirometry. - Eur Respir J.2005 Aug;26(2):319-38.doi: 10.1183/09031936.05.00034805. (0903-1936 (Print); 0903-1936 (Linking)).

(38) Hankinson JL, Odencrantz JR, FAU - Fedan KB, Fedan KB. Spirometric reference values from a sample of the general U.S. population. - Am J Respir Crit Care Med.1999 Jan;159(1):179-87. (1073-449X (Print); 1073-449X (Linking)).

(39) Lozano R, Naghavi M, FAU - Foreman K, Foreman K, FAU - Lim S, Lim S, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. - Lancet.2012 Dec 15;380(9859):2095-128.doi: 10.1016/S0140-6736(12)61728-0. (1474-547X (Electronic); 0140-6736 (Linking)).

(40) Thun GA, Imboden M, FAU - Ferrarotti I, Ferrarotti I, FAU - Kumar A, Kumar A, et al. Causal and synthetic associations of variants in the SERPINA gene cluster with alpha1-antitrypsin serum levels. - PLoS Genet.2013;9(8):e1003585.doi: 10.1371/journal.pgen.1003585.Epub 2013 Aug 22. (1553-7404 (Electronic); 1553-7390 (Linking)).

(41) Pillai SG, Ge D, FAU - Zhu G, Zhu G, FAU - Kong X, Kong X, et al. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. - PLoS Genet.2009 Mar;5(3):e1000421.doi: 10.1371/journal.pgen.1000421.Epub 2009 Mar 20. (1553-7404 (Electronic); 1553-7390 (Linking)).

(42) From the *Global Strategy for the Diagnosis, Management and Prevention of COPD*, Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2016. Available from: http://goldcopd.org/.

(43) Gamble E, Grootendorst DC, FAU - Hattotuwa K, Hattotuwa K, FAU - O'Shaughnessy T, O'Shaughnessy T, et al. Airway mucosal inflammation in COPD is similar in smokers and ex-smokers: a pooled analysis. - Eur Respir J.2007

Sep;30(3):467-71.doi: 10.1183/09031936.00013006.Epub 2007 May 15. (0903-1936 (Print); 0903-1936 (Linking)).

(44) Turato G, Di Stefano A, FAU - Maestrelli P, Maestrelli P, FAU - Mapp CE, Mapp CE, et al. Effect of smoking cessation on airway inflammation in chronic bronchitis. - Am J Respir Crit Care Med.1995 Oct;152(4 Pt 1):1262-7.doi: 10.1164/ajrccm.152.4.7551380. (1073-449X (Print); 1073-449X (Linking)).

(45) Willemse BW, ten Hacken NH, FAU - Rutgers B, Rutgers B, FAU - Lesman-Leegte IGAT, Lesman-Leegte IG, et al. Effect of 1-year smoking cessation on airway inflammation in COPD and asymptomatic smokers. - Eur Respir J.2005 Nov;26(5):835-45.doi: 10.1183/09031936.05.00108904. (0903-1936 (Print); 0903-1936 (Linking)).

(46) Montuschi P. Pharmacological treatment of chronic obstructive pulmonary disease. - Int J Chron Obstruct Pulmon Dis.2006;1(4):409-23. (1176-9106 (Print); 1176-9106 (Linking)).

(47) National Institute for Health and Care Excellence (NICE). **Chronic obstructive pulmonary disease in over 16s: diagnosis and management**. 2018; Available at: https://www.nice.org.uk/guidance/ng115. Accessed 06/14, 2019.

(48) Rice JB, White AG, Scarpati LM, Wan G, Nelson WW. Long-term Systemic Corticosteroid Exposure: A Systematic Literature Review. - Clin Ther.2017 Nov;39(11):2216-2229.doi: 10.1016/j.clinthera.2017.09.011.Epub 2017 Oct 19. (1879-114X (Electronic); 0149-2918 (Linking)).

(49) Carlucci A, Guerrieri A, FAU - Nava S, Nava S. Palliative care in COPD patients: is it only an end-of-life issue? - Eur Respir Rev.2012 Dec 1;21(126):347-54.doi: 10.1183/09059180.00001512. (1600-0617 (Electronic); 0905-9180 (Linking)).

(50) Seamark DA, Seamark CJ, FAU - Halpin DMG, Halpin DM. Palliative care in chronic obstructive pulmonary disease: a review for clinicians. - J R Soc Med.2007 May;100(5):225-33.doi: 10.1177/014107680710000512. (0141-0768 (Print); 0141-0768 (Linking)).

(51) Wilk JB, Chen TH, FAU - Gottlieb DJ, Gottlieb DJ, FAU - Walter RE, Walter RE, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. - PLoS Genet.2009 Mar;5(3):e1000429.doi: 10.1371/journal.pgen.1000429.Epub 2009 Mar 20. (1553-7404 (Electronic); 1553-7390 (Linking)).

(52) Repapi E, Sayers I, FAU - Wain LV, Wain LV, FAU - Burton PR, Burton PR, et al. Genome-wide association study identifies five loci associated with lung function. - Nat Genet.2010 Jan;42(1):36-44.doi: 10.1038/ng.501.Epub 2009 Dec 13. (1546-1718 (Electronic); 1061-4036 (Linking)).

(53) Hancock DB, Eijgelsheim M, FAU - Wilk JB, Wilk JB, FAU - Gharib SA, Gharib SA, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. - Nat Genet.2010 Jan;42(1):45-52.doi: 10.1038/ng.500.Epub 2009 Dec 13. (1546-1718 (Electronic); 1061-4036 (Linking)).

(54) Soler Artigas M, Loth DW, FAU - Wain LV, Wain LV, FAU - Gharib SA, Gharib SA, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. - Nat Genet.2011 Sep 25;43(11):1082-90.doi: 10.1038/ng.941. (1546-1718 (Electronic); 1061-4036 (Linking)).

(55) 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 Jul;46(7):669-77.doi: 10.1038/ng.3011.Epub 2014 Jun 15. (1546-1718 (Electronic); 1061-4036 (Linking)).

(56) 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.doi: 10.1016/S2213-2600(15)00283-0.Epub 2015 Sep 27. (2213-2619 (Electronic); 2213-2600 (Linking)).

(57) 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 Dec 4;6:8658.doi: 10.1038/ncomms9658. (2041-1723 (Electronic); 2041-1723 (Linking)).

(58) Wain LV, AUID- ORCID: http://orcid.org/0000-0003-4951-1867, Shrine N, Artigas MS, Erzurumluoglu AM, AUID- ORCID: http://orcid.org/0000-0003-1322-8138, et al. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. - Nat Genet.2017 Mar;49(3):416-425.doi: 10.1038/ng.3787.Epub 2017 Feb 6. (1546-1718 (Electronic); 1061-4036 (Linking)).

(59) Jackson V, Latourelle J, Wain L, Smith A, Grove M, Bartz T, et al. Meta-analysis of exome array data identifies six novel genetic loci for lung function [version 2; referees: 2 approved]. Wellcome Open Research 2018;3(4).

(60) Wyss AB, Sofer T, Lee MK, Terzikhan N, Nguyen JN, Lahousse L, et al. Multiethnic Meta-analysis Identifies New Loci for Pulmonary Function. bioRxiv 2017.

(61) Cho MH, McDonald ML, 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. - Lancet Respir Med.2014 Mar;2(3):214-25.doi: 10.1016/S2213-2600(14)70002-5.Epub 2014 Feb 7. (2213-2600 (Print); 2213-2600 (Linking)).

(62) Hobbs BD, AUID- ORCID: http://orcid.org/0000-0001-9564-0745, de Jong K, Lamontagne M, Bosse Y, AUID- ORCID: http://orcid.org/0000-0002-3067-3711, et al. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. - Nat Genet.2017 Mar;49(3):426-432.doi: 10.1038/ng.3752.Epub 2017 Feb 6. (1546-1718 (Electronic); 1061-4036 (Linking)).

(63) Wilk JB, Shrine NR, FAU - Loehr LR, Loehr LR, FAU - Zhao JH, Zhao JH, 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.doi:

10.1164/rccm.201202-0366OC.Epub 2012 Jul 26. (1535-4970 (Electronic); 1073-449X (Linking)).

(64) Cho MH, Boutaoui N, FAU - Klanderman BJ, Klanderman BJ, FAU - Sylvia JS, Sylvia JS, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. - Nat Genet.2010 Mar;42(3):200-2.doi: 10.1038/ng.535.Epub 2010 Feb 21. (1546-1718 (Electronic); 1061-4036 (Linking)).

(65) Soler Artigas M, Wain LV, FAU - Repapi E, Repapi E, FAU - Obeidat M, Obeidat M, 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.doi: 10.1164/rccm.201102-0192OC. (1535-4970 (Electronic); 1073-449X (Linking)).

(66) Castaldi PJ, Cho MH, FAU - Litonjua AA, Litonjua AA, FAU - Bakke P, Bakke P, 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.doi: 10.1165/rcmb.2011-0055OC.Epub 2011 Jun 9. (1535-4989 (Electronic); 1044-1549 (Linking)).

(67) Hobbs BD, Auid- Orcid: 0000-0001-9564-0745, Parker MM, Chen H, Lao T, Hardin M, et al. Exome Array Analysis Identifies a Common Variant in IL27 Associated with Chronic Obstructive Pulmonary Disease. - Am J Respir Crit Care Med.2016 Jul 1;194(1):48-57.doi: 10.1164/rccm.201510-2053OC. (1535-4970 (Electronic); 1073-449X (Linking)).

(68) Cho MH, Castaldi PJ, FAU - Wan ES, Wan ES, FAU - Siedlinski M, Siedlinski M, et al. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. - Hum Mol Genet.2012 Feb 15;21(4):947-57.doi: 10.1093/hmg/ddr524.Epub 2011 Nov 11. (1460-2083 (Electronic); 0964-6906 (Linking)).

(69) Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, McDonald ML, et al. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. - BMC Genet.2015 Dec 3;16:138.doi: 10.1186/s12863-015-0299-4. (1471-2156 (Electronic); 1471-2156 (Linking)).

(70) Thorgeirsson TE, Gudbjartsson DF, FAU - Surakka I, Surakka I, FAU - Vink JM, Vink JM, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. - Nat Genet.2010 May;42(5):448-53.doi: 10.1038/ng.573.Epub 2010 Apr 25. (1546-1718 (Electronic); 1061-4036 (Linking)).

(71) Liu JZ, Tozzi F, FAU - Waterworth DM, Waterworth DM, FAU - Pillai SG, Pillai SG, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. - Nat Genet.2010 May;42(5):436-40.doi: 10.1038/ng.572.Epub 2010 Apr 25. (1546-1718 (Electronic); 1061-4036 (Linking)).

(72) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. - Nat Genet.2010 May;42(5):441-7.doi: 10.1038/ng.571.Epub 2010 Apr 25. (1546-1718 (Electronic); 1061-4036 (Linking)).

(73) Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. - PLoS Med.2015 Mar 31;12(3):e1001779.doi: 10.1371/journal.pmed.1001779.eCollection 2015 Mar. (1549-1676 (Electronic); 1549-1277 (Linking)).

(74) Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv 2017.

(75) Munafo MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. - Int J Epidemiol.2018 Feb 1;47(1):226-235.doi: 10.1093/ije/dyx206. (1464-3685 (Electronic); 0300-5771 (Linking)).

(76) McCarthy S, AUID- ORCID: http://orcid.org/0000-0002-2715-4187, Das S, AUID- ORCID: http://orcid.org/0000-0001-6346-1590, Kretzschmar W, AUID- ORCID: http://orcid.org/0000-0002-2575-0807, et al. A reference panel of 64,976 haplotypes for genotype imputation. - Nat Genet.2016 Oct;48(10):1279-83.doi: 10.1038/ng.3643.Epub 2016 Aug 22. (1546-1718 (Electronic); 1061-4036 (Linking)).

(77) Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. - Nat Commun.2015 Sep 14;6:8111.doi: 10.1038/ncomms9111. (2041-1723 (Electronic); 2041-1723 (Linking)).

(78) Fry A, FAU - Littlejohns TJ, Littlejohns TJ, FAU - Sudlow C, Sudlow C, FAU - Doherty N, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. - Am J Epidemiol.2017 Nov 1;186(9):1026-1034.doi: 10.1093/aje/kwx246. (1476-6256 (Electronic); 0002-9262 (Linking)).

(79) Wilk JB, Shrine NR, FAU - Loehr LR, Loehr LR, FAU - Zhao JH, Zhao JH, 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.doi: 10.1164/rccm.201202-0366OC.Epub 2012 Jul 26. (1535-4970 (Electronic); 1073-449X (Linking)).

(80) Tzeng JY, Zhang D, FAU - Pongpanich M, Pongpanich M, FAU - Smith C, Smith C, et al. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. - Am J Hum Genet.2011 Aug 12;89(2):277-88.doi: 10.1016/j.ajhg.2011.07.007. (1537-6605 (Electronic); 0002-9297 (Linking)).

(81) Thomas D. Gene--environment-wide association studies: emerging approaches. - Nat Rev Genet.2010 Apr;11(4):259-72.doi: 10.1038/nrg2764. (1471-0064 (Electronic); 1471-0056 (Linking)).

(82) He JQ, Connett JE, FAU - Anthonisen NR, Anthonisen NR, FAU - Pare PD, Pare PD, et al. Glutathione S-transferase variants and their interaction with smoking on lung

function. - Am J Respir Crit Care Med.2004 Aug 15;170(4):388-94.Epub 2004 Jun 7. (1073-449X (Print); 1073-449X (Linking)).

(83) Sadeghnejad A, Meyers DA, FAU - Bottai M, Bottai M, FAU - Sterling DA, Sterling DA, et al. IL13 promoter polymorphism 1112C/T modulates the adverse effect of tobacco smoking on lung function. - Am J Respir Crit Care Med.2007 Oct 15;176(8):748-52.Epub 2007 Jul 5. (1073-449X (Print); 1073-449X (Linking)).

(84) Hunninghake GM, Cho MH, FAU - Tesfaigzi Y, Tesfaigzi Y, FAU - Soto-Quiros ME, Soto-Quiros ME, et al. MMP12, lung function, and COPD in high-risk populations. - N Engl J Med.2009 Dec 31;361(27):2599-608.doi: 10.1056/NEJMoa0904006.Epub 2009 Dec 16. (1533-4406 (Electronic); 0028-4793 (Linking)).

(85) Hancock DB, Soler Artigas M, FAU - Gharib SA, Gharib SA, FAU - Henry A, Henry 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.doi: 10.1371/journal.pgen.1003098.Epub 2012 Dec 20. (1553-7404 (Electronic); 1553-7390 (Linking)).

(86) Liao SY, FAU - Lin X, Lin X, FAU - Christiani DC, Christiani DC. Gene-environment interaction effects on lung function- a genome-wide association study within the Framingham heart study. - Environ Health.2013 Dec 1;12:101.doi: 10.1186/1476-069X-12-101. (1476-069X (Electronic); 1476-069X (Linking)).

(87) Park B, Koo SM, AUID- ORCID: http://orcid.org/0000-0003-1504-9387, An J, Lee M, Kang HY, et al. Genome-wide assessment of gene-by-smoking interactions in COPD. - Sci Rep.2018 Jun 18;8(1):9319.doi: 10.1038/s41598-018-27463-5. (2045-2322 (Electronic); 2045-2322 (Linking)).

(88) Kilpelainen TO, Qi L, FAU - Brage S, Brage S, FAU - Sharp SJ, Sharp SJ, et al. Physical activity attenuates the influence of FTO variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. - PLoS Med.2011 Nov;8(11):e1001116.doi: 10.1371/journal.pmed.1001116.Epub 2011 Nov 1. (1549-1676 (Electronic); 1549-1277 (Linking)).

(89) Hutter CM, Mechanic LE, FAU - Chatterjee N, Chatterjee N, FAU - Kraft P, Kraft P, et al. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. - Genet Epidemiol.2013 Nov;37(7):643-57.doi: 10.1002/gepi.21756.Epub 2013 Oct 5. (1098-2272 (Electronic); 0741-0395 (Linking)).

(90) Brooks PJ, Enoch MA, FAU - Goldman D, Goldman D, FAU - Li T, Li TK, et al. The alcohol flushing response: an unrecognized risk factor for esophageal cancer from alcohol consumption. - PLoS Med.2009 Mar 24;6(3):e50.doi: 10.1371/journal.pmed.1000050. (1549-1676 (Electronic); 1549-1277 (Linking)).

(91) Cui R, Kamatani Y, FAU - Takahashi A, Takahashi A, FAU - Usami M, Usami M, et al. Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. - Gastroenterology.2009

Nov;137(5):1768-75.doi: 10.1053/j.gastro.2009.07.070.Epub 2009 Aug 19. (1528-0012 (Electronic); 0016-5085 (Linking)).

(92) Hiyama T, Yoshihara M, FAU - Tanaka S, Tanaka S, FAU - Chayama K, Chayama K. Genetic polymorphisms and esophageal cancer risk. - Int J Cancer.2007 Oct 15;121(8):1643-58.doi: 10.1002/ijc.23044. (0020-7136 (Print); 0020-7136 (Linking)).

(93) Wu C, Kraft P, FAU - Zhai K, Zhai K, FAU - Chang J, Chang J, et al. Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. - Nat Genet.2012 Oct;44(10):1090-7.doi: 10.1038/ng.2411.Epub 2012 Sep 9. (1546-1718 (Electronic); 1061-4036 (Linking)).

(94) Wu C, Hu Z, FAU - He Z, He Z, FAU - Jia W, Jia W, et al. Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. - Nat Genet.2011 Jun 5;43(7):679-84.doi: 10.1038/ng.849. (1546-1718 (Electronic); 1061-4036 (Linking)).

(95) Ritz BR, FAU - Chatterjee N, Chatterjee N, FAU - Garcia-Closas M, Garcia-Closas M, FAU - Gauderman WJ, et al. Lessons Learned From Past Gene-Environment Interaction Successes. - Am J Epidemiol.2017 Oct 1;186(7):778-786.doi: 10.1093/aje/kwx230. (1476-6256 (Electronic); 0002-9262 (Linking)).

(96) Hancock DB, Reginsson GW, Gaddis NC, Chen X, Saccone NL, Lutz SM, et al. Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence. - Transl Psychiatry.2015 Oct 6;5:e651.doi: 10.1038/tp.2015.149. (2158-3188 (Electronic); 2158-3188 (Linking)).

(97) Heatherton TF, Kozlowski LT, FAU - Frecker RC, Frecker RC, FAU - Fagerstrom KO, Fagerstrom KO. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. - Br J Addict.1991 Sep;86(9):1119-27. (0952-0481 (Print); 0952-0481 (Linking)).

(98) Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. - Nat Rev Genet.2018 Aug;19(8):491-504.doi: 10.1038/s41576-018-0016-z. (1471-0064 (Electronic); 1471-0056 (Linking)).

(99) 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.doi: 10.1093/bioinformatics/btv240. (1367-4811 (Electronic); 1367-4803 (Linking)).

(100) Rutter M. Biological implications of gene-environment interaction. - J Abnorm Child Psychol.2008 Oct;36(7):969-75.doi: 10.1007/s10802-008-9256-2. (0091-0627 (Print); 0091-0627 (Linking)).

(101) Kraft P, Yen YC, FAU - Stram DO, Stram DO, FAU - Morrison J, Morrison J, et al. Exploiting gene-environment interaction to detect genetic associations. - Hum Hered.2007;63(2):111-9.Epub 2007 Feb 2. (0001-5652 (Print); 0001-5652 (Linking)).

(102) Dai JY, Logsdon BA, FAU - Huang Y, Huang Y, FAU - Hsu L, Hsu L, et al. Simultaneously testing for marginal genetic association and gene-environment interaction. - Am J Epidemiol.2012 Jul 15;176(2):164-73.doi: 10.1093/aje/kwr521.Epub 2012 Jul 6. (1476-6256 (Electronic); 0002-9262 (Linking)).

(103) Piegorsch WW, Weinberg CR, FAU - Taylor JA, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. - Stat Med.1994 Jan 30;13(2):153-62. (0277-6715 (Print); 0277-6715 (Linking)).

(104) Albert PS, Ratnasinghe D, FAU - Tangrea J, Tangrea J, FAU - Wacholder S, Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. - Am J Epidemiol.2001 Oct 15;154(8):687-93. (0002-9262 (Print); 0002-9262 (Linking)).

(105) Yang Q, Khoury MJ. Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. - Epidemiol Rev.1997;19(1):33-43. (0193-936X (Print); 0193-936X (Linking)).

(106) Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. - Biometrics.2008 Sep;64(3):685-94.Epub 2007 Dec 20. (1541-0420 (Electronic); 0006-341X (Linking)).

(107) Li D, Conti DV. Detecting gene-environment interactions using a combined case-only and case-control approach. - Am J Epidemiol.2009 Feb 15;169(4):497-504.doi: 10.1093/aje/kwn339.Epub 2008 Dec 13. (1476-6256 (Electronic); 0002-9262 (Linking)).

(108) Dai JY, Kooperberg C, Leblanc M, Prentice RL. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. Biometrika 2012 revised;99(4):929-944.

(109) Kooperberg C, Leblanc M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. - Genet Epidemiol.2008 Apr;32(3):255-63.doi: 10.1002/gepi.20300. (0741-0395 (Print); 0741-0395 (Linking)).

(110) Murcray CE, Lewinger JP, FAU - Gauderman WJ, Gauderman WJ. Gene-environment interaction in genome-wide association studies. - Am J Epidemiol.2009 Jan 15;169(2):219-26.doi: 10.1093/aje/kwn353.Epub 2008 Nov 20. (1476-6256 (Electronic); 0002-9262 (Linking)).

(111) Murcray CE, Lewinger JP, FAU - Conti DV, Conti DV, FAU - Thomas DC, Thomas DC, et al. Sample size requirements to detect gene-environment interactions in genome-wide association studies. - Genet Epidemiol.2011 Apr;35(3):201-10.doi: 10.1002/gepi.20569.Epub 2011 Feb 9. (1098-2272 (Electronic); 0741-0395 (Linking)).

(112) Hsu L, Jiao S, FAU - Dai JY, Dai JY, FAU - Hutter C, Hutter C, et al. Powerful cocktail methods for detecting genome-wide gene-environment interaction. - Genet

Epidemiol.2012 Apr;36(3):183-94.doi: 10.1002/gepi.21610. (1098-2272 (Electronic); 0741-0395 (Linking)).

(113) Gauderman WJ, Zhang P, FAU - Morrison JL, Morrison JL, FAU - Lewinger JP, Lewinger JP. Finding novel genes by testing G x E interactions in a genome-wide association study. - Genet Epidemiol.2013 Sep;37(6):603-13.doi: 10.1002/gepi.21748.Epub 2013 Jul 19. (1098-2272 (Electronic); 0741-0395 (Linking)).

(114) Pare G, Cook NR, FAU - Ridker PM, Ridker PM, FAU - Chasman DI, Chasman DI. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. - PLoS Genet.2010 Jun 17;6(6):e1000981.doi: 10.1371/journal.pgen.1000981. (1553-7404 (Electronic); 1553-7390 (Linking)).

(115) Patel CJ. Analytical Complexity in Detection of Gene Variant-by-Environment Exposure Interactions in High-Throughput Genomic and Exposomic Research. - Curr Environ Health Rep.2016 Mar;3(1):64-72.doi: 10.1007/s40572-016-0080-5. (2196-5412 (Electronic); 2196-5412 (Linking)).

(116) Allen RJ, Porte J, Braybrooke R, Flores C, Fingerlin TE, Oldham JM, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. - Lancet Respir Med.2017 Nov;5(11):869-880.doi: 10.1016/S2213-2600(17)30387-9.Epub 2017 Oct 20. (2213-2619 (Electronic); 2213-2600 (Linking)).

(117) Noth I, Zhang Y, Ma SF, Flores C, Barber M, Huang Y, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. - Lancet Respir Med.2013 Jun;1(4):309-317.doi: 10.1016/S2213-2600(13)70045-6.Epub 2013 Apr 17. (2213-2600 (Print); 2213-2600 (Linking)).

(118) Fingerlin TE, Murphy E, FAU - Zhang W, Zhang W, FAU - Peljto AL, Peljto AL, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. - Nat Genet.2013 Jun;45(6):613-20.doi: 10.1038/ng.2609.Epub 2013 Apr 14. (1546-1718 (Electronic); 1061-4036 (Linking)).

(119) Adzhubei I, Jordan DM, FAU - Sunyaev SR, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. - Curr Protoc Hum Genet.2013 Jan;Chapter 7:Unit7.20.doi: 10.1002/0471142905.hg0720s76. (1934-8258 (Electronic); 1934-8258 (Linking)).

(120) Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. - Nucleic Acids Res.2003 Jul 1;31(13):3812-4. (1362-4962 (Electronic); 0305-1048 (Linking)).

(121) Shihab HA, Gough J, FAU - Cooper DN, Cooper DN, FAU - Stenson PD, Stenson PD, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. - Hum Mutat.2013 Jan;34(1):57-65.doi: 10.1002/humu.22225.Epub 2012 Nov 2. (1098-1004 (Electronic); 1059-7794 (Linking)).

(122) Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. - Nat Methods.2015 Oct;12(10):931-4.doi: 10.1038/nmeth.3547.Epub 2015 Aug 24. (1548-7105 (Electronic); 1548-7091 (Linking)).

(123) Breiman L. Random forests. Machine Learning, 45(1), 5–32. 2001.

(124) Ritchie MD, Hahn LW, FAU - Roodi N, Roodi N, FAU - Bailey LR, Bailey LR, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. - Am J Hum Genet.2001 Jul;69(1):138-47.Epub 2001 Jun 11. (0002-9297 (Print); 0002-9297 (Linking)).

(125) Strobl C, Boulesteix AL, FAU - Zeileis A, Zeileis A, FAU - Hothorn T, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. - BMC Bioinformatics.2007 Jan 25;8:25. (1471-2105 (Electronic); 1471-2105 (Linking)).

(126) Goldstein BA, Polley EC, FAU - Briggs FBS, Briggs FB. Random forests for genetic association studies. - Stat Appl Genet Mol Biol.2011;10(1):32.doi: 10.2202/1544-6115.1691.Epub 2011 Jul 12. (1544-6115 (Electronic); 1544-6115 (Linking)).

(127) Maenner MJ, Denlinger LC, FAU - Langton A, Langton A, FAU - Meyers KJ, Meyers KJ, et al. Detecting gene-by-smoking interactions in a genome-wide association study of early-onset coronary heart disease using random forests. - BMC Proc.2009 Dec 15;3 Suppl 7:S88. (1753-6561 (Electronic); 1753-6561 (Linking)).

(128) De Lobel L, Geurts P, FAU - Baele G, Baele G, FAU - Castro-Giner F, Castro-Giner F, et al. A screening methodology based on Random Forests to improve the detection of gene-gene interactions. - Eur J Hum Genet.2010 Oct;18(10):1127-32.doi: 10.1038/ejhg.2010.48.Epub 2010 May 12. (1476-5438 (Electronic); 1018-4813 (Linking)).

(129) Calle ML, Urrea V, FAU - Vellalta G, Vellalta G, FAU - Malats N, Malats N, et al. Improving strategies for detecting genetic patterns of disease susceptibility in association studies. - Stat Med.2008 Dec 30;27(30):6532-46.doi: 10.1002/sim.3431. (0277-6715 (Print); 0277-6715 (Linking)).

(130) 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.doi: 10.1016/j.ajhg.2014.06.009. (1537-6605 (Electronic); 0002-9297 (Linking)).

(131) Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. - Genet Epidemiol.2009 Sep;33(6):497-507.doi: 10.1002/gepi.20402. (1098-2272 (Electronic); 0741-0395 (Linking)).

(132) Chapman J, Whittaker J. Analysis of multiple SNPs in a candidate gene or region. - Genet Epidemiol.2008 Sep;32(6):560-6.doi: 10.1002/gepi.20330. (1098-2272 (Electronic); 0741-0395 (Linking)).

(133) 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.doi: 10.1016/j.mrfmmm.2006.09.003.Epub 2006 Nov 13. (0027-5107 (Print); 0027-5107 (Linking)).

(134) 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.doi: 10.1016/j.ajhg.2008.06.024.Epub 2008 Aug 7. (1537-6605 (Electronic); 0002-9297 (Linking)).

(135) Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. - PLoS Genet.2009 Feb;5(2):e1000384.doi: 10.1371/journal.pgen.1000384.Epub 2009 Feb 13. (1553-7404 (Electronic); 1553-7390 (Linking)).

(136) Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. - Genet Epidemiol.2011 Nov;35(7):606-19.doi: 10.1002/gepi.20609.Epub 2011 Jul 18. (1098-2272 (Electronic); 0741-0395 (Linking)).

(137) Ionita-Laza I, Buxbaum JD, FAU - Laird NM, Laird NM, FAU - Lange C, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. - PLoS Genet.2011 Feb 3;7(2):e1001289.doi: 10.1371/journal.pgen.1001289. (1553-7404 (Electronic); 1553-7390 (Linking)).

(138) Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. - Hum Hered.2010;70(1):42-54.doi: 10.1159/000288704.Epub 2010 Apr 23. (1423-0062 (Electronic); 0001-5652 (Linking)).

(139) Derkach A, Lawless JF, FAU - Sun L, Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. - Genet Epidemiol.2013 Jan;37(1):110-21.doi: 10.1002/gepi.21689.Epub 2012 Oct 2. (1098-2272 (Electronic); 0741-0395 (Linking)).

(140) Hoffmann TJ, Marini NJ, FAU - Witte JS, Witte JS. Comprehensive approach to analyzing rare genetic variants. - PLoS One.2010 Nov 3;5(11):e13584.doi: 10.1371/journal.pone.0013584. (1932-6203 (Electronic); 1932-6203 (Linking)).

(141) Bhatia G, Bansal V, FAU - Harismendy O, Harismendy O, FAU - Schork NJ, Schork NJ, et al. A covering method for detecting genetic associations between rare variants and common phenotypes. - PLoS Comput Biol.2010 Oct 14;6(10):e1000954.doi: 10.1371/journal.pcbi.1000954. (1553-7358 (Electronic); 1553-734X (Linking)).

(142) Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. - PLoS Genet.2010 Oct 14;6(10):e1001156.doi: 10.1371/journal.pgen.1001156. (1553-7404 (Electronic); 1553-7390 (Linking)).

(143) Price AL, Kryukov GV, FAU - de Bakker,Paul I.W., de Bakker PI, FAU - Purcell SM, Purcell SM, et al. Pooled association tests for rare variants in exon-resequencing

studies. - Am J Hum Genet.2010 Jun 11;86(6):832-8.doi: 10.1016/j.ajhg.2010.04.005.Epub 2010 May 13. (1537-6605 (Electronic); 0002-9297 (Linking)).

(144) Neale BM, Rivas MA, FAU - Voight BF, Voight BF, FAU - Altshuler D, Altshuler D, et al. Testing for an unusual distribution of rare variants. - PLoS Genet.2011 Mar;7(3):e1001322.doi: 10.1371/journal.pgen.1001322.Epub 2011 Mar 3. (1553-7404 (Electronic); 1553-7390 (Linking)).

(145) Wu MC, Lee S, FAU - Cai T, Cai T, FAU - Li Y, Li Y, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. - Am J Hum Genet.2011 Jul 15;89(1):82-93.doi: 10.1016/j.ajhg.2011.05.029.Epub 2011 Jul 7. (1537-6605 (Electronic); 0002-9297 (Linking)).

(146) Jiao S, Hsu L, FAU - Bezieau S, Bezieau S, FAU - Brenner H, Brenner H, et al. SBERIA: set-based gene-environment interaction test for rare and common variants in complex diseases. - Genet Epidemiol.2013 Jul;37(5):452-64.doi: 10.1002/gepi.21735.Epub 2013 May 29. (1098-2272 (Electronic); 0741-0395 (Linking)).

(147) Chatterjee N, Kalaylioglu Z, FAU - Moslehi R, Moslehi R, FAU - Peters U, Peters U, et al. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. - Am J Hum Genet.2006 Dec;79(6):1002-16.Epub 2006 Oct 20. (0002-9297 (Print); 0002-9297 (Linking)).

(148) Tukey JW. One Degree of Freedom for Non-Additivity. Biometrics 1949;5(3):232-242.

(149) Wang X, Qin H, FAU - Morris NJ, Morris NJ, FAU - Zhu X, Zhu X, et al. Testing gene-environment interactions in gene-based association studies. - BMC Proc.2011 Nov 29;5 Suppl 9:S26.doi: 10.1186/1753-6561-5-S9-S26. (1753-6561 (Electronic); 1753-6561 (Linking)).

(150) Chen H, Meigs JB, FAU - Dupuis J, Dupuis J. Incorporating gene-environment interaction in testing for association with rare genetic variants. - Hum Hered.2014;78(2):81-90.doi: 10.1159/000363347.Epub 2014 Jul 18. (1423-0062 (Electronic); 0001-5652 (Linking)).

(151) Kazma R, Cardin NJ, FAU - Witte JS, Witte JS. Does accounting for gene-environment interactions help uncover association between rare variants and complex diseases? - Hum Hered.2012;74(3-4):205-14.doi: 10.1159/000346825.Epub 2013 Apr 11. (1423-0062 (Electronic); 0001-5652 (Linking)).

(152) Lin X, Lee S, FAU - Christiani DC, Christiani DC, FAU - Lin X, Lin X. Test for interactions between a genetic marker set and environment in generalized linear models. - Biostatistics.2013 Sep;14(4):667-81.doi: 10.1093/biostatistics/kxt006.Epub 2013 Mar 5. (1468-4357 (Electronic); 1465-4644 (Linking)).

(153) Oh S, Lee J, FAU - Kwon M, Kwon MS, FAU - Weir B, Weir B, et al. A novel method to identify high order gene-gene interactions in genome-wide association

studies: gene-based MDR. - BMC Bioinformatics.2012 Jun 11;13 Suppl 9:S5.doi: 10.1186/1471-2105-13-S9-S5. (1471-2105 (Electronic); 1471-2105 (Linking)).

(154) Gauderman WJ, Murcray C, FAU - Gilliland F, Gilliland F, FAU - Conti DV, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. - Genet Epidemiol.2007 Jul;31(5):383-95. (0741-0395 (Print); 0741-0395 (Linking)).

(155) Peng Q, Zhao J, FAU - Xue F, Xue F. A gene-based method for detecting gene-gene co-association in a case-control association study. - Eur J Hum Genet.2010 May;18(5):582-7.doi: 10.1038/ejhg.2009.223.Epub 2009 Dec 23. (1476-5438 (Electronic); 1018-4813 (Linking)).

(156) Zhang K, Cui S, FAU - Chang S, Chang S, FAU - Zhang L, Zhang L, et al. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. - Nucleic Acids Res.2010 Jul;38(Web Server issue):W90-5.doi: 10.1093/nar/gkq324.Epub 2010 Apr 30. (1362-4962 (Electronic); 0305-1048 (Linking)).

(157) Subramanian A, Tamayo P, FAU - Mootha VK, Mootha VK, FAU - Mukherjee S, Mukherjee S, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. - Proc Natl Acad Sci U S A.2005 Oct 25;102(43):15545-50.Epub 2005 Sep 30. (0027-8424 (Print); 0027-8424 (Linking)).

(158) Wang K, Li M, FAU - Bucan M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. - Am J Hum Genet.2007 Dec;81(6):1278-83. (1537-6605 (Electronic); 0002-9297 (Linking)).

(159) Hong MG, Pawitan Y, FAU - Magnusson PKE, Magnusson PK, FAU - Prince JA, Prince JA. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. - Hum Genet.2009 Aug;126(2):289-301.doi: 10.1007/s00439-009-0676-z.Epub 2009 May 1. (1432-1203 (Electronic); 0340-6717 (Linking)).

(160) Chasman DI. On the utility of gene set methods in genomewide association studies of quantitative traits. - Genet Epidemiol.2008 Nov;32(7):658-68.doi: 10.1002/gepi.20334. (1098-2272 (Electronic); 0741-0395 (Linking)).

(161) Wei S, Wang LE, FAU - McHugh MK, McHugh MK, FAU - Han Y, Han Y, et al. Genome-wide gene-environment interaction analysis for asbestos exposure in lung cancer susceptibility. - Carcinogenesis.2012 Aug;33(8):1531-7.doi: 10.1093/carcin/bgs188.Epub 2012 May 27. (1460-2180 (Electronic); 0143-3334 (Linking)).

(162) Wu MC, Lee S, FAU - Cai T, Cai T, FAU - Li Y, Li Y, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. - Am J Hum Genet.2011 Jul 15;89(1):82-93.doi: 10.1016/j.ajhg.2011.05.029.Epub 2011 Jul 7. (1537-6605 (Electronic); 0002-9297 (Linking)).

(163) Mukherjee B, Ahn J, FAU - Gruber SB, Gruber SB, FAU - Chatterjee N, Chatterjee N. Testing gene-environment interaction in large-scale case-control

association studies: possible choices and comparisons. - Am J Epidemiol.2012 Feb 1;175(3):177-90.doi: 10.1093/aje/kwr367.Epub 2011 Dec 22. (1476-6256 (Electronic); 0002-9262 (Linking)).

(164) Wason JM, Dudbridge F. A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. - Am J Hum Genet.2012 May 4;90(5):760-73.doi: 10.1016/j.ajhg.2012.03.007. (1537-6605 (Electronic); 0002-9297 (Linking)).

(165) Ege MJ, Strachan DP. Comparisons of power of statistical methods for gene-environment interaction analyses. - Eur J Epidemiol.2013 Oct;28(10):785-97.doi: 10.1007/s10654-013-9837-4.Epub 2013 Sep 5. (1573-7284 (Electronic); 0393-2990 (Linking)).

(166) Aschard H, Zaitlen N, FAU - Tamimi RM, Tamimi RM, FAU - Lindstrom S, Lindstrom S, et al. A nonparametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes. - Genet Epidemiol.2013 May;37(4):323-33.doi: 10.1002/gepi.21716.Epub 2013 Mar 19. (1098-2272 (Electronic); 0741-0395 (Linking)).

(167) Marigorta UM, Gibson G. A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. - Front Genet.2014 Jul 21;5:225.doi: 10.3389/fgene.2014.00225.eCollection 2014. (1664-8021 (Print); 1664-8021 (Linking)).

(168) Boonstra PS, FAU - Mukherjee B, Mukherjee B, FAU - Gruber SB, Gruber SB, FAU - Ahn J, et al. Tests for Gene-Environment Interactions and Joint Effects With Exposure Misclassification. - Am J Epidemiol.2016 Feb 1;183(3):237-47.doi: 10.1093/aje/kwv198.Epub 2016 Jan 10. (1476-6256 (Electronic); 0002-9262 (Linking)).

(169) Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. - Genome Biol.2017 Apr 27;18(1):77.doi: 10.1186/s13059-017-1212-4. (1474-760X (Electronic); 1474-7596 (Linking)).

(170) Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. - Nat Rev Genet.2010 Jun;11(6):415-25.doi: 10.1038/nrg2779. (1471-0064 (Electronic); 1471-0056 (Linking)).

(171) Gibson G. Rare and common variants: twenty arguments. - Nat Rev Genet.2012 Jan 18;13(2):135-45.doi: 10.1038/nrg3118. (1471-0064 (Electronic); 1471-0056 (Linking)).

(172) Park JH, Gail MH, FAU - Weinberg CR, Weinberg CR, FAU - Carroll RJ, Carroll RJ, et al. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. - Proc Natl Acad Sci U S A.2011 Nov 1;108(44):18026-31.doi: 10.1073/pnas.1114759108.Epub 2011 Oct 14. (1091-6490 (Electronic); 0027-8424 (Linking)).

(173) Palmer LJ, Knuiman MW, FAU - Divitini ML, Divitini ML, FAU - Burton PR, Burton PR, et al. Familial aggregation and heritability of adult lung function: results

from the Busselton Health Study. - Eur Respir J.2001 Apr;17(4):696-702. (0903-1936 (Print); 0903-1936 (Linking)).

(174) Wilk JB, Djousse L, FAU - Arnett DK, Arnett DK, FAU - Rich SS, Rich SS, et al. Evidence for major genes influencing pulmonary function in the NHLBI family heart study. - Genet Epidemiol.2000 Jul;19(1):81-94. (0741-0395 (Print); 0741-0395 (Linking)).

(175) Poole W, Gibbs DL, Shmulevich I, Bernard B, Knijnenburg TA. Combining dependent P-values with an empirical adaptation of Brown's method. - Bioinformatics.2016 Sep 1;32(17):i430-i436.doi: 10.1093/bioinformatics/btw438. (1367-4811 (Electronic); 1367-4803 (Linking)).

(176) Wilson DJ. The harmonic mean p-value for combining dependent tests. bioRxiv 2018 01/01.

(177) Li Q, Hu J, FAU - Ding J, Ding J, FAU - Zheng G, Zheng G. Fisher's method of combining dependent statistics using generalizations of the gamma distribution with applications to genetic pleiotropic associations. - Biostatistics.2014 Apr;15(2):284-95.doi: 10.1093/biostatistics/kxt045.Epub 2013 Oct 29. (1468-4357 (Electronic); 1465-4644 (Linking)).

(178) Thomas DR, Hughes E, Zumbo BD. On Variable Importance in Linear Regression. Soc Indicators Res 1998 11/01;45(1):253-275.

(179) Aschard H. A perspective on interaction effects in genetic association studies. - Genet Epidemiol.2016 Dec;40(8):678-688.doi: 10.1002/gepi.21989.Epub 2016 Jul 7. (1098-2272 (Electronic); 0741-0395 (Linking)).

(180) Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. - Nature.2018 Oct;562(7726):203-209.doi: 10.1038/s41586-018-0579-z.Epub 2018 Oct 10. (1476-4687 (Electronic); 0028-0836 (Linking)).

(181) Anderson CA, Pettersson FH, FAU - Clarke GM, Clarke GM, FAU - Cardon LR, Cardon LR, et al. Data quality control in genetic case-control association studies. - Nat Protoc.2010 Sep;5(9):1564-73.doi: 10.1038/nprot.2010.116.Epub 2010 Aug 26. (1750-2799 (Electronic); 1750-2799 (Linking)).

(182) Gross A, AUID- ORCID: http://orcid.org/0000-0002-5961-5507, Tonjes A, Scholz M. On the impact of relatedness on SNP association analysis. - BMC Genet.2017 Dec 6;18(1):104.doi: 10.1186/s12863-017-0571-x. (1471-2156 (Electronic); 1471-2156 (Linking)).

(183) Manichaikul A, Mychaleckyj JC, FAU - Rich SS, Rich SS, FAU - Daly K, Daly K, et al. Robust relationship inference in genome-wide association studies. - Bioinformatics.2010 Nov 15;26(22):2867-73.doi: 10.1093/bioinformatics/btq559.Epub 2010 Oct 5. (1367-4811 (Electronic); 1367-4803 (Linking)).

(184) Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. - Nat Genet.2015 Mar;47(3):284-90.doi: 10.1038/ng.3190.Epub 2015 Feb 2. (1546-1718 (Electronic); 1061-4036 (Linking)).

(185) Howey R, Cordell H. CASSI: Genome-Wide Interaction Analysis Software. Available at: https://www.staff.ncl.ac.uk/richard.howey/cassi/index.html. Accessed 12/13, 2018.

(186) Bhattacharjee S, Chatterjee N, Han S, Wheeler W. *An R package for analysis of case-control studies in genetic epidemiology*. Available at: https://bioconductor.org/packages/release/bioc/html/CGEN.html. Accessed 12/13, 2018.

(187) Clayton D. *snpStats: SnpMatrix and XSnpMatrix classes and methods*. R package version 1.32.0. Available at: https://bioconductor.org/packages/release/bioc/html/snpStats.html. Accessed 12/13, 2018.

(188) Aulchenko YS, Struchalin MV, FAU - van Duijn,Cornelia M., van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. - BMC Bioinformatics.2010 Mar 16;11:134.doi: 10.1186/1471-2105-11-134. (1471-2105 (Electronic); 1471-2105 (Linking)).

(189) Kutalik Z, Johnson T, FAU - Bochud M, Bochud M, FAU - Mooser V, Mooser V, et al. Methods for testing association between uncertain genotypes and quantitative traits. - Biostatistics.2011 Jan;12(1):1-17.doi: 10.1093/biostatistics/kxq039.Epub 2010 Jun 11. (1468-4357 (Electronic); 1465-4644 (Linking)).

(190) Ehret GB, Lamparter D, FAU - Hoggart CJ, Hoggart CJ, Whittaker JC, FAU - Beckmann JS, et al. A multi-SNP locus-association method reveals a substantial fraction of the missing heritability. - Am J Hum Genet.2012 Nov 2;91(5):863-71.doi: 10.1016/j.ajhg.2012.09.013. (1537-6605 (Electronic); 0002-9297 (Linking)).

(191) Pain O, Dudbridge F, Ronald A. Are your covariates under control? How normalization can re-introduce covariate effects. - Eur J Hum Genet.2018 Aug;26(8):1194-1201.doi: 10.1038/s41431-018-0159-6.Epub 2018 Apr 30. (1476-5438 (Electronic); 1018-4813 (Linking)).

(192) Miceli F, FAU - Soldovieri MV, Soldovieri MV, FAU - Joshi N, Joshi N, FAU - Weckhuysen S, et al. KCNQ2-Related Disorders BTI - GeneReviews((R)).

(193) Poss ZC, Ebmeier CC, FAU - Taatjes DJ, Taatjes DJ. The Mediator complex and transcription regulation. - Crit Rev Biochem Mol Biol.2013 Nov-Dec;48(6):575-608.doi: 10.3109/10409238.2013.840259.Epub 2013 Oct 3. (1549-7798 (Electronic); 1040-9238 (Linking)).

(194) Chen YZ, Matsushita M, FAU - Girirajan S, Girirajan S, FAU - Lisowski M, Lisowski M, et al. Evidence for involvement of GNB1L in autism. - Am J Med Genet B Neuropsychiatr Genet.2012 Jan;159B(1):61-71.doi: 10.1002/ajmg.b.32002.Epub 2011 Nov 16. (1552-485X (Electronic); 1552-4841 (Linking)).

(195) Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. - Heredity (Edinb).2005 Sep;95(3):221-7.doi: 10.1038/sj.hdy.6800717. (0018-067X (Print); 0018-067X (Linking)).

(196) Aschard H, Tobin MD, Hancock DB, Skurnik D, Sood A, James A, et al. Evidence for large-scale gene-by-smoking interaction effects on pulmonary function. - Int J Epidemiol.2017 Jun 1;46(3):894-904.doi: 10.1093/ije/dyw318. (1464-3685 (Electronic); 0300-5771 (Linking)).

(197) Huls A, AUID- ORCID: http://orcid.org/0000-0002-6005-417X, Kramer U, Carlsten C, Schikowski T, Ickstadt K, et al. Comparison of weighting approaches for genetic risk scores in gene-environment interaction studies. - BMC Genet.2017 Dec 16;18(1):115.doi: 10.1186/s12863-017-0586-3. (1471-2156 (Electronic); 1471-2156 (Linking)).

(198) Cho MH, McDonald ML, 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. - Lancet Respir Med.2014 Mar;2(3):214-25.doi: 10.1016/S2213-2600(14)70002-5.Epub 2014 Feb 7. (2213-2600 (Print); 2213-2600 (Linking)).

(199) Yang J, Lee SH, FAU - Goddard ME, Goddard ME, FAU - Visscher PM, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. - Am J Hum Genet.2011 Jan 7;88(1):76-82.doi: 10.1016/j.ajhg.2010.11.011.Epub 2010 Dec 17. (1537-6605 (Electronic); 0002-9297 (Linking)).

(200) Yang J, Ferreira T, FAU - Morris AP, Morris AP, FAU - Medland SE, Medland SE, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. - Nat Genet.2012 Mar 18;44(4):369-75, S1-3.doi: 10.1038/ng.2213. (1546-1718 (Electronic); 1061-4036 (Linking)).

(201) Kraft P, Zeggini E, FAU - Ioannidis JPA, Ioannidis JP. Replication in genome-wide association studies. - Stat Sci.2009 Nov 1;24(4):561-573.doi: 10.1214/09-STS290. (0883-4237 (Print); 0883-4237 (Linking)).

(202) Aschard H, Tobin MD, Hancock DB, Skurnik D, Sood A, James A, et al. Evidence for large-scale gene-by-smoking interaction effects on pulmonary function. - Int J Epidemiol.2017 Jun 1;46(3):894-904.doi: 10.1093/ije/dyw318. (1464-3685 (Electronic); 0300-5771 (Linking)).

(203) Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. - Lancet.2001 Oct 20;358(9290):1356-60.doi: 10.1016/S0140-6736(01)06418-2. (0140-6736 (Print); 0140-6736 (Linking)).

(204) Clayton DG. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. - PLoS Genet.2009 Jul;5(7):e1000540.doi: 10.1371/journal.pgen.1000540.Epub 2009 Jul 3. (1553-7404 (Electronic); 1553-7390 (Linking)).

(205) Delatycki MB. The ethics of screening for disease. - Pathology.2012 Feb;44(2):63-8.doi: 10.1097/PAT.0b013e32834e8496. (1465-3931 (Electronic); 0031-3025 (Linking)).

(206) Sparks JA, AUID- ORCID: http://orcid.org/0000-0002-5556-4618, Iversen MD, Yu Z, Triedman NA, Prado MG, et al. Disclosure of Personalized Rheumatoid Arthritis Risk Using Genetics, Biomarkers, and Lifestyle Factors to Motivate Health Behavior Improvements: A Randomized Controlled Trial. - Arthritis Care Res (Hoboken).2018 Jun;70(6):823-833.doi: 10.1002/acr.23411.Epub 2018 Apr 16. (2151-4658 (Electronic); 2151-464X (Linking)).

(207) ScienceDaily. European Society of Human Genetics. "Individual access to genomic disease risk factors has a beneficial impact on lifestyles.". 2018; Available at: https://www.sciencedaily.com/releases/2018/06/180615185408.htm. Accessed 06/14, 2019.

(208) Lindstrom S, Yen YC, FAU - Spiegelman D, Spiegelman D, FAU - Kraft P, Kraft P. The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. - Hum Hered.2009;68(3):171-81.doi: 10.1159/000224637.Epub 2009 Jun 11. (1423-0062 (Electronic); 0001-5652 (Linking)).

(209) Flegal KM, FAU - Brownie C, Brownie C, FAU - Haas JD, Haas JD. The effects of exposure misclassification on estimates of relative risk. - Am J Epidemiol.1986 Apr;123(4):736-51. (0002-9262 (Print); 0002-9262 (Linking)).

(210) Wu AH. Exposure misclassification bias in studies of environmental tobacco smoke and lung cancer. - Environ Health Perspect.1999 Dec;107 Suppl 6:873-7.doi: 10.1289/ehp.99107s6873. (0091-6765 (Print); 0091-6765 (Linking)).

(211) Murray AL, Molenaar D, Johnson W, Krueger RF. Dependence of Gene-by-Environment Interactions (GxE) on Scaling: Comparing the Use of Sum Scores, Transformed Sum Scores and IRT Scores for the Phenotype in Tests of GxE. - Behav Genet.2016 Jul;46(4):552-72.doi: 10.1007/s10519-016-9783-5.Epub 2016 Feb 1. (1573-3297 (Electronic); 0001-8244 (Linking)).

(212) **NHLBI Trans-Omics for Precision Medicine**. Available at: https://www.nhlbiwgs.org/. Accessed 02/12, 2019.

(213) Wang H, Zhang F, Zeng J, Wu Y, Kemper KE, Xue A, et al. Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. bioRxiv 2019.

(214) Harlid S, Xu Z, FAU - Panduri V, Panduri V, FAU - Sandler DP, Sandler DP, et al. CpG sites associated with cigarette smoking: analysis of epigenome-wide data from the Sister Study. - Environ Health Perspect.2014 Jul;122(7):673-8.doi: 10.1289/ehp.1307480.Epub 2014 Apr 4. (1552-9924 (Electronic); 0091-6765 (Linking)).

(215) Prince C, Hammerton G, Taylor AE, Anderson EL, Timpson NJ, Davey Smith G, et al. Investigating the impact of cigarette smoking behaviours on DNA methylation patterns in adolescence. - Hum Mol Genet.2019 Jan 1;28(1):155-165.doi: 10.1093/hmg/ddy316. (1460-2083 (Electronic); 0964-6906 (Linking)).

(216) Siegfried Z, Simon I. DNA methylation and gene expression. - Wiley Interdiscip Rev Syst Biol Med.2010 May-Jun;2(3):362-371.doi: 10.1002/wsbm.64. (1939-005X (Electronic); 1939-005X (Linking)).

(217) Jang HS, Shin WJ, Lee JE, Do JT. CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. LID - E148 [pii] LID - 10.3390/genes8060148 [doi]. - Genes (Basel).2017 May 23;8(6).pii: genes8060148.doi: 10.3390/genes8060148. (2073-4425 (Print); 2073-4425 (Linking)).