Genetic epidemiology of lung function and chronic obstructive pulmonary disease

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

María Soler Artigas

Department of Health Sciences

University of Leicester

December 2014

Abstract

Chronic obstructive pulmonary disease (COPD) is a leading cause of death worldwide. Lung function measures obtained through spirometry play a key role in the diagnosis of COPD. Both COPD and lung function are affected by genetic factors, and identifying genetic variants that have an effect on lung function or COPD risk has the potential to lead to improved treatment and prevention of COPD.

This thesis is structured in five chapters, an introductory, a concluding chapter and three main chapters which present different approaches that aim to bring insights into the genetics of COPD and lung function. Chapter 2 tests the association with COPD risk of genetic variants previously associated with lung function, and tests their combined effect on lung function and COPD risk, in order to explore the role of risk prediction. Chapter 3 aims to identify new genetic variants associated with lung function and tests the association of genetic variants genome-wide. Chapter 4 focuses on the analysis of low frequency variants using different approaches and methodologies, and includes two studies. One study assesses associations of low frequency variants genome-wide, and the other focuses on genetic regions associated with lung function, in order to improve the localization of association signals that often comprise broad regions and several genes.

These studies overall have identified 16 new genetic variants associated with lung function, have shown the association with COPD of 4 genetic variants previously associated with lung function, and present suggestive evidence of association with COPD for low frequency variants within regions associated with lung function.

Acknowledgements

I would like to start by thanking my supervisors Professor Martin D. Tobin and Dr. Louise V. Wain for their unconditional support and encouragement, for always making time for me and for everything they have taught me.

I acknowledge all of the investigators who have collaborated with me during the time of my PhD, in particular all of the investigators in the SpiroMeta consortium.

Many thanks to my colleagues and friends in Leicester for all the good times we had together. Finally, I would like to thank my parents and Matej for being always there for me.

Statement of originality of the work

I undertook the analyses I present in Chapter 2 on behalf of the SpiroMeta consortium. An analysis plan was drafted and shared with the consortium analysts before I started working on this project. After that, I coordinated analyses across studies, liaising closely with other analysts to clarify any queries. I undertook thorough quality control checks on the results provided by the analysts and again liaised with analysts to solve any issues found. Then, I meta-analysed the results. I undertook individual level analyses in a subset of studies, and designed and carried out sensitivity analyses in subsets of studies with individual level data for the relevant variables available, to aid the interpretation of the results.

The study presented in Chapter 3 consisted of two stages, a discovery stage, where analyses were undertaken in studies within the SpiroMeta and the CHARGE consortia, and a follow-up stage, where analyses were undertaken in an additional set of studies. For both stages I designed the analysis plans, liaised with analysts to clarify any issues, undertook quality control checks on the results from all studies, liaised with analysts to solve any issues found and finally designed the strategy to meta-analyse the results and undertook the meta-analysis. I also undertook the study level analyses for a subset of studies in the follow-up stage, and did some additional analyses to aid the interpretation of the findings. At the time this work was undertaken, it was common practice in large GWAS consortia to have large scale analyses undertaken independently and in parallel to ensure consistency of findings and guard against programming errors influencing the analysis results. This practice, often referred to as “mirroring” of analyses, has also been considered to improve training opportunities for junior researchers. In this case, part of this work was “mirrored” by Dr Daan Loth, who has included the article resulting from this work in his PhD thesis (1). Dr Loth ran the quality control checks in the discovery stage results, and meta-analysed the results of the studies across discovery and follow-up stages following the strategy I designed. We liaised with one another

to ensure the consistency of findings and I led the drafting of the paper (2) resulting from this work and the coordination with consortium members to incorporate appropriate suggestions. I am listed as the first author in the paper and Dr Loth is listed as second author, although these authorships are “starred” to reflect the substantial contribution that is required to undertake an analysis of this scale, and other authors are also “starred” to reflect the substantial role in developing some of the resources that contributed to this effort across two large consortia. Similarly, I am a second-listed starred author for a paper describing a genome-wide association study of forced vital capacity (not included in this thesis) in which Dr Loth led the analysis and I provided “mirroring” of analyses (3).

Chapter 4 presents two studies. The first one was undertaken also within the SpiroMeta consortium. I first piloted this analysis in a study that provided individual level data and then designed an analysis plan to share with consortium members. I coordinated analyses within the consortium, ran quality control checks on the results, liaised with analysts to solve any issues I identified, and then meta-analysed and interpreted the findings. In the second study presented I played a leading role in the design of the study, including defining the genetic regions to be studied, and then processed the data, undertook quality control checks, designed the strategy to analyse the data and follow-up the findings, undertook the analyses and interpreted the results.

List of publications

Publications directly related to the thesis

Soler Artigas, 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. 184(7): p. 786-95.

Soler Artigas, M., et al., Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet, 2011. 43(11): p. 1082-90.

Publications indirectly related to the thesis

Tang, W., Kowgier, M., Loth, D. W., **Soler Artigas, M.** et al., Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. PLoS One, 2014. 9(7): p. e100776.

Loth, D.W., **Soler Artigas, M.,** et al., Genome-wide association analysis identifies six new loci associated with forced vital capacity. Nat Genet, 2014. 46(7): p. 669-77.

Obeidat, M., [9 authors], **Soler Artigas, M.**, [9 authors], GSTCD and INTS12 Regulation and Expression in the Human Lung. PLoS One, 2013. 8(9): p. e74630.

Hancock, D.B., **Soler Artigas, M.**, 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): p. e1003098.

Wilk, J.B., [55 authors], **Soler Artigas, M.**, [22 authors], Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. Am J Respir Crit Care Med, 2012. 186(7): p. 622-32.

Wain, L.V., **Soler Artigas, M.**, and Tobin, M.D., What can genetics tell us about the cause of fixed airflow obstruction? Clin Exp Allergy, 2012. 42(8): p. 1176-82.

Soler Artigas, M., L.V. Wain, and M.D. Tobin, Genome-wide association studies in lung disease. Thorax, 2012. 67(3): p. 271-3, 280.

Obeidat M., Wain L.V., Shrine N., Kalsheker N., **Soler Artigas M.**, et al., A comprehensive evaluation of potential lung function associated genes in the SpiroMeta general population sample. PLoS One, 2011. 6(5): p. e19382.

Contents

List of Tables

List of Figures

Chapter 1: Introductory chapter

This chapter provides an introduction to the genetic epidemiology of lung function and chronic obstructive pulmonary disease (COPD). It provides an introduction to genetic epidemiology in general, explaining the genetic concepts used throughout the thesis and placing a special emphasis on genetic association studies to identify common and rare variant associations. It describes COPD and the lung function measures used for its diagnosis. Finally, it sets the context for the genetic epidemiology studies described in this thesis and presents the outline of the thesis. References to publications where I am a co-author are marked with * in the main text (to aid situations when multiple publications are referenced together, publications where I am a co-author are also marked with * in the bibliography), and the two articles directly related to the thesis are provided in **Appendix A**.

1.1 Genetic epidemiology

Epidemiology studies the distribution and causes of health and disease conditions in populations, and genetic epidemiology focuses on the study of genetic causes and their interactions with environmental factors. Identifying the genetic causes of a disease will lead to an improved understanding of the underlying biological pathways, which may point to molecular targets for the development of new treatments. In addition, understanding the genetic causes

of a disease will enable the development of genetic risk profiles which may be used in disease prevention strategies or stratified approaches to treatment.

1.1.1 Genetic concepts

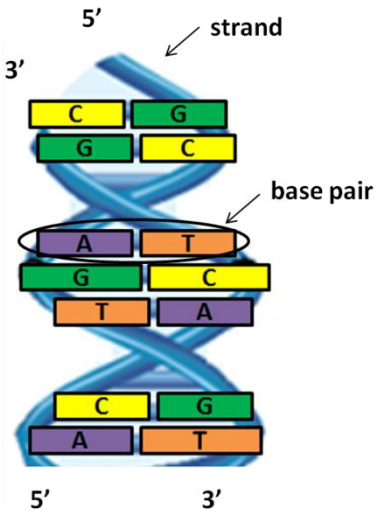
Table 1-1 provides a glossary for genetic terms highlighted in bold throughout this section.

Table 1-1 Genetics glossary box

allele:	alternative DNA sequence in a genomic location
base pair :	DNA nucleotides joined by a hydrogen bond between strands at a given position
chromosome:	structure in which the DNA is tightly packed together
crossover:	exchange of genetic information between chromosome pairs
deletion:	mutation where some DNA sequence is missing
diploid:	cell that has two copies of each chromosome
exons:	DNA sequence in a gene that codes for a functional unit
gene:	stretch of DNA sequence that codes for a protein, or a functional RNA molecule, which includes exons, introns and UTRs
gene expression:	process that uses the information encoded by a gene to create a functional product
genotype:	alleles at a chromosomal position on both chromosomes
haploid:	cell that has one copy of each chromosome
heterozygous:	genotype formed by two different alleles
homozygous:	genotype formed by two copies of the same allele
Hardy Weinberg Equilibrium :	principle which states that under random mating and no major evolutionary influences, allele and genotype frequencies will be stable over generations in a population
insertion:	mutation where some DNA sequence is added
introns:	DNA sequence in a gene that is removed by RNA splicing
linkage disequilibrium:	correlation between alleles of genetic variants produced due to the recombination process
locus:	genomic location
MAF:	Minor Allele Frequency, frequency of the less common allele in a population
mutation:	permanent change in the DNA sequence
nucleotides:	molecules which are the subunits of nucleic acids
recombination:	process by which two DNA molecules exchange information
sex cell (or gamete) :	cell capable of fusing with the sex cell from the opposite sex to form a fertilized egg
SNP:	Single Nucleotide Polymorphism, genetic variation produced by a nucleotide change
somatic cell:	all cells other than sex cells
transcription:	process by which DNA produces RNA
translation:	process by which mRNA produces protein

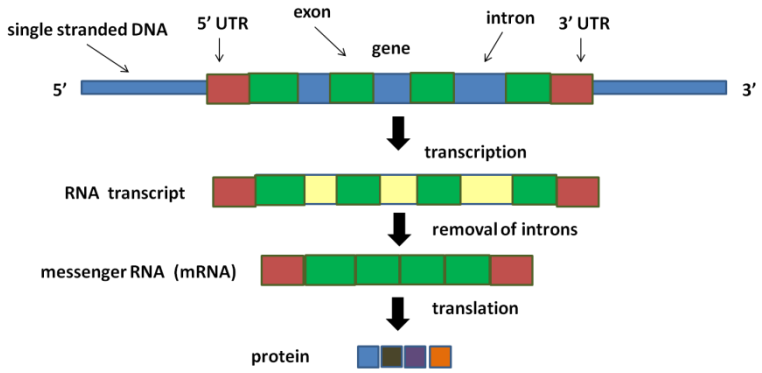
Double-stranded deoxyribonucleic acid (DNA) makes the human genome, and it is formed by four kinds of molecules called **nucleotides**: adenine (A), cytosine (C), guanine (G) and thymine (T). These molecules are joined by strong covalent bonds within a single strand, and by weaker hydrogen bonds with the complementary strand. Adenines are always paired up with thymines and cytosines are always paired up with guanines between strands forming **base pairs** (bp) (**Figure 1-1**). The two ends of each strand are called the 5' (five prime) end and 3' (three prime) end, and the two strands are orientated in opposite directions (**Figure 1-1**). The strand orientated 5' to 3' is called the "forward" strand and the one orientated 3' to 5' is called the "reverse" strand. The DNA is located in the nucleus of the cell and it is organized in **chromosomes**, each **somatic** cell has 22 pairs of autosomal chromosomes and one pair of sex chromosomes. One chromosome of each pair is inherited from the mother and the other is inherited from the father.

Figure 1-1 DNA structure



Genes are the regions of DNA which encode functional molecules such as proteins. In order to produce proteins, bonds between DNA strands are broken and single stranded DNA is used as a template to produce ribonucleic acid (RNA) in a process called **transcription** (**Figure 1-2**). RNA is similar to DNA, but it has uracil (U) instead of thymine. The DNA sequence in genes is formed by alternating regions of **exons** and **introns**. Both exons and introns are transcribed into RNA, but only exons contain the sequence that encodes the proteins. Therefore introns are then removed, and exons are spliced together to produce messenger RNA (mRNA) (**Figure 1-2**). Messenger RNA travels from the nucleus to the cytoplasm of the cell in order to produce proteins in a process called **translation** (**Figure 1-2**). Sometimes exons are spliced in different ways, so that a single gene can encode multiple proteins. At each side of the gene sequence there are untranslated regions (UTRs), the 5' UTR on the 5' end of the gene and the 3' UTR on the 3' end of the gene (**Figure 1-2**). These regions are also transcribed into RNA but do not translate into protein; they may play a role in regulating **gene expression**.

Figure 1-2 From DNA to protein

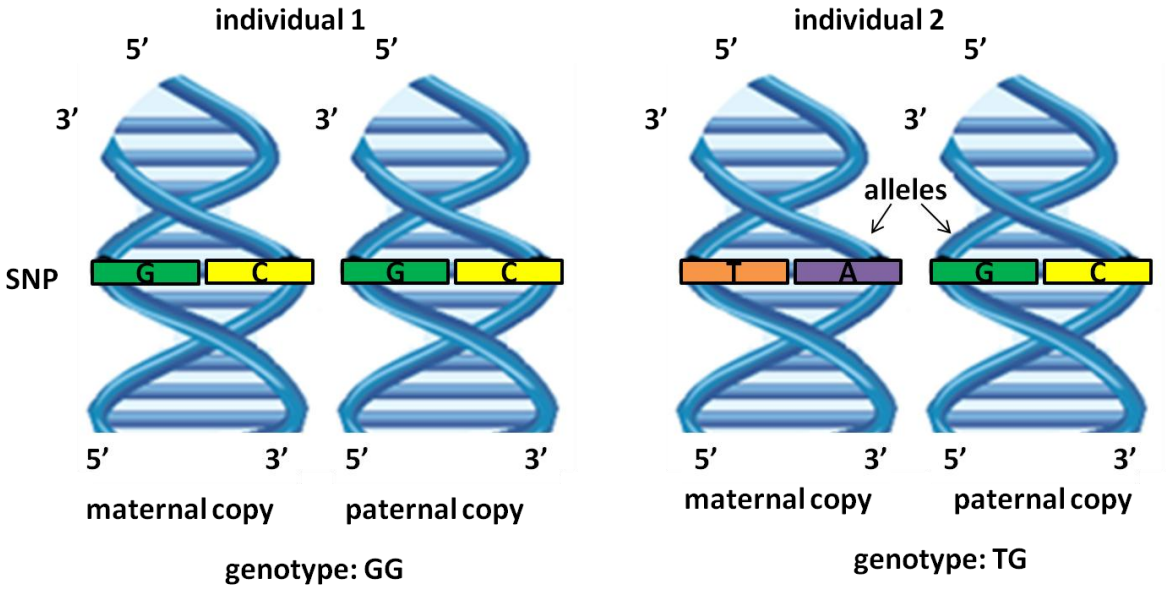


Around 3% of the human genome is comprised of protein-coding genes (4). The function of the remainder of the genome is largely unknown, although efforts like ENCODE (4) are making important advances; the genome outside of the protein-coding gene regions is likely to be involved in regulating gene expression through different mechanisms. Proteins carry out a range of fundamental functions in the human body, and DNA changes that affect protein function or availability can influence health and disease.

Most of the DNA sequence is the same between two individuals. However it may differ at a given location (**locus**), and this difference is called a **mutation**. The alternative sequences of DNA at that location are called **alleles** (**Figure 1-3**). The alleles of an individual at a chromosomal position in both the maternal and paternal chromosomes form the **genotype** (**Figure 1-3**). A genotype can be **heterozygous** if it is made of two different alleles, for example TG, or **homozygous** if it is made of two copies of the same allele (GG). Since information in both strands is complementary, genotypes only use information from one strand; genotypes for the two individuals shown in **Figure 1-3**, could be read as GG and TG, or CC and AC depending on which strand we use, but both give the same information. Alleles on the forward strand are more commonly reported. The term minor allele frequency (**MAF**) refers to the frequency of the “minor”, or less common allele, in a population; for instance in **Figure 1-3** the minor allele is T (or A if we use the other strand). According to the **Hardy Weinberg Equilibrium** principle, allele and genotype frequencies

will be stable over generations, assuming random mating and no major evolutionary influences in a population. Genetic variation produced by a nucleotide change, such as the one shown in **Figure 1-3**, is called a single nucleotide polymorphism (**SNP**). Sometimes a given sequence can be deleted (**deletion**) or repeated (by inserting the same sequence again: **insertion**). Large (for example 1kb and above) deletions or insertions are referred to as copy number variants. The work presented here focuses mainly on SNPs and includes also some analyses of INDELs (small INsertions or DELetions) which are treated in the analysis similarly to SNPs; analyses of larger copy number variants are more complex and are not covered in this thesis.

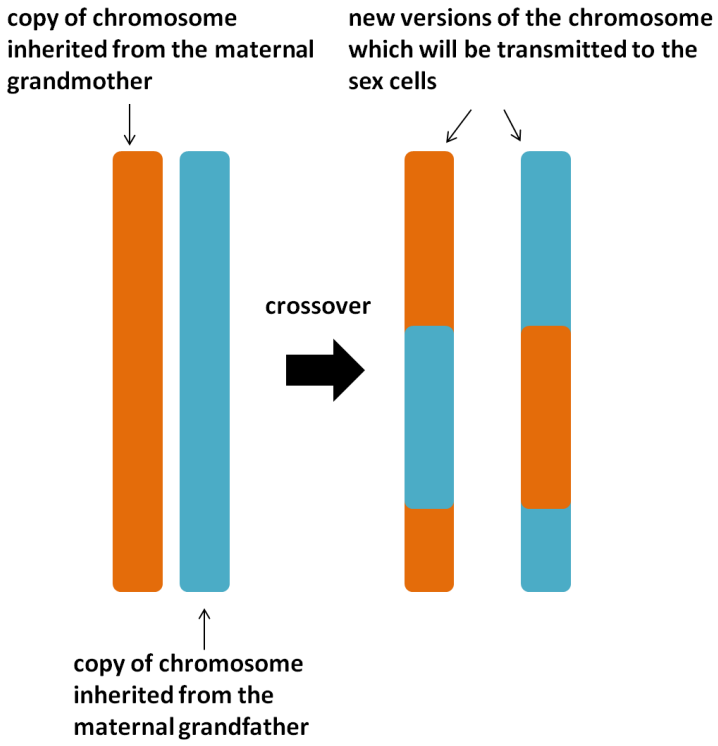
Figure 1-3 Alleles for two individuals at one position



Whilst somatic cells are **diploid**, **sex cells** or gametes (sperm and egg) are **haploid**: they only have 23 chromosomes instead of 23 pairs of chromosomes. Maternal and paternal sex cells fuse to form the zygote, which will become a human embryo with a diploid genome. The chromosomes that make the sex

cells are made from a combination of the two original pairs of chromosomes; for instance the maternal sex cells are formed by a combination of the maternal grandparents' chromosomes. When sex cells are being produced, **crossover** events occur and chromosome pairs exchange information, which leads to **recombination (Figure 1-4**). Thus, new versions of chromosomes, now made from a combination of the two original chromosomes, are assigned randomly to sex cells. Due to this process of recombination, genetic variants that are located close to each other in a chromosome are more likely to be passed on together over many generations. Therefore, in a population the correlation of alleles from variants that are located close to each other tends to be higher than the correlation of alleles from variants located further apart. This correlation is called **linkage disequilibrium** (LD).

Figure 1-4 Crossover and recombination for a maternal chromosome



1.1.2 Genetic epidemiology: an overview

There are several steps in the analysis of a trait in genetic epidemiology. The first step is to assess whether that trait is influenced by genetic factors. This can be done by assessing whether it aggregates within families and in that case, whether the pattern of aggregation is consistent with a genetic effect. At this stage no direct measures of genetic variants are needed.

A trait aggregates within families if there is greater frequency of a disease among close relatives of individuals with the disease than among relatives of individuals without the disease. Once aggregation within families is shown, in order to assess whether the pattern of aggregation is consistent with a genetic effect, variance components models can be used. Variance components modelling takes into account how a gene or group of genes might affect the trait of interest (for example, through an additive effect) and the probability of sharing alleles which have been inherited directly from the same ancestor, identical by descent (IBD), among different classes of relatives. With these models it is possible to estimate the proportion of the variance of the trait that is attributable to genetic effects, called heritability. Narrow sense heritability is the proportion of the trait variance explained by additive genetic effects; and broad sense heritability is the proportion of the trait variance attributable to all genetic effects, including additive and non-additive effects. To estimate broad sense heritability, specific familial structures such as monozygous twins are required.

For quantitative traits, heritability is formally defined, and for binary traits it can be derived using the concept of liability, a quantitative measure assumed to be normally distributed that determines the probability of an individual developing the disease of interest (5), so that an individual is considered diseased if their liability exceeds a certain threshold. Heritability estimates for both quantitative and binary traits might be affected by biases, and in particular the heritability of the liability of a binary trait may be especially hard to interpret (6, 7). With modern advances in technologies, new methodologies to estimate heritability using genomics data have arisen (8). Despite their pitfalls, heritability remains a relevant metric, which can inform decisions about study design. Once it is established that a trait is influenced by genetic factors, different approaches can be used to identify the genomic locations that have an effect on the trait.

Genetic linkage studies are based on a limited number of genetic variants spread throughout the genome and on models that quantify how often alleles are transmitted through a family with the disease status, based on the biology of gamete formation and chromosomal recombination. Genetic linkage studies have been successful in identifying genes with large effects for monogenic disorders (9). However, studying complex traits, affected by the interaction of many genetic variants with small effect sizes and environmental factors, has proven to be more challenging (10).

Candidate gene studies test the association of genes selected in advance (for example due to their suggested biological function) with the trait of interest, and are therefore limited by existing knowledge. These studies have often been underpowered mainly due to their small sample sizes, and their findings have not generally been replicated (10).

The development of chip-based microarray technology, that allows cost-effective assay of a large number of genetic variants genome-wide, made possible the phenomenon of genome-wide association studies (GWAS). In GWAS, individuals are genotyped using microarrays and the association of each variant with the trait of interest is tested. A detailed explanation of genome-wide association studies is provided in section 1.1.3. GWAS have been successful in identifying common genetic variants (MAF > 5%) associated with a number of diseases (11, 12). However, these loci tend to have small effect sizes and explain only a small proportion of the heritability for these traits. A series of hypotheses have been formulated to explain this issue (13, 14). One of them (13) indicates that variants with lower allele frequency are more likely to have larger effects and therefore could explain a larger proportion of the heritability. A large focus in the study of the genetics of complex diseases at the moment is to study rare genetic variation, and different study designs are currently being used. These approaches are discussed in section 1.1.4.

1.1.3 Study of common variants: GWAS

Genome-wide association studies are a powerful tool to study the effect of common genetic variants (MAF > 5%) on complex traits. These studies aim to test the association of common genetic variation genome-wide with a trait of interest, by testing the association of a subset of variants (≥ 300,000) spread throughout the genome. Given the linkage disequilibrium existing between variants across the genome, this limited number of variants tags a large proportion of the common variation genome-wide. In addition, the number of variants being analysed in GWAS can increase to millions after undertaking imputation; this technique makes use of the LD patterns across the genome to infer the genotypes of variants that have not been genotyped by using a reference panel with data on a larger number of variants. Details of this approach are given later in this section. The development of chip-based microarray technology, where a single chip can be used to measure genetic variants in multiple samples in one experiment, has made it possible to assay hundreds of thousands of variants in thousands of individuals at an affordable cost, empowering these studies to detect genetic associations. In addition, GWAS present a hypothesis-free design, which has the potential to identify new biological pathways for the trait of interest.

1.1.3.1 Post-genotyping quality control checks

In order to minimise false positive associations, quality control (QC) checks, both per genetic variant and per individual, are undertaken on the genotype

data (15). Genetic variant QC checks include: (i) estimating variant call rate (for a given variant, proportion of individuals with no missing data), to identify variants with low call rate which may indicate a failure of the assay for that variant; (ii) identifying variants out of Hardy Weinberg Equilibrium (see explanation in **Table 1-1**), which can indicate genotyping or genotype-calling errors; and (iii) depending on the genotyping array used and its coverage for low allele frequency variants, variants with MAF below a certain threshold have often been excluded. Individual sample QC checks include: (i) inferring sex from genomic data and comparing it with the sex information provided in the phenotypic data, in order to identify potential DNA sample mix-up or DNA sample contamination; (ii) estimating heterozygosity rates (the proportion of heterozygous genotypes for a given individual), since individuals with outlying heterozygosity are likely to point to DNA sample contamination; (iii) estimating individual call rate (for a given individual, proportion of genotypes with no missing data), since individuals with low call rate indicate low DNA quality or concentration; and (iv) estimating relatedness for each pair of individuals, in order to identify duplicated individuals, which may have been introduced intentionally as positive controls, or may point to sample mix-ups, or related individuals, which need to be taken into account when choosing association testing methods.

1.1.3.2 Imputation

After undertaking quality control checks on the genotype data, imputation is usually undertaken to increase the number of genetic variants that will be tested for association with the trait of interest, and to fill in any missing genotypes for genotyped variants. Thanks to the LD patterns across the genome, missing genotypes can be inferred using a reference panel with complete data for those variants with missing genotypes, in a process called imputation. Projects like HapMap (16), which sequenced 270 individuals from different ancestries in the first phase and made the data publicly available, allow this inference of missing genotypes. Using HapMap's reference panel (16) the number of variants tested can increase from ~ 300,000 genotyped variants to ~ 2.5 million imputed variants. Well tested methods (17-19) are available to implement this imputation, and they provide imputation quality metrics for each variant (20). Variants with low imputation quality are usually excluded from the analysis.

1.1.3.3 Association testing

Different genetic models can be used to test the effect of a genetic variant on a trait. According different biological scenarios we can use a recessive, dominant or additive genetic model. The effect of one allele on the trait is usually reported, and this allele is referred to as the coded or effect allele. A coded allele is recessive, if two copies are required for it to have an effect on the trait, whereas a coded allele is dominant if only one copy is required to have an effect on the trait and that effect is the same as if there were two copies. In

contrast, a coded allele is additive if the effect on the trait is in equal increments per copy of the coded allele. For instance, given three possible genotypes for one genetic variant, AA, AG and GG, if we want to measure the effect of A (A is coded allele) in a model and A is a recessive allele, we will code the genotypes as 1 for AA and 0 for AG and GG; if A was a dominant allele we would code AA and AG as 1 and GG as 0; and if A had an additive effect we would code AA as 2, AG as 1 and GG as 0. The true genetic model is usually unknown, and additive genetic models are the most commonly employed in GWAS (21, 22)*.

Both quantitative traits and binary disease status (case/control) are studied in GWAS. Case-control studies are generally more prone to being affected by biases; for instance if cases and controls have been genotyped separately, differential biases can arise. In some instances, quantitative traits underlie disease status and a powerful approach in this case is to study the quantitative traits genome-wide and then assess the association with the disease only for the variants that showed association with the quantitative traits (23, 24)*.

Genetic association studies are not as severely affected by confounding as observational epidemiological studies, and often little adjustment is made for covariates. Due to the random allocation of alleles in gamete formation, lifestyle factors are not likely to confound genetic associations. However genetic associations can be confounded by differences in population structure; if a

given allele varies in frequency across strata of a study population, and if the trait of interest also happens to vary also across strata of the study populations, this population structure can confound the association, and potentially cause a "false positive" association. Two main approaches have been developed to deal with population structure in GWAS. The first one is to summarise genetic variation genome-wide across individuals using principal components analysis (25). Principal components are calculated using the covariance matrix of individuals included in the study. These principal components can be used in two ways. They can be calculated jointly for individuals in the study and for individuals of known different ancestries. This way, individuals of different ancestries to the population being studied can be identified and excluded or analysed separately. Alternatively, if all individuals in the study belong to the same ancestry, principal components can be calculated only for individuals included in the study and they can be added as covariates to the model to account for more subtle population structure. The second approach is to assess whether the association test statistics are over-inflated genome-wide, which would be expected in the presence of population structure, and in this case adjust the test statistics. This method is called genomic control (26). In order to assess whether the statistics are over-inflated, the genomic inflation factor (λ) is calculated as the median of the test statistics divided the median of the distribution of these test statistics under the null hypothesis of no association. If λ is greater than one this indicates possible over-inflation, and the statistics are divided by λ in order to correct for this over-inflation.

1.1.3.4 Post-association testing quality control checks

Once the genome-wide association testing has been undertaken, additional checks are usually performed on the most significant variants because associations with very low P-values can be enriched for variants affected by biases. Cluster plots of hybridization intensities for each allele at directly genotyped genetic variants can be examined. Calling algorithms compare the relative strength of hybridization intensities for each allele to call genotypes; if the three genotype clusters (homozygous for one allele, heterozygous and homozygous for the other allele) are not well separated this indicates that the genotype calls might not be reliable. Given the linkage disequilibrium that exists between genetic variants that are located close to each other, it is expected that if a variant is significant, variants in LD with it will also be significant. Manhattan plots, which show P-values genome-wide ordered by chromosomal position, are used to visualise the results genome-wide, and zoomed in versions of the Manhattan plot are produced for any interesting regions. These zoomed in versions of the Manhattan plot are called region plots, and also show the degree of LD for variants in the region with the top variant, represented with different colours; and are used to assess whether variants in LD with the top variant show support for the association.

1.1.3.5 Follow-up studies and replication

Despite undertaking thorough quality control checks throughout the analytic process, some false positive associations might still arise given the large

number of variants tested. For this reason after undertaking the association testing genome-wide, studies seek replication of the GWAS top signals in independent samples. Therefore genome-wide association studies usually have a discovery stage, where associations are tested genome-wide, and a follow-up (or replication) stage, where only a reduced number of variants with the strongest evidence from discovery are tested in a set of independent samples. Effect estimates of the top signals in the discovery stage are likely to be affected by winner's curse bias (27), and overestimate the true effect sizes. For this reason, larger sample sizes for the follow-up stage are required in order to detect the genetic associations identified in discovery; and effect size estimates from the follow-up stage will be more reliable than from discovery.

1.1.3.6 Significance threshold

In GWAS a very large number of hypotheses are being tested, and it is therefore important to use a suitable significance threshold. There is a general consensus on a threshold of ~5 x 10^{-8} for considering a variant genome-wide significant in European populations, after correcting for 1-2 million independent tests (11). In order to increase power, the discovery stage and the follow-up stage are often meta-analysed for the subset of variants selected for follow-up, and a variant is considered genome-wide significant if it meets the $P < 5 \times 10^{-8}$ threshold after discovery and follow-up meta-analysis (11, 28).

1.1.3.7 Meta-analysis

Common genetic variants that affect complex traits tend to have moderate effect sizes, and very large sample sizes are required to detect them. For this reason many studies join their efforts forming large consortia. Sharing individual level data is often challenging, and a common practice in consortia is to develop an analysis plan, which is then followed by each study; and then meta-analyse results across studies centrally. Often studies use different genotyping platforms including different sets of variants, in this context imputation becomes relevant, not just due to the increase in number of variants being analysed, but because it allows the meta-analysis of the same set of variants across studies. It is important that studies included in a meta-analysis have similar characteristics, for instance that the trait being analysed has been produced in the same way. There are different approaches to meta-analyse findings across studies; fixed effect meta-analyses are commonly used, but in the presence of heterogeneity across studies, when for example not all the studies have been undertaken in the same way, random effect meta-analyses might be preferred instead.

Additional quality control checks are undertaken when meta-analysing study level results. Before undertaking the meta-analysis it is important to ensure the quality of the data for all the studies that will take part in the meta-analysis. In general additive genetic models are used in GWAS, so that the effect size of one allele is reported by each study. It is crucial when meta-analysing effect sizes that they all correspond to the same allele, so effect sizes often need to

be flipped for some studies. A comprehensive description of these quality control checks is provided in Chapter 3. Once the meta-analysis has been completed, additional checks can be implemented for the top findings. For instance, evaluating heterogeneity across studies, by formally testing for heterogeneity and by producing forest plots, where the effect sizes across studies are plotted together. In addition, when working with imputed data, a statistic termed “N effective” can be calculated for each variant, by multiplying the imputation quality by the sample size in each study and then summing these across studies. This statistic aids to assess how well imputed a variant is across studies.

1.1.4 Study of rare variants

Despite the success of GWAS in identifying common variants that affect complex traits, variants that meet the genome-wide significance threshold in GWAS tend to explain collectively a small proportion of the heritability (11, 12). In addition, identifying the causal variants that drive GWAS signals is often challenging since GWAS signals often involve several correlated variants and can span several genes. There is evidence in the literature (29, 30) of synthetic associations (associations of common markers as a result of multiple low allele frequency [1% < MAF < 5%] or rare [MAF < 1%] causal variants) explaining some GWAS signals. Although it is not clear how frequently these synthetic associations might occur, it is unlikely that they will explain many GWAS hits (31). There is also evidence in the literature (32) of rare variants in the same

locus identified by a GWAS association for a common variant, which also have an effect on the trait but independently of the common variant. Once a rare variant associated with a trait is identified, it is usually easier to narrow down the association signal, since rare variants are produced by more recent mutations and therefore they are only correlated with a limited number of other variants (33). In addition, rare variants may directly affect function and according to evolutionary theory, deleterious alleles with large effects are more likely to be rare (34). For these reasons rare variants are expected to be more clinically relevant and to be important for explaining the missing heritability.

1.1.4.1 GWAS approaches to identify rare variants

Detecting associations for rare variants is challenging. Tools that have been used in GWAS are not ideal for detecting rare variant associations. GWAS SNP chips mainly contain common variants, and genotype calling algorithms that are based on genotype clustering do not always perform well when there is a small number of variants within a genotype cluster (35), as is the case for rare variants. However, new genotyping chips are being designed with larger content of rare and low allele frequency variants, such as the exome chip which includes coding variants down to low allele frequencies. Imputation for rare variants is also more challenging since imputation uses LD patterns across the genome to infer missing genotypes, and rare variants are only correlated with a limited number of other variants. Larger imputation reference panels are currently being produced by sequencing large numbers of individuals, which

enables the imputation of increasingly lower frequency variants. The 1000 Genomes Project Phase 1(36) sequenced 1,092 individuals from 14 populations from Europe, Africa, East Asia and America in Phase 1 and has made publicly available an imputation reference panel which comprises around 38 million SNPs and 1.4 million INDELs. The UK10K project (http://www.uk10k.org/) is a UK based project that has sequenced 3,781 British individuals and has combined their data with the 1000 Genomes Project (36) to produce a combined reference panel. In addition, the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/) combines data from multiple cohorts and will create a reference panel with more than 30,000 individuals mainly of European ancestry.

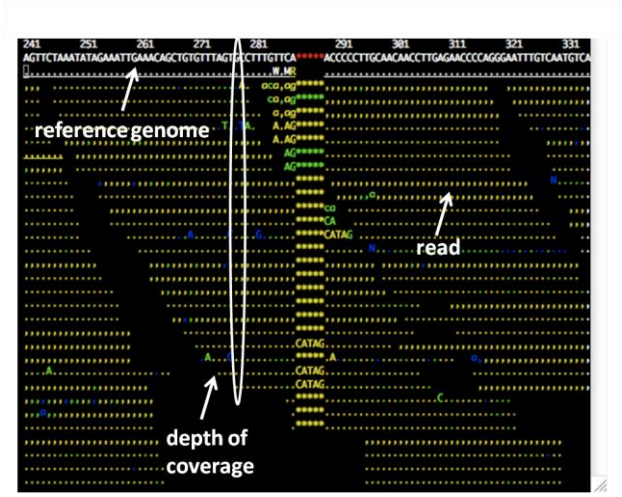
1.1.4.2 Sequencing

DNA sequencing is the process used to determine the nucleotides of a DNA molecule and their order. Therefore sequencing studies have data available for every nucleotide (both those that vary between individuals and those that do not) and not just a subset of genetic variants like in GWAS. For this reason they would be the preferred choice for identifying associations with rare variants and for identifying the causal variants that underlie GWAS findings.

In the sequencing process, the DNA is broken into small fragments which are then sequenced in a large number of parallel reactions. The strings of

nucleotides in each DNA fragment produced after the sequencing process are called reads (**Figure 1-5**). In order to find which part of the genome each read comes from, these reads are mapped to a reference genome in a process called alignment (**Figure 1-5**). Depth of coverage for a given position is the number of reads that align to that position (**Figure 1-5**). When different nucleotides appear in different reads for the same position, variant calling algorithms attempt to distinguish whether this difference is due to a sequencing error or due to a true variant. Reliable variant calling is highly dependent on depth of coverage.

Figure 1-5 Screenshot of read alignment



Although sequencing studies are a promising tool for studying rare variants, they also present challenges. The sources of variation that affect the sequencing process are more varied than those affecting genotyping and are in general less well understood (37). In contrast to the consistent performance of genotyping platforms, sequencing has a higher error rate, which varies across the genome. The use of different sequencing technologies within a study can

introduce biases, and ideally all individuals in an association study based on sequencing should be sequenced using the same technology (37). Even when the same technology and protocol are applied, sequencing reactions theirselves can vary and introduce variation between units of DNA that have been sequenced separately (38). These variation needs to be taken into account in the variant calling process to avoid false positive calls, and variant calling algorithms are gradually incorporating this information (39). In addition, QC and association pipelines for sequencing are not as clear or as well tested as they are for GWAS. Nevertheless, sequencing technologies are improving rapidly, and advances are being made in producing software and guidelines to analyse sequence data (37, 40).

Sequencing prices have dropped in the last few years, but high depth whole genome sequencing remains expensive. For this reason there are sequencing study designs that target specific variation or specific regions, such as exome sequencing studies where only the coding part of the genome is sequenced, or targeted sequencing studies where only candidate regions, or regions with prior evidence of association with the trait are sequenced. In order to maximise sample size, a pooled sample design can be used, where DNA from multiple individuals is pooled, so that DNA from different individuals is sequenced together. This design is especially beneficial for targeted sequencing studies, where only target regions are sequenced instead of whole genomes and this design allows a cost-efficient use of the sequencing capacity. In addition, in

targeted sequencing studies there is high cost involved in capturing the targeted regions. This step consists of enriching the DNA for the regions of interest. In a pooled design this step only needs to be undertaken once per pool, instead of once per individual reducing the cost considerably. However this comes at a price, and variant calling is more challenging in this context. When individuals are sequenced separately, for a variant with a heterozygous genotype an individual is expected to have one allele in around 50% of the reads and the other allele in the remaining 50% of reads, regardless of how common the variant is in the population. However, if a number of individuals are sequenced together in a pool, the number of reads with the minor allele for a rare variant will be small, and depending on depth, could be close to the sequencing error rate, making it much harder to identify a true rare variant. Several methods (41-47) have been developed to account for pooled designs when calling variants, and a selection of them are discussed and applied in Chapter 4.

1.1.4.3 Collapsing methods

Single variant analysis is often underpowered for detecting associations with rare variants. In order to increase power to detect associations with rare variants, their effect can be analysed jointly within a region. Regions can be defined in different ways according to different biological scenarios, for instance regions can be defined using gene coordinates in order to detect the effect of multiple variants in a gene, or can be defined as sliding windows in order to detect the effect of regulatory regions. A large range of methods for analysing

rare variants has been developed in the last few years (48). Some of these methods (49-52) aggregate information across rare variants into a single quantity, which is then tested for association with the trait. These methods are called burden tests, and they assume that all rare variants tested within a region have an effect on the trait on the same direction (all protective or all damaging). Other methods (53, 54) have been developed that model similarities across individuals, and these allow variants within a region to have different direction of effect on the trait. Variants might be weighted according to data quality, allele frequency or functionality. A selection of these methods are discussed and applied in Chapter 4.

1.2 COPD, lung function and spirometry

Chronic obstructive pulmonary disease (COPD) is a major health concern across the world. According to the World Health Organization (http://www.who.int/en/) around 64 million people were estimated to have COPD worldwide in 2004, with more than 3 million deaths from COPD in 2005, 90% of which occurred in low- and middle-income countries; these numbers are expected to increase in the coming decades.

1.2.1 Definition

COPD is a preventable and treatable disease, characterized by chronic airflow limitation that is not fully reversible and by pathological changes in the lung (55).

Symptoms include breathlessness, chronic cough and chronic sputum production. Exacerbations, episodes of acute worsening of symptoms, are common in patients with COPD. Chronic inflammation in response to inhaled irritants leads to the narrowing of the small airways (obstructive bronchitis) and to the destruction of lung functional tissue (emphysema), which produce chronic airflow limitation (55).

COPD prevalence estimates vary from study to study according to differences in the methodologies used (56). COPD often results from accumulation of long term exposure to noxious agents, and is usually developed later in life (56, 57). The prevalence in individuals over 40 years old is estimated to be around 10% (56) and is higher in smokers (around15% (56)) and ex-smokers (around 10% (56)) than in never-smokers (3% to 7% (56, 58)).

1.2.2 Risk factors

The main risk factor for COPD is smoking; cigarette smoking is more common, but other types of tobacco as well as passive smoking are also risk factors (55). However, not all smokers develop COPD and genetics are known to play a role. What is known of the genetics of COPD is discussed in the next section (1.3). Occupational exposures to dusts and fumes (59, 60) and outdoor air pollution are also risk factors for COPD (61), as well as indoor pollution from biomass

cooking and heating (62-64), which is particularly relevant in developing countries.

Aging is associated with increased risk of COPD (55), however it is unclear whether the aging process is itself a risk factor or whether it only reflects the accumulation of exposures through life. The prevalence of COPD in males used to be greater than in females, however recent studies show that prevalence is almost equal (65); this is likely to be due to the increase in women taking up smoking.

Lung growth and development also have an effect on the risk of developing COPD. Since the lungs are not fully developed until late adolescence (66), factors that affect fetal development, childhood or adolescence might also have an effect on an individual's risk of developing COPD; for example, maternal smoking (67), childhood respiratory infections (68) or asthma in childhood (69).

1.2.3 **Diagnosis**

Patients who suffer from breathlessness, chronic cough, or sputum production, or who have a history of exposure to risk factors or a family history of COPD are considered for clinical diagnosis of COPD (55). In order to make a clinical diagnosis of COPD, spirometry is required. Spirometry is a simple test that

measures airflow limitation reproducibly and reliably (55, 70), and should be undertaken following published guidelines (71). The three spirometry measures most commonly used in the diagnosis of COPD are: Forced Vital Capacity (FVC), total volume of air exhaled in one breath; Forced Expiratory Volume in one second (FEV_1), volume of air expired in the first second of a maximal expiration and the ratio of FEV_1 over FVC. A ratio of FEV_1 over FVC after using a bronchodilator, a substance that dilates the bronchi and bronchioles, below 0.7 in adults indicates airflow limitation. Postbronchodilator FEV_1 is used to assess the severity of this airflow limitation. Since FEV_1 is influenced by age, sex, height and ethnicity, a percentage of the predicted normal value is used instead, using normal values for the local populations (72, 73). According to the Global initiative for chronic Obstructive Lung Disease (GOLD) guidelines (55), patients with airflow obstruction ($FEV_1/FVC < 0.7$) and $FEV_1 \geq 80\%$ of predicted are classified as mild or GOLD stage 1, those with $50\% \leq FEV_1 < 80\%$ of predicted are classified as moderate or GOLD stage 2, those with $30\% \leq FEV_1 < 50\%$ of predicted are classified as severe or GOLD stage 3 and those with $FEV_1 < 30\%$ of predicted are classified as very severe or GOLD stage 4.

Until the most recent version of the GOLD guidelines, COPD could be diagnosed only by undertaking spirometry, However, the current version of these guidelines also requires an assessment of the risk of exacerbations and of symptoms according to two questionnaires: the COPD Assessment Test (CAT) or the modified British Medical Research Council (mMRC). Patients are

now divided into four categories: A) GOLD stage 1 or 2, one or less exacerbations per year and a CAT score < 10 or a mMRC score between 0 and 1; B) GOLD stage 1 or 2, one or less exacerbations per year and a CAT score ≥ 10 or a mMRC score ≥ 2; C) GOLD stage 3 or 4, two or more exacerbations per year and a CAT score < 10 or a mMRC score between 0 and 1; and D) GOLD stage 3 or 4, two or more exacerbations per year and a CAT score ≥ 10 or a mMRC score ≥ 2. Additionally an assessment of co-morbidities is also recommended.

Differential diagnoses for COPD include congestive heart failure, bronchiectasis, tuberculosis, etc. Asthma symptoms often overlap with COPD symptoms. However asthma often starts in childhood (55), and is characterised by attacks of breathlessness and wheezing followed by relatively symptom-free periods. In addition, airflow obstruction in asthma is reversible by the use of bronchodilators, whereas in COPD it is not fully reversible. It is sometimes not possible to distinguish chronic asthma from COPD using current testing techniques, and in these cases it is assumed that asthma and COPD coexist (55).

1.2.4 Biological mechanisms and features of COPD

The inhalation of noxious particles triggers an abnormal inflammatory response in the lungs of patients with COPD (55). This abnormal inflammatory response

leads to faulty injury and repair mechanisms that result in structural changes and narrow the airways. Another consequence of the abnormal inflammatory process is the destruction of lung functional tissue (74). The nature of the inflammatory response in patients who do not smoke and have not inhaled other harmful particles is unknown, as are the mechanisms involved in lung inflammation after smoking cessation. The inflammatory response in the lungs leads to the release of proteolytic enzymes as a defence mechanism. However if insufficient antiproteases are produced, an imbalance occurs that might lead to the destruction of elastin, an elastic protein that plays a major role in lung parenchyma. The loss of elasticity in the lungs produced by the destruction of elastin, is a key feature of emphysema (55). Inhalation of harmful particles produces an increase in reactive oxygen species in the lungs, which might produce an imbalance if insufficient antioxidants are produced to counteract their effect. This imbalance is called oxidative stress, and it is thought to contribute to the worsening of COPD through different mechanisms (75).

Airway obstruction in COPD results in air trapping during expiration, so that patients are not able to exhale completely and air is trapped in their lungs (55). This is called hyperinflation, and it also reduces inspiratory capacity, especially during exercise. Gas exchange through the alveolar-capillary membrane is often altered in COPD patients, and worsens with disease progression (55). Hypersecretion of mucus and the consequential chronic cough are present in patients with chronic bronchitis (76). However, they are not always a feature of

COPD, and they are not necessarily associated with airflow limitation (55). Bacterial or viral infections, pollutants or other unknown factors can trigger exacerbations of respiratory symptoms in COPD patients. Exacerbations accelerate disease progression (77) and their frequency varies from patient to patient, although they tend to become more regular when the disease is more severe.

1.3 Genetic epidemiology of lung function and COPD

Family studies have shown that lung function and COPD aggregate within families (78-80), and narrow sense heritability estimates between 40% and 50% for cross sectional lung function (70, 81, 82) and around 60% for COPD susceptibility have been reported (83). COPD is one of the leading causes of death worldwide (55), and lung function measures such as FEV_1 and FEV_1/FVC are used in diagnosis. The understanding of the genetics of lung function and COPD has the potential to lead to the development of new treatment and preventive strategies.

The first gene convincingly associated with COPD was *SERPINA1*. Mutations in this gene lead to alpha1-antitrypsin (AAT) deficiency (84) and cause COPD (AAT deficiency accounts for 1-2% of COPD cases) (85). Given that AAT protects the lung against proteolytic damage, AAT deficiency leads to early-onset emphysema (84). AAT deficiency is produced by mutations in only one

gene (it is a monogenic disorder), however the development and severity of the disease varies among patients (85).

Linkage studies reported linkage with lung function and COPD in several genomic locations; including chromosomes 1, 2, 12 and 19 (86-88). However, this linkage was reported in large genomic regions containing millions of base pairs and hundreds of genes, which made it hard to narrow down the signals. Candidate gene studies selected genes based on their potential connection with COPD, such as those involved in proteinase–antiproteinase and oxidative stress pathways (89). Many candidate gene studies were published for lung function and COPD, but their results were not often replicated (90). Candidate genes that remained significant after meta-analysing findings across different studies include *GSTM1*, *TNF*, *TGFB1* and *SOD3* (91). A comprehensive assessment of the effect of genes reported to be associated with lung function in candidate gene studies on lung function in a GWAS with more than 20,000 individuals did not show strong evidence of association for these genes (92)*. Overall, evidence from the candidate gene literature is hard to interpret (93)*. Differences in study populations, such as different ancestries, as well as differences in adjustments, particularly smoking adjustments, could explain some of the differing results obtained between studies. However, more serious issues are the limited power due to generally small sample sizes, the liberal statistical threshold for significance, failing to properly account for multiple testing, and the severe reporting and publication bias.

Genome-wide association studies overcome many of these problems, given that they present a hypothesis-free approach and use a well-established statistical threshold (the current consensus is $P < 5 \times 10^{-8}$) to define an association. GWAS for lung function and COPD have identified a number of loci to date. Two GWAS published in 2009, identified variants associated with FEV_1/FVC (94) and COPD (95) near *HHIP* (hedgehog interacting protein) at 4q31.21, and variants associated with COPD (95) in the alpha-nicotinic acetylcholine receptor *CHRNA 3/5* at 15q25.1. In 2010 two large consortia, SpiroMeta (23) and CHARGE (96), each with over 20,000 individuals, meta-analysed GWAS results for FEV_1 and FEV_1/FVC. Jointly these two studies identified 10 additional variants that were genome-wide significantly associated with lung function in at least one consortium, in or near: *TNS1*, *FAM13*, *GSTCD*, *HTR4*, *ADAM19*, *AGER*, *GPR126*, *PTCH1*, *THSD4* and *PID1*. *GSTCD* (glutathione S-transferase, C-terminal domain containing) may affect lung function through its involvement in cellular detoxification (97). *HTR4* (5-hydroxytryptamine (serotonin) receptor 4, G protein-coupled) may play a role in mediating air calibre (98). *AGER* is highly expressed in the lung (99), and reduced *AGER* expression has been related to pulmonary fibrosis (100, 101). *THSD4* may have an effect on lung function by playing a role in would healing and inflammation (23, 102). Another GWAS of COPD (103) showed that *FAM13* was also associated with COPD. In Chapter 3 I present a meta-analysis of studies that contributed to the SpiroMeta meta-analysis (23), to the CHARGE meta-analysis (96) and of some additional studies, with a total sample size of 48,201 individuals, the largest meta-analysis of GWAS for lung function (2)* at

the time. In Chapter 5 I discuss the findings of additional GWAS of COPD and lung function undertaken after the studies presented here.

Collectively, all the lung function loci discovered to date only explain a small proportion of the lung function heritability (2)*, similar to findings for other complex traits (13). For this reason, several efforts are in progress with the aim to identify low allele frequency and rare variants associated with lung function and COPD. In Chapter 4 I present two analyses undertaken with the aim of identifying rare variants associated with lung function.

1.4 Outline of the thesis

Chapter 2 describes an analysis of the association with COPD of a subset of variants associated with lung function which were identified in the first SpiroMeta study (23), and assesses their joint effect on lung function and COPD. This study was undertaken by the studies that participated in the initial SpiroMeta study (23). For this study I liaised with analysts from the different studies, carried out thorough quality control checks and meta-analysed the results across studies. I also designed and undertook sensitivity analyses using individual level data from a subset of studies to show the robustness of the findings. This work was published in the American Journal of Respiratory Critical Care Medicine in 2011 (24)*, and it is presented in **Appendix A**.

In Chapter 3 I present a large meta-analysis of GWAS of lung function. For this study I designed the analysis plan and coordinated analyses undertaken by the studies that took part in the meta-analysis. This chapter puts a particular emphasis on the in-depth quality control procedure that I undertook and highlights some of the issues encountered and how they were resolved. I meta-analysed the results across studies and undertook additional analyses to aid the interpretation of the findings. This work was published in Nature Genetics in 2011 (2)* and it is presented in **Appendix A**.

In Chapter 4 I present two different approaches to study the effect of rare variants on lung function. In section 4.2 of Chapter 4 I describe a meta-analysis of the results of a burden test undertaken by a subset of SpiroMeta studies. Since this was a new approach, I first piloted the analysis using individual level data from one study and then designed the analysis plan for the burden test analysis. After that, I applied the same procedure as in previous chapters, ensuring the quality of the data, meta-analysing results across studies and interpreting the findings. In section 4.3, I present a targeted sequencing study, with a pooled design, of the 26 loci associated with lung function in Chapter 3 and previous GWAS (23, 94, 96) undertaken in 300 COPD cases and 300 controls. Here I applied a range of methods to deal with the issues that arose from working with pooled sequence data and designed the final strategy for the analysis. I tested the effect of single variants, and the combined effect of rare variants in a locus using two different approaches.

Finally, Chapter 5 summarises the findings from the different chapters, discusses analytic approaches and limitations, gives an update of additional studies undertaken in the field and presents some potential applications of the findings.

Chapter 2: Associations with COPD and risk scores in the SpiroMeta dataset

This chapter assesses the association with COPD risk of six genetic variants (in or near *TNS1*, *GSTCD*, *HHIP*, *HTR4*, *AGER* and *THSD4*) associated with lung function in the SpiroMeta consortium (23), and investigates the combined effect of risk alleles in these six loci on lung function and COPD risk. This work was published in the American Journal of Respiratory Critical Care Medicine in 2011 (24)* (**Appendix A**).

2.1 Introduction

Chronic obstructive pulmonary disease aggregates in families (78-80) and is a leading cause of morbidity and mortality worldwide (104). The discovery of genetic variants that affect COPD risk could lead to the development of new preventive and treatment strategies. Many studies have investigated the genetics of COPD to date, however only a limited number of loci have been convincingly associated with COPD (92, 105)* (see section 1.3 in Chapter 1 for details). Among others, the reduced statistical power of analyzing a binary outcome (COPD cases vs. controls), has been one of the main limitations of these studies. Genome-wide association studies test the association of a trait with genetic variants across the genome and require a strict correction for multiple testing to avoid false positive associations. This strict correction coupled with the limited power of analyzing a binary trait makes it very challenging to detect genetic associations with COPD risk in a COPD GWAS.

However, if we assume that common genetic variants may have an effect on the risk of developing COPD through their effect on lung function, we could study the quantitative spirometry measures used in the diagnosis of COPD, instead of studying the disease status. Undertaking a GWAS on lung function measures and then assessing the association with COPD risk for the genetic variants with an effect on lung function would reduce the number of multiple tests undertaken and could be a statistically powerful approach.

FEV_1 and FEV_1/FVC play a key role in the diagnosis of COPD (details in section 1.2 in Chapter 1). Reduced FEV_1/FVC indicates airway obstruction and reduced FEV_1 is used to grade the severity of the obstruction. If common genetic variants exert an effect on COPD risk mediated via an effect on lung function, loci associated with FEV_1 and/or FEV_1/FVC will be expected to be also associated with COPD risk. This chapter investigates whether the loci reported to be significantly associated with FEV_1 and/or FEV_1/FVC by the SpiroMeta consortium (23) (*TNS1*, *GSTCD*, *HHIP*, *HTR4*, *AGER* and *THSD4*) are also associated with COPD risk.

As well as to aid development of new treatments to alleviate disease, another potential use of genetic information is to predict disease risk. Although the use of common genetic variants for prediction has been shown to still be of limited use for a range of complex diseases due to their small effect sizes (106, 107), the combined effect of the recently discovered risk alleles on COPD had not been evaluated. This chapter also assesses the combined effect of the genetic

variants associated with lung function reported by SpiroMeta on lung function and COPD risk by constructing risk scores.

2.2 My role in the study

To understand my role in the study, first it is necessary to set the context. The analyses presented in this chapter started just after the meta-analysis of lung function GWAS was completed in the SpiroMeta consortium in 2009 (23). They were undertaken as a follow-up of the findings in Repapi *et al*. (23) in a subset of the studies that took part in the SpiroMeta meta-analysis.

I became involved in this project in September 2009, when the analysis plan for this study had already been agreed and shared with the studies, however I still contributed to the overall strategy of the study. First, my main tasks were to coordinate analyses undertaken by the studies, to liaise with analysts to give advice on analytic issues, as well as interpreting and checking thoroughly the results sent by the studies to make sure no errors had been made. After that, I undertook a meta-analysis of the results and carried out pertinent sensitivity analyses. A subset of studies provided individual level data and this allowed me to design and undertake sensitivity analyses in subsets of studies with the data of interest available. A list of all the analyses included in this chapter and the studies that participated in each analysis, with an indication of which analyses I undertook are given in **Table 2-1**.

Table 2-1 Analyses undertaken

Definitions study abbreviations in section 2.3.1. The analyses that I undertook indicated in bold and the ones I corrected with *.

Study	Primary analyses			COPD risk sensitivity analyses				
	Associations with COPD risk for *TNS1*, *GSTCD*, *HTR4*, *AGER* and *THSD4*	**Associations with COPD risk for *HHIP***	**Lung function and COPD risk scores analyses**	**Effect of pack-years adjustment on the results**	**Effect in of analysing only ever-smokers on the results**	**Effect of the use of bronchodilator on COPD classification**	**Effect of excluding asthma cases on the results**	**Effect of the use of lower limit of normal COPD definition on the results**
EPIC obese cases	Yes	Yes	-	-	**Yes**	-	-	-
EPIC population-based	Yes	Yes	-	-	**Yes**	-	-	-
GS:SFHS	Yes	-	-	-	**Yes** (not for *HHIP*)	-	-	-
KORA F4	Yes (not for *HTR4*)	-	-	-	**Yes** (not for *HTR4* or *HHIP*)	-	-	-
ADONIX	Yes	Yes	Yes	-	**Yes**	-	-	-
BHS	Yes	**Yes**	**Yes**	-	**Yes** (not for *HHIP*)	-	-	-
BRHS	Yes	**Yes**	**Yes**	-	**Yes**	-	-	**Yes**
BWHHS	Yes	Yes	Yes *	-	**Yes**	-	**Yes**	**Yes**
Gedling	Yes	Yes	Yes *	**Yes**	**Yes**	-	**Yes**	**Yes**
HCS	Yes	Yes	Yes	-	**Yes**	-	-	-
Health 2000	Yes	Yes	Yes	-	**Yes** (not for *HHIP*)	-	-	-
Nottingham Smokers	Yes	Yes	Yes *	**Yes**	**Yes**	**Yes**	-	**Yes**
NSHD	Yes	Yes	Yes	-	**Yes** (not for *HHIP*)	-	-	-

2.3 Associations with COPD risk

A meta-analysis undertaken in the SpiroMeta consortium (23), with a discovery stage of 20,288 individuals and a follow-up stage for the top signals of over 50,000 individuals, confirmed the association with lung function of the previously discovered *HHIP* locus (94) and identified five new loci (*TNS1*, *GSTCD*, *HTR4*, *AGER* and *THSD4*) that were associated with lung function. These new loci were followed up in a subset of the studies involved in the SpiroMeta meta-analysis to assess their association with COPD and their results were then pooled together in a meta-analysis. Although the association of *HHIP* with COPD had already been reported (95), it was also included in this analysis for completeness.

2.3.1 Method

Populations, phenotyping and genotyping

The study population was made of 31,422 individuals over the age of 40 years drawn from 12 population-based studies (**Figure 2-1**). The characteristics of the study participants are shown on **Table 2-2**. Only one of these studies (EPIC) was part of the discovery stage that identified the five new loci (23), all the other studies took part in the follow-up stage. These studies included: the European Prospective Investigation into Cancer and Nutrition obese cases cohort (EPIC-obese cases) and population cohort (EPIC population-based), Generation Scotland: Scottish Family Health Study (GS:SFHS), Cooperative Health Research in the Region of

Augsburg (KORA F4), Adult-onset asthma and nitric oxide (ADONIX) study, Busselton Health Study (BHS), British Regional Heart Study (BRHS), British Women's Heart and Health Study (BWHHS), Gedling study (Gedling), Hertfordshire Cohort Study (HCS), Finnish Health 2000 survey (Health 2000), Nottingham Smokers study (Nottingham Smokers) and Medical Research Council National Survey of Health and Development (NSHD, or British 1946 Birth Cohort).

Figure 2-1 Study design

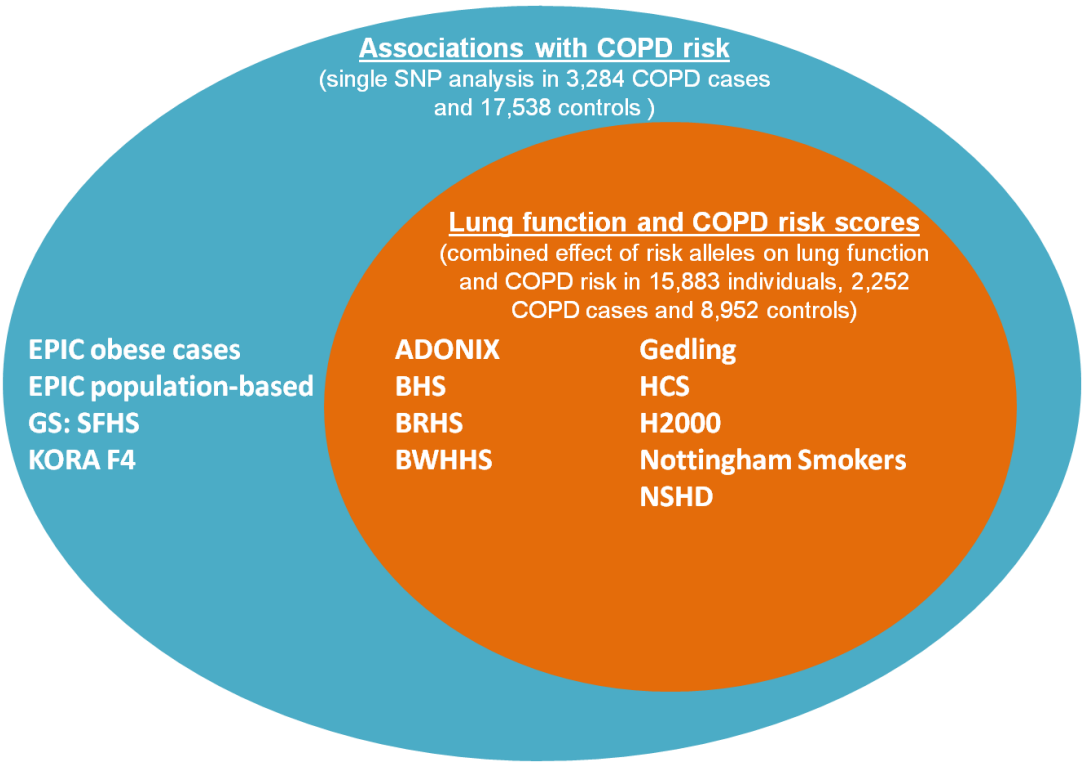


Table 2-2 Study characteristics

Abbreviations: N = number, SD =standard deviation.

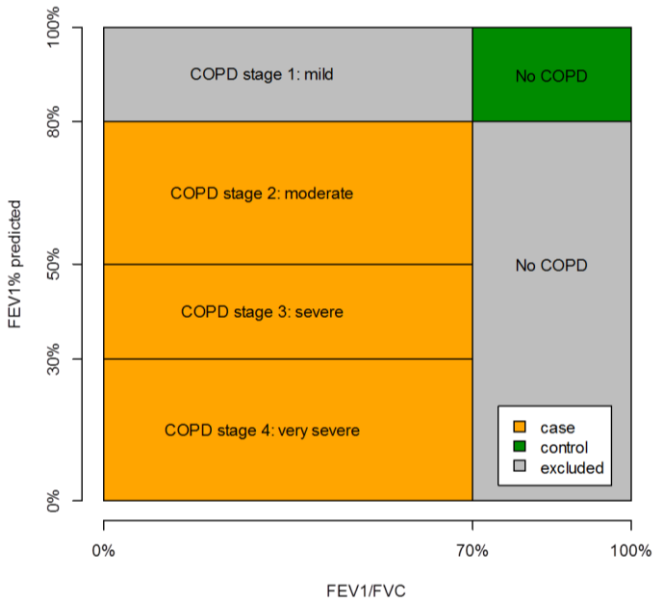
Study name	Subset	N associations with COPD	N risk scores	Male (N); female (N)	Age mean (SD)(years)	FEV_1: mean(SD) (litres)	FEV_1 predicted: mean(SD) (litres)	FVC: mean(SD) (litres)	FEV_1/FVC: mean(SD)	Never-smokers (N); Ever-smokers (N)	% of COPD cases in GOLD stage 3-4	Genotyping
Associations with COPD risk												
EPIC obese cases	All	1104		476;628	59.1 (8.8)	2.35 (0.69)	2.91 (0.62)	2.84 (0.87)	0.82(0.17)	489;615		Affymetrix 500K
	Cases	75		47;28	60.6 (8.45)	1.82 (0.67)	3.09 (0.63)	3.01 (1.05)	0.61 (0.09)	22;53	30.14	
	Controls	599		252;347	58.3 (8.76)	2.67 (0.62)	2.88 (0.61)	3.13 (0.79)	0.86 (0.07)	281;318		
EPIC population-based	All	2336		1100;1236	59.2 (9.0)	2.50 (0.72)	2.95 (0.62)	3.04 (0.90)	0.85 (0.16)	1061;1275		Affymetrix 500K
	Cases	190		105;85	62.3 (8.54)	1.81 (0.62)	2.95 (0.63)	3.00 (0.96)	0.60 (0.09)	72;118	20.11	
	Controls	1442		677;765	58.8 (8.81)	2.78 (0.64)	2.94 (0.62)	3.29 (0.82)	0.85 (0.08)	709;733		
GS:SFHS	All	5474		2254;3220	46.0 (14.3)	3.15 (0.87)	3.32 (0.75)	4.11 (1.03)	0.77 (0.10)	3005;2469		TaqMan
	Cases	335		118;217	58.4 (9.2)	1.89 (0.54)	2.89 (0.59)	3.32 (0.85)	0.58 (0.10)	123;212	11.94	
	Controls	2567		1053:1514	53.2 (8.5)	3.12 (0.72)	3.11 (0.63)	3.99 (0.91)	0.78 (0.07)	1457;1110		
KORA F4	All	1305		610;695	51.6 (5.7)	3.32 (0.81)	3.29 (0.63)	4.28 (1.00)	0.78 (0.06)	499;806		TaqMan
	Cases	59		30;29	53.4 (6.0)	2.14 (0.66)	3.24 (0.64)	3.47 (0.96)	0.61 (0.06)	12;47	10.17	
	Controls	1109		512;597	51.5 (5.7)	3.45 (0.76)	3.28 (0.62)	4.36 (0.96)	0.79 (0.04)	456;653		
Associations with COPD and lung function and COPD risk scores												
ADONIX	All	1423	1282	669;754	49.1 (13.5)	3.34 (0.86)	3.23 (0.66)	4.24 (1.02)	0.79 (0.07)	798;625		KaSPar ‡
	Cases	46	41	27;19	55.7 (9.3)	2.02 (0.57)	3.23 (0.66)	3.35 (0.87)	0.60 (0.07)	12;34	13.04	
	Controls	783	711	361;422	61.4 (8.4)	3.23 (0.73)	3.23 (0.67)	4.08 (0.91)	0.79 (0.04)	448;335		
BHS §	All	4350	787	1793;2557	50.1 (17.0)	3.02 (0.97)	3.18 (0.82)	3.89 (1.16)	0.77 (0.08)	2459;1891		TaqMan
	Cases	200	92	132;68	66.9 (11.6)	1.60 (0.60)	2.85 (0.66)	2.73 (0.91)	0.58 (0.09)	67;133	19.5	
	Controls	2307	386	944;1363	57.9 (12.3)	2.87 (0.83)	2.93 (0.73)	3.66 (1.05)	0.78 (0.05)	1387;920		
BRHS	All	3877	3415	3877;0	68.7 (5.5)	2.57 (0.69)	3.03 (0.4)	3.37 (0.84)	0.77 (0.12)	1125;2752		KaSPar ‡

Study name	Subset	N associations with COPD	N risk scores	Male (N); female (N)	Age mean (SD)(years)	FEV_1: mean(SD) (litres)	FEV_1 predicted: mean(SD) (litres)	FVC: mean(SD) (litres)	FEV_1/FVC: mean(SD)	Never-smokers (N); Ever-smokers (N)	% of COPD cases in GOLD stage 3-4	Genotyping
	Cases	641	572	641;0	69.7 (5.4)	1.76 (0.51)	3 (0.4)	3.01 (0.8)	0.59 (0.09)	111;530	28.39	
	Controls	2168	1905	2168;0	68.3 (5.5)	2.96 (0.48)	3.03 (0.4)	3.65 (0.65)	0.82 (0.07)	760;1408		
BWHHS	All	3644	3319	0;3644	68.8 (5.5)	1.98 (0.52)	2.16 (0.31)	2.82 (0.76)	0.71 (0.09)	2060;1584		
	Cases	659	600	0;659	69.8 (5.4)	1.36 (0.35)	2.14 (0.3)	2.32 (0.54)	0.59 (0.08)	253;406	15.63	KaSPar ‡
	Controls	1808	1653	0;1808	68.2 (5.4)	2.23 (0.41)	2.18 (0.3)	2.93 (0.56)	0.76 (0.05)	1153;655		
Gedling	All	1263	1188	632;631	56.2 (12.3)	2.85 (0.85)	3.07 (0.69)	3.68 (1.02)	0.77 (0.07)	633;630		
	Cases	103	98	67;36	66.2 (9.1)	1.73 (0.61)	2.88 (0.66)	2.82 (0.83)	0.61 (0.09)	21;82	24.27	KaSPar ‡
	Controls	840	789	417;423	57.3 (9.8)	3 (0.73)	3.03 (0.65)	3.80 (0.9)	0.79 (0.04)	431;409		
HCS	All	2850	2343	1511;1339	66.1 (2.8)	2.44 (0.68)	2.80 (0.55)	3.42 (0.92)	0.72 (0.09)	1319;1531		KaSPar ‡
	Cases	536	441	308;228	66.3 (2.8)	1.84 (0.55)	2.87 (0.56)	2.09 (0.85)	0.60 (0.09)	159; 377	15.1	
	Controls	1519	1264	758;761	66.0 (2.9)	2.67 (0.60)	2.75 (0.54)	3.51 (0.82)	0.76 (0.04)	837; 682		
Health 2000	All	888	882	427;456	50.2 (11.0)	3.32 (0.91)	3.14 (0.67)	4.19 (1.08)	0.79 (0.07)	266; 617		Illumina 610K
	Cases	32	32	20;12	60.91 (8.83)	1.78 (0.68)	3.05 (0.66)	3.05 (0.95)	0.58 (0.10)	5;27	37.5	
	Controls	580	580	256;324	53.19 (8.31)	3.28 (0.77)	3.15 (0.67)	4.09 (0.97)	0.80 (0.05)	192;388		
Nottingham Smokers	All	509	466	280;229	59.5 (10.4)	2.00 (0.95)	2.98 (0.61)	3.02 (1.06)	0.64 (0.16)	0;509		
	Cases	242	227	145;97	63.2 (9.5)	1.28 (0.57)	2.87 (0.59)	2.5 (0.87)	0.51 (0.12)	0;242	64.46	KaSPar ‡
	Controls	153	138	70;83	54.8 (8.9)	2.89 (0.61)	3.08 (0.62)	3.69 (0.81)	0.79 (0.05)	0;153		
NSHD	All	2404	2201	1206;1198	53 (0)	2.80 (0.70)	3.20 (0.56)	3.51 (0.89)	0.80 (0.09)	1003;1401		
	Cases	166	149	102;64	53 (0)	2.11 (0.58)	3.35 (0.54)	3.46 (0.89)	0.61 (0.08)	49;117	15.06	KaSPar ‡
	Controls	1663	1526	848;815	53 (0)	3.03 (0.62)	3.20 (0.56)	3.69 (0.81)	0.83 (0.06)	765;898		
Total	All	31422	15883									
	Cases	3284	2252									
	Controls	17538	8952									

‡ KaSPar genotyping (KBiosciences, Hoddesdon, Herts, UK, http://www.kbioscience.co.uk/). § BHS had genotype data for *HHIP* only in a subset of individuals (N = 1168, 131 COPD cases and 565 controls); this is therefore the subset included in the lung function and COPD risk scores calculations.

The spirometry methods used to measure FEV_1 and FEV_1/FVC in each study are detailed in (23). Individuals were defined as COPD cases if they had percent predicted FEV_1 < 80% and FEV_1/FVC < 0.7 (i.e. individuals in stages 2, 3 or 4 of the Global Initiative for Chronic Obstructive Lung Disease [GOLD] (104)) (**Figure 2-2**). Individuals were classified as controls if they had percent predicted FEV_1 > 80% and FEV_1/FVC > 0.7 (**Figure 2-2**). In order to minimize potential misclassification of COPD cases and controls, individuals with percent predicted FEV_1 > 80% and FEV_1/FVC < 0.7 (GOLD stage 1) or with percent predicted FEV_1 < 80% and FEV_1/FVC > 0.7 were excluded from the analysis (**Figure 2-2**). The calculation of percent predicted FEV_1 was undertaken using reference values of FEV_1 that take into account age, sex and height according to previously described equations (72, 73).

Figure 2-2 Selection of COPD cases and controls



Genotyping was undertaken for the sentinel SNP at each of the five loci: *TNS1* (rs2571445), *GSTCD* (rs10516526), *HTR4* (rs3995090), *AGER* (rs2070600) and *THSD4* (rs12899618). KORA F4 failed to genotype *HTR4* (**Table 2-1**). Standard quality control approaches, such as ensuring Hardy Weinberg equilibrium and an adequate call rate, were taken by the studies for all the sentinel SNPs. Data for *HHIP* (rs12504628) were also available in a subset of studies: those that used KASPar genotyping (KBioscience, Hoddesdon, Herts, UK) (**Table 2-2**), EPIC, Health 2000 and a subset of BHS that had *in silico* data (**Table 2-1**).

Statistical analysis

An analysis plan was designed centrally to ensure that each study undertook the same analysis. The analysis plan is provided in **Appendix B**.

Study level analyses

A genetic additive effect was assumed (coding each genotype as 0, 1 or 2 according to the count of the coded allele) and logistic regression was fitted by each study to test the effect of each SNP on COPD risk. Adjustment for additional covariates was not applied, since percent predicted FEV_1 was used in the definition of COPD cases and controls, and this measure takes into account age, sex and height (108).

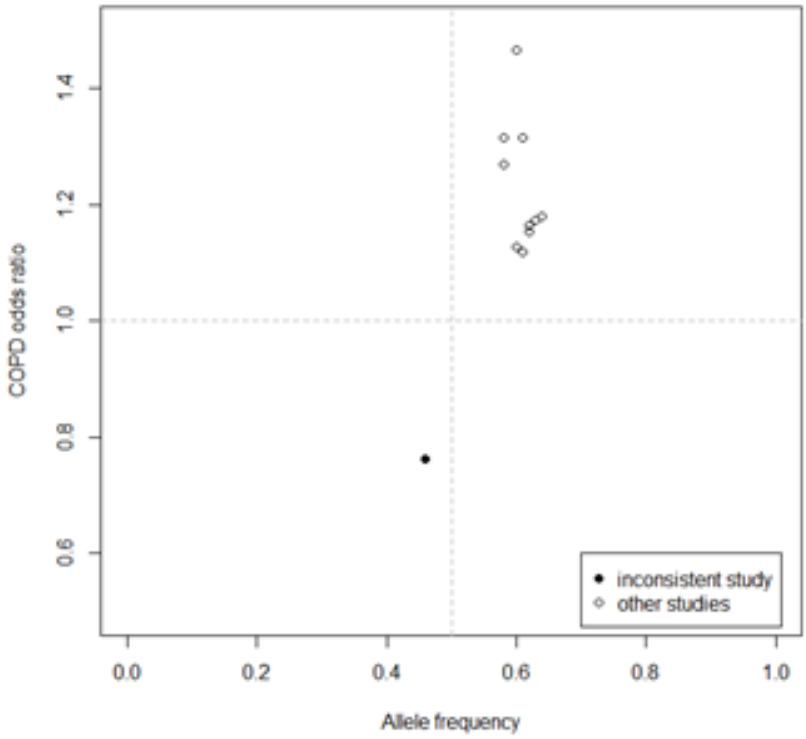
Consortium central analyses

First, the study level results were checked for errors. These checks consisted mainly of identifying any unusual pattern of findings which could indicate inconsistencies between studies and following them up to understand the source; for instance, checking consistency of direction of effects and of allele frequencies across studies. When the quality of the results was ensured, and all the effect estimates were orientated to the forward strand of the National Center for Biotechnology (NCBI) build 36 reference sequence of the human genome using the risk allele (the allele associated with reduced FEV_1 or FEV_1/FVC in the results of the SpiroMeta meta-analysis (23)) as the coded allele, the effect estimates and standard errors were meta-analysed across studies using inverse variance weighting. A Bonferroni correction for 5 tests was used for *TNS1*, *GSTCD*, *HTR4*, *AGER* and *THSD4*, defining statistical significance as P-value < 0.01.

2.3.2 **Results**

The quality control checks undertaken uncovered that one study had reported the wrong coded allele for the SNP rs12504628. **Figure 2-3** shows that the direction of effect for one study was opposite to the direction of effect for all the other studies and that the allele frequency reported for the coded allele for that study was below 0.5, whereas for all the others was above 0.5. This suggested that the coded allele had been wrongly reported. After contacting the analyst for this study the error was corrected.

Figure 2-3 Allele frequencies against odds ratios for rs12504628 in an early stage of the quality control checks



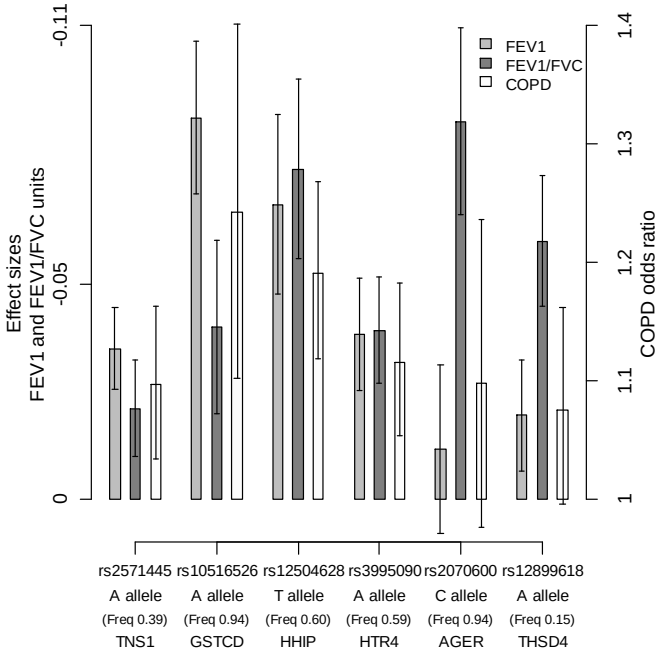
Out of the 31,422 individuals included in this analysis, 3,284 were classified as COPD cases (percent predicted FEV_1 < 80% and FEV_1/FVC < 0.7) and 17,538 were classified as controls (percent predicted FEV_1 > 80% and FEV_1/FVC > 0.7) (**Table 2-2**). Variants at three out of the five new loci associated with lung function (*TNS1*, *GSTCD* and *HTR4*) showed significant associations (P-value < 0.01) with COPD risk (**Table 2-3**). These loci showed consistent direction of effect with the effect estimates reported for lung function (23) (**Figure 2-4**). For the other two loci (*AGER* and *THSD4*), the magnitude and direction of effects were also consistent with the direction of effect estimates reported for lung function (23) (**Figure 2-4**), but they did not reach statistical significance (**Table 2-3**). The previously reported

association of the locus 4q31 near *HHIP* with COPD (94, 95) was confirmed in a subset of 2,890 COPD cases and 13,862 controls (**Table 2-3**).

Table 2-3 COPD results
Abbreviations: OR = odds ratio, CI =confidence intervals, P =P-value.

SNP ID (gene)	Coded allele	OR	95% CI	P	N cases	N controls
rs2571445 (*TNS1)*	A	1.10	1.03-1.16	$1.89x10^{-3}$	3,284	17,538
rs10516526 (*GSTCD)*	A	1.24	1.10-1.40	$3.75x10^{-4}$	3,284	17,538
rs12504628 (*HHIP*)	T	1.19	1.12-1.27	$4.55x10^{-8}$	2,890	13,862
rs3995090 (*HTR4)*	A	1.12	1.05-1.18	$1.79x10^{-4}$	3,225	16,429
rs2070600 (*AGER)*	C	1.10	0.98-1.24	$1.2x10^{-1}$	3,284	17,538
rs12899618 (*THSD4)*	A	1.08	1.00-1.16	$6x10^{-2}$	3,284	17,538

Figure 2-4 Association of six lung function loci with COPD, FEV_1 and FEV_1/FVC
FEV_1 and FEV_1/FVC associations were extracted from the combined discovery and follow-up data reported by Repapi *et al* (23), where the lung function measures were inverse normally transformed. The bars indicate the point estimates of the effect sizes and the whiskers the 95% confidence intervals. Abbreviations: freq. = frequency.



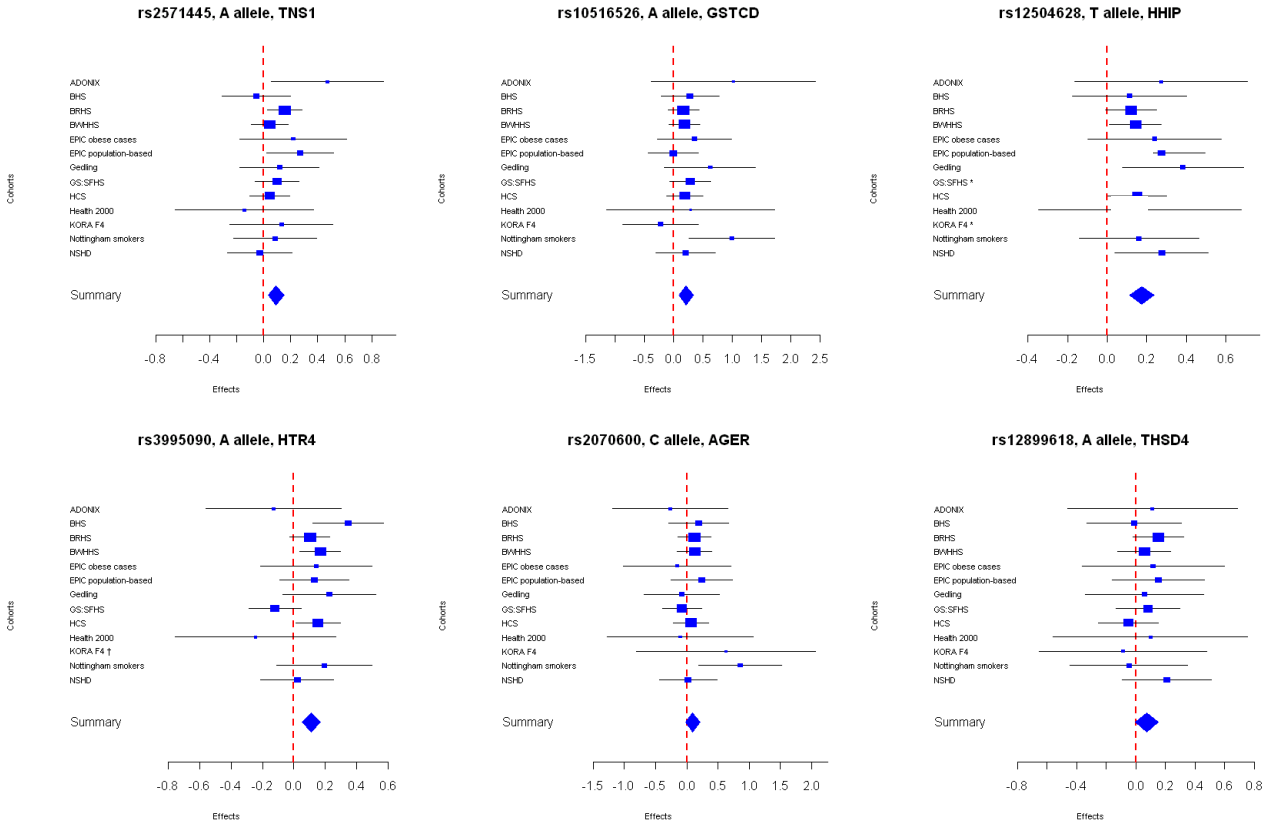
Heterogeneity of effect sizes across studies was tested for each SNP with a chi-square heterogeneity test and was not statistically significant for any of the six SNPs (P > 0.1) (**Table 2-4**, **Figure 2-5**). This indicates that the results illustrate a common trend across studies and they are not driven by just one or two studies.

Table 2-4 Heterogeneity test
Abbreviations: P = P-value, d.f. = degrees of freedom.

SNP ID (gene)	Chi-square statistic	P	d.f.
rs2571445 (*TNS1)*	10.541	$5.69x10^{-1}$	12
rs10516526 (*GSTCD)*	10.190	$5.99x10^{-1}$	12
rs3995090 (*HTR4)*	17.165	$1.03x10^{-1}$	11
rs2070600 (*AGER)*	8.729	$7.26x10^{-1}$	12
rs12899618 (*THSD4)*	4.242	$9.79x10^{-1}$	12
rs12504628 (*HHIP*)	4.799	$9.04x10^{-1}$	10

Figure 2-5 Forest plots of the meta-analysis of association tests with COPD for the 6 loci

The logarithm of the odds ratios are presented for each study and for the meta-analysis results ("summary" in the plots).



Sensitivity analyses

Smoking behaviour

Smoking is a major risk factor for developing COPD and as expected a much higher proportion of individuals are ever-smokers among the cases (72%) than among the controls (49%) (**Table 2-2**). This might lead us to think that the associations found with COPD are mediated via smoking behaviour. In order to explore this hypothesis, several analyses were undertaken. First, the association of the six genetic variants with two smoking-related traits was assessed in the Oxford-GlaxoSmithKline (Ox-GSK) consortium dataset (109). None of the SNPs was significantly associated ($P > 0.1$) with either ever smoking status (18,598 ever-smokers vs. 15,041 never-smokers) or number of cigarettes smoked per day (15,574 individuals) (**Table 2-5**).

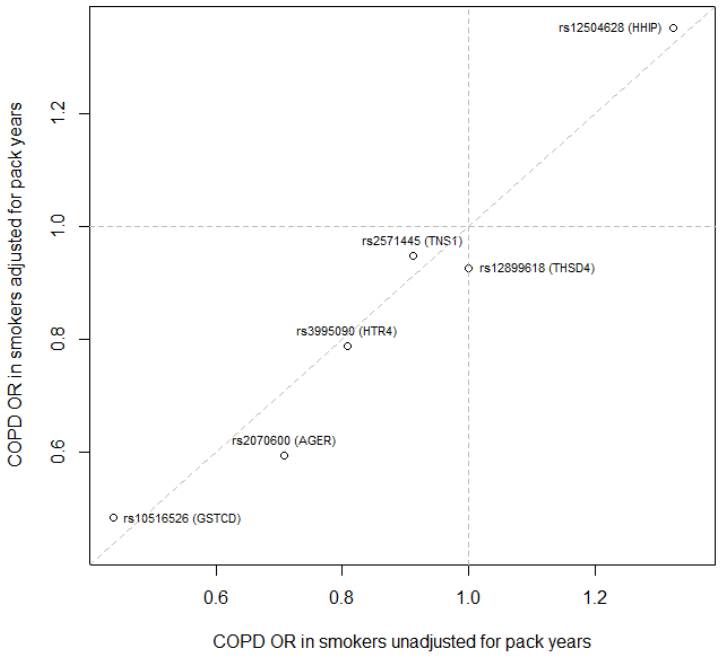
Table 2-5 Six loci associated with lung function looked up for smoking related traits

Abbreviations: Chr. = chromosome, P = P-value, SE = standard error.

Chr.	SNP ID (NCBI36 position), function	Coded allele	Lung measure	Cigarettes per day			Ever vs. never-smokers		
				Beta	SE	P	Beta	SE	P
2	rs2571445 (218391399), *TNS1* (ns)	G	FEV_1	-0.014	0.012	$2.4x10^{-1}$	-0.011	0.019	$5.55x10^{-1}$
4	rs10516526 (106908353), *GSTCD* (intron)	G	FEV_1	0.026	0.022	$2.35x10^{-1}$	-0.001	0.034	$9.79x10^{-1}$
4	rs1032296 (145654138), *HHIP* (upstream)	T	FEV_1	0.007	0.011	$5.61x10^{-1}$	-0.014	0.018	$4.43x10^{-1}$
4	rs11100860 (145698589), *HHIP* (upstream)	G	FEV_1/FVC	0.018	0.011	$1.03x10^{-1}$	0.006	0.017	$7.32x10^{-1}$
5	rs3995090 (147826008), *HTR4* (intron)	C	FEV_1/FVC	-0.007	0.011	$5.4x10^{-1}$	0.001	0.018	$9.7x10^{-1}$
6	rs2070600 (32259421), *AGER* (ns)	T	FEV_1/FVC	0.043	0.028	$1.23x10^{-1}$	-0.01	0.043	$8.13x10^{-1}$
15	rs12899618 (69432174), *THSD4* (intron)	G	FEV_1/FVC	-0.019	0.015	$2.4x10^{-1}$	-0.03	0.024	$2.07x10^{-1}$

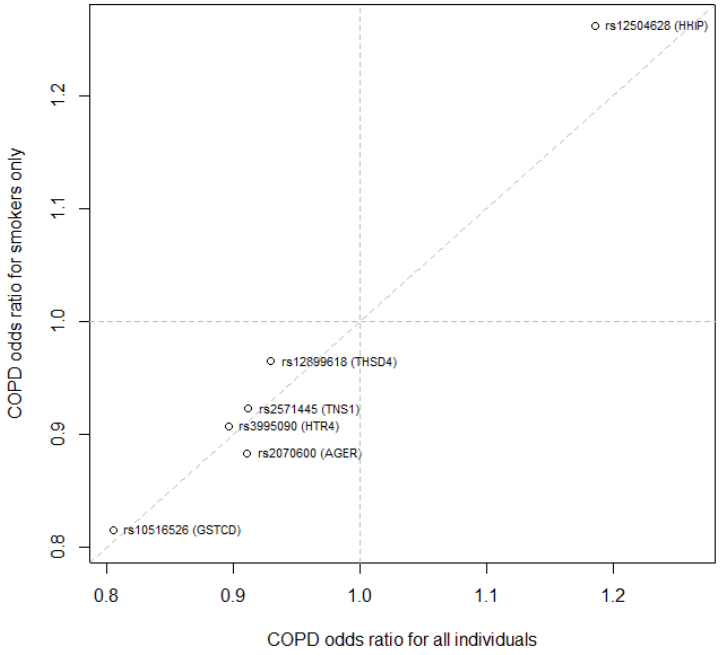
Secondly, the effect of a pack-years adjustment in two of the studies (Gedling and Nottingham Smokers) with data on pack-years available was assessed, and the effect sizes in ever-smokers with and without pack-years adjustment were compared and there was no substantial difference (**Figure 2-6**).

Figure 2-6 SNPs associations unadjusted against associations adjusted for pack-years
Data from Gedling and Nottingham Smokers studies only. Effect is for the alphabetically higher allele on the forward strand.



Finally, the effect sizes of the six genetic variants in all individuals and in ever-smokers only were compared, and again no substantial difference was found (**Figure 2-7**). These results suggest that the effect of each of these genetic variants on COPD risk is independent of smoking. However, further insights could be gained from analyses in populations with more detailed smoking phenotypes.

Figure 2-7 SNP associations in all individuals against associations in ever-smokers only
Data from all the studies was used for all the SNPs except for rs3995090 (*HTR4*) that was not available in KORA F4 and for rs12504628 (*HHIP*) that was not available for ever-smokers in Generation Scotland, KORA F4, BHS, NSHD and Health 2000 (**Table 2-1**). Effects are shown for the effect alphabetically higher allele on the forward strand.



Misclassification of cases and controls

The GOLD guidelines (104) recommend the use of post-bronchodilator FEV_1 /FVC (measure taken after inhaling a short-acting bronchodilator) in the diagnosis of COPD to minimize variability. However, post-bronchodilator FEV_1 /FVC was not available in most of the studies included in this analysis and for that reason the selection of COPD cases and controls was based on pre-bronchodilator FEV_1/FVC. Individuals recruited in the Nottingham Smokers study had both pre

and post bronchodilator spirometry measures taken, making possible a comparison of the classification of COPD cases and controls using both criteria. As previously shown (110), the comparison undertaken here showed that if the definition of COPD cases had included individuals with mild COPD (GOLD stage 1) the number of individuals misclassified would have been substantial (**Table 2-6**), however when individuals with mild COPD (GOLD stage 1) were excluded, the number of individuals misclassified was minimal (**Table 2-7**). This illustrates the relevance of the exclusion criteria used in this study excluding individuals with GOLD stage 1.

Table 2-6 Effect of misclassification using pre-bronchodilator spirometry had GOLD stage 1 individuals been included in our COPD case definition

The table illustrates the number of individuals that would be classified as COPD cases and controls had cases been defined as GOLD stage 1-4. Data from Nottingham Smokers of individuals with no missing values for pre and post-bronchodilator FEV_1 were used for this table.

		Post		
		COPD cases	controls	Total
Pre	COPD cases	227	27	254
	controls	16	182	198
	Total	243	209	452

Number of COPD cases defined on pre-bronchodilator reclassified as controls on post-bronchodilator = 27

Positive predictive value = 89%

Number of controls defined on pre-bronchodilator reclassified as COPD cases on post-bronchodilator = 16

Negative predictive value = 92%

Table 2-7 Number of COPD cases and controls defined using pre and post-bronchodilator FEV_1

Cases were defined as GOLD stage 2-4. Data from Nottingham Smokers of individuals with no missing values for pre and post-bronchodilator FEV_1 were used for this table.

		Post		
		COPD cases	controls	Total
Pre	COPD cases	201	4	205
	controls	2	120	122
	Total	203	124	327

Number of COPD cases defined on pre-bronchodilator reclassified as controls on post-bronchodilator = 4

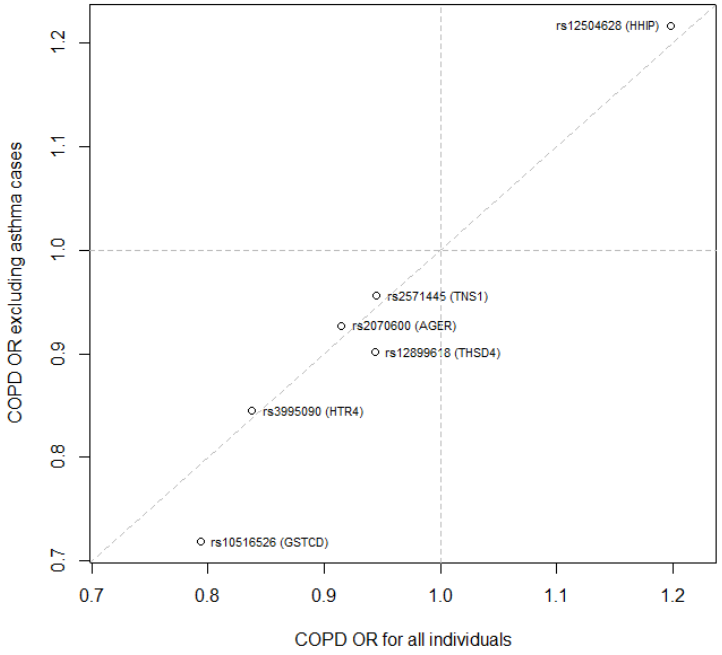
Positive predictive value = 98%

Number of controls defined on pre-bronchodilator reclassified as COPD cases on post-bronchodilator = 2

Negative predictive value = 98%

COPD patients are characterized by airway obstruction that is not fully reversible, whereas asthma patients often have fully reversible airway obstruction. For this reason, another consequence of using pre-bronchodilator spirometry in the definition of COPD cases might be the misclassification of some asthma patients as COPD cases. This misclassification would have the potential to overestimate genetic effects on the risk of COPD if the SNPs analysed had an effect on asthma. In order to investigate the impact of a possible misclassification of asthma cases on the results, the effect sizes obtained for all the individuals and the effect sizes obtained excluding diagnosed asthma subjects from the cases were compared for all the SNPs in a subset of the data (BWHHS and Gedling) with asthma diagnosis available to us. For rs10516526, the odds ratio estimated in individuals without asthma seemed to indicate a greater risk of developing COPD for individuals with the risk allele than the odds ratio estimated for all individuals. This would be inconsistent with an effect of the SNP on asthma that could have overinflated the COPD odds ratio, and it would be consistent with the asthma cases adding noise to the data instead for this SNP. Overall this comparison showed that the results from both analyses were consistent (**Figure 2-8**), and the potential misclassification of asthma cases did not seem to have a substantial impact on the results.

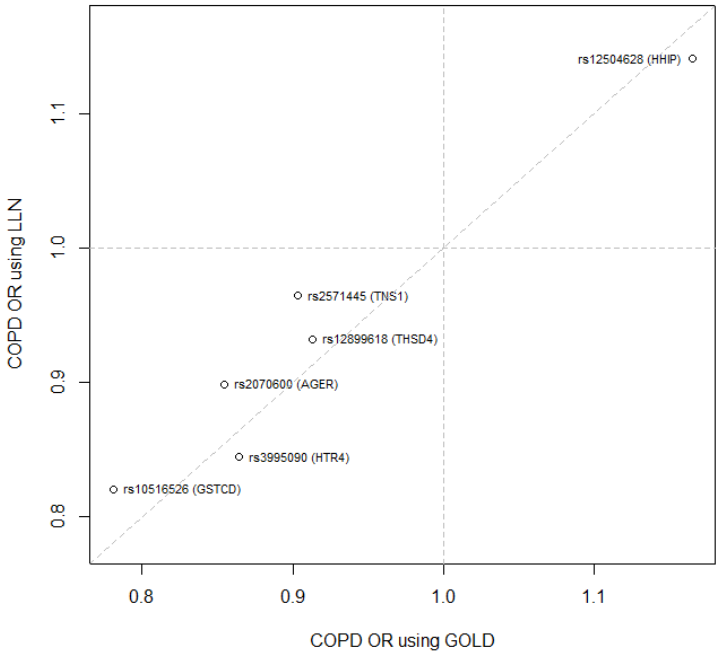
Figure 2-8 SNP associations for all individuals against associations excluding patients with known asthma from the cases
Data from BWHHS and Gedling studies only. Effect is for the alphabetically higher allele on the forward strand.



The Global Initiative for Chronic Obstructive Lung Disease (104) suggests the use of a fixed FEV_1/FVC ratio (< 0.7) to define airflow obstruction and then classify severity according to FEV_1 predicted into mild, moderate, severe and very severe. The use of a fixed FEV_1/FVC ratio in this definition will tend to increase the prevalence of COPD in the elderly, while reducing it in adults younger than 45 years old, especially the diagnosis of mild disease (104). Another way to define airway obstruction that overcomes these issues is to use a cutoff based on the lower limit of normal (LLN) values for FEV_1/FVC (108). This definition classifies the bottom 5% of a healthy population distributed normally as cases. However, this

method is highly dependent on the choice of reference equations (104). Although in this study only individuals aged over 40 years and with at least a moderate stage of COPD were included in the analysis, to assess the consistency of the results, the analysis was repeated using the LLN definition of COPD cases. This analysis was performed in a subset of the studies (BRHS, BWHHS, Gedling and Nottingham Smokers) that provided individual level data to test the associations with COPD, and it showed that effect size estimates for both analyses were of similar magnitudes (**Figure 2-9**).

Figure 2-9 SNP associations using GOLD against associations using LLN
Data from BRHS, BWHHS, Gedling and Nottingham Smokers,studies only. Effect is for the alphabetically higher allele on the forward strand.



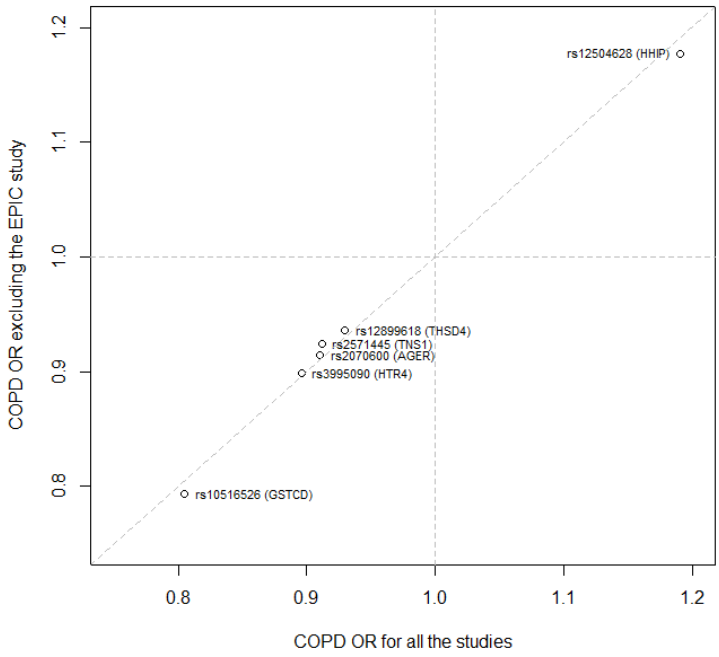
Misclassification of cases and controls due to random spirometry measurement error is also possible, although it would be minimized by the exclusion of mild

COPD cases. This misclassification would lead to an underestimation of SNPs effect on COPD risk.

Exclusion of discovery samples

Effect sizes estimated in discovery samples might overestimate the real effect of a genetic variant due to winner's curse bias (27). For that reason, most of the studies included in this analysis belong to the follow-up stage of the study that discovered the five genetic variants analysed here (23). However, in order to increase the power to detect the small effect of genetic variants on the binary COPD status, the EPIC study, part of the discovery stage, was also included in this analysis. To assess the effect of this study on the results, the effect sizes with and without the EPIC study were compared, and their magnitudes did not differ substantially (**Figure 2-10**).

Figure 2-10 SNPs associations excluding EPIC studies compared with all the studies
Effect is for the alphabetically higher allele on the forward strand.



2.3.3 Discussion

Three of the five new loci associated with lung function in the SpiroMeta dataset showed a significant association with COPD risk in this study, and the remaining two variants showed consistent magnitude and direction of effect but their associations were not statistically significant. An explanation for this might be the lack of power of this study to detect the small effect of these loci on a binary outcome. Post-hoc power calculations showed that this study was well powered to detect association with COPD for *TNS1*, *GSTCD* and *HTR4*, however it was underpowered to detect potential associations for *AGER* and *THSD4* (**Table 2-8**). Post-hoc calculations were undertaken just to illustrate a

point, since they do not really add new information to the results. Variants with less significant P-values will always correspond to decreased power to detect them, as discussed by J.M. Hoenig and D.M. Heisey (111).

Table 2-8 Post-hoc power calculations
Abbreviations: N = number, OR = odds ratio.

N cases	N controls	SNP ID (gene)	OR	Coded allele	Allele frequency	Alpha	Power
3,284	17,538	rs2571445 (*TNS1*)	1.1	A	0.39	0.01	0.82
3,284	17,538	rs10516526 (*GSTCD*)	1.24	A	0.94	0.01	0.84
3,225	16,429	rs3995090 (*HTR4*)	1.12	A	0.59	0.01	0.93
3,284	17,538	rs2070600 (*AGER*)	1.1	C	0.94	0.01	0.17
3,284	17,538	rs12899618 (*THSD4*)	1.08	A	0.15	0.01	0.31

The association of *HHIP* with COPD was confirmed in this study. However the effect size estimate for the *HHIP* locus presented here (OR = 1.19, 95% CI 1.12-1.27) is modest compared with some of the estimates obtained in previous studies, such as the Bergen case-control and the US National Emphysema Treatment Trial (NETT)/Normative Aging Study (NAS) (95), with estimated odds ratios of around 1.4. Other studies, such as the Rotterdam (112) and Framingham (94) studies estimated odds ratios of 1.25 and 1.1, more in line with the estimates obtained here. The variability of the magnitude of odds ratio estimates might be caused by differences in the characteristics of the study populations, such as age distribution. Yet another factor that can contribute to this variation is an upwards bias of the effect sizes estimated in the discovery samples known as winner's curse bias (27). Odds ratios estimated in large populations independent of discovery, such as SprioMeta, are needed in order to obtain more accurate estimates of real effect sizes.

2.4 Lung function and COPD risk scores

In order to assess the combined effect of the genetic variants associated with lung function in the SpiroMeta dataset (23) (in or near *TNS1*, *GSTCD*, *HHIP*, *HTR4*, *AGER* and *THSD4*) risk scores were constructed in studies that took part in the follow-up stage of the SpiroMeta meta-analysis. Effects of these risk scores on lung function and COPD risk were assessed in each study separately and then their results were pooled together in a meta-analysis.

2.4.1 Methods

Populations, phenotyping and genotyping

The populations included in this analysis are a subset of the populations that took part in the COPD analysis (section 2.3). Only those studies that (i) had genotype data available in all six SNPs and (ii) did not take part in the discovery stage of the SpiroMeta meta-analysis were included. Details of these studies are given in **Figure 2-1** and **Table 2-2**. Phenotyping and genotyping information for these populations are already included in section 2.3.

Statistical analyses

Centrally, an analysis plan was designed to ensure that all the studies undertook the exact same analysis. This analysis plan gave instructions on how to construct risk scores and on how to assess their association with FEV_1, FEV_1/FVC and COPD risk. The analysis plan is provided in **Appendix B**. Only

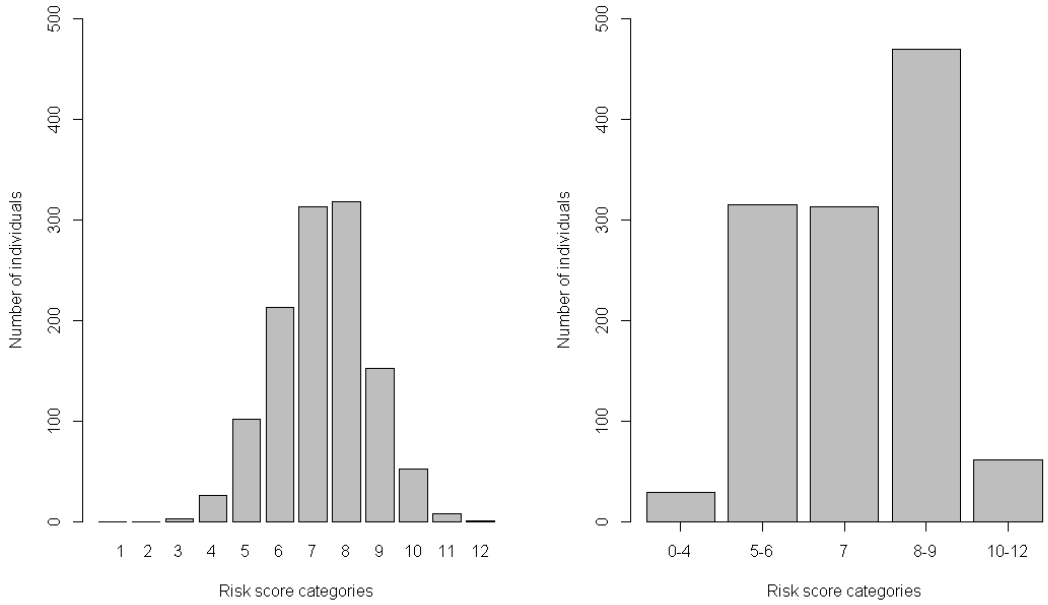
individuals with (i) complete data for all of the six SNPs and (ii) with complete data for both FEV_1 and FEV_1/FVC were included in the analysis.

Study level analyses: construction of unweighted risk scores

To create unweighted risk scores, the number of risk alleles per individual within the six loci associated with lung function in the SpiroMeta dataset (23) were summed, with risk allele as the allele associated with reduced FEV_1 or reduced FEV_1/FVC in the results of the SpiroMeta meta-analysis (23). Thus, the risk score for an individual could range from 0 risk alleles, if the individual was homozygous for the non risk allele at all six loci, to 12 risk alleles, if the individual was homozygous for the risk allele at all six loci.

This definition of unweighted risk score would group individuals into twelve categories (individuals having 0 to 12 risk alleles), however some of these categories included a very small number of individuals or no individuals at all (**Figure 2-11**). For this reason, individuals were grouped instead into five categories: 0-4 risk alleles, 5-6 risk alleles, 7 risk alleles, 8-9 risk alleles and 10-12 risk alleles (**Figure 2-11**). The group with 7 risk alleles was used as the baseline group for comparisons, since this was estimated to be the mean and median number of risk alleles per person, using summaries provided by the studies. Twenty-eight percent of all individuals carried 7 risk alleles.

Figure 2-11 Number of individuals per risk score category in the Gedling dataset



Study level analyses: risk score association analyses

Association of risk score categories with FEV_1, FEV_1/FVC and COPD risk were undertaken. First, the lung function measures (FEV_1 and FEV_1/FVC) were adjusted for age, age^2, sex and height using linear regression, and the residuals obtained here were then used for the subsequent association analyses. The COPD analysis was not adjusted for any covariates as the selection of COPD cases and controls was based on percent predicted FEV_1, which takes into account age, sex and height (section 2.3).

To assess the effect of the unweighted risk score on the lung function measures, indicator variables were created for the four non baseline risk categories, and they were added as covariates to linear regressions with residuals for FEV_1 and FEV_1/FVC as the outcome variables. COPD cases and

controls were defined as in section 2.3. Individuals over 40 years old with $FEV_1/FVC < 0.7$ and percent predicted $FEV_1 < 80\%$ were classified as COPD cases; individuals over 40 years old with $FEV_1/FVC > 0.7$ and percent predicted $FEV_1 > 80\%$ were classified as controls and; individuals that did not fall in either category were excluded (**Figure 2-2**). The effect of the unweighted risk score on COPD risk was assessed using logistic regression with indicator variables for the four non baseline risk categories as covariates and COPD status as the outcome variable.

Consortium central analyses

Quality control checks were carried out centrally on the study level results to ensure that no errors were included in the analysis. Checks included ensure that range of measurements given were biologically plausible, that units reported by all the studies were consistent, and that any inconsistency across study results could be explained. Once the quality of the results was ensured, effect estimates and standard errors were pooled together using an inverse variance weighted meta-analysis.

Sensitivity analyses

Construction of weighted risk scores (only for FEV_1 and FEV_1/FVC)

In order to assess the effect on the results of assigning weights to each variant according to the magnitude of their effects on lung function, weighted risk scores were calculated as follows. The number of risk alleles for each individual in each particular locus was multiplied by the weight of that locus, and the risk

score for that individual was obtained by adding up these products across the six loci. Weights for each locus were obtained separately for FEV_1 and FEV_1/FVC, using untransformed effect sizes estimated by linear regressions of FEV_1 and FEV_1/FVC assuming an additive effect for each locus and adjusting for age, age^2, sex and height in the SpiroMeta discovery dataset, results shown in **Table 2-9**, derived from Repapi *et al.* (23).

Table 2-9 Effect sizes and weights for FEV_1 and FEV_1/FVC used to obtain the weighted risk score
Weights for FEV_1 and FEV_1/FVC were obtained using untransformed effect sizes estimated by linear regressions of FEV_1 and FEV_1/FVC, assuming an additive effect for each locus and adjusting for age, age^2, sex and height in the SpiroMeta discovery dataset (23).

SNP ID (gene)	FEV_1		FEV_1/FVC	
	Beta (ml)	Weights	Beta (%)	Weights
rs2571445 (*TNS1*)	23.088	1.014	0.163	0.345
rs10516526 (*GSTCD*)	52.475	2.304	0.325	0.687
rs12504628 (*HHIP*)	26.233	1.152	0.553	1.167
rs3995090 (*HTR4*)	18.783	0.825	0.347	0.733
rs2070600 (*AGER*)	7.392	0.325	0.817	1.724
rs12899618 (*THSD4*)	8.654	0.380	0.636	1.344

Weighted risk scores were constructed so they add up to 12 for individuals who were homozygous for the risk allele in all the six loci. Again, due to the small numbers in some of the categories, individuals were grouped into five categories instead of 12: risk score < 5, 5 <= risk score < 7, 7 <= risk score < 8, 8 <= risk score < 10 and 10 <= risk score < 12; and the middle group (7 <= risk score < 8) was used as baseline for comparisons.

Weighted risk scores were not calculated for COPD because there were no data on previously estimated odds ratios for all the six loci that could be used as weights.

COPD per allele risk score approximations

In order to compare the effect of a COPD risk score in all individuals and in ever-smokers only, an approximated risk score was calculated from the single SNP results in a subset of the studies included in section 2.3 that had ever/never smoking data available (Adonix, BWHHS, Gedling, Nottingham Smokers, EPIC, HCS and BRHS). A method developed by Toby Johnson and colleagues (113, 114) that uses summary data from single variant analyses to estimate risk scores was adapted for this purpose. This method approximates the effect of an m-SNP risk score on a trait in a testing dataset, given the effects of m SNPs on a different trait in a discovery dataset, using only betas and standard errors from single SNP association tests. A simplification of this approximation (with only one trait and one dataset) was used here to obtain the effect of a 6-SNP risk score on COPD risk given their single SNP results.

A brief description of the method is shown here.

For an m-SNPs genetic risk score,

$$s_j = s_0 + w_1 x_1 + \cdots + w_m x_m$$

$$where \;\; s_0 \; is \; a \; constant,$$

$$w_1, .., w_m \; are \; the \; effects \; of \; the \; SNPs \; on \; the \; discovery \; dataset \; and$$

$$x_1, .., x_m \; are \; the \; number \; of \; risk \; alleles \; per \; SNP$$

the following regression model for individual j,

$$y_j = y_0 + as_j + e_j \text{ (1)}$$

$$where\ y_0 is\ the\ intercept, e_j is\ the\ error\ term, s_j\ is\ the\ m - SNPs\ risk\ score$$

$$and\ a\ the\ effect\ of\ the\ risk\ score\ on\ the\ trait\ of\ interest\ (y_j)$$

can be written as

$$y_j = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + e_j \text{ (2)}$$

$$for\ \beta_i = aw_i\ with\ i \geq 1$$

$$and\ \beta_0 = y_0 + as_0$$

We were interested in calculating the maximum likelihood estimate of a (the effect of the risk score on the trait of interest). To obtain the maximum likelihood estimate of a, we needed to maximize the likelihood function of model (1) or equivalently of model (2).

For large sample sizes, the likelihood function for the parameters of a regression model is approximately Gaussian, and when the explanatory variables are uncorrelated and the fraction of the variance explained by them is small, the likelihood function for a multi-SNP regression model is well approximated by the sum of the likelihood functions for the corresponding single SNP regression models (113). Given that these assumptions hold we could use the single SNPs likelihood functions in the testing dataset (specified in terms of single SNPs effect sizes and standard errors) to approximate the likelihood function for model (2) and then find the value of a that maximizes that likelihood function.

For the simplified analysis carried out here, there is only one trait (COPD risk) and one dataset (the testing dataset). The same rationale explained above was followed but the effects of the SNPs on the discovery dataset ($w_1, .., w_m$) were assumed to be one, which would be equivalent to estimating an unweighted risk score, where all the SNPs were assumed to have the same magnitude of effect on the trait. The sample size for all individuals is 11,804 (2,492 cases and 9,312 controls) and for smokers only was 6,535 (1,842 cases and 4,693 controls); the six SNPs were not in LD with each other; and their effect sizes are small, we could therefore presume that the assumptions hold.

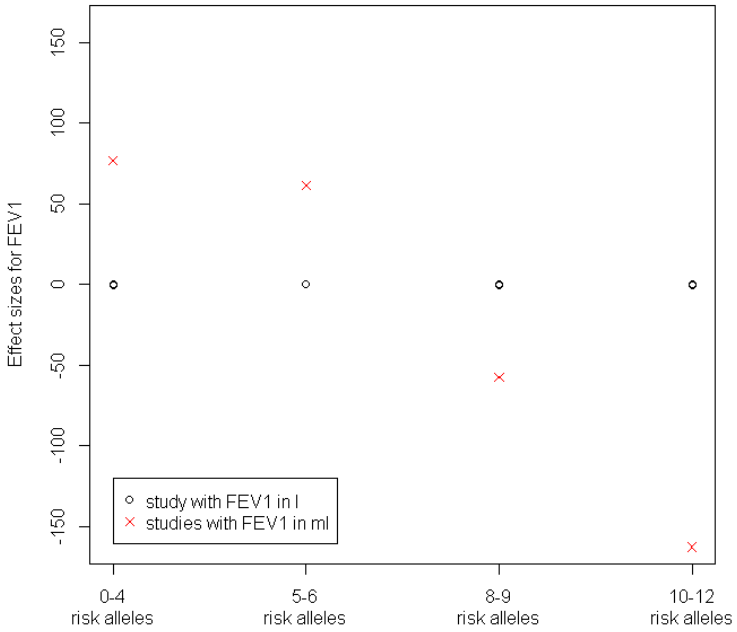
2.4.2 **Results**

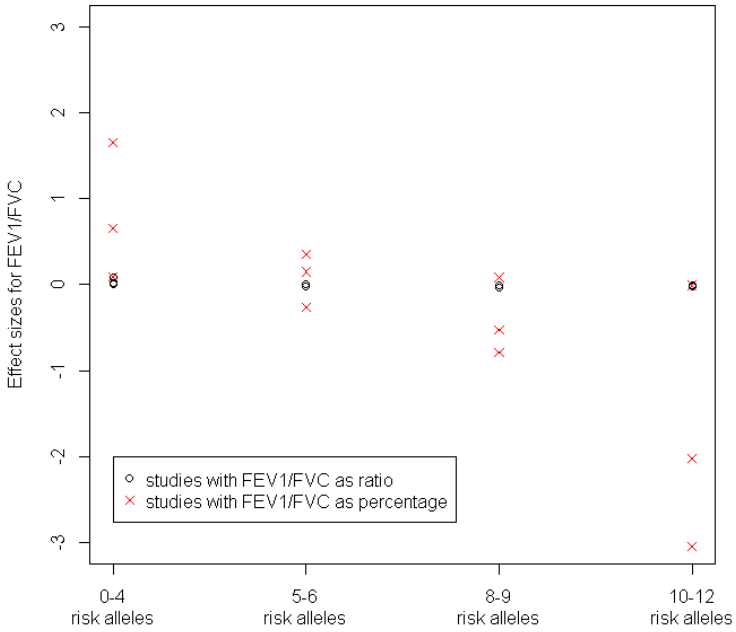
The analyses of the combined effect of risk alleles on lung function measures included a total of 15,883 individuals and on COPD included 2,252 cases and 8,952 controls. Individuals were sampled from nine population-based studies (**Table 2-2**, **Figure 2-1**) from the follow-up stage of the SpiroMeta meta-analysis (23), and had complete data on all six SNPs analysed (rs2571445 in *TNS1*, rs10516526 in *GSTCD*, rs3995090 in *HTR4*, rs2070600 in *AGER* and rs12899618 in *THSD4*).

The issues encountered in the quality control stage are described here. The inclusion in one study of individuals with outlying values for FEV_1/FVC was identified by an abnormal range of values for FEV_1/FVC (from 0.43 to 1.57). Analysts for that study were contacted and the analysis was repeated after

excluding these two outliers. Results were reported in different units to those requested, in one study for FEV_1 and in three studies for FEV_1/FVC (**Figure 2-12**). Estimates for these studies were converted to the correct units before meta-analysing the results.

Figure 2-12 Effect sizes for the unweighted genetic risk scores for FEV_1 and FEV_1/FVC in a subset of studies in an early stage of the quality control checks

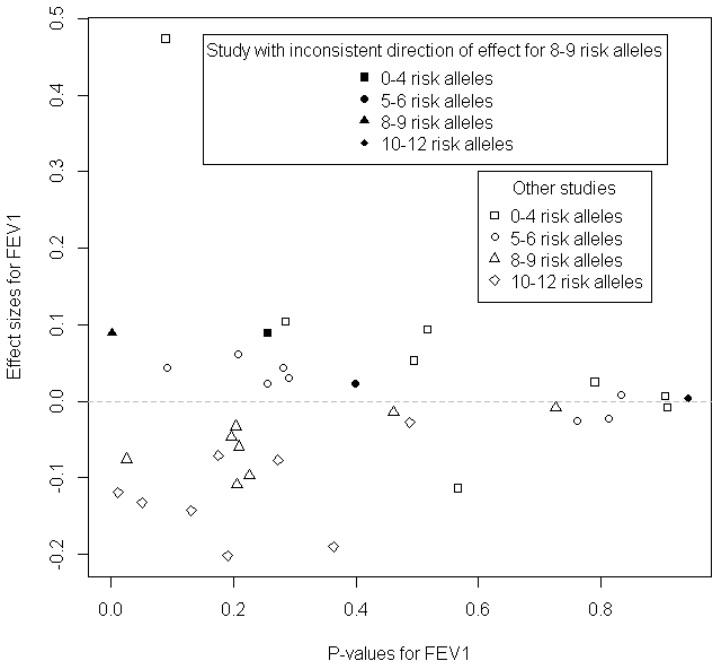




When checking for consistency of direction of effect across studies, an issue was identified for one study. **Figure 2-13** shows that the direction of effect for the 8-9 risk alleles category, which had a small P-value, is opposite to the estimates for all the other studies. After contacting the analyst for this study, they identified a problem with allele coding and some outlying values for the lung function measures, which were then corrected, and results that were consistent across studies were provided.

Figure 2-13 P-values against effect sizes for the unweighted risk scores for FEV_1 in an early stage of the quality control checks

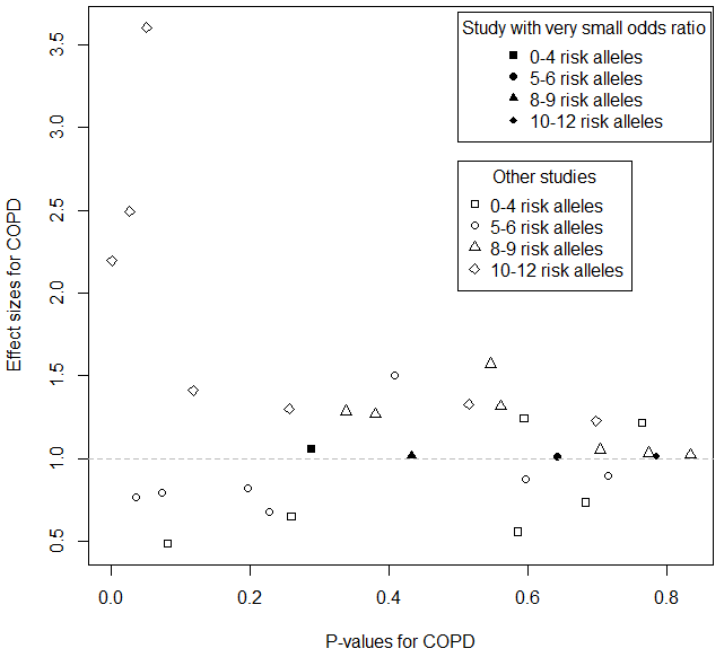
Different symbols represent different risk score categories, and the filled symbols point out one study with an inconsistent direction of effect for the 8-9 risk alleles category.



A programming error in the COPD analysis made in one study was found by comparing magnitudes of odds ratios across studies (**Figure 2-14**). The binary

disease status had been treated as a continuous trait instead in the modelling process. Corrected analyses were provided after contacting the analyst for this study.

Figure 2-14 P-values against odds ratios for the unweighted risk scores for COPD in an early stage of the quality control checks
Different symbols represent different risk score categories, and the filled symbols point out one study with very small odds ratios compared to the other studies.



After resolving all quality control issues the meta-analysis was undertaken. Associations of the unweighted risk scores with both FEV_1 and FEV_1/FVC showed a clear trend, with positive effects on lung function for groups with less risk alleles than the baseline (7 risk alleles) and negative effects on lung function for groups with more risk alleles than the baseline (**Table 2-10**). Having 10 to 12 risk alleles in comparison to having 7 risk alleles was strongly associated with a reduction in FEV_1 and FEV_1/FVC (P-values < $4x10^{-4}$), with a

magnitude of effect for FEV_1 equivalent to the physiological average ageing decline in FEV_1 over approximately four years in a non-smoking population (115).

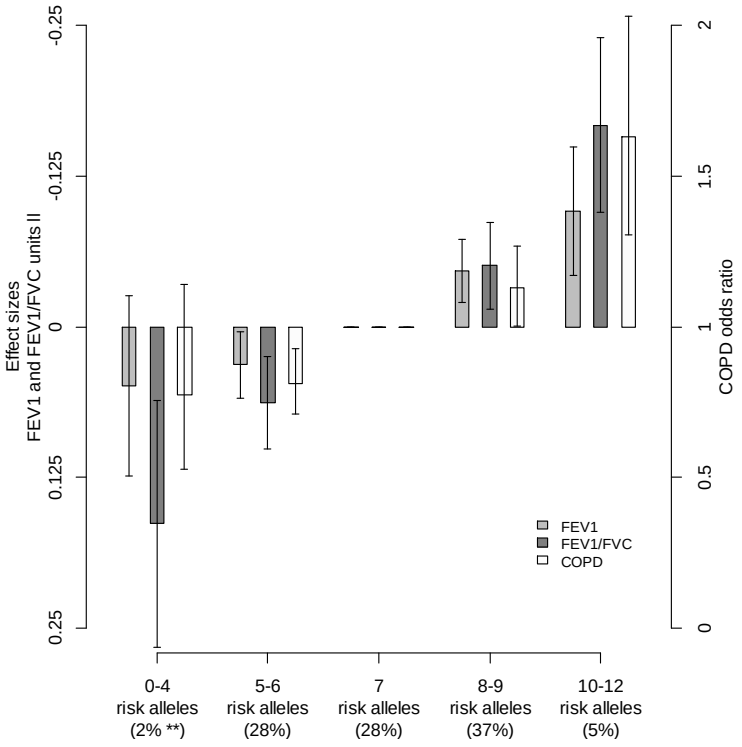
Table 2-10 Statistics of association of unweighted risk scores with lung function and COPD
Abbreviations: SE = standard error, CI = confidence interval, P = P-value.

Risk alleles	FEV_1(ml)			FEV_1/FVC (%)			COPD		
	Beta (SE)	**95% CI**	**P**	**Beta (SE)**	**95% CI**	**P**	**OR(SE)**	**95% CI**	**P**
0-4	36.722 (28.742)	(-19.612, 93.057)	$2.01x10^{-1}$	1.498 (0.478)	(0.561, 2.435)	$1.73x10^{-3}$	0.776 (0.196)	(0.528, 1.14)	$1.97x10^{-1}$
5-6	23.614 (10.616)	(2.806, 44.421)	$2.6x10^{-2}$	0.581 (0.179)	(0.229, 0.932)	$1.19x10^{-3}$	0.813 (0.067)	(0.712, 0.927)	$1.96x10^{-3}$
7	0			0			1		
8-9	-35.021 (9.94)	(-54.504, -15.539)	$4.26x10^{-4}$	-0.465 (0.168)	(-0.794, -0.136)	$5.64x10^{-3}$	1.127 (0.06)	(1.002, 1.268)	$4.55x10^{-2}$
10-12	-72.213 (20.361)	(-112.122, -32.305)	$3.90x10^{-4}$	-1.532 (0.339)	(-2.197, -0.867)	$6.35x10^{-6}$	1.628 (0.112)	(1.306, 2.03)	$1.46x10^{-5}$

The results of the COPD risk score analysis also showed a clear trend across risk categories (**Table 2-10**), consistent with the trend observed for lung function (**Figure 2-15**). Again the strongest association ($P = 1.46x10^{-5}$) was for the group with the highest risk score, with a 1.6 fold increased risk for individuals with 10-12 risk alleles (5% in our population) compared to individuals with 7 risk alleles (28% in our population).

Figure 2-15 Association of unweighted risk scores with lung function and COPD

The bars indicate the point estimates of the effect sizes and the whiskers indicate the 95% confidence intervals.



[II]To facilitate the plotting of the effect size estimates for FEV_1 and FEV_1/FVC on the same axes, effect sizes are given in terms of the proportion of a standard deviation of FEV_1 and FEV_1/FVC; A standard deviation of 754 ml for FEV_1 and 0.092 for FEV_1/FVC (obtained as weighted averages across studies) were used. [**]Proportions of individuals within each risk score category are given on the x-axis.

To assess the effect of weighting the loci according to the single effect of the sentinel SNP on lung function, weighted risk scores were constructed in a subset of studies that provided individual level data (BHS, BRHS, BWHHS, Gedling and Nottingham Smokers). The results obtained were broadly consistent with the effect of unweighted risk scores constructed for these studies, although stronger effects where shown for FEV_1 the when using weighted risk scores (**Table 2-11**).

Table 2-11 Statistics of association of unweighted and weighted risk scores with lung function

Data was available for: BHS, BRHS, BWHHS, Gedling and Nottingham Smokers. Abbreviations: ml = millilitres, SE = standard error, CI = confidence interval, P = P-value.

Risk alleles	FEV_1(ml)			FEV_1/FVC (%)		
	Beta (SE)	95% CI	P	Beta (SE)	95% CI	P
Unweighted analysis						
0-4	33.111 (38.655)	(-42.65,108.88)	$3.92x10^{-1}$	2.075 (0.693)	(0.72,3.44)	$2.77x10^{-3}$
5-6	30.040 (14.469)	(1.68,58.40)	$3.79x10^{-2}$	0.689 (0.262)	(0.18,1.20)	$8.58x10^{-3}$
7	0			0		
8-9	-29.842 (13.467)	(-56.24,-3.45)	$2.67x10^{-2}$	-0.344 (0.244)	(-0.82,0.13)	$1.58x10^{-1}$
10-12	-61.890 (27.284)	(-115.37,-8.41)	$2.33x10^{-2}$	-1.357 (0.49)	(-2.32,-0.40)	$5.58x10^{-3}$
Weighted analysis						
0-4	58.619 (36.419)	(-12.76,13)	$1.08x10^{-1}$	1.483 (0.513)	(0.48,2.49)	$3.85x10^{-3}$
5-6	35.105 (17.809)	(0.12,70.01)	$4.87x10^{-2}$	0.641 (0.249)	(0.15,1.13)	$1.01x10^{-2}$
7	0			0		
8-9	-30.434 (14.072)	(-58.02,-2.85)	$3.06x10^{-2}$	-0.473 (0.246)	(-0.95,0.01)	$5.4x10^{-2}$
10-12	-81.828 (20.089)	(-121.20,-42.45)	$4.64x10^{-5}$	-1.566 (0.635)	(-2.81,-0.32)	$1.36x10^{-2}$

Per allele risk scores estimated for COPD risk for all individuals (OR = 1.145) and for ever-smokers only (OR = 1.132) in a subset of studies with ever smoking data available (Adonix, BWHHS, Gedling, Nottingham Smokers, EPIC, HCS and BRHS) were consistent, suggesting that the effect of these loci on COPD is similar in the general population and in ever-smokers only.

2.4.3 Discussion

This study has estimated combined effects of risk alleles in six genetic variants on lung function and COPD risks and has shown significantly reduced lung

function and increased risk of COPD for individuals with 10 to 12 risk alleles in comparison to individuals carrying 7 risk alleles.

The studies that took part in this analysis were all part of the follow-up stage of the study that discovered five out of the six variants. This becomes especially relevant when estimating combined effects across variants, since the potential over estimation of effect sizes due to winner's cure bias (27) would be cumulative if discovery samples were included.

The choice of an adequate baseline for comparisons in this context is also important. As discussed by Goddard and Lewis (116), comparisons with a group of average risk are more meaningful than comparisons between groups with more extreme risks. Comparisons with a baseline of 0 risk alleles for instance would probably give a misleading impression of the real predictive value of the genetic variants. For this reason the average number of risk alleles per person across studies (7 risk alleles) was chosen as the baseline group for comparisons.

Sensitivity analyses undertaken previously (section 2.3.2) to assess the effect of smoking behaviour on the lung function and COPD associations of these six loci, support the hypothesis of an effect on lung function independent of smoking behaviour. However, to investigate this hypothesis further, approximations of risk scores were constructed for COPD both for all individuals and for ever-smokers only in a subset of studies included in section 2.3 with

ever smoking status data available. Per allele risk scores for all individuals and ever-smokers only were very similar, supporting a genetic effect on COPD risk that is not mediated via smoking behaviour.

The computation of unweighted risk scores is simpler and the results easier to interpret than if using weighted risk scores, however they do not account for the fact that not all risk alleles at the six SNPs exert the same effect on lung function. For this reason, weighted risk scores weighting risk alleles in each SNP by their estimated effect sizes were also constructed in a subset of the studies that provided their individual level data, as a sensitivity analysis. Overall, results from the weighted and unweighted analyses were consistent in this study (**Table 2-11**). However, there are greater differences between weighted and unweighted analyses for FEV_1 than for FEV_1/FVC; this might be explained by the larger variation in effect size magnitudes for FEV_1 in comparison to FEV_1/FVC (**Figure 2-4**) for the variants included in the risk scores, which is taken into account in the weighted analysis. As more comprehensive risk scores are constructed including a wider range of effect sizes, the inclusion of weights will likely become more relevant.

New studies undertaken afterwards have identified additional regions associated with lung function and COPD (details are given in the following chapters). Fine mapping of some of these regions where signals are not well localized might lead to the discovery of multiple causal variants within a locus.

Incorporating all the genetic variants known to be associated with lung function and their potential multiple causal variants will lead to improved risk scores.

2.5 Conclusion

This chapter has shown that three genetic variants (*GSTCD*, *TNS1* and *HTR4*) out the five associated with lung function in the SpiroMeta dataset are also associated with COPD, and that the remaining two variants (*AGER* and *THSD4*) have the magnitude and direction of effect expected although they do not reach statistical significance. Another study undertaken by Castaldi and colleagues (117), subsequently confirmed the association of *GSTCD* with COPD and also showed associations with COPD for *AGER* and *ADAM19*. Variants near *HHIP* and in *FAM13* had previously been reported to be associated with COPD (95, 103). This illustrates that studying the quantitative spirometry measures in order to detect common genetic variants that can ultimately shown to be associated with COPD is a promising approach.

Individuals with 10 to 12 risk alleles (5% of our population) had significantly reduced lung function and 1.6 fold increase in their risk of developing COPD in comparison to individuals with 7 risk alleles (28% of our population). Improved risk scores will be obtained after incorporating the remaining loci known to be associated with lung function, and their utility in comparison with existing risk predictions using age, sex, height, family history, etc will need to be assessed. In particular, risk scores incorporating genetic information might become

relevant for smokers since their absolute risk of developing COPD is already high (118).

2.6 Extension of this work: Royal Society Summer Science Exhibition

The effect of *TNS1*, *GSTCD*, *HHIP*, *HTR4*, *AGER* and *THSD4* on COPD risk presented in this chapter was used to develop "The Risky Gene Machine", an activity that was part of the "Breathless genes: the lung and the short of it" exhibition at the Royal Society Summer Science Exhibition in London 2012 (http://sse.royalsociety.org/2012/exhibits/breathless-genes/). This is a week long exhibition attended by over 10,000 members of the public and 2,000 school students. Using the method described in *COPD per allele risk score approximations*, in section 2.4.1, a per allele COPD risk score was approximated using the single SNP effects on COPD risk for the six loci analysed in this chapter. Baseline COPD risk for ever and never-smokers was taken from a study published in 2006 on 8,045 individuals followed up for 25 years (118). This allowed the estimation of COPD absolute risk for an individual given their smoking status and their number of risk alleles. The Risky Gene Machine looks like a "fruit machine", but instead of a random combination of fruits, a random combination of risk and non risk alleles is provided for the six loci; the user starts by being a smoker and then changes to being a non-smoker and sees the difference that the number of risk alleles and the smoking status makes on their risk of developing COPD.

Chapter 3: Analysis of common genetic variants: GWAS of lung function

This chapter describes a meta-analysis of genome-wide association studies of lung function measures. It includes a detailed explanation of the quality control checks undertaken on the study level results and the kind of issues found in this process. It also presents the 16 novel loci discovered in this study and discusses their effect on other relevant traits. This study was published in Nature Genetics in 2011 (2)* (**Appendix A**). I was the lead analyst for this study (as reflected by my first author status on the publication) and I independently generated all results that appear in the paper and in this thesis, unless otherwise stated. As this study was a collaboration between the pre-existing SpiroMeta and CHARGE cohorts, a second analyst (Daan Loth, representing the CHARGE consortium) independently analysed the data, and more detailed description of his role and the outputs from this work is described in the "Statement of originality of the work" in the introduction to this thesis

3.1 Introduction

Lung function measures predict morbidity and mortality (119-121), and are used in the diagnosis of chronic obstructive pulmonary disease (COPD). As discussed in Chapter 1 section 1.3, lung function measures are known to aggregate in families, and heritability studies estimate narrow sense heritability of lung function to be between 40% and 50% (70, 81, 82). Detecting genetic

variants associated with lung function measures might provide insights into the molecular pathways involved in lung function and lung disease.

Previous genome-wide association studies (23, 96) undertaken by the SpiroMeta and the CHARGE consortia separately, each one in over 20,000 individuals, have confirmed the association of a locus (4q31) known to affect lung function (94) and reported 10 additional genetic variants associated with lung function which reached genome-wide significance in at least one of the consortia. However, these variants only explain a very small proportion of the heritability of the lung function measures (23). Due to the modest effects of genetic variants, large sample sizes are required in order to detect them. This chapter presents a joint meta-analysis of GWAS including studies from the SpiroMeta and CHARGE consortia and studies new to both consortia with a discovery stage of 48,201 individuals and with the top signals followed up in up to 46,411 individuals.

Often, sharing individual level data is a complicated process, and a meta-analysis of study level results according to a shared analysis plan is a common approach to overcome this issue in genetic studies. Meta-analysing study level results has been shown to be as efficient as analysing individual level data jointly (122). Concerns that may arise when meta-analysing published findings, where studies are selected on the basis of their findings, such as publication

bias, are not an issue in the context of a meta-analysis of study level results according to a common analysis plan, where the studies are selected before undertaking the analysis. However, challenges are still faced when meta-analysing results from different studies, particularly if many studies and therefore many analysts are involved in analyses of very large datasets where the consequences of programming errors may be harder to detect. This chapter presents the procedure I adopted to ensure the quality of the results and discusses the issues found.

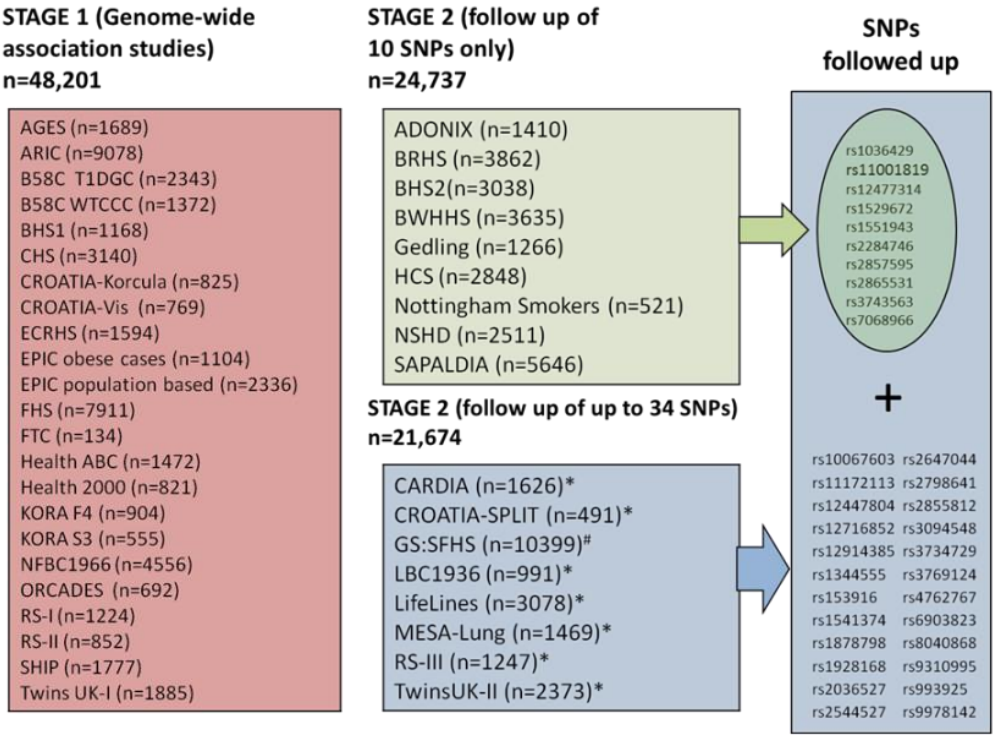
New genomic regions associated with lung function discovered in this study are also presented, as well as the effect of variants in these regions, and in regions previously reported to affect lung function, on other traits of interest, such as smoking, height or lung cancer.

3.2 SpiroMeta-CHARGE meta-analysis of GWAS: methods

3.2.1 Study design

This study consisted of two stages (**Figure 3-1**). A discovery stage (stage 1), where around 2.5 million SNPs were analysed in 23 studies and 48,201 individuals of European ancestry; and a follow-up stage (stage 2), where the 10 strongest signals were analysed in up to 17 studies and up to 46,411 individuals of European ancestry and a further 24 SNPs in a subset of up to 21,674 individuals.

Figure 3-1 Study design
Definitions of all study abbreviations are given in section 3.2.2.



* Studies with genome-wide data which provided results for the ten top SNPs and up to an additional 24 SNPs; # Study which undertook genotyping on a 32-SNP multiplex genotyping platform including the ten top SNPs and an additional 12 SNPs

3.2.2 **Stage 1 samples**

Stage 1 was formed of 23 studies, 17 from the SpiroMeta consortium and 6 from the CHARGE consortium. The studies were: AGES, Age, Gene/Environment Susceptibility; ARIC, Atherosclerosis Risk in Communities; B58C T1DGC, British 1958 Birth Cohort Type 1 Diabetes Genetics Consortium; B58C WTCCC, British 1958 Birth Cohort Wellcome Trust Case Control Consortium; BHS1, Busselton Health Study 1; CHS, Cardiovascular Health Study; the CROATIA- Korcula study; the CROATIA-Vis study; ECRHS, the European Community Respiratory Health Survey; EPIC obese cases, European Prospective Investigation into Cancer and Nutrition, Obese Cases; EPIC population based, European Prospective Investigation into Cancer and Nutrition

Cohort; FHS, Framingham Heart Study; FTC, Finnish Twin Cohort incorporating FinnTwin16 and FITSA; Health ABC, Health, Aging, and Body Composition; H2000, Finnish Health 2000 survey; KORA F4, Cooperative Health Research in the Region of Augsburg; KORA S3, Cooperative Health Research in the Region of Augsburg; NFBC1966, Northern Finland Birth Cohort of 1966; ORCADES, Orkney Complex Disease Study; RS-I and RS-II, Rotterdam Studies; SHIP, Study of Health in Pomerania; the TwinsUK-I study. **Table 3-1** gives descriptive information of the studies. Spirometry measurements were undertaken in each study as described in (2)*.

Table 3-1 Sample population characteristics for each study in stage 1
Abbreviations: N = number, y. =years, s.d. =standard deviation, smk = ever-smokers, nonsmk = never-smokers.

Study	N total	N male	N female	Age range (y) at FEV_1/FVC measurement	Mean age, y (s.d.)	Mean FEV_1, L (s.d.)	Mean FVC, L (s.d.)	Mean FEV_1/ FVC (s.d.)	N smk	N nonsmk	Genomic inflation factor (λ) FEV_1		Genomic inflation factor (λ) FEV_1/FVC	
											Smk	Nonsmk	Smk	Nonsmk
AGES	1689	686	1003	66-95	76.19 (5.63)	2.13 (0.69)	2.86(0.85)	0.74 (0.11)	886	803	1.009	1.012	1.003	1.003
ARIC	9078	4279	4799	44-66	54.27 (5.70)	2.94 (0.78)	3.99 (0.98)	0.74 (0.08)	5458	3620	1.034	1.007	1.019	1.019
B58C T1DGC	2343	1131	1212	44–45	44.5 (0)	3.31 (0.78)	4.19 (0.96)	0.79 (0.08)	1651	692	1.009	0.999	1.023	1.009
B58C WTCCC	1372	691	681	44–45	44.5 (0)	2.93 (0.75)	4.18 (0.96)	0.79 (0.08)	978	394	0.999	0.996	1.007	0.99
BHS1	1168	455	713	17-91	52.98 (17.07)	2.81 (0.97)	3.68 (0.11)	0.76 (0.09)	515	653	1.02	1.034	1.02	1.015
CHS	3140	1226	1914	65-95	72.3 (5.4)	2.12 (0.66)	3.00 (0.87)	0.71 (0.11)	1597	1543	1.035	1.021	1.02	1.033
CROATIA-Korcula	825	300	525	18–90	55.5 (13.5)	2.84 (0.81)	3.37 (0.93)	0.84 (0.09)	428	397	1.039	1.014	0.999	1.041
CROATIA-Vis	769	323	446	18–88	56.3 (15.3)	3.39 (1.22)	4.38 (1.43)	0.77 (0.09)	441	328	1.019	1.002	1.066	1.027
ECRHS	1594	784	810	19-48	33.90 (7.17)	3.78 (0.82)	4.59 (1.03)	0.83 (0.07)	895	699	1.018	1.014	1.005	1.024
EPIC obese cases	1104	476	628	39–76	59.1 (8.8)	2.35 (0.69)	2.84 (0.87)	0.82 (0.17)	615	489	1.005	1.02	1.02	1.014
EPIC population based	2336	1100	1236	39–77	59.2 (9.0)	2.50 (0.72)	3.04 (0.90)	0.85 (0.16)	1275	1061	1.013	1.008	1.002	1.018

Study	N total	N male	N female	Age range (y) at FEV_1/FVC measurement	Mean age, y (s.d.)	Mean FEV_1, L (s.d.)	Mean FVC, L (s.d.)	Mean FEV_1/ FVC (s.d.)	N smk	N nonsmk	Genomic inflation factor (λ) FEV_1		Genomic inflation factor (λ) FEV_1/FVC	
											Smk	Nonsmk	Smk	Nonsmk
FHS	7911	3650	4261	19-92	52.2 (14.6)	3.03 (0.94)	4.02 (1.14)	0.75 (0.08)	4355	3556	1.02	1.032	1.007	1.026
FTC	134	13	121	23–76	57.4 (19.3)	2.69 (0.94)	2.93 (0.61)	0.79 (0.09)	30	104	1.054	1.011	1.005	1.003
Health ABC	1472	786	686	70-79	73.7 (2.8)	2.31 (0.66)	3.11 (0.81)	0.74 (0.08)	831	641	0.997	1.001	0.998	1.012
Health 2000	821	394	427	30-75	50.47(10.91)	3.29 (0.90)	4.16 (1.07)	0.79 (0.07)	572	249	1.001	1.023	1.007	1
KORA F4	904	426	478	42-61	53.82(4.39)	3.25 (0.79)	4.20 (0.97)	0.77 (0.06)	560	344	1.051	1.013	1.032	1.017
KORA S3	555	261	294	29–73	47.6 (9.0)	3.43 (0.78)	4.18 (0.99)	0.83 (0.07)	289	266	1.014	1.019	1.012	1.029
NFBC1966	4556	2182	2374	31–31	31.0 (0)	3.96 (0.79)	4.73 (0.99)	0.84 (0.06)	2908	1648	1.022	1.011	1.023	1.004
ORCADES	692	322	370	19–93	54.9 (15.3)	2.88 (0.84)	3.58 (0.98)	0.80 (0.09)	288	404	1.014	1.046	1.015	1.051
RS-I	1224	556	668	65-97	74.5 (5.6)	2.31 (0.73)	3.16 (0.92)	0.73 (0.08)	863	361	1.029	1.023	1.026	1.017
RS-II	852	381	471	58-88	67.2 (6.3)	2.71 (0.78)	3.61 (1.08)	0.76 (0.09)	565	287	1.027	1.006	1.009	1.016
SHIP	1777	870	907	25–85	52.3 (13.7)	3.28 (0.89)	3.87 (1.03)	0.87 (0.06)	1004	773	1	0.996	1.016	0.991
TwinsUK-I	1885	0	1,885	18–79	48.4 (12.2)	2.73 (0.56)	3.40 (0.61)	0.80 (0.08)	942	943	0.998	1.012	1.009	1.005
Stage 1 sample size	48201													

3.2.3 Association analyses of stage 1

I designed a common analysis plan to ensure that the same analyses were undertaken across all studies. This was circulated between the studies and after discussion it was agreed and adopted by all studies. The analysis plan can be found in **Appendix B**. After that, I carried out quality control checks on the data uploaded and meta-analysed the results once all the QC issues had been resolved.

3.2.3.1 Study level analyses

The genotyping platforms and quality control criteria implemented by each study are given in **Appendix C**. Each study carried out imputation of non-genotyped SNPs against the European subset of samples in HapMap NCBI build 35 or 36 (**Appendix C**). The software programs used for imputation differed between studies and included: MACH (18), IMPUTE (17) or BIMBAM (123). MACH and IMPUTE are two software implementations that share similar underlying population genetic models (124), and BIMBAM has been shown to perform similarly to MACH and IMPUTE in contrast with other imputation methods (125, 126).

Forced expiratory volume in one second (FEV_1) and the ratio of FEV_1 over forced vital capacity (FEV_1/FVC) were the traits studied. Only individuals with no missing data for ever smoking status and with complete data on both FEV_1 and

FEV_1/FVC were included in the analysis. Both traits were adjusted for age, age^2, sex, height and ancestry principal components using linear regression. To ensure the normality of the data, the residuals obtained in the linear regressions were then transformed to ranks and to normally distributed *Z*-scores. These transformed residuals were then used as the dependent variables for association testing assuming additive genetic effects, separately for ever-smokers and never-smokers. The software used for association testing is specified in **Appendix C**. Appropriate tests for association in related individuals were applied where necessary, as described in (2)*.

3.2.3.2 Consortium central analyses

Quality control checks

A series of quality control checks were carried out on the study level results to make sure that no analytic errors were included in the meta-analysis. A series of plots were generated genome-wide for each dataset in each study to assess the quality of the data and to identify any irregularities.

SNPs with low imputation quality indicate that there is not enough information from surrounding SNPs in LD with the SNPs being imputed to reliably infer their genotype. These SNPs may give erroneous results in the association tests. SNPs with low minor allele frequencies (MAF) are more prone to errors in the variant calling process, since most clustering-based algorithms do not perform

well when there is a small number of samples within a genotype cluster (35). Moreover, an error in the calling of a SNP with low minor allele frequency will have a substantial influence in the overall allele frequency for that SNP, and this may also influence the association test results. For these reasons, some of the checks performed across studies were mainly focused on the SNPs remaining after applying an imputation quality filter (SNPs with imputation quality < 0.3 were removed) and a minor allele frequency filter (SNPs with minor allele frequency < 0.05 were removed) using study specific information. These SNPs were only excluded for the quality control checks, when meta-analysing the results only SNPs with imputation quality < 0.3 were removed, but no minor allele frequency filter was applied.

File formatting

Files uploaded by each study were checked to make sure that they were formatted as requested in the analysis plan and they were re-formatted when that was not the case. The inclusion of wrongly formatted files in the analysis may have serious consequences, producing errors in the meta-analysis process, or giving incorrect results in the meta-analysis. When managing large data files automated pipelines are often used and “eyeballing” the data is not a common practice. However, simple checks such as manually going through all the column names in the files to make sure they are as requested, or checking that a study has not uploaded the same file twice with a different name, by making sure all the file sizes are different, proved to be useful.

Consistency across studies

As described below, plots were generated genome-wide for each study and then compared in order to identify inconsistencies between studies. Inconsistencies might arise due to genuine differences in the study population or they could also reveal systematic biases; they could for instance be introduced due to a programming error during the analysis or due to an error in reporting coded alleles or strands. Some systematic errors, such as a programming error that leads to underestimated standard errors and therefore inflated statistics, can be very influential in the meta-analysis results.

Plots generated include: (i) plots of effect sizes and standard errors for all SNPs after applying imputation quality (< 0.3) and minor allele frequency (< 0.05) filters; (ii) plots of the density of weights (the inverse of the standard error squared was used as the weight) used in the meta-analysis: this might highlight systematic differences in the way the results were estimated; and (iii) plots of study allele frequencies against HapMap allele frequencies: to ensure that the allele coding was consistent across studies, the effect sizes were flipped so that the effect of the alphabetically higher allele on the forward strand of the NCBI build 36 reference sequence of the human genome was reported by each study, and then their allele frequencies were plotted against HapMap frequencies for the same alleles.

Data quality

Quantile quantile plots (QQ plots) are used to compare two probability distributions, by plotting the quantiles of one distribution against the quantiles of the other distribution. If the two distributions are the same, the quantiles would be expected to follow the line of correlation with slope equal to 1. In genome-wide association studies QQ plots are used to compare the distribution of the observed P-values with the distribution of the expected P-values if the null hypothesis holds, that is in case of no association. The –log10 of the P-values are usually plotted instead of the P-values to facilitate visualization. If the observed P-values show no more associations than expected by chance, we would expect the dots representing SNPs on the plot to follow the line of correlation with slope equal to 1. If there are some real associations we would expect to have more significant P-values on the observed set of P-values and therefore we would expect to see a deviation at the upper right end of the plot.

QQ plots are also a useful tool to detect genomic inflation; overinflated statistics would produce smaller P-values than expected by chance all along the distribution of P-values distribution. In order to assess the overinflation of the statistics in a GWAS, the genomic inflation factor (λ) is calculated. The test statistic used to test the association of a SNP with the trait in a linear regression $((Beta/SE)^2)$ should follow a Chi-square distribution with one degree of freedom, and overinflated statistics will deviate from it by a factor λ. The genomic inflation factor is obtained as the median of the test statistics over the

median of a Chi-square distribution with one degree of freedom. The expected value of λ if there is no overinflation is around 1. Genomic inflation can arise for example due to population structure or relatedness that have not been appropriately adjusted for in the modelling process.

For each study, QQ plots were generated after applying imputation quality (< 0.3) and minor allele frequency (< 0.05) filters and genomic inflation factors generated. Imputation quality metrics r2.hat (MACH), .info (IMPUTE) or OEvar (BIMBAM), were also plotted across studies.

Meta-analysis

SNPs with imputation quality below 0.3 were excluded from the analysis. Genomic control (26) (explained in section 1.1.3 in Chapter 1) was applied to the ever-smokers and never-smokers datasets separately for each study using the genomic inflator factors (λ) given in **Table 3-1**. For each study the effect sizes were flipped so that the effect of the alphabetically higher allele in the forward strand of the NCBI build 36 reference sequence of the human genome was reported. Then, effect size estimates and standard errors for ever-smokers and never-smokers for each study were meta-analysed using inverse variance weighting (the inverse of the standard error squared was used as the weight), and genomic control was applied to the pooled estimates at study level. Finally, effect sizes and standard errors were meta-analysed across studies using

inverse variance weighting and genomic control was applied one last time to the final estimates.

3.2.4 Selection of SNPs for stage 2

First, a list of the strongest associations was produced, selecting only the most significant SNP (or sentinel SNP) for each independent region, using a P-value threshold of $P < 3 \times 10^{-6}$ for either FEV_1 or FEV_1/FVC. Independent regions were defined as those with sentinel SNPs more than 500kb apart. Only novel regions were selected for follow-up, hence SNPs in previously reported regions (23, 94, 96) were removed from the list. The criteria to select SNPs from the novel regions took into account their effective sample size (obtained as the sum across studies of the product of the imputation quality metric for each SNP and the sample size), the association shown by surrounding SNPs assessed by examining region plots and the consistency of the direction of effect across studies assessed by examining forest plots. Only SNPs with effective sample sizes ≥ 70% of the total stage 1 sample size, with association signals for surrounding SNPs that were consistent with their correlation (or LD) with the sentinel SNP, and with consistent direction of effect across studies were selected. SNPs in twenty-nine regions met these criteria. In two regions, the sentinel SNP had effective sample size ≥ 70% but < 80% of the total stage 1 sample size and, for that reason, a proxy SNP from each of these two regions ($r^2 = 1$ and $r^2 = 0.97$) with effective sample size > 80% was also selected. In three regions there were different sentinel SNPs showing association with FEV_1

and FEV_1/FVC and all of them were taken forward. In total, 29 regions and 34 SNPs were selected for follow-up in stage 2.

Table 3-2 presents the list of SNPs followed up in stage 2 and includes the reason why each SNP was selected: “sentinel SNP” if it was the sentinel SNP with $P < 3 \times 10^{-6}$, “proxy for …” if it was a proxy with effective sample size >= 80% for a sentinel SNP with $P < 3 \times 10^{-6}$ and N effective < 80%, and “sentinel SNP (different trait)” if it was the second SNP selected in a region with different sentinel SNPs for each trait. **Table 3-2** also provides a ranking (“Ranking for follow-up”) by P-value for association with the trait (“Measure”) that had the strongest association, used to prioritise SNPs for follow-up in a larger number of samples.

Table 3-2 List of SNPs selected for follow-up

Definitions of all study abbreviations are given in section 3.2.5. Abbreviations: N = effective sample sizes, Chr. = chromosome.

Chr.	Measure	SNP_ID (NCBI36 position), function	N stage1	N stage2	Why selected for follow-up?	Ranking for follow- up	Direct genotyping follow-up			*In-silico* follow-up	
							In ADONIX, BRHS, BWHHS, Gedling, HCS, Nottingham Smokers, NSHD and SAPALDIA?	In BHS2?	In GS: SFHS ?	In CARDIA, CROATIA-SPLIT, LBC1936, LifeLines, MESA-Lung and RS-III?	In TwinsUK -II?
10	FEV_1/FVC	rs7068966 (12317998), *CDC123* (intron)	47085	45892	sentinel SNP	1	Yes	Yes	Yes	Yes	Yes
3	FEV_1/FVC	rs1529672 (25495586), *RARB* (intron)	40624	45386	sentinel SNP	2	Yes	Yes	Yes	Yes	Yes
1	FEV_1/FVC	rs2284746 (17179262), *MFAP2* (intron)	45944	35310	sentinel SNP	3	Yes	Yes	-	Yes	Yes
10	FEV_1	rs1878798 (12283489), *CDC123* (intron)	46164	21086	sentinel SNP (different trait)	4	-	-	Yes	Yes	Yes
2	FEV_1/FVC	rs12477314 (239542085), *HDAC4* (downstream)	45585	45704	sentinel SNP	5	Yes	Yes	Yes	Yes	Yes
5	FEV_1/FVC	rs1551943 (52230790), *ITGA1* (intron)	43787	45914	sentinel SNP	6	Yes	Yes	Yes	Yes	Yes
12	FEV_1/FVC	rs1036429 (94795559), *CCDC38* (intron)	47814	46183	sentinel SNP	7	Yes	Yes	Yes	Yes	Yes
10	FEV_1	rs11001819 (77985230), *C10orf11* (intron)	45546	45677	sentinel SNP	8	Yes	Yes	Yes	Yes	Yes
16	FEV_1/FVC	rs2865531 (73947817), *CFDP1* (intron)	47594	46286	sentinel SNP	9	Yes	Yes	Yes	Yes	Yes
16	FEV_1/FVC	rs12447804 (56632783), *MMP15* (intron)	35123	23693	sentinel SNP	10	-	Yes	Yes	Yes	Yes

Chr.	Measure	SNP_ID (NCBI36 position), function	N stage1	N stage2	Why selected for follow-up?	Ranking for follow- up	Direct genotyping follow-up			*In-silico* follow-up	
							In ADONIX, BRHS, BWHHS, Gedling, HCS, Nottingham Smokers, NSHD and SAPALDIA?	In BHS2?	In GS: SFHS ?	In CARDIA, CROATIA-SPLIT, LBC1936, LifeLines, MESA-Lung and RS-III?	In TwinsUK -II?
16	FEV_1/FVC	rs3743563 (56636666), *MMP15* (missense)	47179	43190	proxy for rs12447804	11	Yes	-	Yes	Yes	Yes
6	FEV_1/FVC	rs2857595 (31676448), *NCR3* (upstream)	45540	45657	sentinel SNP	12	Yes	Yes	Yes	Yes	Yes
6	FEV_1	rs2855812 (31580699), *MICB* (intron)	46921	21190	sentinel SNP (different trait)	13	-	-	Yes	Yes	Yes
6	FEV_1/FVC	rs1928168 (22125717), *AK026189* (intron)	47936	21323	sentinel SNP	14	-	-	Yes	Yes	Yes
2	FEV_1/FVC	rs2544527 (15843619), *DDX1* (downstream)	45352	21115	sentinel SNP	15	-	-	Yes	Yes	Yes
6	FEV_1	rs6903823 (28430275), *ZKSCAN3* (intron)	47057	21428	sentinel SNP	16	-	-	Yes	Yes	Yes
1	FEV_1/FVC	rs993925 (216926691), *TGFB2* (downstream)	42402	21162	sentinel SNP	17	-	-	Yes	Yes	Yes
2	FEV_1	rs3769124 (239014101), *ASB1* (intron)	44924	10579	sentinel SNP	18	-	-	-	Yes	Yes
6	FEV_1/FVC	rs2798641 (109374743), *ARMC2* (intron)	46369	20999	sentinel SNP	19	-	-	Yes	Yes	Yes
16	FEV_1	rs12716852 (76746239), *WWOX* (intron)	47510	21228	sentinel SNP	20	-	-	Yes	Yes	Yes
6	FEV_1	rs3094548 (29463181), *OR12D2* (upstream)	42516	20733	sentinel SNP	21	-	-	Yes	Yes	Yes
21	FEV_1/FVC	rs9978142 (34574109), *KCNE2* (upstream)	44577	20693	sentinel SNP	22	-	-	Yes	Yes	Yes

Chr.	Measure	SNP_ID (NCBI36 position), function	N stage1	N stage2	Why selected for follow-up?	Ranking for follow- up	Direct genotyping follow-up			*In-silico* follow-up	
							In ADONIX, BRHS, BWHHS, Gedling, HCS, Nottingham Smokers, NSHD and SAPALDIA?	In BHS2?	In GS: SFHS ?	In CARDIA, CROATIA-SPLIT, LBC1936, LifeLines, MESA-Lung and RS-III?	In TwinsUK -II?
6	FEV_1	rs3734729 (150612560), *PPP1R14C* (untranslated-3)	43680	20998	sentinel SNP	23	-	-	Yes	Yes	Yes
15	FEV_1/FVC	rs8040868 (76698236), *CHRNA3* (synonymous)	35121	21131	sentinel SNP	24	-	-	Yes	Yes	Yes
15	FEV_1/FVC	rs12914385 (76685778), *CHRNA3* (intron)	47226	21327	proxy for rs8040868	25	-	-	Yes	Yes	Yes
3	FEV_1	rs9310995 (32904119), *TRIM71* (intron)	44835	21070	sentinel SNP	26	-	-	Yes	Yes	Yes
12	FEV_1/FVC	rs11172113 (55813550), *LRP1* (intron)	45387	20256	sentinel SNP	27	-	-	Yes	Yes	Yes
5	FEV_1/FVC	rs10067603 (131831767), *C5orf56* (downstream)	44134	21167	sentinel SNP	28	-	-	Yes	Yes	Yes
3	FEV_1	rs1344555 (170782913), *MECOM* (intron)	46067	21104	sentinel SNP	29	-	-	Yes	Yes	Yes
5	FEV_1/FVC	rs153916 (95062456), *SPATA9* (upstream)	47530	21428	sentinel SNP	30	-	-	Yes	Yes	Yes
15	FEV_1	rs2036527 (76638670), *CHRNA5* (uptream)	45038	20874	sentinel SNP (different trait)	31	-	-	Yes	Yes	Yes
12	FEV_1/FVC	rs4762767 (19757396), *AEBP2* (downstream)	48016	21324	sentinel SNP	32	-	-	Yes	Yes	Yes
4	FEV_1	rs1541374 (106267809), *TET2* (upstream)	45221	20516	sentinel SNP	33	-	-	Yes (reserve list)	Yes	Yes

Chr.	Measure	SNP_ID (NCBI36 position), function	N stage1	N stage2	Why selected for follow-up?	Ranking for follow- up	Direct genotyping follow-up			*In-silico* follow-up	
							In ADONIX, BRHS, BWHHS, Gedling, HCS, Nottingham Smokers, NSHD and SAPALDIA?	In BHS2?	In GS: SFHS ?	In CARDIA, CROATIA-SPLIT, LBC1936, LifeLines, MESA-Lung and RS-III?	In TwinsUK -II?
6	FEV_1/FVC	rs2647044 (32775888), *HLA-DQB1* (upstream)	44610	8381	sentinel SNP	34	-	-	-	Yes	-

3.2.5 Stage 2 samples

Studies that contributed to stage 2 with *in-silico* data (studies that already had GWAS data) were: CARDIA, Coronary Artery Risk Development in Young Adults; the CROATIA-Split study; LBC1936, Lothian Birth Cohort 1936; the LifeLines study; MESA-Lung, Multi-Ethnic Study of Atherosclerosis; RS-III, Rotterdam Study and; the TwinsUK-II study. Studies that contributed to stage 2 with direct genotyping (studies that did not have GWAS data and undertook de novo genotyping for a selection of the variants) were: ADONIX, Adult-Onset Asthma and Nitric Oxide; BHS2, Busselton Health Study 2; BRHS, British Regional Heart Study; BWHHS, British Women's Heart and Health Study; the Gedling study; GS:SFHS, Generation Scotland: Scottish Family Health Study; HCS, Hertfordshire Cohort Study; the Nottingham Smokers study; NSHD, Medical Research Council National Survey of Health and Development (also known as the British 1946 Birth Cohort) and; SAPALDIA, Swiss study on Air Pollution and Lung Disease in adults. **Table 3-3** gives descriptive information of these studies.

Table 3-3 Sample population characteristics for each study in stage 2
Abbreviations: N = number, y = years, s.d = standard deviation, L = litres, smk = ever-smokers, nonsmk = never-smokers.

Study	N total	N male	N female	Age range (y) at FEV_1/FVC measurement	Mean age, y (s.d.)	Mean FEV_1, L (s.d.)	Mean FVC, L (s.d.)	Mean FEV_1/ FVC (s.d.)	N nonsmk	N smk
Studies with *in-silico* data										
CARDIA	1626	768	858	17-32	25.6 (3.33)	3.68 (0.81)	4.70 (1.00)	0.82 (0.06)	932	694
CROATIA-SPLIT	491	209	282	18-85	49.07 (14.60)	3.19 (0.91)	3.80 (1.06)	0.84 (0.08)	239	252
LBC1936	991	501	490	67-71	69.55 (0.84)	2.38 (0.67)	3.04 (0.87)	0.79 (0.10)	437	554
LifeLines	3078	1232	1846	21-88	54.94 (9.75)	3.15 (0.81)	4.21 (1,01)	0.75 (0.08)	1075	2003
MESA-Lung	1469	737	732	48-90	66.1 (9.7)	2.57 (0.76)	3.44 (0.99)	0.73 (0.09)	636	833
RS-III	1247	549	698	46-89	56.59 (5.58)	3.15 (0.85)	4.06 (1.14)	0.78 (0.09)	425	822
TwinsUK-II	2373	0	2373	17-85	53.5(14.3)	2.62(0.61)	3.27(0.65)	0.80 (0.08)	1230	1143
Studies with direct genotyping										
ADONIX	1410	660	750	25-75	49.08 (13.54)	3.34 (0.86)	4.24(1.02)	0.79 (0.07)	792	618
BHS2	3038	1368	1670	18-97	50.0 (16.7)	3.05 (0.951)	3.97 (1.15)	0.78 (0.08)	1633	1405

Study	**N total**	**N male**	**N female**	**Age range (y) at FEV_1/FVC measurement**	**Mean age, y (s.d.)**	**Mean FEV_1, L (s.d.)**	**Mean FVC, L (s.d.)**	**Mean FEV_1/ FVC (s.d.)**	**N nonsmk**	**N smk**
BRHS	3862	3862	0	58-80	68.72 (5.48)	2.58 (0.69)	3.38 (0.84)	0.77 (0.17)	1121	2741
BWHHS	3635	0	3635	59-80	68.83 (5.49)	1.98 (0.52)	2.82 (0.76)	0.71 (0.09)	2055	1580
Gedling	1266	633	633	27-80	56.14 (12.29)	2.85 (0.85)	3.68 (1.01)	0.77 (0.07)	634	632
GS:SFHS	10399	4304	6095	18-93	46.37 (14.61)	3.11 (0.87)	4.05 (1.02)	0.77 (0.09)	5674	4725
HCS	2848	1509	1339	59-73	66.14 (2.84)	2.44 (0.68)	3.42 (0.92)	0.72 (0.09)	1318	1530
Nottingham Smokers	521	236	285	36-89	59.60 (10.48)	1.10 (0.95)	3.02 (1.05)	0.64 (0.16)	0	521
NSHD	2511	1258	1253	53	53	2.79 (0.70)	3.50 (0.90)	0.80 (0.09)	1045	1466
SAPALDIA	5646	2753	2893	18-62	42.0 (11.4)	3.58 (0.84)	4.53 (1.04)	0.79 (0.07)	2653	2993
Stage 2 sample size	46411									

All 34 SNPs were followed up in up to 11,275 individuals from seven studies with *in silico* data: CARDIA, CROATIA-SPLIT, LBC1936, LifeLines, MESA-Lung, RS-III and TwinsUK-II. The SNP rs2647044 was not available in TwinsUK-II (**Table 3-2**). The top ten ranking SNPs (**Table 3-2**) were selected for follow-up by direct genotyping in up to 35,136 individuals from ADONIX, BHS2, BRHS, BWHHS, Gedling, GS:SFHS, HCS, Nottingham Smokers, NSHD and SAPALDIA. If a SNP in the top ten had an *N* effective < 80%, only the proxy SNP (with *N* effective >= 80%) was included in the top ten for follow-up. For regions that showed association with both FEV_1 and FEV_1/FVC, only the leading SNP with the lowest P-value for either trait was included if it was within the top ten SNPs. rs3743563 (proxy for rs12447804 with *N* effective >= 80%) could not be genotyped in BHS2, so rs12447804 was genotyped instead (**Table 3-2**). The genotyping in GS: SFHS was undertaken using a 32-SNP multiplex genotyping platform, so the top ten SNPs were included plus an additional 22: the 32 top ranking SNPs, including proxies and both SNPs from regions that showed association with both FEV_1 and FEV_1/FVC. This assay failed for one SNP (rs3769124), which was subsequently replaced with the thirty-third ranking SNP (rs1541374); and rs2284746 was excluded because of poor clustering (**Table 3-2**).

3.2.6 Association analyses of stage 2

The analysis plan that I drafted was circulated to all stage 2 studies for comments. The final agreed version was then re-sent to all studies as included

in **Appendix B**. I undertook study level association analyses for BRHS, BWHHS, Gedling, GS:SFHS and Nottingham Smokers. Then I undertook quality control checks on the uploaded results, and once all issues were resolved I meta-analysed the results.

3.2.6.1 Study level analyses

Each study undertook the same association analysis as in stage 1 for up to 34 SNPs as described in section 3.2.3.1. In two follow-up studies (BHS2 and GS: SFHS), which had family data, ever-smokers and never-smokers were analysed together in order to account for familial correlations, and ever smoking status was included as a covariate in the model. The Nottingham Smokers study only included smokers, so the analysis was only done in one dataset.

3.2.6.2 Consortium central analyses

Quality control checks

A series of quality control checks once study-level results were provided was undertaken to make sure that no analytic errors were included in the meta-analysis. In the context of a follow-up analysis when only a small number of SNPs are analysed, checking data quality and consistency across studies was more challenging. Systematic differences that can appear obvious genome-wide, such as reporting the wrong coded allele, are much harder to identify

when only looking at a small subset of SNPs where the role of chance is harder to rule out.

File formatting

Files uploaded by each study were checked to make sure that they were formatted as requested in the analysis plan and they were re-formatted if necessary.

Consistency across studies and data quality

Imputation quality metrics for the SNPs analysed were requested and assessed for those studies with *in-silico* data. The consistency of direction of effect size estimates across studies was examined by generating forest plots and the consistency of the allele frequency for the coded allele was also examined by plotting coded allele frequencies across studies. When inconsistencies were found they were followed up contacting the study analysts.

Meta-analysis

After the quality of the results was ensured, the effect sizes were flipped so they were all reported for the alphabetically higher allele on the forward strand of the NCBI build 36 reference sequence of the human genome. For each study with stratified results effect sizes and standard errors for ever-smokers and never-

smokers were meta-analysed using an inverse variance weighted meta-analysis. Those studies which undertook the analysis genome-wide, although they only reported results for up to 34 SNPs, provided the genomic inflation factor (λ), so the pooled estimates for those studies were then corrected for genomic control. Finally, pooled effect sizes and standard errors across studies were obtained using inverse variance weighted meta-analysis.

3.2.7 Combined analysis of stage1 and stage 2 samples

Results from stage 1 and stage 2 were meta-analysed for the 34 top SNPs followed up in stage 2, using inverse variance weighted meta-analysis. Statistical significance was defined as equivalent to a Bonferroni correction for one million tests ($P < 5 \times 10^{-8}$) (11).

3.2.8 Additional analyses

Associations in stage 1 of SNPs previously associated with lung function

Associations in stage 1 of 13 previously reported regions were investigated. Regions included were: (i) 11 regions (with signals in or near *TNS1, PID1, FAM13A, GSTCD-NPNT, HHIP, HTR4, ADAM19, AGER, GPR126, PTCH1* and *THSD4*) reported as showing genome-wide significant association ($P < 5 \times 10^{-8}$) with lung function (23, 94, 96), (ii) *CHRNA3-CHRNA5-IREB2-LOC123688* reported as showing genome-wide significant association with COPD with additional evidence of association with lung function (95), and (iii) *DAAM2*,

which reached borderline genome-wide significance in the SpiroMeta consortium (23). If multiple SNPs had been reported for these regions, results for all SNPs were extracted, as well as the SNP that showed the strongest association in the stage 1 of this study.

Association of lung-function-associated SNPs with other traits

The association with other traits of sentinel SNPs in the new regions that met genome-wide significance after meta-analysing stage 1 and stage 2, and of sentinel variants in 12 previously discovered regions associated ($P < 5\times10^{-8}$) with lung function or COPD (and also associated with lung function) was tested. The following related traits were assessed: (i) lung function in children; (ii) height in the GIANT consortium (127) dataset; (iii) smoking amount and ever smoking status in the Ox-GSK consortium (109) dataset; and (iv) lung cancer in the International Lung Cancer Consortium (ILCCO) dataset (128). For regions with multiple SNPs reported, all the SNPs were included, except those having $r^2 > 0.9$ with another SNP already selected which were excluded. The most significant SNP in stage 1 of this study was also included for each region.

Gene x smoking interaction analysis

The analysis carried out to test for gene x smoking interaction was a Z-test comparing the effect of a given SNP in ever-smokers and in never-smokers.

Proportion of variance explained by loci discovered to date

The number of putative undiscovered variants with similar effects on lung function to those associated with lung function in the SpiroMeta-CHARGE dataset was estimated, and then the proportion of the variance that they collectively explain was calculated. The approach used was based on the method developed by Park *et al.* (129). Winners' curse bias free effect sizes were estimated, and then the number of undiscovered variants were estimated using the statistical power in the discovery dataset (discovery power) to detect the unbiased effect sizes.

To calculate the unbiased (winners' curse bias free) effect sizes for the 26 genome-wide significant ($P < 5 \times 10^{-8}$) variants (including both new and previously discovered variants) discovery data were excluded for each variant. Effect size estimates for the new variants that were genome-wide significant after meta-analysing stage1 and stage2 were obtained using SpiroMeta-CHARGE stage 2 data. The first study to report the association of *HHIP* with lung function was undertaken using data from FHS (94), therefore this study was excluded when estimating the effect size for *HHIP*. For the remaining 9 loci previously discovered, effect sizes were calculated excluding studies involved in the discovery GWAS of Repapi *et al*. 2010 (23), or in the discovery GWAS of Hancock *et al*. 2010 (96), or studies from both discovery GWAS, depending on which studies originally reported the loci.

To estimate the power for discovered associations the approach used by the ICBP consortium for systolic and diastolic blood pressure (113) was followed. This approach takes into account that two phenotypes were analysed in parallel and it also takes into account uncertainty about true effect sizes of the discovered variants. The discovery power is expressed as a function of the true effect sizes and it is then integrated with respect to a joint probability distribution for true effect sizes on FEV_1 and FEV_1/FVC. This joint probability distribution is a bivariate normal distribution with mean the unbiased (winners' curse bias free) effect size estimates for FEV_1 and FEV_1/FVC and variance-covariance matrix formed by their corresponding standard errors and the phenotypic correlation between FEV_1 and FEV_1/FVC.

The approach in Park *et al*. 2010 (129) was followed to obtain the proportion of variance explained by the inferred number of variants. First, the number of variants of a given effect size was obtained as the inverse of the power to detect them. Then, the proportion of variance explained by the i-th variant (i ranging from 1 to 26 in this case), with effect size β_i and allele frequency p_i was calculated as $\frac{2\,p_i(1-p_i){\beta_i}^2}{V}$ where V is the phenotypic variance. Finally, the proportion of variance explained by the inferred number of variants was obtained by summing the product of the number of variants of a given effect size by the proportion of variance explained by one of them, over each of the 26

genome-wide significant ($P < 5 \times 10^{-8}$) variants. Heritability of 40% (70, 81, 82) was assumed to estimate the proportion of the additive polygenic variance of each trait. The confidence interval for the total number of variants was obtained using bootstrapping, as in Park *et al*. 2010 (129).

3.3 SpiroMeta-CHARGE meta-analysis of GWAS: results

3.3.1 Results of the quality control checks in stage 1

Before meta-analysing the stage 1 study level results, a series of quality control checks were carried out. Details about the method and rationale for each check are given in *Quality control checks*, section 3.2.3.2.

Only plots for a selection of studies and datasets are presented here for simplicity. Study names in this section are not given; a random number has been allocated to all the stage 1 and stage 2 studies, so they are referred to by their number. A different dataset (for example, all individuals or ever-smokers only) for the same study may be presented to illustrate different issues in different sections.

3.3.1.1 File formatting

Files were re-formatted when they did not follow the guidance provided in the analysis plan. For instance, column names had to be changed to agree with

those requested in the analysis plan for a subset of studies. In addition, one study uploaded the same file for the never-smokers and ever-smokers results for FEV_1 and for FEV_1/FVC. This was identified by checking the file sizes. The analyst was contacted and the appropriate files were provided.

3.3.1.2 Consistency across studies

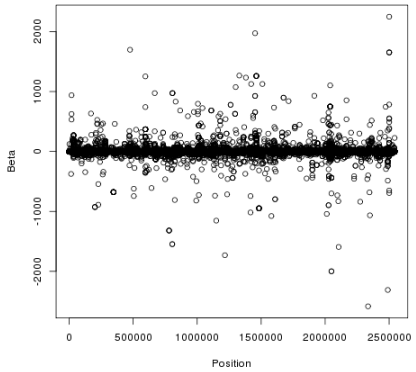
Plots of effect sizes and standard errors

Effect sizes (beta) and standard errors (SE) for all ~2.5 million SNPs for each study were plotted to look for inconsistencies across studies.

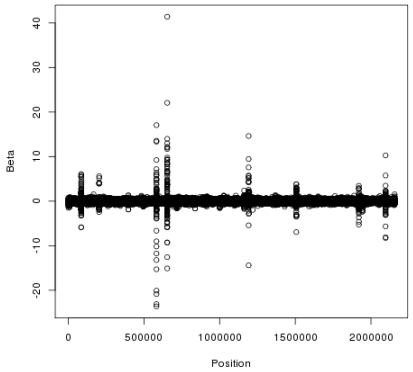
Figure 3-2 illustrates the effect of using minor allele frequency (< 0.05) and imputation quality (< 0.3) filters on the effect size estimates. The magnitude of the betas after the filtering appears to range between -0.5 and 0.5 in study 23 (N < 400) and between -0.15 and 0.15 in study 24 (N > 3000), while clear outliers can be observed before the filtering (**Figure 3-2**). A similar pattern can be seen for the SE (**Figure 3-3**). The plots for these two studies also illustrate that the magnitude of betas and SE differs between studies according to sample size. Larger studies have more accurate beta estimates and thus smaller SE (such as study 24), while estimates from smaller studies tend to be more variable and have larger SE (such as study 23).The same trend for variation according to sample size was seen across all studies.

Figure 3-2 Beta plots for study 23 and study 24 using different filtering strategies

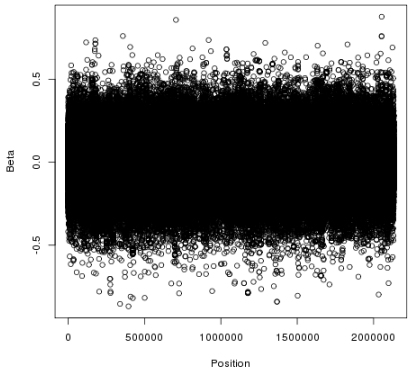
a) Study 23 unfiltered



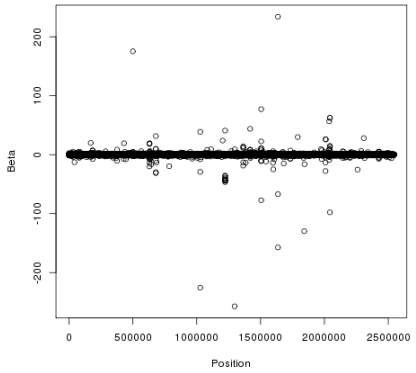
b) Study 23 MAF filtered



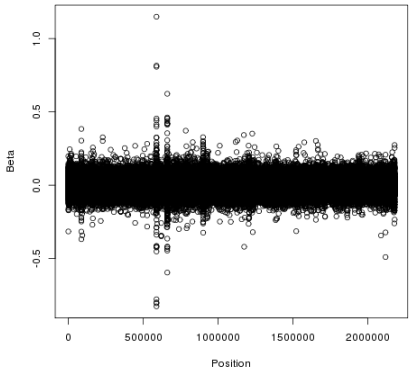
c) Study 23 MAF and imputation quality filtered



a) Study 24 unfiltered



b) Study 24 MAF filtered



c) Study 24 MAF and imputation quality filtered

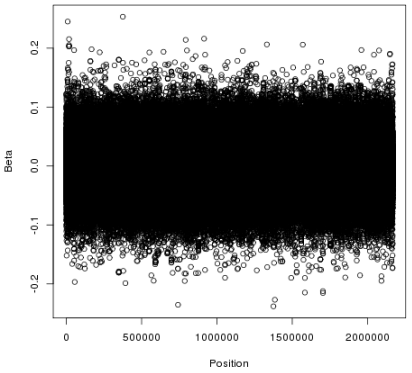
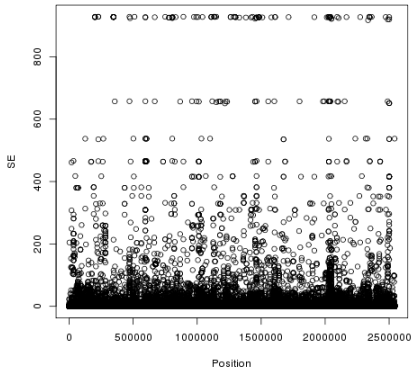
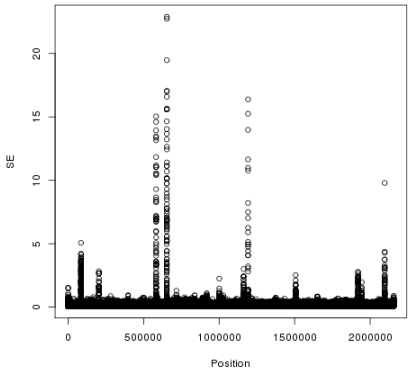


Figure 3-3 Standard error plots for study 23 and study 24 using different filtering strategies

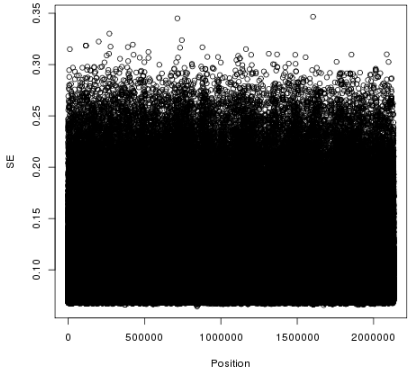
a) Study 23 unfiltered



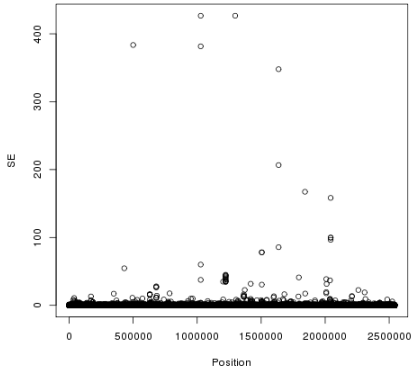
b) Study 23 MAF filtered



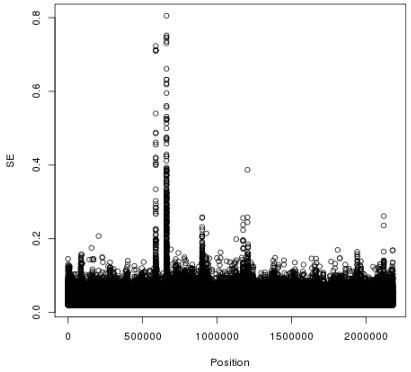
c) Study 23 MAF and imputation quality filtered



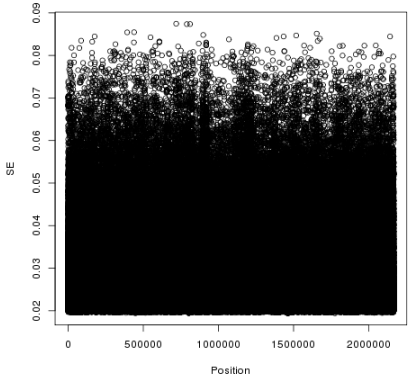
a) Study 24 unfiltered



b) Study 24 MAF filtered



c) Study 24 MAF and imputation quality filtered

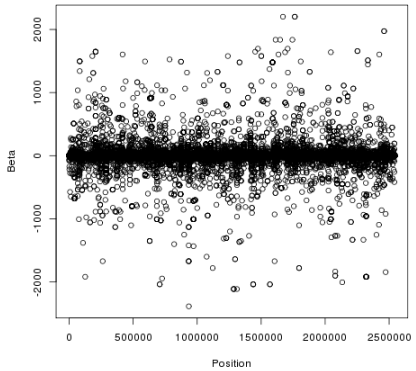


A remarkably high number of outliers were observed for study 20 (N < 300) after using minor allele frequency (< 0.05) and imputation quality (< 0.3) filters (**Figure 3-4 a)** and **b)**). The imputation quality for SNPs with outlying betas and standard errors after applying the filters ranged from 0.3 to 0.94, however their minor allele frequencies were all between 5% and 10% with most of them were between 5% and 7% (**Figure 3-4 d)**). This seemed to indicate that their outlying values could still be related to their allele frequency. Study 20 is a southern Finnish study, and allele frequencies with another Finnish study, study 19, in this case a northern Finnish study with a larger sample size (N >1000) were compared. Most of the outlying variants in study 20 had minor allele frequencies < 5% in study 19 (**Figure 3-4 c)**). This discrepancy in allele frequencies could be explained by south Finnish specific variation or due to the reduced sample size of this subset of study 20. Nevertheless, it seemed that the overall allele frequency for these variants across studies would be < 5% and since low allele frequency variants are not well imputed, this set of variants would be filtered out at the meta-analysis due to low N effective anyway, therefore this issue did not seem problematic.

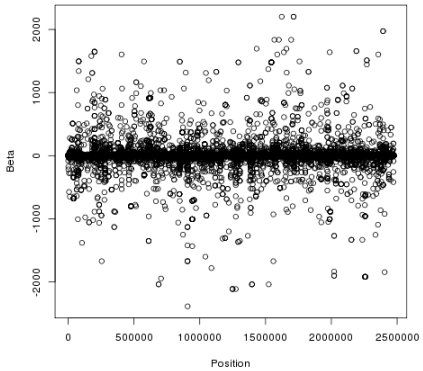
There were no issues with beta or standard error plots for any other study.

Figure 3-4 Beta and SE plots for study 20 and allele frequency distribution for SNPs with outlying values

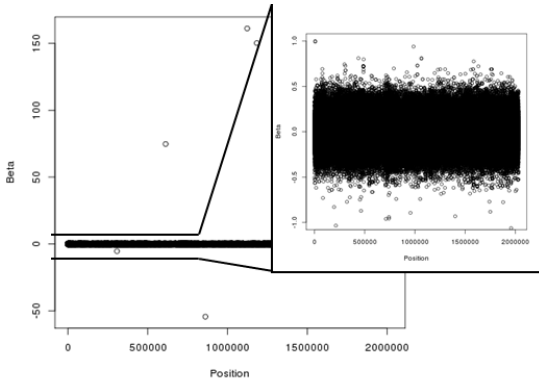
a) Betas unfiltered



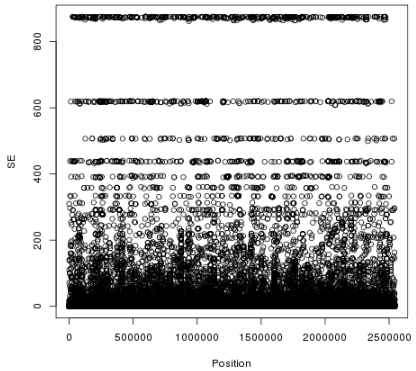
b) Betas MAF and imputation quality filtered



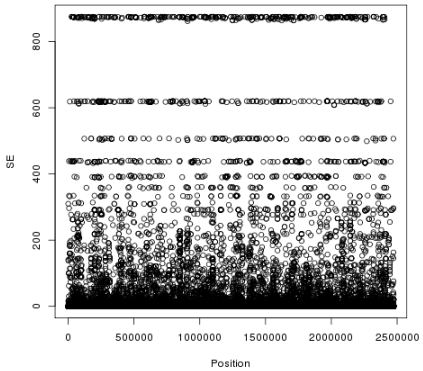
c) Betas MAF (using study 19 and 20 MAF) and imputation quality filtered



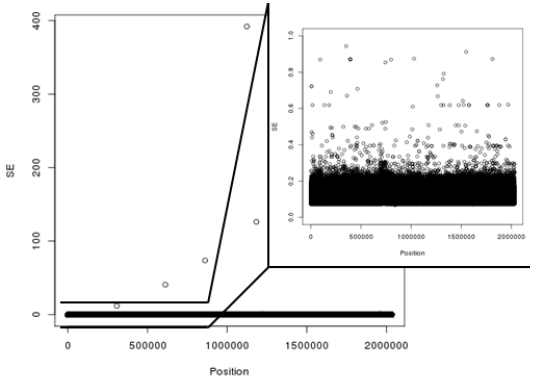
a) SE unfiltered



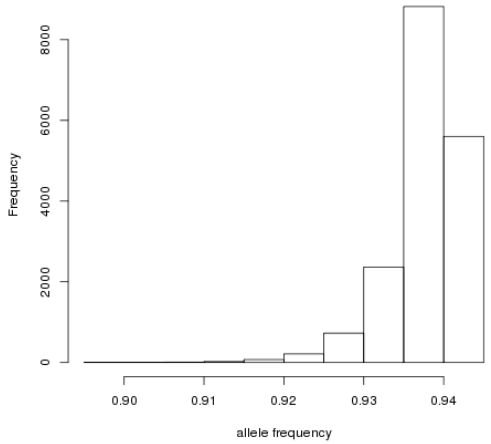
b) SE MAF and imputation quality filtered

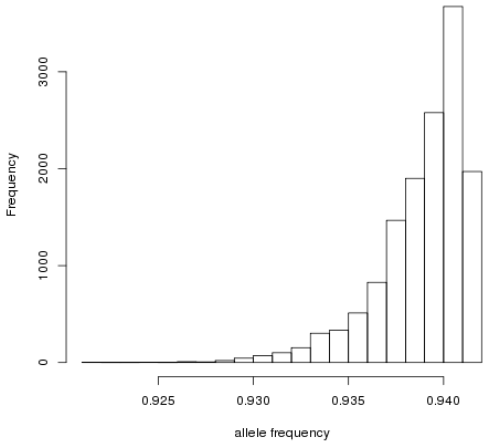


c) SE MAF (using study 19 and 20 MAF) and imputation quality filtered



d) Major allele frequency distribution for SNPs with Betas with absolute value > 1 (on the left) and for SNPs with SE > 1 (on the right)

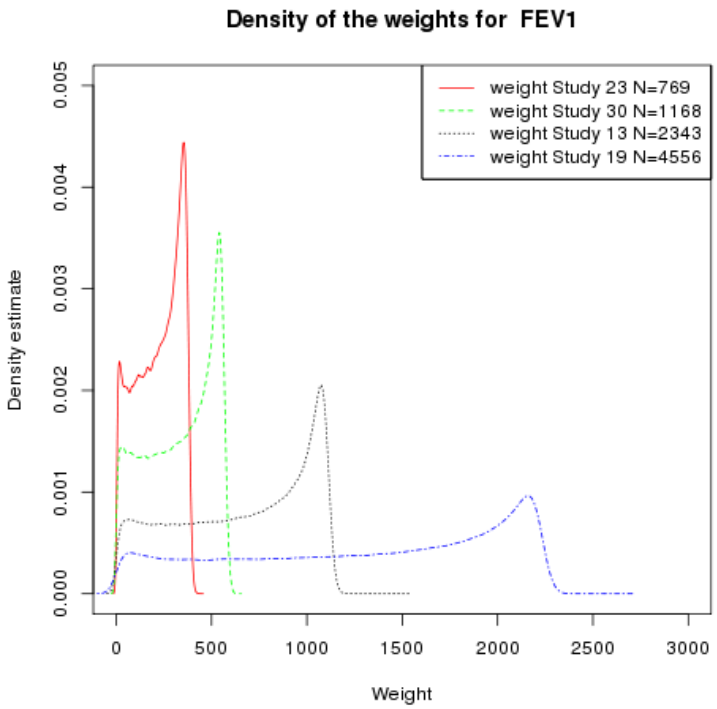




Plot of density of weights

Inverse variance weighted meta-analysis was used to produce pooled estimates of effect sizes and standard errors across studies. The density of the weights (inverse of the standard errors squared) was plotted across studies in order to identify possible systematic differences. However, they all seemed consistent. **Figure 3-5** shows the density of weights for a subset of studies with different sample sizes. Only a subset of studies are plotted here to facilitate visualization. Since standard errors scale inversely with sample size, we expect the weights to scale with sample size, as can be observed in **Figure 3-5** .

Figure 3-5 Density plot of FEV1 weights



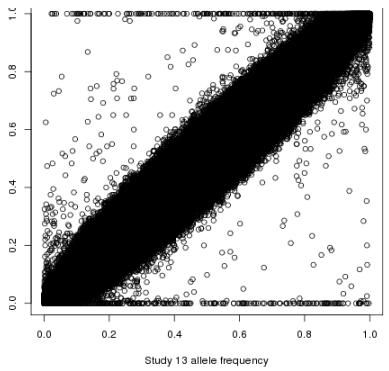
Plot of study allele frequencies against HapMap allele frequencies

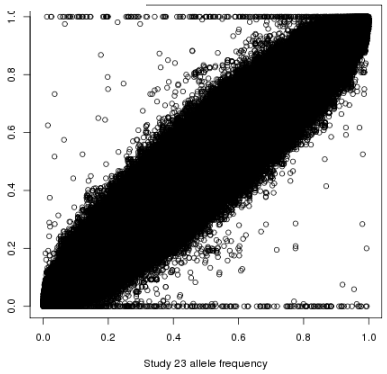
In order to meta-analyse effect sizes across studies it is necessary to ensure that all the effect size estimates for a given SNP corresponded to the same allele. To do that, effect sizes were flipped so that the effect of the alphabetically higher allele on the forward strand of the NCBI build 36 reference sequence of the human genome was reported by each study. Allele frequencies for the coded alleles in each study versus the allele frequency for the same allele in HapMap were plotted as an additional check. **Figure 3-6 a)** shows these plots for three studies and they illustrate that the allele frequencies of the studies were highly consistent with the HapMap allele frequencies. **Figure 3-6 b)** shows an inconsistency between the allele frequencies plotted for HapMap and another study, which seemed to correspond to different alleles, highlighting an error in how the alleles were reported in that study. The analysts for this study were contacted regarding this issue and the correct information for the allele coding was provided.

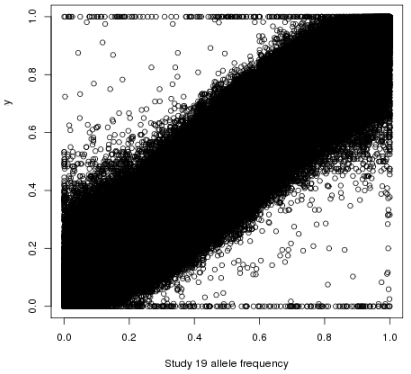
There were no issues with allele coding for any other study.

Figure 3-6 Study specific allele frequencies (x-axis) plotted against HapMap allele frequencies (y-axis)

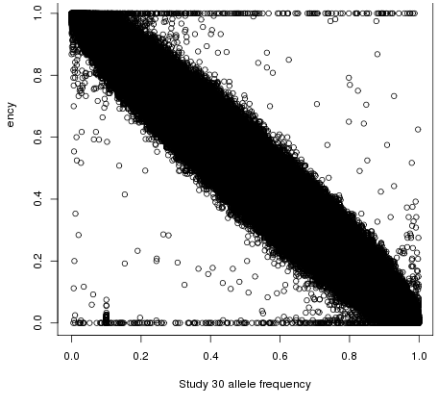
a) Regular allele frequencies for three studies







b) Irregular allele frequencies for one study



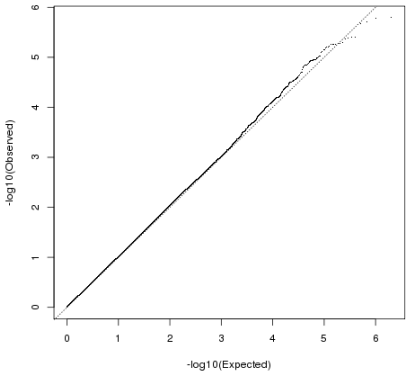
3.3.1.3 Data quality

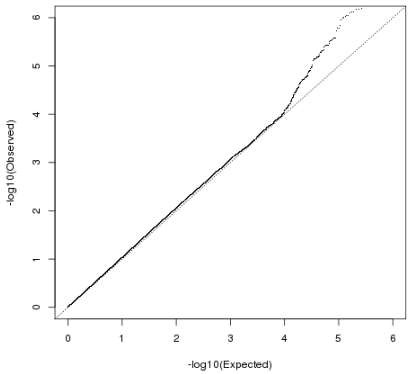
Quantile-quantile plots

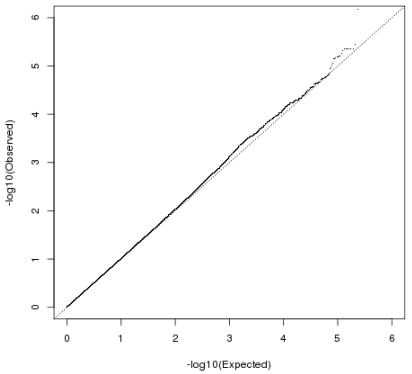
For each study, QQ plots were generated after applying imputation quality (< 0.3) and minor allele frequency (< 0.05) filters. **Figure 3-7 a)** shows regular QQ plots for a subset of studies, whereas **Figure 3-7 b)** shows an abnormal QQ plot, as well as plots for Betas and SEs, for another study (N < 500). The Betas for the study with abnormal QQ plots seemed consistent with the Betas for other studies of similar sample size (see study 23 in **Figure 3-2**), however the standard errors were noticeably higher (compared with study 23 in **Figure 3-3**) leading to the very non-significant P- values shown in the QQ plot. The genomic inflation factor for this study was 0.014. The analyst of this study was contacted, but no explanation was found for this abnormal results. For this reason this study had to be excluded from the analysis. QQ plots for the remaining studies looked satisfactory. Genomic inflation factors were also calculated for all studies and are presented in **Table 3-1**.

Figure 3-7 Quatile quantile plots

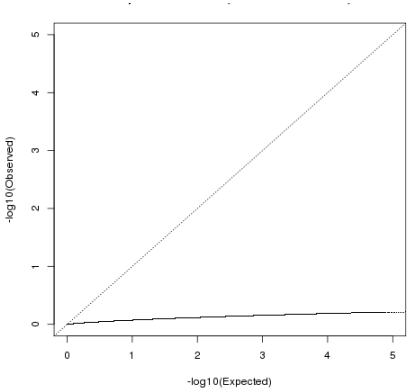
a) Regular QQ plots

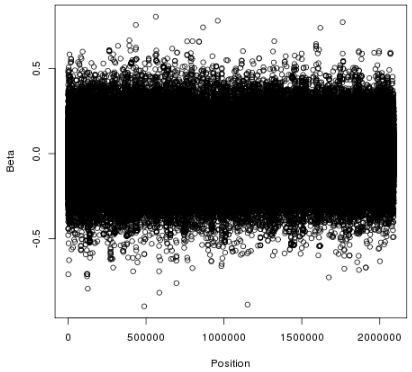


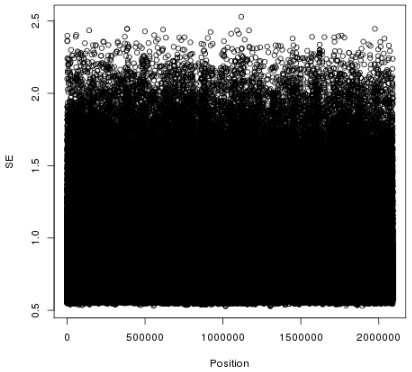




b) Irregular QQ plot for one study, Beta and Standard error plots for the same study







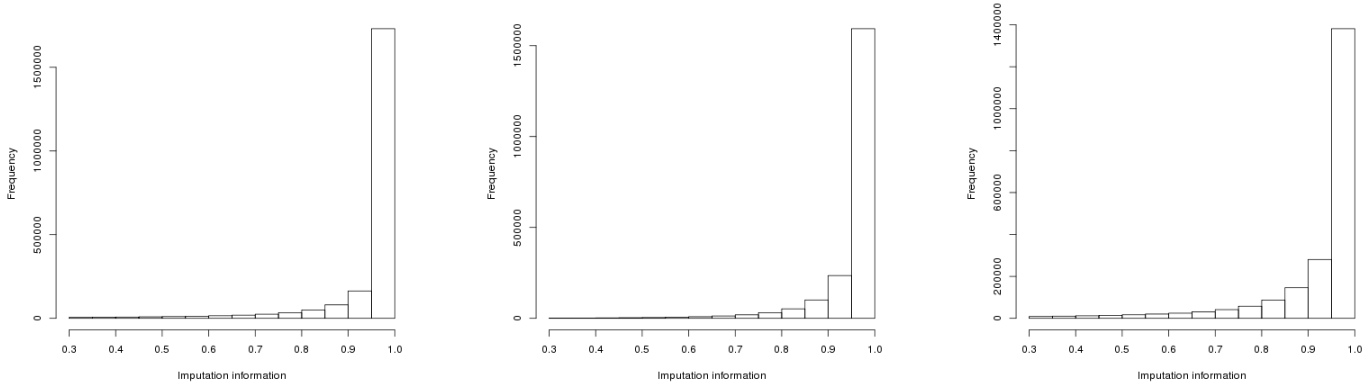
Plots of imputation quality

Imputation quality metrics: r2.hat (MACH), .info (IMPUTE) or OEvar (BIMBAM), were plotted across studies. **Figure 3-8 a)** shows regular histograms of imputation quality for three studies, with a clear peak for imputation quality around 1. **Figure 3-8 b)** shows a histogram of imputation quality for another study, with a peak around 1, but also another peak at around 0.1. The analysts of this study were contacted and they reported a bug in the association software (mach2qtl) which had rounded values of imputation quality 0.99 to 0.1 instead of 1. The association was undertaken again with a more recent version of the software with the bug fixed and the new results were uploaded.

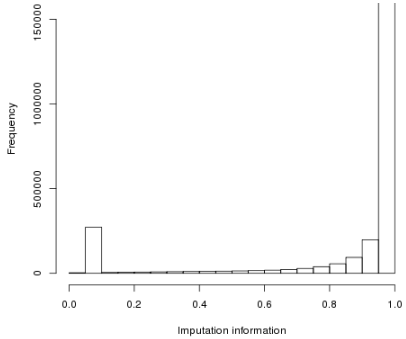
There were no issues with imputation quality metrics for any other study.

Figure 3-8 Histograms of imputation quality

a) Regular imputation quality



b) Irregular imputation quality

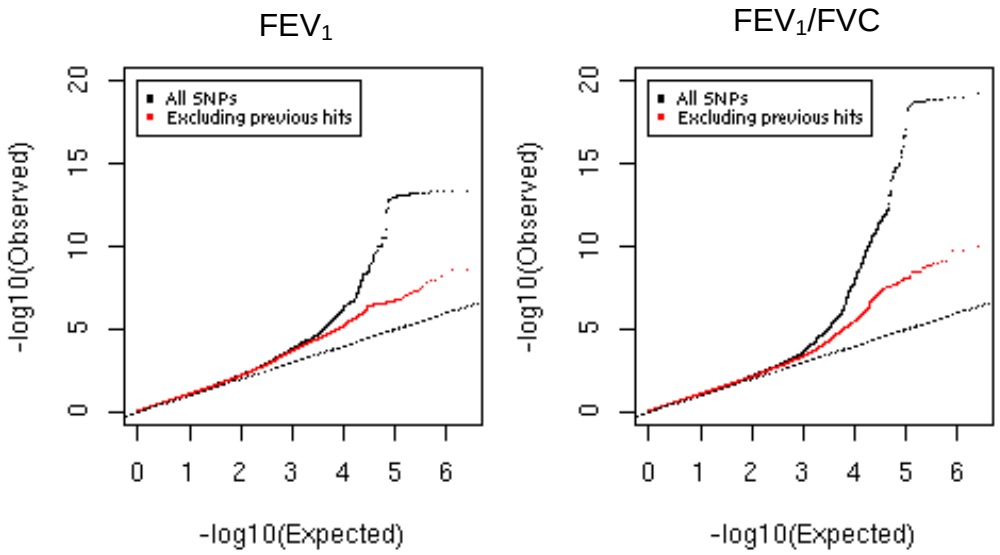


3.3.2 Association analyses of stage 1

The stage 1 (discovery stage) of this analysis consisted of 48,201 individuals of European ancestry in 23 studies, where the association of FEV_1 and FEV_1/FVC with 2.5 million SNPs was tested adjusting for age, age^2, sex, height and ancestry principal components, and stratifying by ever smoking status (details on the method are given in section 3.2.3.1). Once all the issues found in the quality control process were solved, stage 1 study level results were meta-analysed using inverse variance weighted meta-analysis. Details on the strategy followed for the meta-analysis are given in *Meta-analysis*, section 3.2.3.2.

QQ plots of the –log10 P-values expected under the null hypothesis against the –log10 P-values observed in stage 1 meta-analysis (**Figure 3-9)** showed a large deviation from the expected line of slope 1 at the right end of the plot, suggesting that an increased number of significant associations were detected both for FEV_1 and FEV_1/FVC. The deviation from the expected line was still apparent in the QQ plots after the exclusion of regions previously reported (23, 94-96), suggesting that new associations were also identified.

Figure 3-9 QQ plots for FEV_1 and FEV_1/FVC
Black dots represent all SNPs and red dots represent only the SNPs remaining after the exclusion of SNPs in regions previously reported (23, 94, 96) to be associated with lung function (in or near *TNS1*, *PID1*, *FAM13A*, *GSTCD*/*NPNT*, *HHIP*, *HTR4*, *ADAM19*, *AGER*, *DAAM2*, *GPR126*, *PTCH1* and *THSD4*). Regions were defined as 500kb either side of the lead reported SNP.



The genomic inflation factors (λ_{GC}) were 1.12 for FEV_1 and 1.09 for FEV_1/FVC after applying genomic control twice at study level (before and after meta-analysing ever and never-smokers). A final correction for genomic control was also applied at the meta-analysis level. Genomic inflation is known to increase with sample size, as the power to detect genetic associations also increases (130). This has been observed for other traits (127, 131, 132). Genomic inflation factors scaled to a sample size of a 1000 individuals (λ_{GC_1000}) both for FEV_1 and FEV_1/FVC were 1.002, indicating that there was no over inflation of the test statistics.

3.3.3 Results of the quality control checks in stage 2

After the stage 1 meta-analysis, a selection of 34 SNPs in 29 independent regions were taken forward for follow-up in seven studies with *in-silico* data and 10 studies that undertook direct genotyping. Details on the SNP selection and the stage 2 samples are given in sections 3.2.4 and 3.2.5 respectively. Results from these studies were also subject to a thorough quality control process (details on the method followed can be found in section 3.2.6.2).

Study names in this section are not given; a random number has been allocated to all the stage 1 and stage 2 studies, so they are referred to by their number.

3.3.3.1 File formatting

Files were re-formatted when they did not follow the guidance provided in the analysis plan. For one study the column names did not match the column values: the column names included two “Markerid” columns and did not include the “coded allele” column, however the data had the correct fields. A query was sent to the analyst and the correct column names were provided.

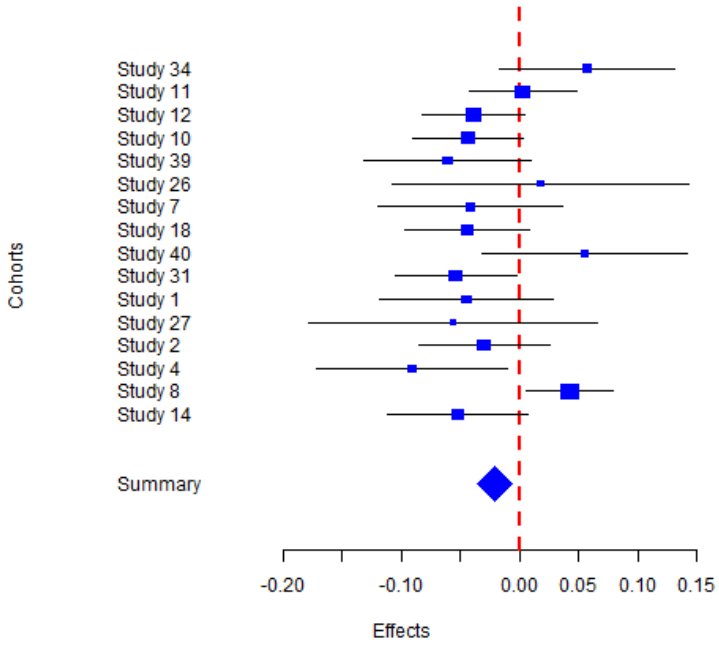
3.3.3.2 Consistency across studies and data quality

Imputation quality for all imputed SNPs included in stage 2 was > 0.6.

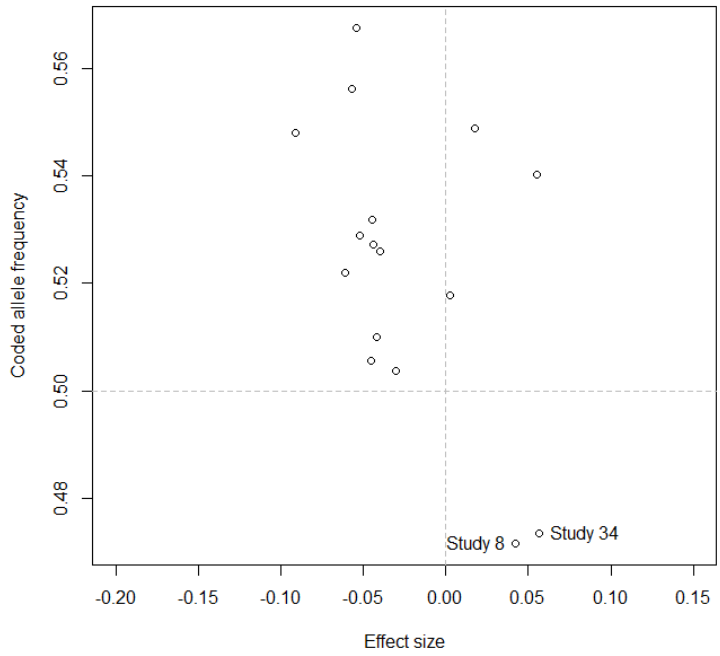
Consistency of direction of effect and of allele frequency across studies was examined, and two issues were found. **Figure 3-10 a)** shows a forest plot for rs2284746, where most of the effect sizes were negative, except for four studies that had positive effect. Two of these four studies (studies 8 and 34) also had allele frequencies < 0.5, whereas the rest had allele frequencies > 0.5 (**Figure 3-10 b)**). Overall, the allele frequencies for this SNP were all around 0.5, so that made harder to rule out that the variation seen was just due to chance. Also, the alleles of this SNP were G and C which meant that it was not possible to detect whether the wrong strand had been reported just by looking at the alleles. The analysts of these two studies were contacted to enquire about these issues.

Figure 3-10 Forest plot and plot of betas vs. allele frequencies for rs2284746 in stage 2

a) Forest plot for rs2284746



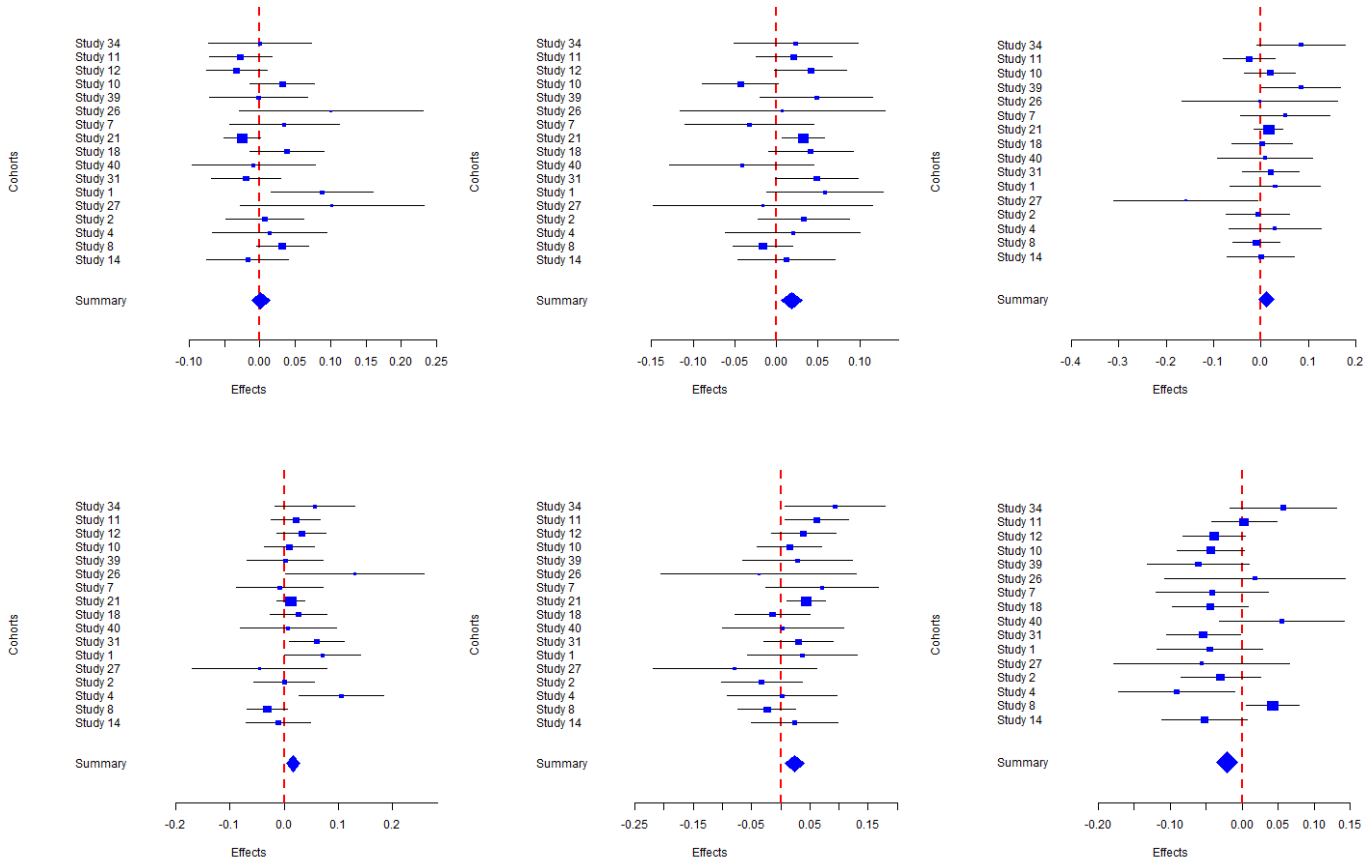
b) Effect sizes vs. allele frequencies for rs2284746

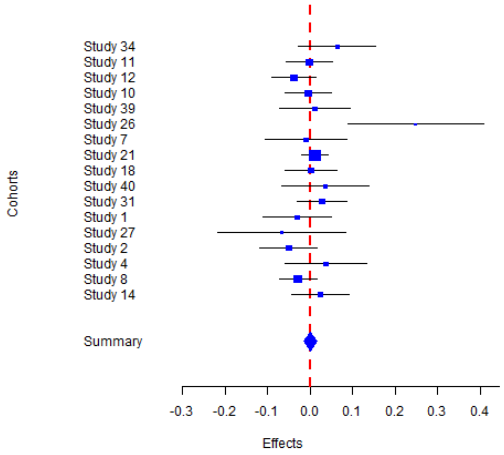


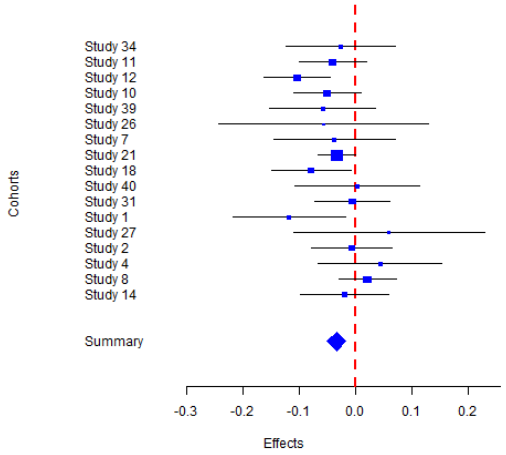
Although study 34 had reported that alleles for all the SNPs were on the forward strand, alleles for rs2284746 were found to be on the reverse strand, so the effect size had to be flipped. Study 8 however, did not find any issue with the allele coding or with the strand. **Figure 3-11** shows forest plots for the SNPs that were analysed in study 8 (results shown just for FEV_1/FVC for simplicity) and the direction of effect goes in the opposite direction to most of the studies in most instances. After sharing this information with the study 8 analysts, they noticed that the ranking of the phenotype values undertaken in the transformation had been done in the wrong direction and that explained the pattern seen in the results. Corrected results were then provided.

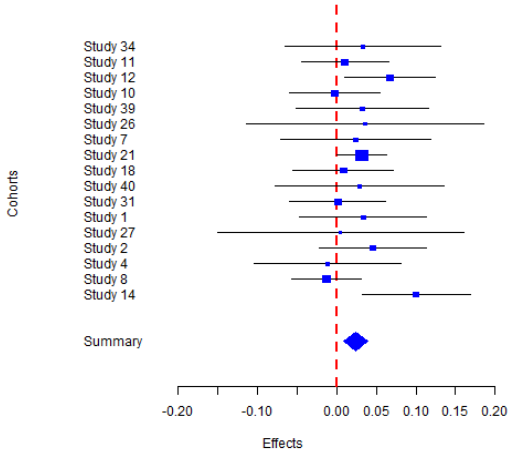
Consistent results were found for the remaining studies.

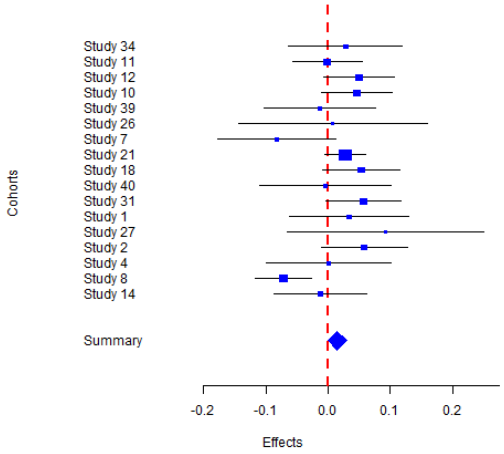
Figure 3-11 Forest plots for FEV_1/FVC for 10 SNPs in stage 2 at an early stage of the quality control process




Cohorts
Study 34
Study 11
Study 12
Study 10
Study 39
Study 26
Study 7
Study 21
Study 18
Study 40
Study 31
Study 1
Study 27
Study 2
Study 4
Study 8
Study 14
Summary
-0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4
Effects


Cohorts
Study 34
Study 11
Study 12
Study 10
Study 39
Study 26
Study 7
Study 21
Study 18
Study 40
Study 31
Study 1
Study 27
Study 2
Study 4
Study 8
Study 14
Summary
-0.3 -0.2 -0.1 0.0 0.1 0.2
Effects


Cohorts
Study 34
Study 11
Study 12
Study 10
Study 39
Study 26
Study 7
Study 21
Study 18
Study 40
Study 31
Study 1
Study 27
Study 2
Study 4
Study 8
Study 14
Summary
-0.20 -0.10 0.00 0.05 0.10 0.15 0.20
Effects


Cohorts
Study 34
Study 11
Study 12
Study 10
Study 39
Study 26
Study 7
Study 21
Study 18
Study 40
Study 31
Study 1
Study 27
Study 2
Study 4
Study 8
Study 14
Summary
-0.2 -0.1 0.0 0.1 0.2
Effects

3.3.4 Combined analysis of stage1 and stage 2 samples

Thirty-four SNPs in 29 independent regions with P-values < 3 x 10^{-6} in stage 1 were followed up in stage 2 in up to 46,411 individuals. After the stage 2 study level results for the 34 SNPs passed the quality control process and the issues found were resolved (clean results for the 34 SNPs are shown in **Appendix C**), stage 2 results were meta-analysed across studies (method in section 3.2.6.2), and then stage1 and stage 2 results were also meta-analysed.

SNPs in 16 regions achieved stage 1 and stage 2 combined P-values below 5 x 10^{-8} (**Table 3-4** and **Figure 3-12**), and 9 of these 16 also showed independent replication in stage 2, reaching a Bonferroni corrected threshold for 34 tests (P < 1.47 × 10^{-3}) (**Table 3-4**). Out of these 16 loci, three of the sentinel SNPs showed the strongest association with FEV_1: in *MECOM* (intron), *ZKSCAN3* (intron)/ *ZNF323* (intron) and *C10orf11* (intron); one locus showed genome-wide significance for FEV_1 and FEV_1/FVC, in *CDC123* (intron); and the remaining 12 loci showed strongest association with FEV_1/FVC in or near: *MFAP2* (intron), *TGFB2* (downstream), *HDAC4* (downstream), *RARB* (intron), *SPATA9* (upstream), *NCR3* (upstream), *ARMC2* (intron), *LRP1* (intron), *CCDC38* (intron), *MMP15* (intron), *CFDP1* (intron) and *KCNE2* (upstream). Region plots and forest plots for these 16 regions are presented in **Appendix D**.

Table 3-4 Results for the 16 new regions associated with lung function

Abbreviations: Chr. =chromosome, freq. = frequency, SE = standard error, P = P-value, N = effective sample size.

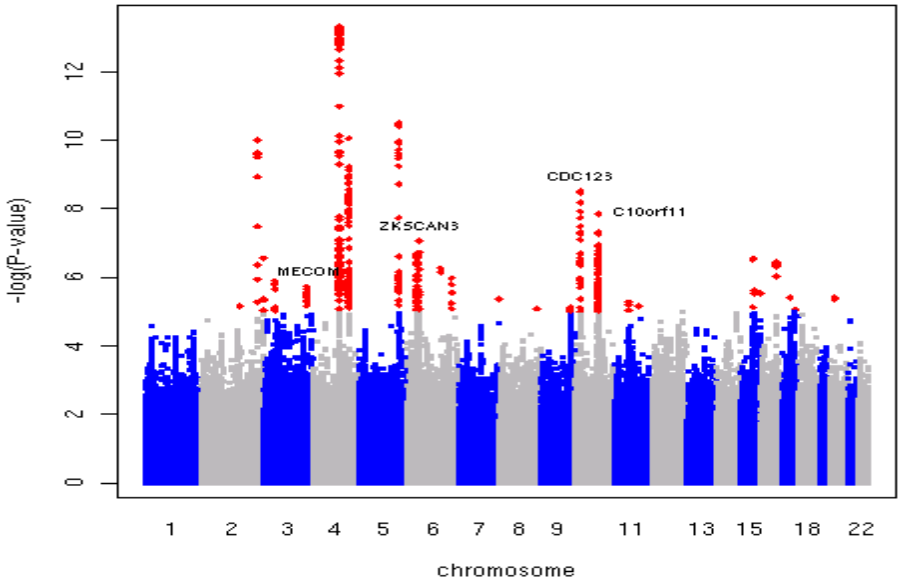
Chr.	SNP_ID (NCBI36 position), Function	Coded allele	Measure	Stage 1				Stage 2				Joint meta-analysis of all stages	
				Beta (SE)	P	Coded allele freq.	N	Beta (SE)	P	Coded allele freq.	N	Beta (SE)	P
1	rs2284746 (17179262), *MFAP2* (intron)	G	FEV_1/FVC	-0.042 (0.007)	$2.47x10^{-9}$	0.516	45944	-0.038 (0.007)	$2.64x10^{-7}$	0.522	35371	-0.04 (0.005)	$7.5x10^{-16}$
			FEV_1	0.008 (0.007)	$2.78x10^{-1}$			0.006 (0.007)	$3.7x10^{-1}$			0.007 (0.005)	$1.48x10^{-1}$
1	rs993925 (216926691), *TGFB2* (downstream)	T	FEV_1/FVC	0.04 (0.007)	$2.54x10^{-7}$	0.308	42402	0.023 (0.01)	$1.76x10^{-2}$	0.348	21414	0.034 (0.006)	$1.16x10^{-8}$
			FEV_1	0.025 (0.007)	$1.51x10^{-3}$			0.003 (0.007)	$7.29x10^{-1}$			0.014 (0.005)	$8.71x10^{-3}$
2	rs12477314 (239542085), *HDAC4* (downstream)	T	FEV_1/FVC	0.052 (0.008)	$4.48x10^{-9}$	0.202	45585	0.031 (0.008)	$8.41x10^{-5}$	0.206	45821	0.041 (0.006)	$1.68x10^{-12}$
			FEV_1	0.032 (0.008)	$2.77x10^{-4}$			0.025 (0.007)	$1.82x10^{-4}$			0.028 (0.005)	$1.02x10^{-7}$
3	rs1529672 (25495586), *RARB* (intron)	C	FEV_1/FVC	-0.06 (0.009)	$7.75x10^{-10}$	0.829	40624	-0.038 (0.009)	$1.16x10^{-5}$	0.831	45466	-0.048 (0.006)	$3.97x10^{-14}$
			FEV_1	-0.037 (0.009)	$1.78x10^{-4}$			-0.011 (0.007)	$9.33x10^{-2}$			-0.02 (0.006)	$2.16x10^{-4}$
3	rs1344555 (170782913), *MECOM* (intron)	T	FEV_1/FVC	-0.019 (0.008)	$2.61x10^{-2}$	0.205	46067	-0.017 (0.012)	$1.55x10^{-1}$	0.209	21313	-0.018 (0.007)	$6.65x10^{-3}$
			FEV_1	-0.042 (0.008)	$1.91x10^{-6}$			-0.025 (0.009)	$6.44x10^{-3}$			-0.034 (0.006)	$2.65x10^{-8}$
5	rs153916 (95062456), *SPATA9* (upstream)	T	FEV_1/FVC	-0.033 (0.007)	$2.06x10^{-6}$	0.552	47530	-0.025 (0.009)	$6.67x10^{-3}$	0.535	21647	-0.031 (0.005)	$2.12x10^{-8}$
			FEV_1	-0.001 (0.007)	$8.91x10^{-1}$			0.004 (0.007)	$6.22x10^{-1}$			0.001 (0.005)	$8.2x10^{-1}$
6	rs6903823 (28430275), *ZKSCAN3* (intron) /	G	FEV_1/FVC	-0.027 (0.008)	$2.28x10^{-3}$	0.209	47057	-0.013 (0.011)	$2.34x10^{-1}$	0.246	21489	-0.021 (0.007)	$1.19x10^{-3}$

Chr.	SNP_ID (NCBI36 position), Function	Coded allele	Measure	Stage 1				Stage 2				Joint meta-analysis of all stages	
				Beta (SE)	P	Coded allele freq.	N	Beta (SE)	P	Coded allele freq.	N	Beta (SE)	P
	ZNF323 (intron)		FEV_1	-0.046 (0.008)	$2x10^{-7}$			-0.029 (0.008)	$4.75x10^{-4}$			-0.037 (0.006)	$2.18x10^{-10}$
6	rs2857595 (31676448), *NCR3* (upstream)	G	FEV_1/FVC	0.049 (0.009)	$7.86x10^{-8}$	0.809	45540	0.028 (0.008)	$5.36x10^{-4}$	0.796	46107	0.037 (0.006)	$2.28x10^{-10}$
			FEV_1	0.04 (0.009)	$1.46x10^{-5}$			0.017 (0.007)	$9.41x10^{-3}$			0.025 (0.005)	$1.3x10^{-6}$
6	rs2798641 (109374743), *ARMC2* (intron)	T	FEV_1/FVC	-0.047 (0.009)	$2.81x10^{-7}$	0.183	46369	-0.03 (0.012)	$1.57x10^{-2}$	0.179	21173	-0.041 (0.007)	$8.35x10^{-9}$
			FEV_1	-0.046 (0.009)	$5.39x10^{-7}$			-0.009 (0.01)	$3.35x10^{-1}$			-0.03 (0.006)	$4.69x10^{-6}$
10	rs7068966 (12317998), *CDC123* (intron)	T	FEV_1/FVC	0.045 (0.007)	$1.28x10^{-10}$	0.519	47085	0.023 (0.006)	$3.86x10^{-4}$	0.518	46067	0.033 (0.005)	$6.13x10^{-13}$
			FEV_1	0.04 (0.007)	$1.19x10^{-8}$			0.022 (0.005)	$3.56x10^{-5}$			0.029 (0.004)	$2.82x10^{-12}$
10	rs11001819 (77985230), *C10orf11* (intron)	G	FEV_1/FVC	-0.019 (0.007)	$6.5x10^{-3}$	0.522	45546	-0.006 (0.006)	$3.17x10^{-1}$	0.506	45932	-0.012 (0.005)	$7.58x10^{-3}$
			FEV_1	-0.041 (0.007)	$1.42x10^{-8}$			-0.022 (0.005)	$3.1x10^{-5}$			-0.029 (0.004)	$2.98x10^{-12}$
12	rs11172113 (55813550), *LRP1* (intron)	T	FEV_1/FVC	-0.035 (0.007)	$1.36x10^{-6}$	0.607	45387	-0.026 (0.01)	$5.83x10^{-3}$	0.59	20509	-0.032 (0.006)	$1.24x10^{-8}$
			FEV_1	-0.021 (0.007)	$3.55x10^{-3}$			-0.003 (0.007)	$6.94x10^{-1}$			-0.013 (0.005)	$1.19x10^{-2}$
12	rs1036429 (94795559), *CCDC38* (intron)	T	FEV_1/FVC	0.049 (0.008)	$1.24x10^{-8}$	0.2	47814	0.028 (0.008)	$3.35x10^{-4}$	0.214	46311	0.038 (0.006)	$2.3x10^{-11}$
			FEV_1	0.01 (0.008)	$2.67x10^{-1}$			0.004 (0.006)	$5.38x10^{-1}$			0.006 (0.005)	$2.26x10^{-1}$
16	rs12447804 (56632783), *MMP15* (intron)	T	FEV_1/FVC	-0.053 (0.009)	$7.12x10^{-8}$	0.208	35123	-0.021 (0.01)	$4.2x10^{-2}$	0.222	24398	-0.038 (0.007)	$3.59x10^{-8}$
			FEV_1	-0.017 (0.009)	$8.02x10^{-2}$			0.004 (0.007)	$5.71x10^{-1}$			-0.004 (0.006)	$4.73x10^{-1}$

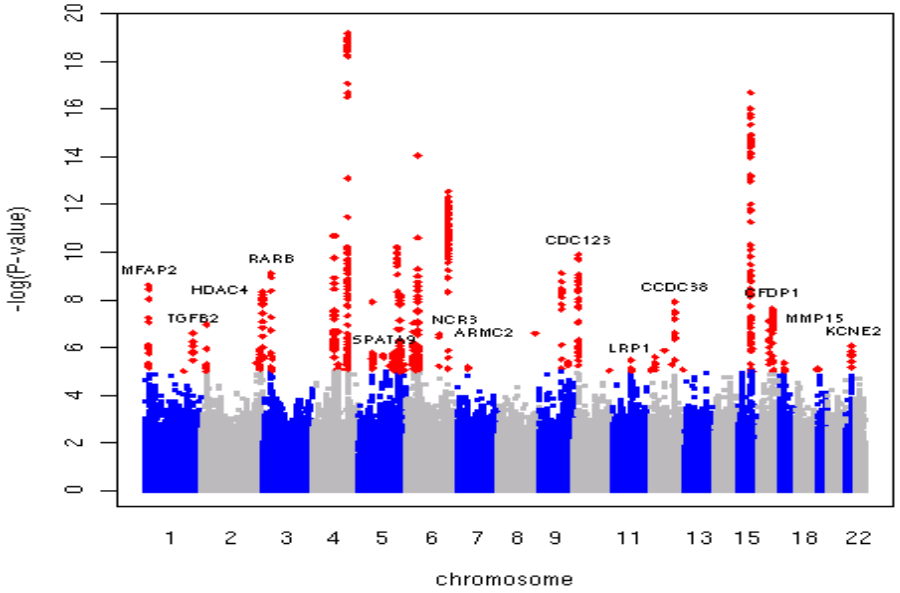
Chr.	SNP_ID (NCBI36 position), Function	Coded allele	Measure	Stage 1				Stage 2				Joint meta-analysis of all stages	
				Beta (SE)	P	Coded allele freq.	N	Beta (SE)	P	Coded allele freq.	N	Beta (SE)	P
16	rs2865531 (73947817), *CFDP1* (intron)	T	FEV_1/FVC	0.039 (0.007)	$2.3x10^{-8}$	0.418	47594	0.024 (0.006)	$1.94x10^{-4}$	0.409	46304	0.031 (0.005)	$1.77x10^{-11}$
			FEV_1	0.024 (0.007)	$6.3x10^{-4}$			0.011 (0.005)	$3.89x10^{-2}$			0.016 (0.004)	$1.09x10^{-4}$
21	rs9978142 (34574109), *KCNE2* (upstream)	T	FEV_1/FVC	-0.048 (0.009)	$8.23x10^{-7}$	0.156	44577	-0.031 (0.013)	$1.75x10^{-2}$	0.149	20944	-0.043 (0.008)	$2.65x10^{-8}$
			FEV_1	-0.012 (0.009)	$2.47x10^{-1}$			-0.015 (0.01)	$1.35x10^{-1}$			-0.013 (0.007)	$5.57x10^{-2}$

Figure 3-12 Manhattan plots for FEV_1 and FEV_1/FVC
SNPs with −log10 P > 5 are indicated in red. Newly associated regions that reached genome-wide significance after meta-analysis of stages 1 and 2 are labelled.

a) Manhattan plot for FEV_1



b) Manhattan plot for FEV_1/FVC



Chi-square heterogeneity tests were undertaken for effect sizes across stage 1 and stage 2 studies for the 16 novel SNPs. None of the SNPs reached statistical significance for heterogeneity after applying a Bonferroni correction for 16 tests ($P = 0.05 / 16 = 3.13 \times 10^{-3}$), although a limited number of SNPs had P < 0.05 (rs11001819 in *C10orf11* for FEV_1; rs153916 in *SPATA9* and rs9978142 in *KCNE2* for FEV_1/FVC) (**Table 3-5**). The forest plots presented in **Appendix D** illustrate however, that most of the studies have consistent direction of effect for these SNPs and that the associations described here do not seem to be driven just by a small number of studies.

Table 3-5 Chi-square heterogeneity test results for the 16 new regions associated with lung function
Abbreviations: P = P-value, d.f. = degrees of freedom

SNP ID	Gene(function)	Measure	Stage1 & stage2		
			Chi squared	P	d.f.
rs1036429	*CCDC38* (intron)	FEV_1/FVC	39.334	$4.55x10^{-1}$	39
rs11001819	*C10orf11* (intron)	FEV_1	56.046	$3.8x10^{-2}$	39
rs11172113	*LRP1* (intron)	FEV_1/FVC	26.303	$6.6x10^{-1}$	30
rs12447804	*MMP15* (intron)	FEV_1/FVC	19.209	$7.41x10^{-1}$	24
rs12477314	*HDAC4* (downstream)	FEV_1/FVC	41.187	$3.75x10^{-1}$	39
rs1344555	*MECOM* (intron)	FEV_1	28.977	$5.19x10^{-1}$	30
rs1529672	*RARB* (intron)	FEV_1/FVC	42.577	$3.2x10^{-1}$	39
rs153916	*SPATA9* (upstream)	FEV_1/FVC	50.318	$1.1x10^{-2}$	30
rs2284746	*MFAP2* (intron)	FEV_1/FVC	35.254	$5.97x10^{-1}$	38
rs2798641	*ARMC2* (intron)	FEV_1/FVC	31.217	$4.05x10^{-1}$	30
rs2857595	*NCR3* (upstream)	FEV_1/FVC	40.503	$4.04x10^{-1}$	39
rs2865531	*CFDP1* (intron)	FEV_1/FVC	40.387	$4.09x10^{-1}$	39
rs6903823	*ZKSCAN3* (intron)/*ZNF323* (intron)	FEV_1	23.72	$7.85x10^{-1}$	30
rs7068966	*CDC123* (intron)	FEV_1/FVC	39.482	$4.48x10^{-1}$	39
rs7068966	*CDC123* (intron)	FEV_1	31.199	$8.09x10^{-1}$	39
rs993925	*TGFB2* (downstream)	FEV_1/FVC	35.13	$2.38x10^{-1}$	30
rs9978142	*KCNE2* (upstream)	FEV_1/FVC	45.681	$3.3x10^{-2}$	30

3.3.5 Plausible pathways for lung function involving new loci

The most significant signal was an intronic SNP (rs2284746) in *MFAP2*. This gene encodes a major antigen of elastin-associated microfibrils (133) that might be involved in the causation of inherited connective tissue diseases (134). Another potential candidate to influence lung function in this region is an intronic SNP (rs7513616) with an r^2 of 0.42 with the sentinel SNP located in *CROCC*, which encodes rootletin, a component of cilia (135). Experiments suggest impaired mucociliary clearance in *CROCC* knockout mice (136).

The next most significant signal was found in Retinoic Acid Receptor Beta (*RARB*), which is a vitamin A metabolite receptor. This receptor also controls cell proliferation and differentiation. Retinoic acid has been implicated in embryonic lung branching morphogenesis (137); and the epigenetic regulation of the *RARB* gene promoter has been linked to various cancers including non-small cell lung cancer (138).

The third most significant signal, the only one genome-wide significant both for FEV_1 and FEV_1/FVC, was found in *CDC123*, a cell division cycle protein 123 homolog. Its homolog in yeast plays a role in regulating eukaryotic initiation factor 2 in times of cell stress (139).

The following strongest association was for a SNP (rs12477314) downstream of *HDAC4*, a gene that codes for histone deacetylase 4, and could possibly repress gene transcription. It has been shown that COPD patients have a reduction in histone deacetylase activity (140).

3.3.6 Additional analyses

Associations in stage 1 of SNPs previously associated with lung function

Effects in stage 1 of loci previously reported to be associated with FEV_1, FEV_1/FVC, or COPD (23, 94-96) providing that they also showed association with lung function, were assessed for both lung function measures (**Table 3-6**). Details of the selection of these SNPs are shown in section 3.2.8. Ten regions (*TNS1*, *FAM13A*, *GSTCD-NPNT*, *HHIP*, *HTR4*, *ADAM19*, *AGER*, *GPR126*, *PTCH1* and *TSHD4*) previously reported to be associated with lung function (23, 94-96) reached genome-wide significance in stage 1 (**Table 3-6**). *DAAM2*, *PID1* and *CHRNA3/5* reprted in (23, 95, 96), did not reach genome-wide significance (**Table 3-6**).

Table 3-6 Lung function associations (FEV_1 and FEV_1/FVC) in stage 1 for all previously reported loci
Abbreviations: ns = nonsynonymous, s = synonymous., SE = standard error, P = P-value, N = effective sample sizes.

Chr.	Paper reported	Measure	SNP ID (NCBI36 position), function	Gene reported	Coded allele	FEV_1		FEV_1/FVC		N
						Beta (SE)	P	Beta (SE)	P	
2	Repapi *et al.*	FEV_1	rs2571445 (218391399), *TNS1* (ns)	*TNS1*	G	0.047 (0.007)	$9.83x10^{-11}$	0.033 (0.007)	$4.46x10^{-6}$	45839
2	Hancock *et al.*	FEV_1/FVC	rs10498230 (229210747), *PID1* (downstream)	*PID1*	T	0.03 (0.014)	$3.6x10^{-2}$	0.068 (0.014)	$1.13x10^{-6}$	44957
4	Hancock *et al.*	FEV_1/FVC	rs2869967 (90088355), *FAM13A* (intron)	*FAM13A*	T	0.012 (0.007)	$9.38x10^{-2}$	0.047 (0.007)	$2.08x10^{-11}$	47710
4	SpiroMeta-CHARGE	FEV_1/FVC	rs2045517 (90089987), *FAM13A* (intron)	*FAM13A*	T	-0.012 (0.007)	$8.93x10^{-2}$	-0.047 (0.007)	$2x10^{-11}$	47675
4	Cho *et al.*	COPD	rs7671167 (90103002), *FAM13A* (intron)	*FAM13A*	T	-0.017 (0.007)	$1.64x10^{-2}$	-0.042 (0.007)	$1.27x10^{-9}$	47723
4	Repapi *et al.*	FEV_1	rs10516526 (106908353), *GSTCD* (intron)	*GSTCD-NPNT*	G	0.108 (0.014)	$4.75x10^{-14}$	0.039 (0.014)	$6.17x10^{-3}$	47970
4	Hancock *et al.*	FEV_1	rs17331332 (107027556), *NPNT* (upstream)	*GSTCD-NPNT*	G	-0.102 (0.014)	$1.11x10^{-12}$	-0.057 (0.014)	$5.3x10^{-5}$	39503
4	SpiroMeta-CHARGE	FEV_1/FVC	rs6823809 (107048244), *NPNT* (intron)	*GSTCD-NPNT*	T	0.050 (0.011)	$4.82x10^{-6}$	0.056 (0.011)	$2.2x10^{-7}$	23656
4	SpiroMeta-CHARGE	FEV_1	rs1032296 (145654138), *HHIP* (upstream)	*HHIP*	T	-0.047 (0.007)	$8.74x10^{-11}$	-0.050 (0.007)	$3.42x10^{-12}$	45318
4	Repapi *et al.*	FEV_1/FVC	rs12504628 (145655774), *HHIP* (upstream)	*HHIP*	T	-0.044 (0.007)	$1.03x10^{-9}$	-0.063 (0.007)	$5.54x10^{-19}$	46204
4	Wilk *et al.*	FEV_1/FVC	rs11100860 (145698589), *HHIP* (upstream)	*HHIP*	G	0.041 (0.007)	$4.27x10^{-9}$	0.064 (0.007)	$6.81x10^{-20}$	47876
4	Hancock *et al.*	FEV_1/FVC	rs1980057 (145705188), *HHIP* (upstream)	*HHIP*	T	0.042 (0.007)	$4.07x10^{-9}$	0.063 (0.007)	$1.06x10^{-19}$	47865
5	Hancock *et al.*	FEV_1/FVC	rs11168048 (147822546), *HTR4* (intron)	*HTR4*	T	-0.046 (0.007)	$2.43x10^{-10}$	-0.047 (0.007)	$5.97x10^{-11}$	44976
5	Repapi *et al.*	FEV_1	rs3995090 (147826008), *HTR4* (intron)	*HTR4*	C	0.045 (0.007)	$3.33x10^{-10}$	0.046 (0.007)	$1.04x10^{-10}$	47607

Chr.	Paper reported	Measure	SNP ID (NCBI36 position), function	Gene reported	Coded allele	FEV_1		FEV_1/FVC		N
						Beta (SE)	P	Beta (SE)	P	
5	SpiroMeta-CHARGE	FEV_1	rs1985524 (147827981), *HTR4* (intron)	*HTR4*	G	-0.048 (0.007)	$3.06x10^{-11}$	-0.045 (0.007)	$2.9x10^{-10}$	45623
5	Hancock *et al.*	FEV_1/FVC	rs2277027 (156864954), *ADAM19* (intron)	*ADAM19*	C	-0.026 (0.007)	$3.1x10^{-4}$	-0.042 (0.007)	$6.65x10^{-9}$	48023
5	SpiroMeta-CHARGE	FEV_1/FVC	rs11134779 (156869344), *ADAM19* (intron)	*ADAM19*	G	-0.027 (0.007)	$2.4x10^{-4}$	-0.042 (0.007)	$6.01x10^{-9}$	48075
6	Hancock *et al.* and Repapi *et al.*	FEV_1/FVC	rs2070600 (32259421), *AGER* (ns)	*AGER*	T	0.025 (0.016)	$1.27x10^{-1}$	0.126 (0.016)	$9.07x10^{-15}$	46314
6	Repapi *et al.*	FEV_1/FVC	rs2395730 (39892343), *DAAM2* (intron)	*DAAM2*	C	-0.004 (0.007)	$5.95x10^{-1}$	0.022 (0.007)	$1.39x10^{-3}$	47256
6	SpiroMeta-CHARGE	FEV_1/FVC	rs11756622 (39898021), *DAAM2* (intron)	*DAAM2*	T	0.047 (0.019)	$1.23x10^{-2}$	0.064 (0.019)	$5.48x10^{-4}$	28276
6	Hancock *et al.*	FEV_1/FVC	rs3817928 (142792209), *GPR126* (intron)	*GPR126-LOC153910*	G	0.023 (0.009)	$8.63x10^{-3}$	0.059 (0.008)	$2.27x10^{-12}$	46730
6	SpiroMeta-CHARGE	FEV_1/FVC	rs262129 (142894837), *LOC153910* (unknown)	*GPR126-LOC153910*	G	0.031 (0.008)	$5.44x10^{-5}$	0.056 (0.008)	$2.91x10^{-13}$	47014
9	SpiroMeta-CHARGE	FEV_1/FVC	rs16909859 (97244613), *PTCH1* (downstream)	*PTCH1*	G	-0.014 (0.013)	$2.93x10^{-1}$	0.08 (0.013)	$7.45x10^{-10}$	43353
9	Hancock *et al.*	FEV_1/FVC	rs16909898 (97270829), *PTCH1* (intron)	*PTCH1*	G	0.015 (0.012)	$2.21x10^{-1}$	-0.072 (0.012)	$3.94x10^{-9}$	42486
15	Repapi *et al.*	FEV_1/FVC	rs12899618 (69432174), *THSD4* (intron)	*THSD4*	G	0.036 (0.01)	$1.57x10^{-4}$	0.076 (0.01)	$1.86x10^{-15}$	46657
15	SpiroMeta-CHARGE	FEV_1/FVC	rs8033889 (69467134), *THSD4* (intron)	*THSD4*	T	-0.044 (0.009)	$3.01x10^{-7}$	-0.072 (0.008)	$2.03x10^{-17}$	46995
15	DeMeo *et al* (2009)	COPD	rs2568494 (76528019), *IREB2* (intron)	*CHRNA3-CHRNA5-IREB2-LOC123688*	G	0.023 (0.007)	$1.64x10^{-3}$	0.029 (0.007)	$5.25x10^{-5}$	47919
15	Pillai *et al.* (2009)	COPD	rs8034191 (76593078), *LOC123688* (intron)	*CHRNA3-CHRNA5-IREB2-LOC123688*	T	0.031 (0.007)	$2.07x10^{-5}$	0.032 (0.007)	$9.65x10^{-6}$	47954
15	SpiroMeta-CHARGE	FEV_1	rs2036527 (76638670), *CHRNA5* (upstream)	*CHRNA3-CHRNA5-IREB2-LOC123688*	G	0.036 (0.008)	$2.4x10^{-6}$	0.032 (0.007)	$1.19x10^{-5}$	45038

Chr.	Paper reported	Measure	SNP ID (NCBI36 position), function	Gene reported	Coded allele	FEV_1		FEV_1/FVC		N
						Beta (SE)	P	Beta (SE)	P	
15	SpiroMeta-CHARGE	FEV_1/FVC	rs8040868 (76698236), *CHRNA3* (s)	*CHRNA3-CHRNA5-IREB2-LOC123688*	T	0.039 (0.008)	$2.98x10^{-6}$	0.04 (0.008)	$1.14x10^{-6}$	35121

Association with lung function in children

The effects of the 16 new variants, as well as the 12 previously reported variants for lung function and COPD (23, 95, 96) (details of the selection of these variants in section 3.2.8), were assessed in two children's cohorts: the Avon Longitudinal Study of Parents and Children (ALSPAC) (141) and the Raine Study (142-144), with a joint sample size of 6,281 individuals aged between 7 and 9.

Out of the 16 new loci associated with lung function, 11 showed consistent direction of effect in children (**Appendix C**), and out of the 12 regions previously discovered 11 had consistent direction of effects (**Appendix C**). To compare direction of effects for a locus the direction of effect of the most significant SNP in the SpiroMeta-CHARGE dataset across both traits was used.

Association of lung function loci with height

The effect of the 16 new loci and of 12 previously discovered loci (23, 95, 96) on height was assessed by looking up their association results in the GIANT consortium dataset (N = 183,727) (127). After applying a Bonferroni correction for 28 tests ($P < 1.8 \times 10^{-3}$) the sentinel SNP in *MFAP2* (rs2284746) showed a significant association with height ($P = 5.64 \times 10^{-15}$), although with different direction of effect; the allele associated with increased height was also associated with decreased FEV_1/FVC (**Appendix C**). Three of the previously

discovered loci known to be associated with height (127, 145) (*HHIP*, *PTCH1* and *GPR126*) also showed significant P-values after applying the Bonferroni correction for 28 tests (**Appendix C**). The direction of effect for *HHIP* was the same for height and FEV_1/FVC, however for *PTCH1* and *GPR126* it was in the opposite direction (**Appendix C**).

Association of lung function loci with smoking

Smoking is a major risk factor for developing COPD and it is known to severely affect lung function. For this reason the analyses undertaken here were stratified by ever smoking status, however it was not possible to adjust for amount smoked, since there was not enough information available in all studies. To further investigate whether the association of the 16 novel regions or any of the 12 previously reported regions (23, 95, 96) might be mediated via smoking behaviour, the association of these variants with two smoking phenotypes was assessed in the Oxford-GlaxoSmithKline (Ox-GSK) study: ever smoking status (N = 33,639) and number of cigarettes smoked per day (N = 15,574) (109). None of the 16 novel regions showed a significant association for either trait using a Bonferroni correction for 28 tests ($P < 1.8 \times 10^{-3}$); the lowest P-value for either trait was 3.7×10^{-2} (**Appendix C**). None of the previously reported regions showed even nominal significance (P < 0.05) in their association with ever smoking status (**Appendix C)**. The *CHRNA3/5* locus, associated with nicotine dependence (146, 147), was the top signal in the Ox-GSK meta-analysis (109), and as expected showed a strong association with number of cigarettes per day

($P < 3 \times 10^{-15}$) (**Appendix C**). The *PID1* locus showed evidence of association with number of cigarettes per day ($P = 1.6 \times 10^{-3}$) (**Appendix C**), which would pass a Bonferroni corrected threshold for 28 tests ($P < 1.8 \times 10^{-3}$).

To assess whether the 16 new associations might have arisen due to a gene by smoking interaction, effect sizes were calculated separately in ever and in never-smokers and compared by testing whether they differed substantially. None of the loci showed significant interaction with ever smoking status after applying a Bonferroni correction for 16 tests ($P < 3.1 \times 10^{-3}$) (**Table 3-7**).

Table 3-7 Associations in never-smokers and ever-smokers in the joint meta-analysis of stage 1 and 2 data, and tests for interaction with smoking
Abbreviations: Chr. = chromosome, SE = standard error, P = P-value.

Chr.	SNP_ID (NCBI36 position), function	Measure	Joint meta-analysis of all stages				Interaction
			Ever-smokers		Never-smokers		
			Beta	SE	Beta	SE	P
1	rs2284746 (17179262), *MFAP2* (intron)	FEV_1/FVC	-0.043	0.007	-0.036	0.007	$5.12x10^{-1}$
1	rs993925 (216926691), *TGFB2* (downstream)	FEV_1/FVC	0.041	0.008	0.026	0.009	$1.91x10^{-1}$
2	rs12477314 (239542085), *HDAC4* (downstream)	FEV_1/FVC	0.048	0.008	0.032	0.008	$1.88x10^{-1}$
3	rs1529672 (25495586), *RARB* (intron)	FEV_1/FVC	-0.059	0.009	-0.033	0.009	$4.29x10^{-2}$
3	rs1344555 (170782913), *MECOM* (intron)	FEV_1	-0.040	0.009	-0.029	0.009	$3.81x10^{-1}$
5	rs153916 (95062456), *SPATA9* (upstream)	FEV_1/FVC	-0.035	0.008	-0.024	0.008	$3.28x10^{-1}$
6	rs6903823 (28430275), *ZKSCAN3* (intron)/ *ZNF323* (intron)	FEV_1	-0.038	0.008	-0.037	0.008	$9.64x10^{-1}$
6	rs2857595 (31676448), *NCR3* (upstream)	FEV_1/FVC	0.043	0.008	0.031	0.009	$3.11x10^{-1}$
6	rs2798641 (109374743), *ARMC2* (intron)	FEV_1/FVC	-0.050	0.010	-0.030	0.010	$1.67x10^{-1}$
10	rs7068966 (12317998), *CDC123* (intron)	FEV_1/FVC	0.041	0.006	0.024	0.007	$7.15x10^{-2}$
10	rs11001819 (77985230), *C10orf11* (intron)	FEV_1	-0.026	0.006	-0.031	0.006	$5.56x10^{-1}$
12	rs11172113 (55813550), *LRP1* (intron)	FEV_1/FVC	-0.035	0.008	-0.029	0.008	$5.97x10^{-1}$
12	rs1036429 (94795559), *CCDC38* (intron)	FEV_1/FVC	0.044	0.008	0.033	0.008	$3.45x10^{-1}$
16	rs12447804 (56632783), *MMP15* (intron)	FEV_1/FVC	-0.045	0.010	-0.030	0.010	$2.71x10^{-1}$
16	rs2865531 (73947817), *CFDP1* (intron)	FEV_1/FVC	0.034	0.006	0.028	0.007	$5.42x10^{-1}$
21	rs9978142 (34574109), *KCNE2* (upstream)	FEV_1/FVC	-0.052	0.011	-0.032	0.011	$1.94x10^{-1}$

Association of lung function loci with lung cancer

Effects of the 16 novel SNPs and the 12 previously reported SNPs associated with lung function or COPD (23, 95, 96) on lung cancer were assessed in the International Lung Cancer Consortium (ILCCO) GWAS meta-analysis (128). The ILCCO GWAS meta-analysis (13,300 cases and 19,666 controls) only had data on directly genotyped SNPs, for this reason proxy SNPs were given when the top SNP was not included in their data. No proxy SNPs were available for *TGFB2*, therefore only the associations of 27 loci with lung cancer were tested.

Out of the 15 new regions tested, SNPs in two loci (*ZKSCAN3*/ *ZNF323* and *NCR3*) that are in linkage disequilibrium ($r^2 > 0.6$) with the sentinel SNP in each region were significantly associated with lung cancer ($P < 4 \times 10^{-5}$) after applying a Bonferroni correction for 28 tests ($P < 1.8 \times 10^{-3}$), and had consistent direction of effect with lung function, the alleles associated with reduced lung function were also associated with increased risk of developing lung cancer (**Appendix C**). Out of the previously reported regions, *CHRNA3/5* known to be associated with lung cancer (147-149) showed a very strong association ($P < 2.2 \times 10^{-46}$) in the ILCCO dataset, and also showed consistent direction of effects for the lung function (**Appendix C**).

Proportion of variance explained by loci discovered to date

The association of 10 previously discovered regions (*TNS1*, *FAM13A*, *GSTCD-NPNT*, *HHIP*, *HTR4*, *ADAM19*, *AGER*, *GPR126*, *PTCH1* and *THSD4*) with lung function was confirmed in stage1 of the SpiroMeta-CHARGE meta-analysis (**Table 3-6**). In addition to these 10 variants, 16 further variants reached genome-wide significance in stage 1 and stage 2 combined, bringing the total number of loci associated with lung function at genome-wide significant levels in the SpiroMeta-CHARGE dataset to 26. Jointly these 26 variants explained 1.5% of the additive polygenic variance of FEV_1 and 3.2% of the additive polygenic variance of FEV_1/FVC. Methods have been developed in order to estimate the number of undetected variants with similar effect sizes to those identified in a GWAS and calculate the proportion of the variance of a given trait that both the discovered and the estimated number of undetected variants would explain (129). This method (details in section 3.2.8) was applied and it estimated that there were a total of 102 (95% confidence interval 57-155) independent variants, including the 26 reported and 76 putative additional variants of similar effect sizes. In aggregate the 102 variants would explain 3.4% of the additive polygenic variance for FEV_1 and 7.5% of the additive polygenic variance for FEV_1/FVC (**Table 3-8**).

Table 3-8 Estimated number of undiscovered variants and proportion of variance explained
Effect sizes and standard errors estimated using non-discovery data are shown for genome-wide significant loci in SpiroMeta-CHARGE stage 1 or stage1 + 2 data. Abbreviations: Chr. = chromosome, N = effective sample sizes, SE = standard deviation.

Chr.	SNP ID (NCBI36 position), function	FEV_1 excluding winners' curse bias		FEV_1/FVC excluding winners' curse bias		N	Power	Estimated number of variants of similar effect	R2 (%) FEV_1	R2 (%) FEV_1/FVC
		Beta	SE	Beta	SE					
1	rs2284746 (17179262), *MFAP2* (intron)	0.006	0.007	-0.038	0.007	35371[n2]	0.707	1.4	0.002	0.072
1	rs993925 (216926691), *TGFB2* (downstream)	0.003	0.007	0.023	0.01	21414[n2]	0.214	4.7	0	0.024
2	rs2571445 (218391399), *TNS1* (ns)	0.041	0.009	0.034	0.009	29130[s]	0.863	1.2	0.082	0.055
2	rs12477314 (239542085), *HDAC4* (downstream)	0.025	0.007	0.031	0.008	45821[n2]	0.341	2.9	0.02	0.031
3	rs1529672 (25495586), *RARB* (intron)	-0.011	0.007	-0.038	0.009	45466[n2]	0.376	2.7	0.003	0.041
3	rs1344555 (170782913), *MECOM* (intron)	-0.025	0.009	-0.017	0.012	21313[n2]	0.207	4.8	0.021	0.01
4	rs2045517 (90089987), *FAM13A* (intron)	-0.006	0.009	-0.037	0.009	25736[c]	0.654	1.5	0.002	0.067
4	rs10516526 (106908353), *GSTCD* (intron)	0.07	0.034	0.035	0.033	7587[sc]	0.627	1.6	0.062	0.016
4	rs11100860 (145698589), *HHIP* (upstream)	0.042	0.008	0.058	0.007	40202[f]	0.996	1	0.085	0.163
5	rs153916 (95062456), *SPATA9* (upstream)	0.004	0.007	-0.025	0.009	21647[n2]	0.252	4	0.001	0.031
5	rs1985524 (147827981), *HTR4* (intron)	-0.047	0.017	-0.052	0.017	7204[sc]	0.961	1	0.107	0.134
5	rs11134779 (156869344), *ADAM19* (intron)	-0.03	0.01	-0.023	0.01	25917[c]	0.495	2	0.04	0.024
6	rs6903823 (28430275), *ZKSCAN3* (intron)/ZNF323 (intron)	-0.029	0.008	-0.013	0.011	21489[n2]	0.248	4	0.031	0.006
6	rs2857595 (31676448), *NCR3* (upstream)	0.017	0.007	0.028	0.008	46107[n2]	0.129	7.8	0.009	0.025
6	rs2070600 (32259421), *AGER* (ns)	0.012	0.04	0.093	0.04	7226[sc]	0.695	1.4	0.001	0.084

Chr.	SNP ID (NCBI36 position), function	FEV_1 excluding winners' curse bias		FEV_1/FVC excluding winners' curse bias		N	Power	Estimated number of variants of similar effect	R2 (%) FEV_1	R2 (%) FEV_1/FVC
		Beta	SE	Beta	SE					
6	rs2798641 (109374743), *ARMC2* (intron)	-0.009	0.01	-0.03	0.012	21173[n2]	0.214	4.7	0.002	0.026
6	rs262129 (142894837), *LOC153910* (unknown)	0.014	0.01	0.045	0.01	25317[c]	0.319	3.1	0.008	0.081
9	rs16909859 (97244613), *PTCH1* (downstream)	-0.021	0.018	0.062	0.017	22923[c]	0.539	1.9	0.006	0.058
10	rs7068966 (12317998), *CDC123* (intron)	0.022	0.005	0.023	0.006	46067[n2]	0.209	4.8	0.024	0.026
10	rs11001819 (77985230), *C10orf11* (intron)	-0.022	0.005	-0.006	0.006	45932[n2]	0.108	9.3	0.024	0.002
12	rs11172113 (55813550), *LRP1* (intron)	-0.003	0.007	-0.026	0.01	20509[n2]	0.292	3.4	0	0.033
12	rs1036429 (94795559), *CCDC38* (intron)	0.004	0.006	0.028	0.008	46311[n2]	0.204	4.9	0.001	0.026
15	rs8033889 (69467134), *THSD4* (intron)	-0.039	0.011	-0.072	0.011	28974[s]	0.996	1	0.05	0.174
16	rs12447804 (56632783), *MMP15* (intron)	0.004	0.007	-0.021	0.01	24398[n2]	0.059	16.8	0.001	0.015
16	rs2865531 (73947817), *CFDP1* (intron)	0.011	0.005	0.024	0.006	46304[n2]	0.175	5.7	0.006	0.028
21	rs9978142 (34574109), *KCNE2* (upstream)	-0.015	0.01	-0.031	0.013	20944[n2]	0.253	3.9	0.006	0.024
Total variants								101.5		
Total % variance explained by estimated variants									1.355	3.016
Total % polygenic variance explained by estimated variants									3.388	7.538

[n2] no exclusions in SpiroMeta-CHARGE stage 2
[s] excluding SpiroMeta consortium discovery GWAS of Repapi et al. (2010)
[c] excluding CHARGE consortium discovery GWAS of Hancock et al. (2010)
[sc] excluding SpiroMeta consortium discovery GWAS of Repapi et al. (2010) and CHARGE consortium discovery GWAS of Hancock et al. (2010)
[f] excluding FHS from SpiroMeta-CHARGE stage 1

3.4 SpiroMeta-CHARGE meta-analysis of GWAS: discussion

This chapter presents a meta-analysis of 23 genome-wide association studies of lung function including 48,201 individuals, with a follow-up stage of 17 studies and up to 46,411 individuals undertaken for 29 independent regions. Previously reported associations with lung function were evaluated in the genome-wide results and the association of 10 regions (*TNS1*, *FAM13A*, *GSTCD-NPNT*, *HHIP*, *HTR4*, *ADAM19*, *AGER*, *GPR126*, *PTCH1* and *THSD4*) was confirmed with P-values reaching genome-wide significance ($P < 5 \times 10^{-8}$). Out of the 29 independent regions that were followed up, 16 achieved genome-wide significance after meta-analysing stage 1 and stage 2. All these novel loci are located in genomic regions that had not been previously related to lung function and they have the potential to highlight new molecular pathways that would improve our understanding of lung health.

The 16 new variants bring the total number of loci with strong evidence of association with lung function to 26. However, these 26 variants combined with another 76 putative (as yet undiscovered) variants estimated (129) to have similar effect sizes, only explain a small proportion (7.5% for FEV_1/FVC) of the additive polygenic variance of the lung function measures. This is consistent with findings for other complex traits (13). It is likely that part of the heritability not yet explained is accounted for by additional common variants of small effect sizes. A recently developed approach estimates that around 45% of the variance for height would be explained by all common genetic variants captured

in GWAS chips (150). It would be of interest to apply the same approach to estimate the proportion of the variance of the lung function traits that would be explained by common variants, however it requires individual level data genome-wide which were not available in this study. Studies with larger sample sizes will be required in order to identify these common variants with small effect sizes. Rare variants are expected to have larger effect sizes and the discovery of these would also add to the proportion of the variance explained. In addition, the study of rare variants in known regions where the signals are not well localized could help fine mapping these regions. For other traits it has been shown that the same region could host more than one independent signal (151); conditional analyses within discovered regions conditioning on the genome-wide significant variants could add to the proportion of the variance explained if additional signals were identified. Structural variation has not yet been studied so widely due to its more complex nature, however it might also play an role in explaining the missing heritability.

Three of the 26 regions associated with lung function (*AGER* (rs2070600), *NCR3* (rs2857595) and *ZKSCAN3*/ *ZNF323* (rs6903823)) are within a 3.8Mb window in the major histocompatibility complex (MHC), known to be a long range linkage disequilibrium region. LD estimated using HapMap data indicates that these three regions are independent (r^2 ~ 0.01 between rs2070600 and the other two SNPs; and r^2 ~ 0.31 between rs2857595 and rs6903823). However the HapMap sample size is limited and estimates in larger populations would be

required to obtain more accurate estimates of LD patterns, especially for variants with comparatively low allele frequency, such as rs2070600 (MAF = 0.05). In addition, a joint analysis of these three variants conditioning on each other would bring more insights into their dependency or lack thereof.

Population structure can lead to increased type I error. For this reason the studies that took part in the meta-analysis were asked to adjust their models using ancestry principal components, and genomic control was applied three times in the meta-analysis process, twice at study level (for ever-smokers and never-smokers separately and after meta-analysing them) and once at the meta-analysis level. The use of genomic control in this study is likely to be overly conservative, given that genomic inflation factors are known to increase with sample size, as the power to detect real associations for polygenic traits also increases (130).

Inadequate adjustment for relatedness in the data is also known to increase type I error. Most studies included in these analyses had data on unrelated individuals, but a subset of studies included related individuals. All the studies with data on related individuals undertook their analysis taking proper account of relatedness, although in some instances related individuals were split into ever smoking and never smoking categories and their association results were meta-analysed. This could have led to inflated statistics, however the

conservative approach taken when applying genomic control probably accounted for this and the final stage lambdas were not overinflated (λ_{GC_1000} = 1.002).

Heterogeneity of the results across studies was tested using a Chi-square test and no significant results were found when applying a Bonferroni corrected threshold. This test however is of limited value, since it is underpowered when there are few studies (for example < 20) (152) and it can be too sensitive if there are many studies. Also, a Bonferroni corrected threshold is a conservative approach to account for multiple testing. Forest plots show that the results are broadly consistent across studies and not just driven by just a small number of studies (**Appendix D**). However, applying an additional approach, such as estimating I^2, which measures the amount of heterogeneity that is not due to chance, would have provided a more complete analysis to assess heterogeneity.

This study focused on lung function in adult individuals, however the effect of the 26 variants shown to influence lung function were also assessed in two cohorts of children (141-144). Overall the direction of effects seemed to be consistent between children and adults for the majority of the variants. This would suggest that either there is an overlap of pathways involved in the development and decline of lung function or that the pathways detected so far

are mainly involved in developmental processes. To investigate specific pathways influencing the decline of lung function, longitudinal analysis of lung function measures in adults would be required.

In order to assess whether an inappropriate adjustment for smoking behaviour might have led to some of the findings, since quantitative information on amount smoked was not available for many studies, I assessed the association of the 16 new regions and other previously discovered regions with ever smoking status and with number of cigarettes per day in the Oxford-GlaxoSmithKline (Ox-GSK) study (109). No evidence of association was found for any of the 16 novel loci with either of the traits. However, the sentinel SNP in *PID1*, a locus that was genome-wide significantly associated with FEV_1/FVC in the CHARGE consortium (96), but which did not replicate in the SpiroMeta consortium (96), and which did not reach genome-wide significance in the SpiroMeta-CHARGE results ($P = 1.13 \times 10^{-6}$) shows evidence of association with number of cigarettes per day ($P = 1.6 \times 10^{-3}$). This could indicate that the association seen with lung function might be mediated via an effect on smoking behaviour. To assess whether the novel findings might have occurred in part due to a gene by smoking interaction, the genetic effects were evaluated separately in ever and never-smokers. Effect sizes in ever and never-smokers were consistent overall. These results show that the associations of the 26 regions associated with lung function do not seem to be mediated via a smoking behaviour or a gene by smoking interaction.

Some of the loci previously associated with lung function (*HHIP*, *GPR126*) have an effect on height (127), and one of the novel loci (*MFAP2*) also shows a significant association with height in the GIANT consortium dataset (127). For the lung function sentinel SNPs in *GPR126* and *MFAP2*, the allele associated with increased height is associated with reduced FEV_1/FVC. For *HHIP* the direction of effect is the same for the lung function sentinel SNP, although the SNP with the strongest association with height is not associated with FEV_1 or FEV_1/FVC ($P > 0.3$). This suggests that the association of these regions with lung function is not simply through an incomplete adjustment of height in the lung function analysis.

This study presented 16 novel regions associated with lung function and evaluated the effect of these variants and other previously discovered (23, 94-96) on other traits, with the aim of providing additional insights on the mechanisms by which they influence lung function. However, little is known about the molecular pathways involved, and additional analyses within these genomic regions are required. Understanding the molecular pathways by which these regions affect lung function should provide new insights into the regulation of lung function and could lead to the development of new therapeutic targets.

Chapter 4: Analyses of rare genetic variants

This chapter focuses on the study of rare (MAF < 1%) and low allele frequency (1% < MAF < 5%) variants and presents two different approaches undertaken in order to identify rare or low allele frequency variants that have an effect on lung function. The first approach presented is a collapsing method that was undertaken by a subset of SpiroMeta studies; the studies provided summary statistics according to a central analysis plan that I developed and I undertook quality control checks, meta-analysed and interpreted their results. The second approach is a targeted sequencing study of the 26 genomic regions known to affect lung function (2, 23, 94-96)* at the time of the study in 300 COPD cases and 300 controls. The sequencing was outsourced and I undertook the data processing, quality control checks, analysis and interpretation of the data. The results obtained and the challenges that these data presented are discussed here.

4.1 Introduction

Genome-wide association studies (2, 23, 94-96)* discussed in previous chapters have collectively identified 26 loci that have an effect on lung function. However the proportion of the variance of the lung function measures that these variants explain is very limited (around 3.2% of the additive polygenic variance of FEV_1/FVC, see Chapter 3 for more details). Genome-wide association studies of lung function undertaken to date have focused on identifying common

variants (MAF > 5%). It is hypothesized that variants with lower allele frequency might have larger effect sizes and therefore they might play an important role in explaining the missing heritability (13). In addition, many of the association signals for the loci known to affect lung function are not well localized, and identifying rare or low allele frequency variants within these regions can aid identification of the causal variants.

Detecting new associations for rare or low allele frequency variants is challenging. Meta-analyses of GWAS undertaken to date were designed to detect common variants, however they did not have enough power to detect single low frequency variants with small to moderate effect sizes. Collapsing methods, which pool the effect of rare or low allele frequency variants within a locus, increase the power to detect new associations. I hypothesized that loci with rare or low allele frequency variants that have a moderate effect on lung function could be identified by applying a collapsing method to the SpiroMeta studies (23). Therefore the collapsing method implemented by the software QuTie (153) was applied to a subset of studies within the SpiroMeta consortium with a total sample size of 20,941 individuals.

Given that COPD can be diagnosed using spirometry measures, I hypothesized that loci associated with lung function measures would also be associated with COPD. The association to date of 12 out of the 26 lung function regions with

COPD risk or airflow obstruction (*TGFB2*, *TNS1, RARB*, *FAM13, GSTCD*, *HHIP, HTR4, ADAM19, AGER*, *GPR126*, *C10orf11* and *THSD4*) (24, 95, 103, 117, 154, 155)*, support this hypothesis. In order to fine map or identify new signals within the loci known to affect lung function, the 26 regions (2, 23, 94-96)* were sequenced in 300 COPD cases and 300 controls, assuming that loci associated with lung function also affect COPD susceptibility. To maximize the sample size of this study a cost-effective pooled design was chosen. This consisted of pooling individual DNA across individuals, separately for cases and controls. The design was made of 24 pools (12 case pools and 12 control pools) with DNA from 25 individuals in each pool.

The challenges that these analyses presented and the strategies chosen to deal with them as well as the results obtained are reported in this chapter.

4.2 Collapsing method to detect rare variants in the SpiroMeta dataset

The collapsing method described in (153) was undertaken by a subset of SpiroMeta studies according to a central analysis plan. This method, implemented by the software QuTie v4 (153), examines the accumulation of low frequency and rare variants in a given locus either using a gene-based or sliding window approach. In order to develop the analysis plan I piloted the analysis in the Busselton dataset (BHS1, n = 1168) to familiarize myself with the

method and the possible complications that could arise while running it. I coordinated these analyses within the consortium and meta-analysed their results.

4.2.1 Methods and quality control

Based on the QuTie manual an analysis plan was developed and shared with the SpiroMeta studies. The analysis plan can be found in **Appendix B**.

Study level analysis

Samples

SpiroMeta studies that undertook these analyses were: ALSPAC, B58C, BHS1, ECRHS, the EPIC studies (obese cases and population-based studies), the EUROSPAN studies (CROATIA-Korcula, CROATIA-Vis and ORCADES), FTC (incorporating the FinnTwin16 and Finnish Twin Study on Aging), Health 2000, KORA F4, KORA S3, NFBC1966, SHIP and TwinsUK-I. The total sample size was 20,941. Details on phenotype and genotype can be found in (2, 23)*.

Trait preparation

Only unrelated individuals were included in this analysis, hence studies with family data only undertook the analysis in a subset of unrelated individuals. All

individuals included in the analysis had complete data for ever smoking status, age, sex, height, FEV_1 and FVC. FEV_1 and FEV_1/FVC were regressed against the following covariates: age, age^2, sex, height, ever smoking status and ancestry principal components. The residuals were then transformed to ranks and to normally distributed Z-scores.

Collapsing method

The analysis was only applied to directly genotyped SNPs, within genes, that passed the study level quality control process without any MAF filter. In order to assess genotyping quality of any interesting signal, cluster plots could be examined at a later stage.

The command line for the perl script that runs QuTie v4 (153) was provided to the studies, as well as a file with the gene coordinates required by the software in order to undertake the gene-based analysis.

Algorithm:

- QuTie v4 (153) filters variants with MAF greater than a given threshold
- It tests whether the means of a quantitative trait are significantly different between individuals with and without rare or low allele frequency variants

for each locus using linear regression and Student's T-test; the loci are defined by the gene coordinates file provided +/- certain distance

- It then runs permutations for loci with P-values below a certain threshold in order to generate empirical P-values

For these analyses the quantitative traits used were the inverse normal transformed residuals for FEV_1 and FEV_1/FVC, the loci were defined as the gene coordinates +/- 50kb either side; only SNPs with MAF < 5% in each study were included, and loci with P-values < 10^{-5} were subject to 100,000 permutations. The version of QuTie v4 used was last updated on the 7th of May 2009.

Consortium central analyses

QC issues

Each study provided results and plots (Manhattan and histograms of the phenotype) produced by the software. These plots together with plots of betas, standard errors, and QQ plots produced centrally were examined as well as the genomic inflation factor (λ) for each study, in order to detect studies with irregular results. One study was removed from the meta-analysis due to a high λ (λ for FEV_1 = 1.59 and λ for FEV_1/FVC = 1.31). No other issues were found.

Meta-analysis

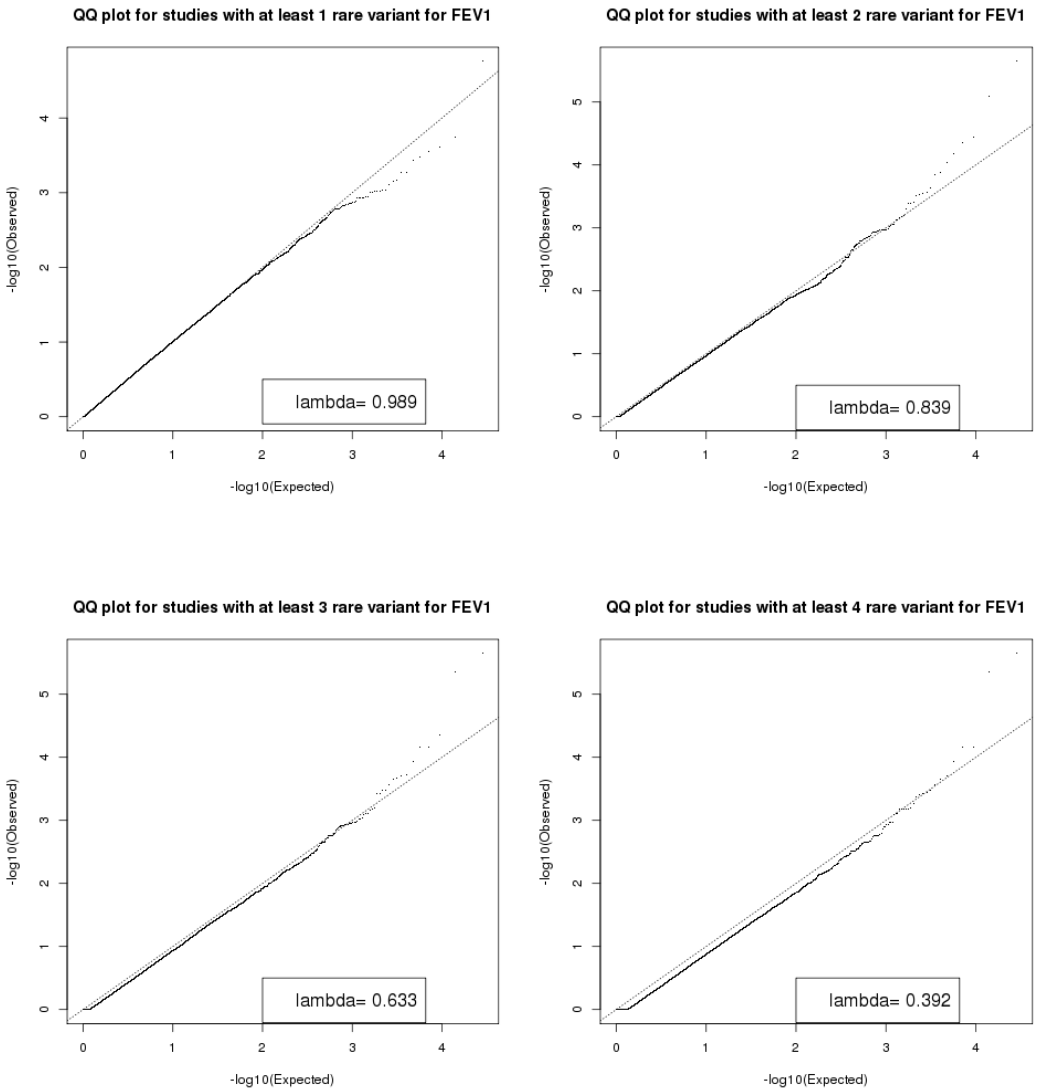
After the quality control checks, the P-values were meta-analysed across studies using the standard weighted Stouffeer's method weighting the P-values by the square root of the sample size. Genomic control was applied at study level and after the meta-analysis of the findings. Four meta-analyses were undertaken for different scenarios depending on the number of rare variants that each study had for a given gene: 1) only studies with at least 1 rare variant, 2) only studies with at least 2 rare variants, 3) only studies with at least 3 rare variants included and 4) only studies with at least 4 rare variants.

4.2.2 Results

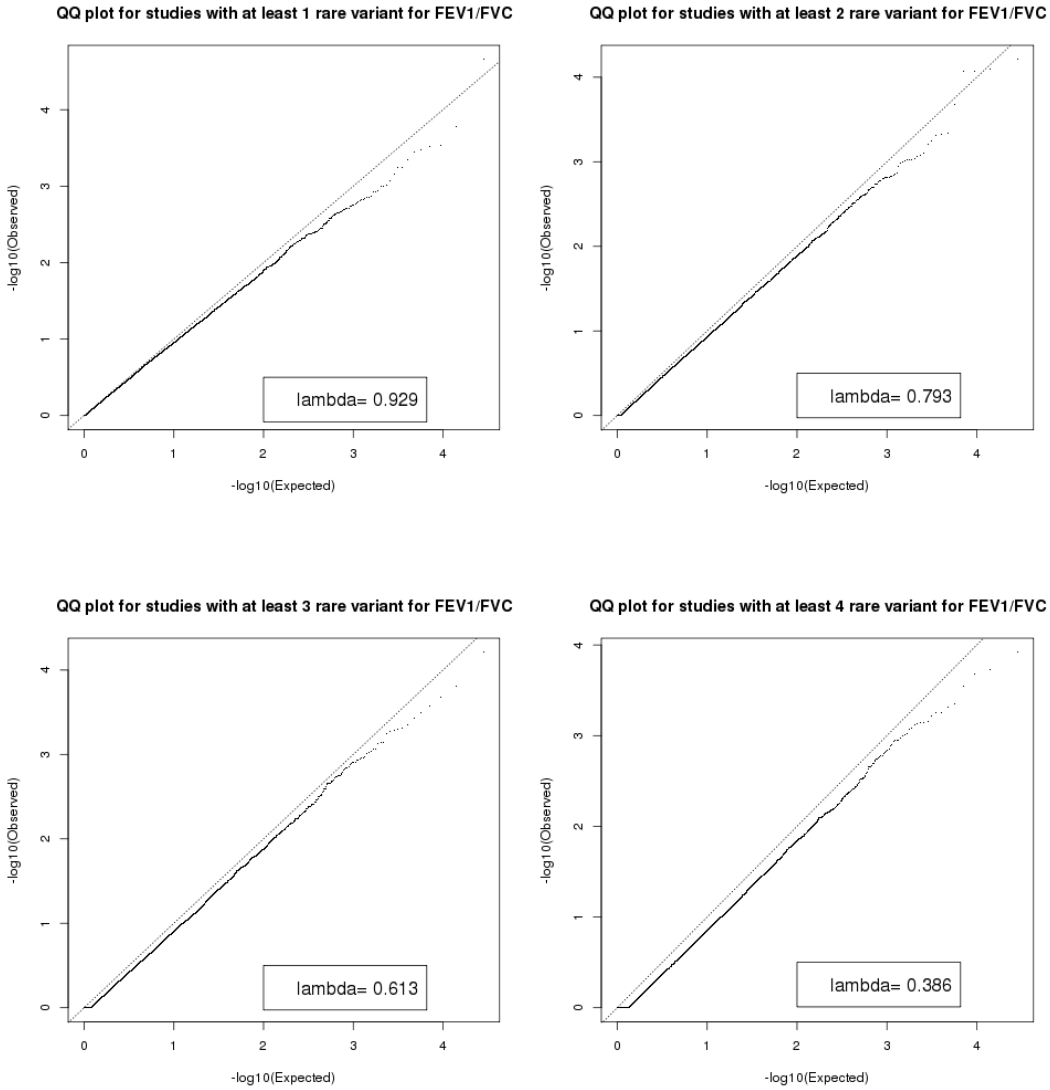
QQ plots with genomic inflation values are presented in **Figure 4-1** both for FEV_1 and FEV_1/FVC for the four meta-analyses. The lambdas decreased as the criteria for inclusion in the meta-analysis became stricter, which reflects a reduction in power when the criteria were stricter and a smaller number of studies were included in the meta-analysis for each locus.

Figure 4-1 QQ plots for the collapsing method applied to SpiroMeta studies

a) for FEV_1



b) for FEV_1/FVC



Lower allele frequency variants are known to have poor clustering properties that could lead to spurious associations, for this reason in order to select a locus for follow-up, it was required that the association signal was driven by more than two variants and that the association was seen in more than one study. To select regions for follow-up they needed to meet both of the following conditions: (i) to have P-values below 10^{-5} in the meta-analyses with at least 3 or 4 rare variants, and (ii) to have rare variants in more than one study.

Table 4-1 presents results for genes with P-values below 10^{-5}, and none of them met the criteria for follow-up. Three genes had P-values < 10^{-5} for FEV_1 and no regions for FEV_1/FVC exceeded this threshold. The two most significant genes for FEV_1 (*LOC402116* and *CRYGFP1*) had rare variants only present in one study. The third most significant gene for FEV_1 only met the significance threshold in the meta-analysis of studies with at least 2 rare variants, but not in the meta-analyses of at least 3 or 4 rare variants.

Table 4-1Top results for the collapsing method applied to SpiroMeta studies

Results for genes with P-values < 10^{-5} are shown. Only results for FEV_1 are presented, since no gene reached the threshold for FEV_1/FVC. P-values < 10^{-5} are presented in bold. Abbreviations: Chr. = chromosome.

Gene	Chr.	Start	End	Minimum number of rare variants per study	Mean number of rare variants across studies	Number of studies	Sample size	Number of individuals with rare alleles	Z meta-analysis	P meta-analysis
LOC402116	2	209701737	209803983	1	2.67	3	4587	396	-1.886	$5.93x10^{-2}$
				2	6	1	1953	391	-4.728	**$2.26x10^{-6}$**
				3	6	1	1953	391	-4.728	**$2.26x10^{-6}$**
				4	6	1	1953	391	-4.728	**$2.26x10^{-6}$**
CRYGFP1	2	209668096	209770774	1	4.33	3	4545	510	-1.118	$2.64x10^{-1}$
				2	4.33	3	4545	510	-1.118	$2.63x10^{-1}$
				3	9	1	1923	401	-4.587	**$4.49x10^{-6}$**
				4	9	1	1923	401	-4.587	**$4.49x10^{-6}$**
FLJ45966	4	8509455	8615237	1	3.44	9	11651	2088	4.295	$1.74x10^{-5}$
				2	3.75	8	9942	1487	4.460	**$8.21x10^{-6}$**
				3	4.8	5	3566	740	3.703	$2.13x10^{-4}$
				4	6	3	2235	545	1.438	$1.5x10^{-1}$

4.2.3 Discussion

The rationale for applying this collapsing method was to increase the power to detect rare or low allele frequency variants with modest effect sizes that would not have been picked up by previously employed GWAS analyses. This study had a large sample size (N = 20,941), however no convincing association signals were detected. A possible explanation for this is that the GWAS platforms used by the studies included in these analyses were primarily designed to capture common variation and do not tag lower frequency variants adequately. In addition different platforms were used by different studies, and therefore a different set of uncommon variants was genotyped in each study, which reduced the power of the meta-analysis. Denser genotyping arrays and imputation panels that enable reliable imputation of low frequency variants, such as the 1000 Genomes Project (36) or UK10K (http://www.uk10k.org/), and the decreased cost of sequencing will help to overcome these issues in future studies.

The approach followed here was to meta-analyse P-values using the standard weighted Stouffeer's method weighting the P-values by the square root of the sample size, instead of the inverse variance weighted meta-analysis used for common variants in the previous chapters. This is because the standard weighted Stouffeer's approach is more robust and therefore more suitable for the meta-analysis of rare variants results.

A limitation of the collapsing method used here is that it assumes the direction of effect of all the rare alleles to be the same within a gene, either deleterious or protective. This method is underpowered to detect loci with rare alleles that affect lung function in opposite directions. In recent years collapsing methods have improved dramatically (34, 44, 156); an additional method that collapses effects of both deleterious and protective variants (53) is presented in the next section (4.3).

An additional challenge of this study was not having access to the individual level data. If the individual level data had been available, the findings could have been examined further. Cluster plots could have been assessed for the variants in the most significant loci. If the same variants were driving the associations in all the studies and these variants had acceptable cluster plots, that would have given more confidence in the findings.

4.3 Targeted sequencing in COPD cases and controls

Twenty six regions known to affect lung function (2, 23, 94-96)* were sequenced in 300 COPD cases and 300 controls in order to identify rare and low allele frequency variants that could aid to fine map the association signals in these regions or that might point to new signals. In order to maximise sample size, a pooled design was used here, with DNA pooled for sets of 25 individuals and sequenced together, separately within cases and controls. This made the

variant calling step particularly challenging since sequencing error rates and minor allele frequencies for rare variants can be very similar when using a pooled design. This section starts with a brief introduction to the pooled sequencing design used, then provides a detailed description of the methods and presents and discusses the findings.

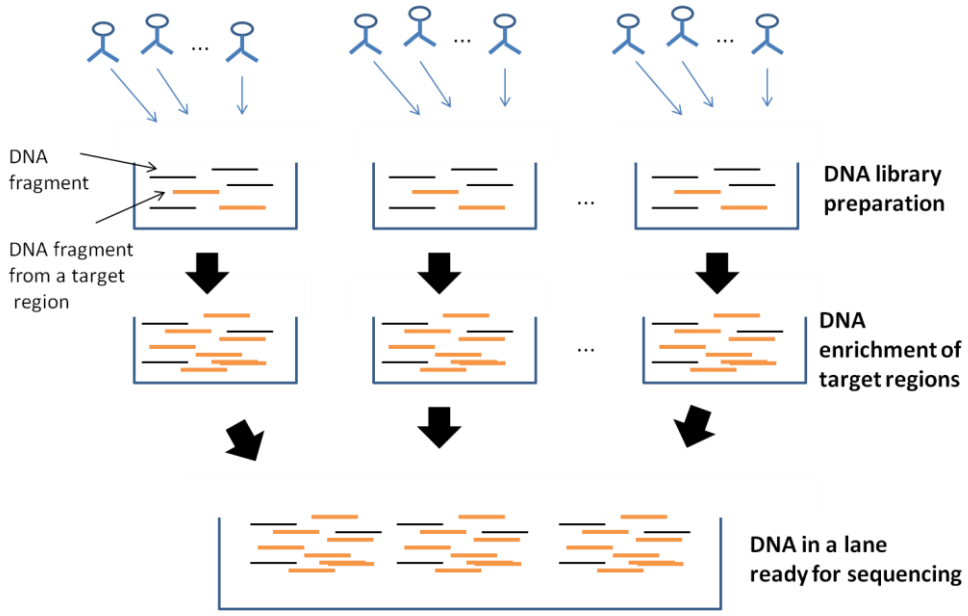
4.3.1 Introduction to the pooled sequencing design

In order to sequence DNA, it must first be fragmented and prepared in DNA libraries. In a pooled experiment, DNA from several individuals is combined in a single DNA library (**Figure 4-2**). In the experiment presented in this chapter, DNA from different libraries was sequenced together in a lane (**Figure 4-2**), therefore the DNA in each library was indexed so that at the end of the sequencing experiment we could track the library each DNA fragment came from. Since it was a pooled design, DNA from all the individuals in the same library (or pool) had the same index and it was not possible to know which individual a specific fragment of DNA belonged to, only which library they belonged to. In a targeted sequencing experiment, only the sequence in the target regions is of interest, and an enrichment step is necessary, where the DNA in the relevant regions is copied many times, in order to increase its amount (**Figure 4-2**). For this enrichment step it is necessary to design a probe library, which defines the regions we want to sequence. This design takes into account local characteristics of the genome that can influence the performance of the sequencing and alignment, and these include repetitive sequence and

GC content. A large proportion of the genome is made of highly repetitive DNA sequences (157), which can create ambiguities in the alignment of reads to a reference genome, especially when sequencing small DNA fragments. GC content bias is the dependence between the GC content (guanine-cytosine content) of a DNA fragment and the coverage of this fragment (158). In particular this affects the ability to call copy number variants.

More general concepts about sequencing are explained in Chapter 1 section 1.1.4.2.

Figure 4-2 Pooled sequencing diagram



4.3.2 Methods

4.3.2.1 Study design

Samples

Individuals from three studies were included in this analysis: Gedling, Nottingham Smokers and the Leicester COPD cases. Spirometry procedures for Gedling and Nottingham Smokers can be found in (2)* and for the Leicester COPD cases in (159).

Individuals were excluded if: (i) they were younger than 40 years old, (ii) they had pack-years of smoking < 5, or > 100, or (iii) if they had DNA concentration ≤ 20ng/ul (minimum concentration required for quality sequencing). Additionally, in the Leicester COPD cases study individuals with asthma were also excluded. This left a sampling frame of 965 individuals (403 from Gedling, 468 from Nottingham Smokers and 96 from the Leicester COPD cases).

COPD cases were defined as spirometric GOLD stage 2 (104)(55) and above (percent predicted FEV_1 < 80% and FEV_1/FVC < 0.7) and controls as individuals with percent predicted FEV_1 > 80% and FEV_1/FVC > 0.7, based on pre-bronchodilator spirometry. Individuals with percent predicted FEV_1 > 80% and FEV_1/FVC < 0.7 (GOLD stage 1 (104)) or with percent predicted FEV_1 < 80% and FEV_1/FVC > 0.7 were excluded from the analysis to minimize misclassification. The calculation of percent predicted FEV_1 was undertaken

using reference values of FEV_1 that take into account age, sex and height according to previously described equations (72, 73). In order to select the most extreme 300 COPD cases and controls, COPD cases and controls were ranked according to their percent predicted FEV_1. In addition, to remove extremely healthy individuals from the controls, individuals were excluded if: (i) they had percent predicted FEV_1 > 120.26 (the 99th percentile of percent predicted FEV_1) or (ii) if they had FEV_1/FVC > 0.85 (the 95th percentile of FEV_1/FVC). Characteristics of individuals included in the study are presented in **Table 4-2**. **Figure 4-3 a)** illustrates how cases and controls were selected. Individuals were grouped into pools of 25 (separately for cases and controls), following the percent predicted FEV_1 ranking, so that individuals with more similar phenotype would be grouped together (**Figure 4-3 b)**).

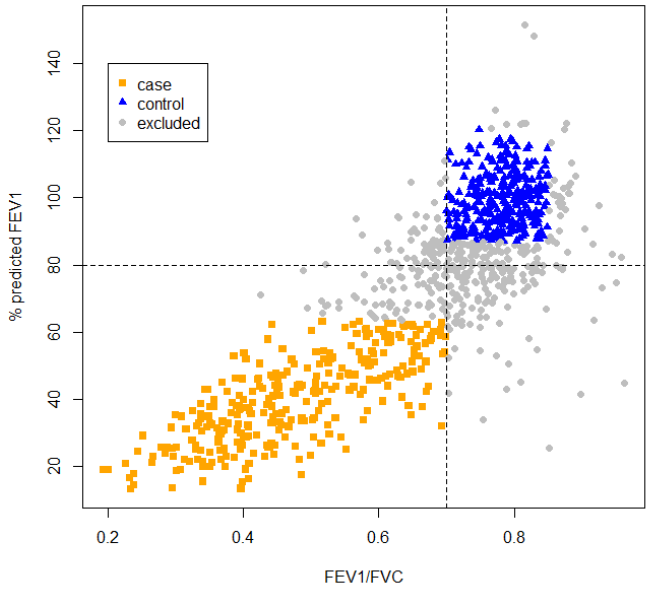
Table 4-2 Study characteristics

Abbreviations:N = number, s.d. = standard deviation, y = years, L = litres.

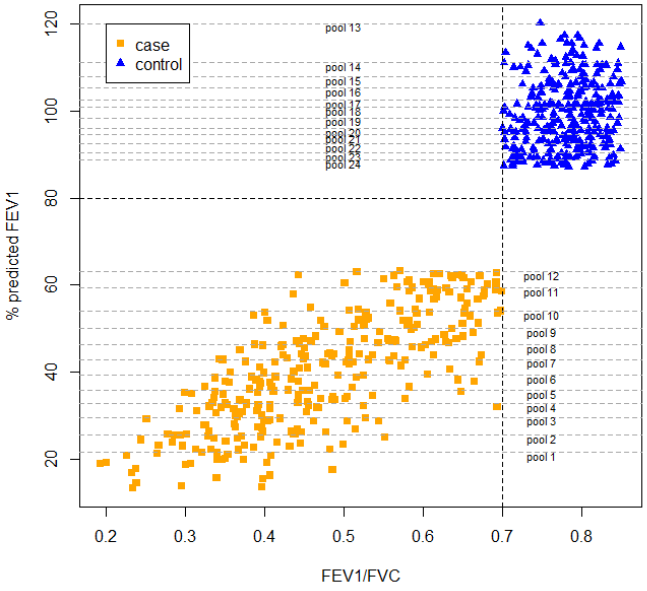
Status	**N Total**	**N male**	**N female**	**Age range (y)**	**Mean age, y (s.d.)**	**Mean FEV_1, L (s.d.)**	**Mean % predicted FEV_1 (s.d.)**	**Mean FVC, L (s.d.)**	**Mean FEV_1/ FVC (s.d.)**	**Mean Pack-years (s.d.)**
Case	300	192	108	40-86	65.35 (9.61)	1.15 (0.48)	39.8 (12.96)	2.42 (0.78)	0.47 (0.12)	41.94 (19.18)
Control	300	162	138	40-79	56.89 (9.97)	3.04 (0.68)	99.27 (8.02)	3.90 (0.86)	0.78 (0.04)	24.69 (16.37)

Figure 4-3 Selection of COPD cases and controls

a) All individuals



b) Cases and controls allocated to pools



Definition of regions

Region plots produced with GWAS data from the SpiroMeta-CHARGE meta-analysis results for FEV_1 and FEV_1/FVC (2)* (Chapter 3) for the 26 regions associated with lung function (2, 23, 94-96)* were examined to define the association regions. SNPs with $-log_{10}(P-value) > 2.5$ and not further than 50kb away from the next SNP moving away from the sentinel SNP, were selected. Any gene intersecting the association region was added to the region +/-10kb. If the association region did not include or intersect the closest gene, the association region was enlarged to include the closest gene +/- 10kb. If the enlarged regions also intersected other genes, the regions were not enlarged again, so they included small portions of genes. Regions were selected using the $-log_{10}(P-value)$ for the most significant trait only, except for *CDC123* which was genome-wide significant for FEV_1 and FEV_1/FVC and the sentinel SNP was the same for both traits. For *CDC123* the association region was defined so it included the association regions for both traits. Region plots generated with data from the SpiroMeta-CHARGE meta-analysis (2)* with the association region highlighted can be found in **Appendix E**. The regions covered a total of 10.3Mb (**Table 4-3**).

Table 4-3 Regions summary

The columns “GWAS sentinel” and “GWAS gene” present the lung function GWAS sentinel SNP and the closest gene to the sentinel respectively (2)*. Abbreviations: Chr. = chromosome.

Chr.	GWAS.sentinel	GWAS.gene	Start	End	Length	Number of genes
1	rs2284746	*MFAP2*	17238444	17455948	217504	5
1	rs993925	*TGFB2*	218508675	218885482	376807	2
2	rs2571445	*TNS1*	218627794	218818796	191002	1
2	rs12477314	*HDAC4*	239839616	240332643	493027	2
3	rs1529672	*RARB*	25459833	25649422	189589	2
3	rs1344555	*MECOM*	168791286	169391563	600277	1
4	rs2045517	*FAM13A*	89637105	90077431	440326	2
4	rs10516526	*GSTCD*	106280233	106902828	622595	5
4	rs11100860	*HHIP*	145227600	145669881	442281	1
5	rs153916	*SPATA9*	94984019	95038027	54008	2
5	rs1985524	*HTR4*	147682118	148026624	344506	4
5	rs11134779	*ADAM19*	156597906	157139503	541597	7
6	rs6903823	*ZKSCAN3*	27982152	28415572	433420	14
6	rs2857595	*NCR3*	30584612	31959223	1374611	75
6	rs2070600	*AGER*	31996092	32205942	209850	14
6	rs2798641	*ARMC2*	109159618	109305352	145734	1
6	rs262129	*LOC153910*	142613055	142968973	355918	2
9	rs16909859	*PTCH1*	98153197	98313032	159835	1
10	rs7068966	*CDC123*	12170174	12335588	165414	4
10	rs11001819	*C10orf11*	77532518	78643886	1111368	1
12	rs11172113	*LRP1*	57472676	57617125	144449	4
12	rs1036429	*CCDC38*	96041582	96400071	358489	6
15	rs8033889	*THSD4*	71423787	72085722	661935	1
16	rs12447804	*MMP15*	57906243	58143392	237149	5
16	rs2865531	*CFDP1*	75252927	75538926	285999	5
21	rs9978142	*KCNE2*	35595821	35753440	157619	2

Sequencing method

The enrichment and the sequencing were outsourced. The enrichment kit was produced by Agilent (http://www.agilent.com/) and the sequencing was undertaken by Source BioScience (http://www.sourcebioscience.com/).

I used Agilent's eArray for the first draft of the design probe library for enrichment of the target regions from genomic DNA prior to sequencing. I then liaised with Agilent and they finalized the design. After applying a correction for GC content and applying repeat-masking filters, a total of 7.7Mb of sequence was covered by probes in the final design.

Sequencing was undertaken using Illumina HiSeq2000 with 100 bp paired-end reads (both ends of a DNA fragment are sequenced forming a paired-end read) and 8 lanes, each with 3 pools. Pools were assigned to lanes sequentially, so that pool1 to pool3 were allocated to lane 1, pool4 to pool6 were allocated to lane 2 and so on; in total there were four case lanes and four control lanes.

Coverage per individual of around 40x was expected, assuming 50% on-target capture (proportion of the reads that overlap the target sequence).

4.3.2.2 Data processing

The data were provided in FASTQ format. Since the sequencing was undertaken using paired-end reads, two files were provided per pool, one with the forward strand derived reads and other with the reverse strand derived reads. These files include base quality scores for each base in each read. The base quality scores are presented as phred quality scores ($-log_{10}(Probability\ that\ the\ base\ called\ is\ wrong)$); for instance a base quality of 30 would indicate that the probability of the base called being wrong is 0.001.

The data were aligned against 1000 Genomes Project Phase 1 data (36) (GRCh37; h19) using BWA.6.2 (160), with –q 15 for read trimming, to remove the 3' end of the reads which tend to have lower quality. BWA (160) generates a mapping quality score, also presented as a phred quality score ($-log_{10}(Probability\ that\ the\ mapping\ is\ wrong)$).

Alignments were then sorted and PCR duplicates were removed using SAMtools (161). PCR duplicates are artefacts from the sequencing technology that are exact copies of each other and do not add new information, therefore they are usually removed. After the removal of duplicates, coverage summaries were produced using SAMtools (161) and BEDtools (162). Given that the alignment of INDELs is particularly challenging, local realignment around INDELs was undertaken with GATK (39) using known INDEL coordinates from

1000 Genomes Project Phase 1 data (36) and Mills *et al.* (163) as reference. Also, in order to obtain more accurate base quality scores than those produced in the sequencing process, an empirical recalibration of base quality scores was undertaken using GATK (39). This recalibration of base quality scores takes into account the reported quality score, the position within the read and the preceding and current nucleotide observed by the sequencing machine.

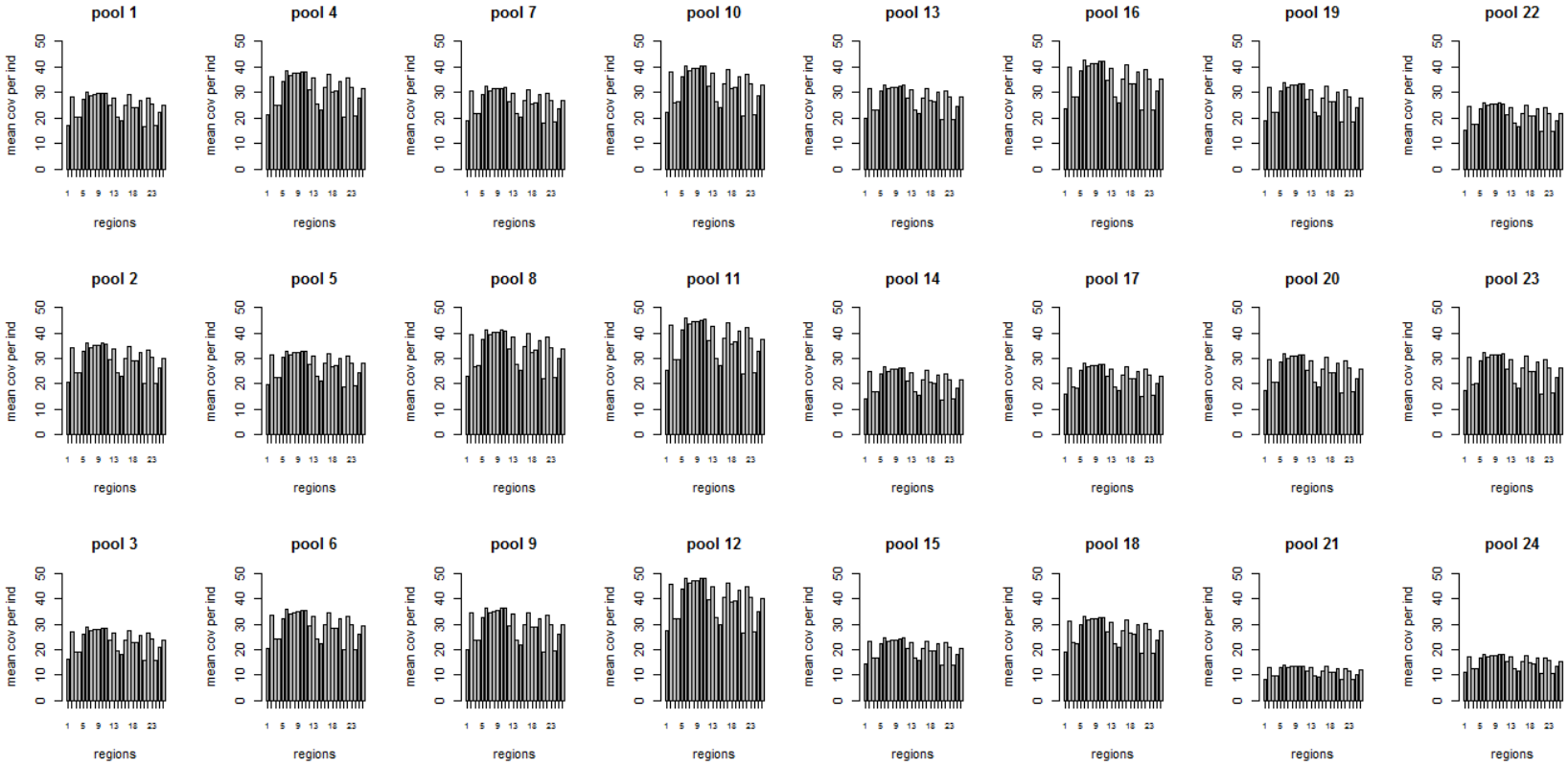
In order to assess the quality of the data, coverage summaries were produced. Coverage per individual per pool calculated after removing unmapped reads, duplicated reads and reads outside of the regions of interest ranged from 12.28x to 42.23x with an average of 28.74x (**Table 4-4**).

Table 4-4 Number of reads per pool and coverage

Pool	Number of reads				On-target capture (%)	Coverage per individual
	In total	Mapped	Mapped after removing duplicates	Mapped in regions of interest after removing duplicates		
1	109,826,166	109,516,893	66,490,002	50,852,447	76.48	26.42
2	102,470,681	102,132,135	79,072,130	60,697,022	76.76	31.53
3	82,185,922	81,864,111	62,490,941	48,364,746	77.39	25.12
4	121,167,978	120,842,295	82,980,615	64,349,712	77.55	33.43
5	134,842,243	134,497,208	72,815,031	56,264,867	77.27	29.23
6	137,317,797	136,991,796	78,031,571	60,221,653	77.18	31.28
7	186,147,801	185,658,558	71,068,003	54,682,451	76.94	28.41
8	124,843,995	124,427,602	91,418,655	69,630,567	76.17	36.17
9	93,129,017	92,770,408	81,055,677	61,021,017	75.28	31.7
10	112,581,871	112,166,404	88,035,910	67,479,703	76.65	35.05
11	139,217,132	138,677,742	100,177,543	76,376,749	76.24	39.68
12	189,602,678	188,962,411	107,617,615	81,346,912	75.59	42.26
13	138,857,591	138,517,962	73,573,204	56,301,930	76.53	29.25
14	66,785,328	66,608,716	57,138,920	43,783,474	76.63	22.74
15	75,171,767	75,008,024	53,430,349	41,506,061	77.68	21.56
16	122,464,563	122,134,268	93,657,958	71,209,717	76.03	36.99
17	110,294,762	110,017,214	60,638,846	46,864,360	77.28	24.35
18	182,786,012	182,316,562	72,460,650	55,730,089	76.91	28.95
19	155,797,230	155,419,610	74,221,765	56,702,253	76.4	29.46
20	106,332,429	106,060,303	67,605,216	52,577,023	77.77	27.31
21	157,062,037	156,698,524	30,232,596	23,648,595	78.22	12.28
22	141,975,890	141,629,730	57,062,737	44,022,937	77.15	22.87
23	122,808,781	122,485,636	67,981,471	52,943,898	77.88	27.5
24	151,301,579	150,916,877	40,203,619	31,102,507	77.36	16.16

Pools 21 and 24 had lower coverage (12.28x and 16.16x respectively) than the rest and a higher proportion of duplicated reads (80.7% and 73.4% respectively) (**Table 4-4**). Enquiries with Agilent (http://www.agilent.com/) and Source BioScience (http://www.sourcebioscience.com/) indicated that the DNA quality for these two pools was lower than for the rest. For this reason these two pools were excluded from the analysis. The average coverage per individual after excluding these two pools was 30x. There was some variation in coverage per region, and the same pattern of variation between regions was observed in all pools (**Figure 4-4**).

Figure 4-4 Mean coverage per individual per region and per pool



4.3.2.3 Variant calling

In order to distinguish true calls from sequencing error, three different calling algorithms specific for pooled data were used. These were vipR (41), SNVer (42) and Syzygy (43). A description of the three algorithms is presented here and the notation used in this section is given in Table 4-5. Throughout this section “allele read count” at a given position refers to the number of reads with a certain allele, and “allele chromosome count” refers to the number of chromosomes with a certain allele. In a pool made of 25 individuals, at a given position a variant may have an allele read count up to the coverage at that position, however it can only have an allele chromosome count up to 50.

Table 4-5 Notation for variant calling algorithms

N	Number of haploid individuals per pool (50)
M	Number of pools (24)
rc	Read counts
MAF	Minor allele frequency
C	Coverage
ε	Sequencing error rate
SEc	Sequencing error count
$expSEc$	Expected sequencing error count
MAc (or $ALTrc$)	Minor allele count (or alternative allele read count)
$obsMAc$ (or $obsALTrc$)	Observed minor allele count (or observed alternative allele read count)
$ALTchrc$	Alternative allele chromosome count
$REFrc$	Reference allele read count
$REFchrc$	Reference allele chromosome count

vipR

The principle behind this algorithm (41) is that data from multiple DNA pools can be used to compensate for differences in sequencing error rates along genomic regions assuming that sequence-dependent error rate is conserved across pools. If the sequence-dependent error rate is conserved across pools, we would expect to see the same sort of variation due to sequencing error at a given position across pools. vipR (41) simply tests whether minor allele frequencies differ significantly between pools, and calls a variant when its minor allele frequency in at least 2 pools is significantly different and this difference is unlikely to be due to sequencing error. This idea had been put in practice previously by the software CRISP (44), however vipR presents a more computationally efficient implementation.

To test whether variation seen for a base is a variant or sequencing error, vipR (41) uses the Skellam distribution. The Skellam distribution is a discrete probability distribution that models the difference between two statistically independent variables following Poisson distributions with different expected values. The number of sequencing errors, the sequencing error count (SEc_i), for a given base that occurs in the i^{th} DNA pool can be thought as following a Poisson distribution with the expected number of sequencing errors in that pool ($expSEc_i$), obtained as the product of the sequencing error rate (ε) and the coverage in that pool (C_i), as its expected value.

$$SEc_i \sim Pois(expSEc_i)\ and\ SEc_j \sim Pois(expSEc_j)$$

$$SEc_i - SEc_j \sim Skellam(expSEc_i, expSEc_j)$$

$$for\ expSEc_i = \varepsilon\ C_i,$$

$$expSEc_j = \varepsilon\ C_j,$$

$$i, j = 1, \dots, M\ and\ i \neq j$$

Therefore, by using the Skellam distribution we can test whether the observed difference of minor allele counts between the i^{th} pool ($obsMAc_i$) and the j^{th} pool ($obsMAc_j$) is produced by sequencing errors in both pools. If this hypothesis is rejected for any combination of i^{th} and j^{th} pools, a variant is called.

$$H_0: obsMAc_i - obsMAc_j \sim Skellam(expSEc_i, expSEc_j)$$

$$H_1: obsMAc_i - obsMAc_j \nsim Skellam(expSEc_i, expSEc_j)$$

$$for\ i, j = 1, \dots, M\ and\ i \neq j$$

Algorithm:

- Use SAMtools to produce allele counts for each strand and to filter positions with base quality below 10 or mapping quality below 20.
- Estimate sequencing error rate for each strand across pools as: q^{th} percent quantile of the minor allele frequencies.

- Test whether variation seen for a given base is a variant or is sequencing error. For each variant one sided P-values are computed based on the Skellam distribution for all possible pairs of pools, using minor allele count differences between pools, error rate estimates and coverage (coverage is unified between pools before estimating the number of expected sequencing errors). This is done both for the forward and the reverse strands. For each base in each pool the most significant P-value (out of all the pair-wise comparisons) is kept for both strands. A base in a pool is considered a variant if the P-value for both strands are significant after a Bonferroni correction for 2 x number of sequencing positions. If the coverage of one strand is below a pre-defined threshold, a significant P-value on the other strand (with good enough coverage) is enough to call a variant.

The parameters used for this analysis were: 97.5th percentile of minor allele frequencies to estimate the sequencing error rate and 50x as the minimum coverage in a pool to call a variant. These parameters were chosen empirically looking for consistency with the other calling algorithms.

This algorithm does not detect insertions, only deletions, and this is done in a separate execution of the algorithm, where the deletion is treated as a fifth base ("-" alongside A, C, G and T). The sequencing error rate is set to $1/(1.5\ x\ Number\ of\ haplotypes\ in\ DNA\ pool)$ in both strands for deletions. vipR

only calls 1bp deletions, thus long deletions are called as a series of 1bp deletions. For this reason long deletion were re-formatted and for each pool the minimum $REFchrc$ and the minimum $ALTchrc$ across the positions that form the long deletion were selected as the allele counts for the long deletion. All deletions were also left aligned in order to have comparable results across algorithms.

vipR provides allele counts for all pools for each variant that was successfully called in at least one pool and provides a list of pools where the variant was successfully called. Only data in pools where the variant was successfully called and in pools where the $ALTchrc$ (alternative allele chromosome count) was 0 were included in the allele frequency calculation and in the analyses. If only pools with $ALTchrc > 0$ were included, minor allele frequency for rare variants would be overestimated.

SNVer

This algorithm (42) uses a binomial-binomial model to test whether the variation observed for a base is a variant in each pool. Then, it applies the Simes method (164) to calculate a pooled P-value across pools for each variant.

For each pool, SNVer (42) tests the null hypothesis that the population minor allele frequency (MAF_{pop}) for a base is smaller than or equal to a threshold (ρ) against the alternative hypothesis that $MAF_{pop} > \rho$, in which case it calls a variant.

$$H_0: MAF_{pop} \leq \rho$$

$$H_1: MAF_{pop} > \rho$$

Given a base with minor allele frequency MAF_{pop} in a population, if we sample N haploid individuals and n of them carry the minor allele, then we can assume that n follows a binomial distribution with parameters N and MAF_{pop}.

$$n \sim B(N, MAF_{pop})$$

If the coverage for that base is C and if there is no error rate we can assume that the minor allele count (MAc) belongs to a binomial distribution with parameters C and n/N.

$$MAc \sim B(C, n/N)$$

If there is error rate ε, under which the minor allele can be flipped to one of the other three nucleotides we can assume that MAc belongs to a binomial distribution with parameters C and $\frac{n}{N}(1-\varepsilon) + \frac{N-n}{N}\frac{\varepsilon}{3}$.

$$MAc \sim B\left(C, \frac{n}{N}(1-\varepsilon) + \frac{N-n}{N}\frac{\varepsilon}{3}\right)$$

To test the null hypothesis (H_0: $MAF_{pop} \leq \ \rho$) SNVer (42) estimates the probability that the minor allele count in a pool (MAc) is greater than or equal to the number of minor alleles observed ($obsMAc$), given a minor allele frequency (MAF_{pop}) equal to the threshold (ρ). If this probability is smaller than a given threshold, it rejects the null hypothesis and classifies the base as a variant.

$$P\left(MAc \geq obsMAc;\ MAF_{pop} = \rho\right) = 1 - P(MAc < obsMAc;\ MAF_{pop} = \rho)$$

To calculate this probability we have the distribution of MAc ($MAc \sim B\left(C, \frac{n}{N}(1 - \varepsilon) + \frac{N-n}{N}\frac{\varepsilon}{3}\right)$) and we know all the parameters except n, so we can sum over all the values of n using $n \sim B(N, MAF_{pop})$.

$$P(MAc; MAF_{pop}) = \sum_{n=0}^{N} P(MAc|n)P(n; MAF_{pop})$$

After testing the null hypothesis in each pool for a given base, we have a P-value per pool. SNVer (42) calls a variant if the null hypothesis is rejected in at least one pool. To do this, it applies the Simes method (164), which orders all the P-values ($P_1, P_2, \ldots, P_M$), so that $P_1$is the smallest P-value and P_M is the largest P-value, and then it estimates a pooled P-value as $\min(\frac{M}{j}P_j, for\ j = 1, \ldots, M)$.

Algorithm:

- Use SAMtools to obtain allele counts for each strand and to filter base quality below 17 or mapping quality below 20.
- Estimate sequencing error as coming from two sources of error, mapping error, which is set up to be 0.01 and base error, estimated as a weighted mean of the base quality.
- Test for strand bias and remove potential false positives using a one-sided binomial test for the alternative forward allele count and the alternative reverse allele count with a threshold of 10^{-4}.
- Tests for allele imbalance and remove potential false positives using a one-sided Fisher's exact test with a threshold of 10^{-4}.
- Inactivate strand bias and allele imbalance tests if more than a certain number of alternative allele counts are observed (30 by default).
- Removes bases with less than a minimum number of reads with the alternative allele for both strands (1 by default).
- Test whether variation seen for a base is a variant using a binomial-binomial model for each pool, and then apply the Simes method (23) to calculate a pooled P-value across pools. If this pooled P-value is significant after a Bonferroni correction for the number of tests a variant is called.

This algorithm is applied to detect SNPs but also insertions and deletions. All INDELs were left aligned in order to have comparable results across algorithms. Default settings, as described above, were used for this analysis.

SNVer estimates allele frequencies per pool and then estimates an overall allele frequency by calculating the average across pools of allele frequencies which are > 0. This calculation would overestimate allele frequencies for rare variants. Allele frequencies were obtained by dividing the number of $ALTchrc$ by the number of total chromosome counts across pools. For each variant SNVer only provides data from pools that pass all the QC steps.

Syzygy

Syzygy (43) calls variants by computing a logarithm of odds (LOD) score for each base that compares the likelihood of obtaining the data if there are no alternative allele chromosome counts ($ALTchrc$) with the likelihood of obtaining the data if there is at least one alternative allele chromosome count.

$$H_0: ALTchrc = 0$$

$$H_1: ALTchrc > 0$$

The likelihood computations use Bayes' Rule, and Watterson's theta (165) to generate prior probabilities.

For each variant we can classify the read counts (rc) into three categories: reference allele read counts ($REFrc$), alternative allele read counts ($ALTrc$) and sequencing error counts (SEc), and we can classify the chromosome counts into two categories: reference allele chromosome count ($REFchrc$) and alternative allele chromosome count ($ALTchrc$).

$$rc = \begin{pmatrix} REFrc \\ ALTrc \\ SEc \end{pmatrix}$$

The probability of the observed read counts (rc) if there are n alternative allele chromosomes counts ($P(rc \mid ALTchrc = n)$) is calculated using a multinomial distribution. A multinomial distribution is a generalization of the binomial distribution. Each of a number of independent trials leads to a success for one of a number of categories (instead of one of two categories for the binomial distribution), each category having a fixed probability of success. This distribution gives the probability for a combination of successes for the various categories. In this case the three categories are the number of $REFrc$, the number of $ALTrc$ and the number of SEc; the number of independent trials is the coverage (C) and their probabilities are computed taking into account the allele chromosome counts ($REFchrc$ and $ALTchrc$), the total number of haploid individuals in the pool (N) and error rate (ε) as follows:

$$rc \mid ALTchrc = n \sim Mulitnomial(C, \begin{pmatrix} P(REFrc = 1) \\ P(ALTrc = 1) \\ P(SEc = 1) \end{pmatrix})$$

$$P(REFrc = 1) = \frac{REFchrc}{N} - \frac{REFchrc}{N}\ x\ \varepsilon$$

$$P(ALTrc = 1) = \frac{n}{N} - \frac{n}{N}\ x\ \varepsilon$$

$$P(SEc = 1) = \ \varepsilon$$

In order to reduce computational cost, Syzygy (43) obtains the probability of the observed read counts (rc) if there are no alternative allele chromosome counts ($P(rc \mid ALTchrc = 0)$) using a binomial distribution with parameters the coverage C and $1 - \varepsilon$, which is equivalent but computationally more efficient.

The LOD score is the base 10 logarithm of the likelihood of obtaining the data if there is at least 1$ALTchrc$ (computed summing over the probabilities of obtaining the data if there are n $ALTchrc$ with $n = 1, \ldots, 50$ using a prior probability of 0.005 for having $ALTchrc > 0$), over the likelihood of obtaining the data if there are 0 $ALTchrc$ (computed using a prior probability of 0.99 for having 0 $ALTchrc$). The LOD score is computed for both strands separately ($LOD\ score\ fwd$ and $LOD\ score\ rev$) and also jointly ($LOD\ score\ joint$). If the $LOD\ score\ joint$ is $\geq\ 3$, which would indicate 1000 to 1 odds of obtaining the data observed if the alternative chromosome count is at least 1, a variant is called at that base.

$$LOD\ score\ fwd\ = \log_{10}(\frac{0.005\ P(rc_f \mid ALTchrc = 1) + \cdots + 0.005\ P(rc_f \mid ALTchrc = 50)}{0.99\ P(rc_f \mid ALTchrc = 0)})$$

$$LOD\ score\ joint\ = \log_{10}(\frac{P(rc_f \mid ALTchrc = 1) + \cdots + P(rc_f \mid ALTchrc = 50)}{P(rc_f \mid ALTchrc = 0)}$$

$$x\frac{P(rc_r \mid ALTchrc = 1) + \cdots + P(rc_r \mid ALTchrc = 50)}{P(rc_r \mid ALTchrc = 0)}x\frac{0.005}{0.99})$$

$$for\ rc_f\ the\ read\ counts\ in\ the\ forward\ strand$$

$$and\ rc_r\ the\ read\ counts\ in\ the\ reverse\ strand$$

Syzygy (43) also estimates variant allele dosages within pools and allele frequencies. To do this it performs an expectation –maximization (EM) algorithm. An EM algorithm is an iterative method to estimate parameters in statistical models where the model depends on observed and unobserved latent variables. In this case we have allele read counts for each variant in each pool (our observed variables), we want to estimate allele frequencies (our parameters), but in order to do that we need the number of $ALTchrc$ (our unobserved latent variables). This method starts by assigning random values to the set of parameters, then it computes the expected values for the latent variables using the random values assigned to the parameters, after that it estimates the parameters using the expected values for the latent variables, and so on. It carries on iterating between these two steps until it reaches convergence.

For a given variant in a given pool the algorithm starts by assigning a value of 0.4 to the minor allele frequency (MAF)

$$\widehat{MAF}_{iteration=1} = 0.4$$

Then, the posterior probability of having $n\ ALTchrc$ given the observed $ALTrc$ ($P(ALTchrc = n \mid ALTrc = obsALTrc)$) with $n = 0, \dots, N$ are obtained using Bayes' Rule and the allele frequency estimate from the previous step as follows:

$$P(ALTchrc = n \mid ALTrc = obsALTrc)_{iteration=1}$$
$$= P(ALTrc = obsALTrc \mid ALTchrc = n)\ x\ P(ALTchrc = n)$$

$$for\ ALTrc \mid ALTchrc = n \sim B(coverage, 0.001), \qquad if\ n = 0,$$

$$ALTrc \mid ALTchrc = n \sim B\left(coverage, \frac{n}{N}\right), \qquad if\ n > 0,$$

$$ALTchrc \sim B\left(N, \widehat{MAF}_{iteration=1}\right),$$

$$and\ n = 0, \dots, N$$

With these posterior probabilities the expected number of $ALTchrc$ in a pool are calculated as

$$E(ALTchrc)_{iteration=1} = \sum_{n=0}^{N} n\ x\ P(ALTchrc = n \mid ALTrc = obsALTrc)_{iteration=1}$$

After that, using the expected $ALTchrc$ the allele frequency is estimated as

$$\widehat{MAF}_{iteration=2} = \frac{E(ALTchrc)_{iteration=1}}{N}$$

These two steps are iterated as many times as required for convergence.

Algorithm:

- Filter positions with base quality below 22 or mapping quality below 1, and undertake additional filters on mapping quality according to alignment types as given by the CIGAR variable in the file obtained after alignment.
- Use SAMtools to obtain allele counts for each strand.
- Estimate sequencing error rate. Sequencing error rate is estimated by modelling the miscall rate ($(C - REFrc)/C$) assuming that the factors that explain base to base variation in the miscall rate are: strand, sequence context and coverage around a base. Details of the error rate estimation are given in **Appendix F**.
- If the LOD score calculated using both strands ($LOD\ score\ joint$) is ≥ 3 in at least one pool and the coverage is ≥ 50 a variant is called at that base.
- Test for strand bias using Fisher's Exact test when $LOD\ score\ joint \geq 3$ and $LOD\ score\ rev \leq -1.5$ or $LOD\ score\ fwd \leq -1.5$.
- Perform an EM algorithm to estimate variant allele frequencies and allele dosages in each pool.
- Syzygy (43) undertakes an additional test for strand bias, it constructs a strand logarithm of odds (LOD) score ($strand\ LOD\ score$) comparing the maximum of the likelihood of obtaining the data if the overall allele frequency is the same as the allele frequency for one strand and the allele frequency for the other strand is zero, against the likelihood of

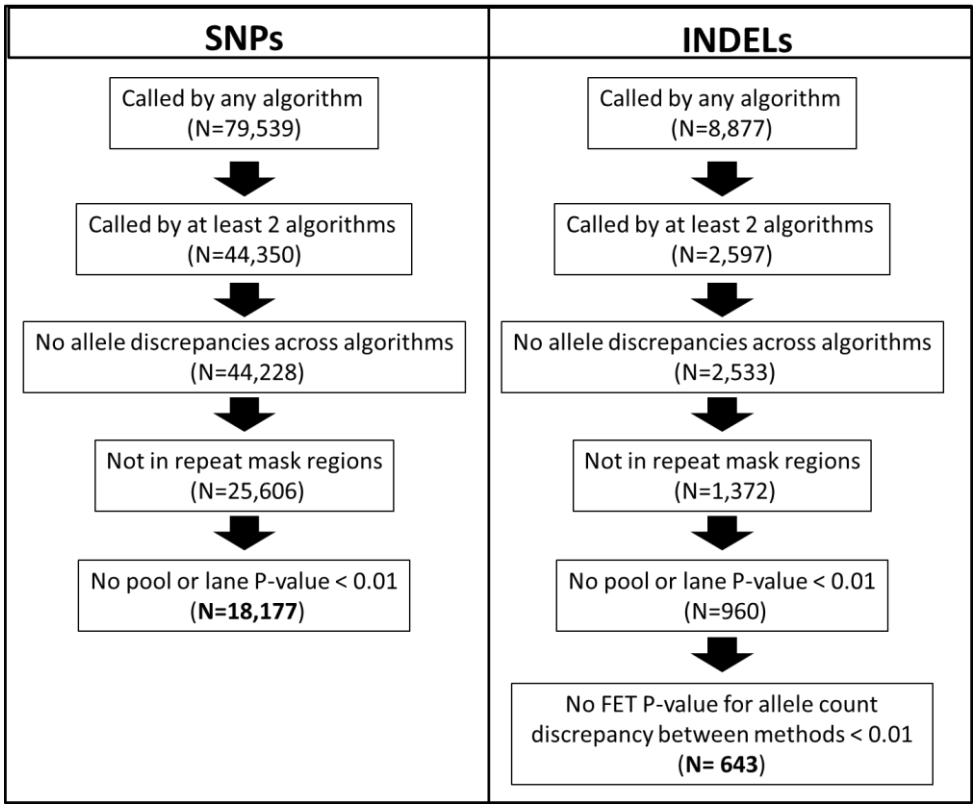
obtaining the data if the overall allele frequency is the same as the allele frequency calculated using information from either strand. Details of the method are given in **Appendix F**.

Syzygy classifies variants as high quality variants if: (i) $median\ (Fisher\ Pval\ to\ test\ for\ strand\ bias\ across\ pools) > 0.1$, (ii) there is no other variant within a 4 bp window and (iii) their $strand\ LOD\ score < 0$. For the analysis only high quality variants were selected. In addition, within the variants selected, data from pools which had a P-value < 0.05 for the Fisher's Exact test to test for strand bias were removed.

4.3.2.4 Quality control and selection of high quality variants

In order to select a subset of high quality variants out of those called by the different calling algorithms, to take forward for association testing, a number of additional quality control checks and filtering strategies were performed. The steps of this process are described in this section and illustrated with the number of variants kept in each step in **Figure 4-5**. Data quality for INDELs was poorer than for SNPs, for this reason an additional QC step for INDELs was undertaken (**Figure 4-5**).

Figure 4-5 Flow chart of the variant selection process

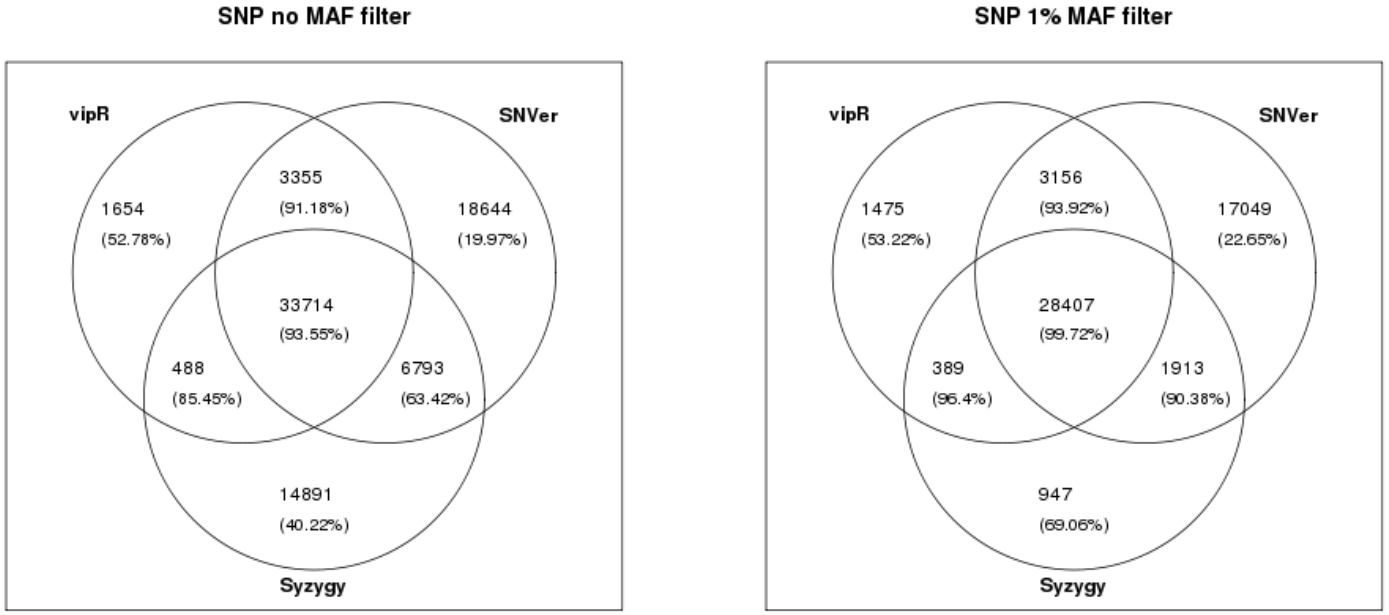


Variant overlap across algorithms

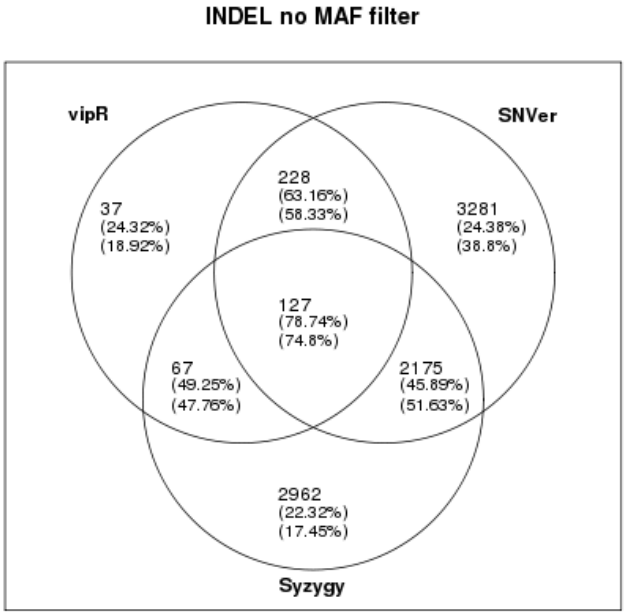
The overlap of variants called by each algorithm was examined (**Figure 4-6**). The assumption that most variants with MAF > 1% would already be included in public databases was used to assess which subset of variants was called more reliably. More than 90% of SNPs with MAF > 1% called by at least two algorithms were in dbNSP137 (166) (**Figure 4-6 a)**) and more than 46% of INDELs with MAF > 1% called by at least two algorithms were in 1000 Genomes Project Phase 1 (36) and more than 51% in Mills et al. (163) (**Figure 4-6 b)**). For this reason variants called by at least two algorithms were taken forward (**Figure 4-5**).

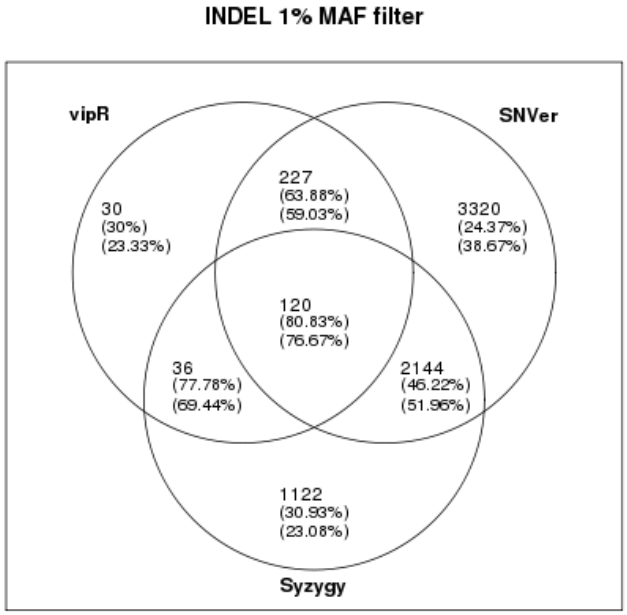
Figure 4-6 Venn diagrams of variants called by vipR, SNVer or Syzygy

a) Venn diagrams of SNPs called by any of the three algorithms with and without a 1% MAF filter. The proportion of SNPs included in dbSNP137 (166) for each section is presented in brackets.



b) Venn diagrams of INDELs called by any of the three algorithms with and without 1% MAF filter. The proportion of INDELs included in 1000 Genomes Project Phase 1 data (36) and the proportion included in Mills et al. (163) for each section are presented in brackets in this order.





Allele overlap across algorithms

A small number of variants (122 SNPs and 64 INDELs) were called by at least two algorithms but their alleles did not match across algorithms, so they were excluded (**Figure 4-5**).

Repeat mask regions

Despite applying repeat-masking filters at the probe design stage, some repeat mask regions were sequenced. Variants within repeat mask regions were removed (**Figure 4-5**). Repeat mask regions were extracted from UCSC table browser (167).

Pool and lane tests

The sequencing experiment consisted of 12 case pools and 12 control pools, grouped into 4 case lanes (with 3 pools each) and 4 control lanes (with 3 pools each). Note that 2 control pools were excluded due to low DNA quality. In order to remove variants which could generate false associations due to a lane or pool effect, two additional tests were run.

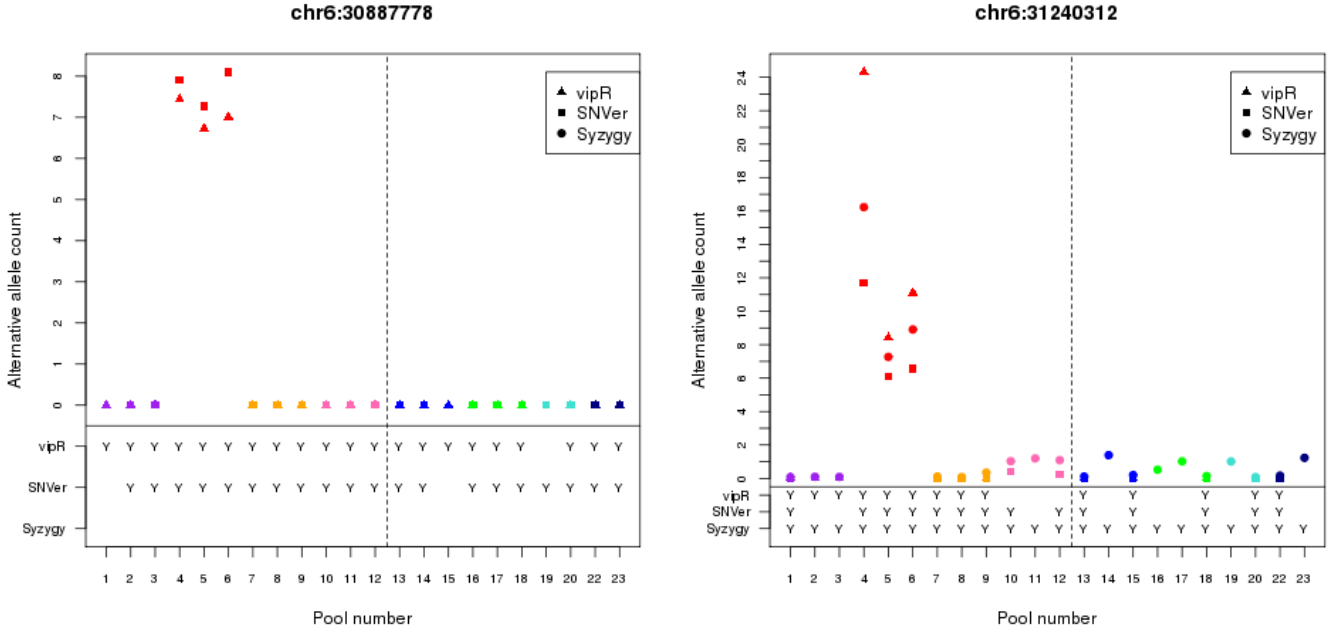
The lane test was designed to detect variants affected by lane effects; given that lanes included only case pools or only control pools, a sequencing artefact in a lane could lead to a false association. **Figure 4-7 a)** illustrates two

examples where alternative allele counts in pools 4, 5 and 6, all of them in lane 2, are higher than the rest, and would lead to a significant association, with no support from pools in other lanes; these two associations could be the result of a sequencing artefact in lane 2. A chi-square test with three degrees of freedom was run for the four case lanes and the four control lanes for each variant. The pool test assessed whether the data were consistent between case pools and between control pools, and was designed to detect sequencing artefacts in pools which could lead to false associations. **Figure 4-7 b)** illustrates two examples where significant associations would be driven by allele counts in one pool only (pool 4 for chr4:106565917 and pool 14 for chr6:32077690); and therefore a sequencing artefact in pool 4 or pool 14 could have driven these associations. A Binomial test was performed for each pool, testing whether the observed allele count would be expected given the number of chromosomes in the pool and the allele frequency observed across case or control pools (allele frequency was calculated separately for case pools and control pools). Variants with either a P-value in the lane test (lane P-value) for case pools or control pools, or a P-value in the pool test (pool P-value) for any pool < 0.01 were excluded (**Figure 4-5**).

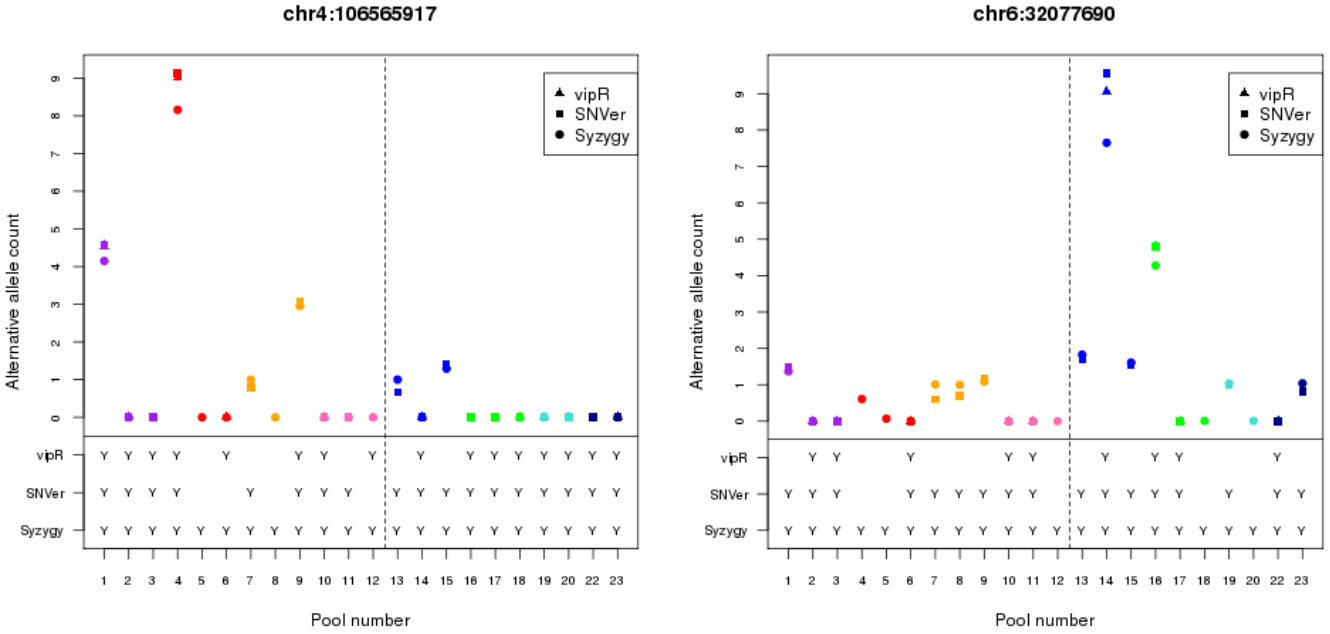
Figure 4-7 Allele count plots

The plots show alternative allele counts per pool on the y axis and pool numbers on the x axis. Counts obtained for each algorithm are represented with different symbols, as indicated in the legend. At the bottom of the plot it is indicated which of the three algorithms called the variant in each pool, with a “Y” from “Yes” if the variant was called and nothing if the variant was not called. A vertical dashed line separates case pools (1 to 12) from control pools (13 to 20, 22 and 23). Pools within the same lane are presented with the same color.

a) Example of variants that would be excluded due to lane P-value < 0.01



b) Example of variants that would be excluded due to pool P-value < 0.01



Allele frequency comparisons

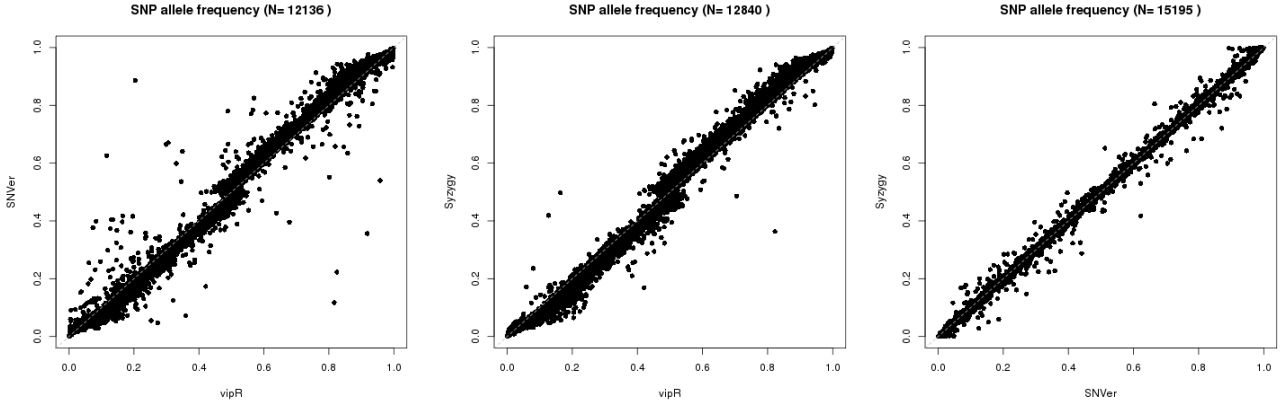
As a quality control check allele frequencies were compared between algorithms and with 1000 Genomes Project Phase 1 data (36). For SNPs, allele frequencies were consistent overall (**Figure 4-8**).

Allele frequency comparisons for INDELs showed more discrepancies, especially when comparing SNVer vs. Syzygy (**Figure 4-9 a)**). There was a subset of INDELs (55 variants in the top left corner and bottom right corner of the SNVer vs. Syzygy plot in **Figure 4-9 a)**), for which allele frequencies seemed to be flipped between SNVer and Syzygy. The two calling algorithms provided the same alleles for these variants (otherwise they would have been removed in a previous step), however the reference allele for one algorithm was the alternative allele for the other. These INDELs seemed to be in repetitive regions (not removed by filtering out repeat mask regions) where variants are harder to call, and it seemed that where one algorithm called an insertion the other algorithm called a deletion. Most of these variants were not present in 1000 Genomes Project Phase 1 data (36), probably indicating that they were calling artefacts rather than actual variants since they were not especially rare (only one variant with MAF < 1% for SNVer or Syzygy). The other subset of variants (247 INDELs) for which there was an allele frequency discrepancy between SNVer and Syzygy was present in 1000 Genomes Phase 1 data (36). The comparison with 1000 Genomes (36) allele frequencies (**Figure 4-9 b)**) showed that 1000 Genomes (36) and SNVer allele frequencies were consistent

for this set of variants, whereas there was a discrepancy between 1000 Genomes (36) and Syzygy allele frequencies. Looking at some of these variants in more detail it seemed that the discrepancy was caused by higher coverage estimated by Syzygy than by SNVer for these variants; probably due to Syzygy using a more liberal approach to filtering reads at an early stage of the pipeline, since this discrepancy appeared in the SAMtools output of allele counts. In order to remove variants with allele frequency discrepancies across algorithms Fisher's exact tests (FET) were performed for allele counts obtained by the different algorithms. INDELs with FET P-values < 0.01 using any allele frequency comparison between algorithms are indicated in red in **Figure 4-9** and were removed (**Figure 4-5**).

Figure 4-8 Allele frequency comparisons for SNPs

a) Between algorithms



b) With 1000 Genomes

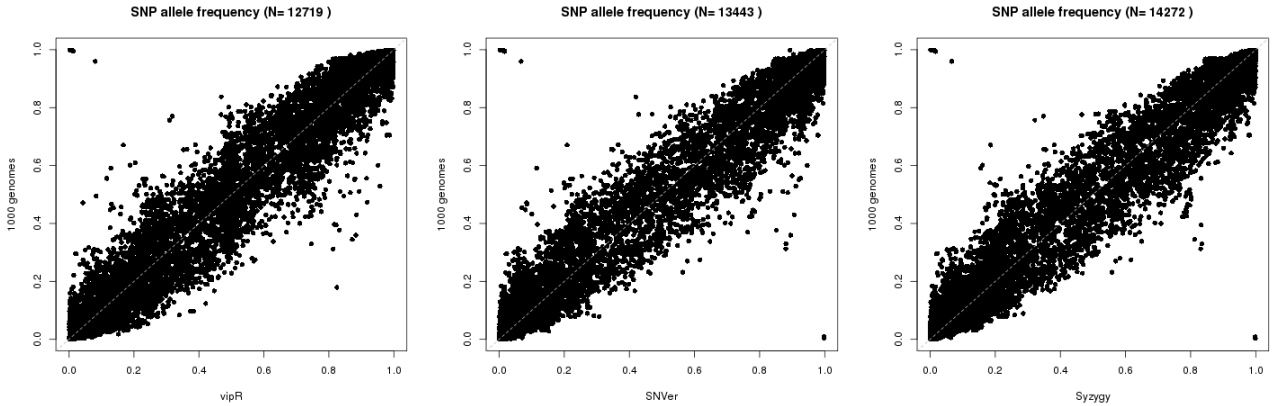
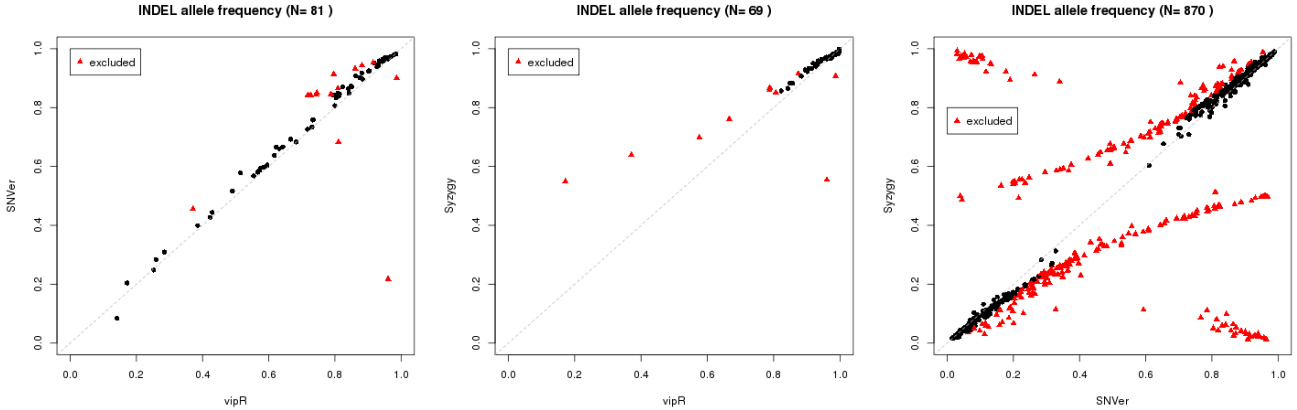
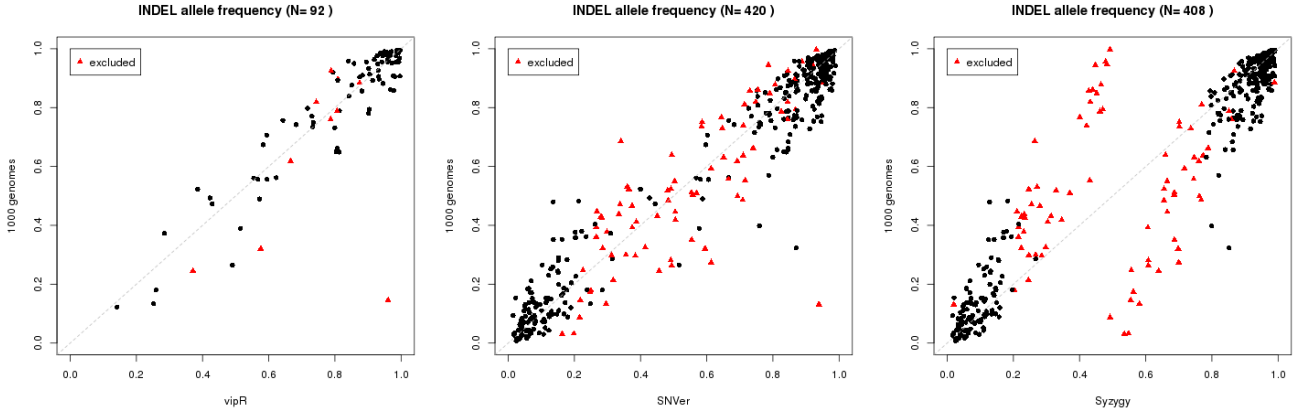


Figure 4-9 Allele frequency comparisons for INDELs

a) Between algorithms



b) With 1000 Genomes



The quality control checks undertaken illustrate that the quality of the data for INDELs was not as good as for SNPs. For this reason SNPs and INDELs were analysed separately and burden test analyses were only undertaken for SNPs.

4.3.2.5 Association testing

Single variant associations were tested as well as combined effects across variants using two different collapsing methods.

Single variant

Fisher's exact tests on allele counts were run in order to test whether the variants were associated with COPD risk. Tests were run both for SNPs and INDELs using allele counts produced by the different calling algorithms. Only variants that met the criteria described in section 4.3.2.4 were included.

Collapsing method: burden test

The burden test applied here is a modification of CCRaVAT (Case-Control Rare Variant Analysis Tool) (153), a method similar to QuTie (153) applied to quantitative traits in the SpiroMeta consortium in section 4.2, but for case-control analyses. CCRaVAT (153) tests whether accumulation of rare variants in a locus (number of individuals with at least one rare allele) is associated with COPD risk. This method assumes that all the variants included in the test will

exert their effect on COPD risk in the same direction, they will be either all protective or all detrimental. Only variants with MAF < 1% were included in this analysis. In order to infer how many individuals had at least one rare allele it was assumed that individuals with the alternative allele would always be heterozygous (rather than homozygous, since only overall allele count per pool was available). This seemed a sensible assumption, since the probability of having a homozygous individual out of 600 individuals for a variant with MAF = 1% is 10^{-4}. In addition, it was assumed that rare variants within a locus were independent, so that for example if there were 2 variants with 1 alternative allele count each in the same pool it was assumed they belonged to 2 different individuals. Fisher's exact test was used to test whether accumulation of rare variants in a locus was associated with COPD risk.

Collapsing method: C-alpha

In order to test whether a locus was associated with COPD risk allowing for variants to be protective or detrimental, the C-alpha test (53) was also applied. This test compares the variance of the observed distribution of alternative allele counts in cases relative to controls for a set of variants in a region, with the variance of their expected distribution in the case of no association. If variants in the region are either protective or detrimental, it would be expected that their alternative allele counts in cases would be either decreased (if protective) or increased (if detrimental); in both cases this would lead to over-dispersion.

The C-alpha test statistic T compares the variance of each observed alternative allele count in a locus with m variants, with the expected variance in the case of no association assuming a binomial distribution.

$$T = \sum_{i=1}^{m} [(y_i - n_i p_0)^2 - n_i p_0 (1 - p_0)]$$

$$for\ y_i\ number\ of\ alternative\ allele\ counts\ in\ cases, and\ y_i \sim B(n_i, p_i),$$

$$n_i number\ of\ alternative\ allele\ counts\ in\ cases\ and\ controls,$$

$$with\ i = 1, \dots, m$$

$$and\ p_0\ probability\ of\ observing\ an\ alternative\ allele\ count$$

$$in\ a\ case\ under\ the\ null\ hypothesis$$

$$(p_0 = \frac{Number\ of\ cases}{Number\ of\ individuals})$$

The variance of T is:

$$c = \sum_{n=2}^{\max n} m(n) \sum_{u=0}^{n} [(u - np_0)^2 - np_0(1 - p_0)]^2 f(u|n, p_0)$$

$$for\ m(n)\ the\ number\ of\ variants\ with\ n\ number\ of\ alternative\ allele\ counts$$

$$and\ f(u|n, p_0)\ the\ probability\ of\ u, assuming\ u \sim B(n, p_0)$$

Under the null hypothesis $\frac{T}{\sqrt{c}} \sim N(0,1)$; and we reject the null hypothesis if $\frac{T}{\sqrt{c}}$ is greater than expected using a one-tailed standard normal distribution for reference.

Singletons do not provide information on over-dispersion, and this test accounts for them by collapsing their allele accounts into one and then treating this allele count as coming from a single variant. In this sense the test assumes that all singletons will have an effect on the phenotype in the same direction, as does the burden test. The test does not run for loci with only singletons, as it would not run if only data for a single variant was present. This test works best for larger number of variants in a region and for sets of variants with similar allele frequencies (53). It also assumes that variants within a region are independent (53).

Collapsing method: region definition

Loci boundaries were defined in three different ways, in order to detect associations given different biological scenarios: (i) sliding window: 3kb sliding windows with an overlap of 1.5kb to detect the effect of regulatory variants, (ii) gene based: gene coordinates to detect gene effects, and (iii) exon based: exons, 5' UTR and 3' UTR for each gene to detect the effect of functional or regulatory variants within a gene. Tests were run separately for these three definitions using chromosome counts from each of the three algorithms used for variant calling. For each algorithm only SNPs that met the criteria described in section 4.3.2.4 and had MAF < 1% were included. Gene, UTR and exon coordinates were extracted from UCSC table browser (167) using the RefSeq Genes track.

Collapsing method: sensitivity analysis

In order to assess the effect on the results of the assumption that variants with MAF < 1% in a locus were independent, the collapsing tests were run again for the top hits after removing variants in LD ($r^2 > 0.2$) with each other. Within a group of variants in LD the one with the smallest P-value across methods was chosen. LD was calculated using the combined UK10K (http://www.uk10k.org/) and 1000 Genomes Project Phase 1 (36) (UK10K+1000G) reference panel. The effect on the results of assuming that variants not present in UK10K+1000G were independent or were in LD with any of the other variables in the region was assessed, by running the collapsing methods with and without variants not in UK10K+1000G.

4.3.2.6 Significance thresholds

Significance thresholds to account for multiple testing were defined for each of the 26 regions separately. As there is already strong prior evidence for the association of the 26 regions with lung function (2, 23, 94-96)*, no multiple testing adjustment for the number of regions was undertaken. Significance thresholds for the single variant and collapsing methods are described below and presented in **Table 4-6**.

Single variant

For each region the effective number (M_{eff}) of independent variants tested (equivalent to the number of independent tests) was estimated using the approach developed by Li and Ji (168), and then a Bonferroni correction for the number of independent tests was applied. The Li and Ji method (168) calculates the eigenvalues ($\lambda_i \; for \; i = 1, \ldots, M$) of a correlation matrix for the variants included in the region, and then calculates the number of independent variants using this formula:

$$M_{eff} = \sum_{i=1}^{M} f\left(|\lambda_i|\right)$$

$$f(x) = I(x \geq 1) + (x - [x]), x \geq 0$$

$$for \; \lambda_i \; the \; i^{th} \; eigenvalue,$$

$$I(x \geq 1) \; the \; indicator \; function \; which \; gives \; 1 \; when \; x \geq 1 \; and \; 0 \; otherwise,$$

$$[x] the \; floor \; function \; which \; gives \; the \; largest \; integer \leq x$$

UK10K+1000G data were used to estimate the correlation matrix for each region, and the variants not included in UK10K+1000G were assumed to be independent.

Collapsing method: sliding window

A Bonferroni correction was applied for the number of independent tests within each region. A test was undertaken for each sliding window. However, given the overlap between windows the number of independent tests was defined as the number of sliding windows divided by 2.

Collapsing method: gene based and exon based

A Bonferroni correction was applied for the number of independent tests within each region. A test was undertaken for each gene, so the number of independent tests was defined as the number of genes.

Table 4-6 Significance thresholds

Significance thresholds for each region are presented for SNPs and INDELs for the single variant analysis and for SNPs for the collapsing methods (no INDELs were included in the collapsing methods analysis).The column “GWAS gene” presents the gene reported in the lung function GWAS (2)* for each region. Abbreviations: Chr. = chromosome, N: number.

Chr.: start-end	GWAS gene	Variant type	Collapsing methods thresholds			N variants	N variants in UK10K+ 1000G	N tests in UK10K+1000G	N tests final	Single variant thresholds
			Gene based	Exon based	Slide window					
chr1:17238444-17455948	*MFAP2*	SNP	$1.25x10^{-2}$	$1.67x10^{-2}$	$1.72x10^{-3}$	338	291	99	146	$3.43x10^{-4}$
		INDEL				9	5	5	9	$5.56x10^{-3}$
chr1:218508675-218885482	*TGFB2*	SNP	$5x10^{-2}$	$5x10^{-2}$	$4.24x10^{-4}$	691	612	248	327	$1.53x10^{-4}$
		INDEL				36	23	18	31	$1.61x10^{-3}$
chr2:218627794-218818796	*TNS1*	SNP	$5x10^{-2}$	$5x10^{-2}$	$1.09x10^{-3}$	337	309	146	174	$2.87x10^{-4}$
		INDEL				5	5	5	5	$1x10^{-2}$
chr2:239839616-240332643	*HDAC4*	SNP	$5x10^{-2}$	$5x10^{-2}$	$3.68x10^{-4}$	1243	1159	396	480	$1.04x10^{-4}$
		INDEL				34	26	21	29	$1.72x10^{-3}$
chr3:25459833-25649422	*RARB*	SNP	$2.5x10^{-2}$	$2.5x10^{-2}$	$5.81x10^{-4}$	471	414	200	257	$1.95x10^{-4}$
		INDEL				24	18	14	20	$2.5x10^{-3}$
chr3:168791286-169391563	*MECOM*	SNP	$5x10^{-2}$	$5x10^{-2}$	$1.87x10^{-4}$	1326	1152	460	634	$7.89x10^{-5}$
		INDEL				54	37	29	46	$1.09x10^{-3}$
chr4:89637105-90077431	*FAM13A*	SNP	$5x10^{-2}$	$5x10^{-2}$	$5x10^{-4}$	666	591	200	275	$1.82x10^{-4}$
		INDEL				31	24	18	25	$2x10^{-3}$
chr4:106280233-106902828	*GSTCD*	SNP	$1x10^{-2}$	$1.25x10^{-2}$	$2.69x10^{-4}$	1031	922	328	437	$1.14x10^{-4}$
		INDEL				58	37	20	41	$1.22x10^{-3}$
chr4:145227600-145669881	*HHIP*	SNP	$5x10^{-2}$	$5x10^{-2}$	$2.76x10^{-4}$	802	686	248	364	$1.37x10^{-4}$

Chr.: start-end	GWAS gene	Variant type	Collapsing methods thresholds			N variants	N variants in UK10K+ 1000G	N tests in UK10K+1000G	N tests final	Single variant thresholds
			Gene based	Exon based	Slide window					
		INDEL				25	15	12	22	$2.27x10^{-3}$
chr5:94984019-95038027	*SPATA9*	SNP	$2.5x10^{-2}$	NA	$2.5x10^{-3}$	77	69	42	50	$1x10^{-3}$
		INDEL				2	2	2	2	$2.5x10^{-2}$
chr5:147682118-148026624	*HTR4*	SNP	$1.25x10^{-2}$	$1.67x10^{-2}$	$4.9x10^{-4}$	468	414	197	251	$1.99x10^{-4}$
		INDEL				17	13	7	11	$4.55x10^{-3}$
chr5:156597906-157139503	*ADAM19*	SNP	$8.33x10^{-3}$	$1.25x10^{-2}$	$5.88x10^{-4}$	670	615	236	291	$1.72x10^{-4}$
		INDEL				25	20	17	22	$2.27x10^{-3}$
chr6:27982152-28415572	*ZKSCAN3*	SNP	$3.85x10^{-3}$	$1x10^{-2}$	$4.39x10^{-4}$	520	459	162	223	$2.24x10^{-4}$
		INDEL				26	23	10	13	$3.85x10^{-3}$
chr6:30584612-31959223	*NCR3*	SNP	$1.22x10^{-3}$	$3.13x10^{-3}$	$2.79x10^{-4}$	3507	3307	647	847	$5.9x10^{-5}$
		INDEL				115	98	54	71	$7.04x10^{-4}$
chr6:31996092-32205942	*AGER*	SNP	$7.14x10^{-3}$	$2.5x10^{-2}$	$2.17x10^{-3}$	283	270	87	100	$5x10^{-4}$
		INDEL				14	11	10	13	$3.85x10^{-3}$
chr6:109159618-109305352	*ARMC2*	SNP	$5x10^{-2}$	$5x10^{-2}$	$1.11x10^{-3}$	213	189	97	121	$4.13x10^{-4}$
		INDEL				9	7	6	8	$6.25x10^{-3}$
chr6:142613055-142968973	*GPR126*	SNP	$2.5x10^{-2}$	$5x10^{-2}$	$4.27x10^{-4}$	443	388	196	251	$1.99x10^{-4}$
		INDEL				20	9	6	17	$2.94x10^{-3}$
chr9:98153197-98313032	*PTCH1*	SNP	$5x10^{-2}$	NA	$9.8x10^{-4}$	226	200	96	122	$4.1x10^{-4}$
		INDEL				13	8	6	11	$4.55x10^{-3}$
chr10:12170174-12335588	*CDC123*	SNP	$1.67x10^{-2}$	$5x10^{-2}$	$1.09x10^{-3}$	226	192	91	125	$4x10^{-4}$
		INDEL				11	4	4	11	$4.55x10^{-3}$
chr10:77532518-78643886	*C10orf11*	SNP	$5x10^{-2}$	NA	$1.71x10^{-4}$	1513	1336	535	712	$7.02x10^{-5}$

Chr.: start-end	GWAS gene	Variant type	Collapsing methods thresholds			N variants	N variants in UK10K+ 1000G	N tests in UK10K+1000G	N tests final	Single variant thresholds
			Gene based	Exon based	Slide window					
		INDEL				31	17	15	29	$1.72x10^{-3}$
chr12:57472676-57617125	*LRP1*	SNP	$1.25x10^{-2}$	$2.5x10^{-2}$	$1.56x10^{-3}$	169	155	87	101	$4.95x10^{-4}$
		INDEL				2	2	2	2	$2.5x10^{-2}$
chr12:96041582-96400071	*CCDC38*	SNP	$8.33x10^{-3}$	$1.67x10^{-2}$	$5.05x10^{-4}$	651	586	216	281	$1.78x10^{-4}$
		INDEL				26	21	18	23	$2.17x10^{-3}$
chr15:71423787-72085722	*THSD4*	SNP	$5x10^{-2}$	$5x10^{-2}$	$2.66x10^{-4}$	1266	1151	446	561	$8.91x10^{-5}$
		INDEL				32	19	18	31	$1.61x10^{-3}$
chr16:57906243-58143392	*MMP15*	SNP	$1x10^{-2}$	$1.67x10^{-2}$	$1.61x10^{-3}$	310	288	130	152	$3.29x10^{-4}$
		INDEL				9	4	4	9	$5.56x10^{-3}$
chr16:75252927-75538926	*CFDP1*	SNP	$1x10^{-2}$	$2.5x10^{-2}$	$5.88x10^{-4}$	517	481	177	213	$2.35x10^{-4}$
		INDEL				7	2	2	7	$7.14x10^{-3}$
chr21:35595821-35753440	*KCNE2*	SNP	$2.5x10^{-2}$	$5x10^{-2}$	$1.25x10^{-3}$	213	190	108	131	$3.82x10^{-4}$
		INDEL				8	6	6	8	$6.25x10^{-3}$

4.3.2.7 Selection of top hits

In order to minimise false positive associations, the criteria to select the top hits required that a variant met the significance threshold using allele counts for one calling algorithm and that it also showed supporting evidence (described below) when using allele counts from another calling algorithm. The motivation for this was that the most significant ($P = 5.8 \times 10^{-10}$) association across regions and algorithms was achieved using allele counts obtained by Syzygy (43), however when using SNVer allele counts it was not significant (P = 0.13). When looking in more detail at this signal, Syzygy had classified it as very high quality, however the $LOD\ score$ for the pools that seemed to drive the associations was high for one strand but rather low (around -1) for the other, suggesting evidence of strand bias, although none of them made the -1.5 threshold to test for strand bias in Syzygy.

Single variant

Variants were selected if their FET P-value met the significance threshold using the allele counts for at least one calling algorithm and showed supportive evidence (P-value < threshold x 2) when using allele counts from another calling algorithm.

Collapsing method

Loci were selected if: (i) their P-value met the significance threshold using the allele counts for at least one calling algorithm after the sensitivity analysis (either assuming that variants not in UK10K+1000G were independent or were in LD with other variants in the region) and (ii) supportive evidence (P-value < threshold x 2) was shown when using allele counts from another calling algorithm after the sensitivity analysis (either assuming that variants not in UK10K+1000G were independent or were in LD with other variants in the region).

Alignments were visually inspected for all the single variant top hits and a random sample of the variants in the collapsing method top hits.

4.3.2.8 Follow-up (stage 2)

Follow-up resource

Variants and loci selected were followed-up in UK BiLEVE, a subset of ~50,000 individuals from UK Biobank (http://www.ukbiobank.ac.uk/) sampled from the extremes of the % predicted FEV_1 distribution separately in never-smokers and heavy-smokers.

An Affymetrix custom array was designed for the genome-wide genotyping of the UK BiLEVE project. This array includes 130K rare missense and loss of function variants (selected to be polymorphic in UK populations based on currently available “exome chip” data), 642K variants selected for optimal imputation of common variation and improved imputation of low frequency variation (MAF 1-5%), and 9000 variants selected for improved coverage of known and candidate respiratory regions. These data had been imputed against the UK10K+1000G joint reference panel using SHAPEIT (169) and IMPUTE2 (125).

Selection of cases and controls

The sampling frame was made of 41,260 individuals over 40 years old, with no asthma (diagnosed or self-reported) who smoked between 5 and 100 pack years. Case-control status was defined as in the COPD case-control sequencing study (stage 1). COPD cases were individuals with COPD spirometric GOLD stage 2 (104) or above (percent predicted FEV_1 < 80% and FEV_1/FVC < 0.7) and controls were individuals with percent predicted FEV_1 > 80% and FEV_1/FVC > 0.7, based on pre-bronchodilator spirometry. Percent predicted FEV_1 was obtained using reference values from healthy (no respiratory diseases diagnosed) never-smokers (N= 81,719) from UK Biobank. In total there were 4,249 COPD cases and 11,916 controls.

Association testing: single variant

The association of single variants with COPD risk in UK BiLEVE was tested using logistic regression on allele dosages obtained from the imputation output. IMPUTE2 (125) provides probabilities for each of three possible genotypes for each variant ($P0, P1, P2$) and allele dosages were obtained as $0xP0 + 1xP1 + 2xP2$. An adjustment for 5 principal components of ancestry was included; and an adjustment for pack-years of smoking was undertaken as a sensitivity analysis.

Association testing: burden test

The same method as in stage 1 was used. The most likely genotype for each individual was used for the analysis, only including genotypes with probability (as given by IMPUTE2 (125)) > 0.9. Sensitivity analyses were undertaken for the top hits including only independent variants ($r^2 < 0.2$) within each locus.

Association testing: C-alpha test

The same method as in stage 1 was used. The most likely genotype for each individual was used for the analysis using a threshold of 0.9. Sensitivity analyses were undertaken for the top hits including only independent variants ($r^2 < 0.2$) within each locus. In addition, 10,000 permutations (permuting case

control status) were run for the top hits, only including independent variants, in order to obtain more accurate P-values.

Association testing: collapsing methods region definition

The same region boundaries as in stage 1 were used and only variants with MAF < 1%, HWE P-value > 10^{-6} and imputation quality ≥ 0.8 were included.

Association testing: significance thresholds

Significance thresholds per region were defined by a Bonferroni corrected threshold for the number independent tests undertaken in each region. For the single variant analysis the number of independent tests were the number of variants followed up. For the gene based and exon based analyses a gene was considered as an independent test, and for the sliding window analysis, two overlapping windows were counted as 1.5 tests.

4.3.3 Results

After undertaking quality control checks a total of 18,177 SNPs and 643 INDELs across the 26 regions were selected to be tested for association with COPD risk in 300 COPD cases and 250 controls (individuals from two control pools were excluded given low DNA quality). A subset of these variants were novel: 1,429 SNPs (93% with MAF < 1%) were not in dbSNP137 (166) or UK10K+1000G

and 216 INDELs (4% with MAF < 1%) were not in 1000 Genomes Project Phase 1 data (36) or Mills et al. (163).

4.3.3.1 Single variant association testing

A total of 8 SNPs and 3 INDELs (**Table 4-7**), with minor allele frequencies ranging from 1% to 40% for SNPs and from 3% to 5% for INDELs met the significance thresholds described in **4.3.2.7**, and were taken forward for follow up in UK BiLEVE. All these variants were present in the UK10K+1000G imputation reference panel and had imputation quality > 0.8.

Table 4-7 Single variants associated with COPD risk: stage 1

Single variants results for stage 1 are presented for variants that met the criteria for follow-up. The columns “Threshold” and “Threshold support” present the threshold and the threshold for supporting evidence for each region. The column “GWAS gene” presents the gene reported in the lung function GWAS (2)* for each region. Abbreviations: chr. = chromosome, Ref. = reference, Alt. = alternative, M.A.F. = minor allele frequency and a.c. = allele count.

Rs number (chr.: position), function	GWAS gene	Ref. allele	Alt. allele	Threshold	Threshold support	Calling algorithm	M.A.F.	Alt. a.c. cases	Alt. a.c. controls	P-value
rs11678706 (chr2:239908773), intergenic	*HDAC4*	A	C	$1.04x10^{-4}$	$2.08x10^{-4}$	vipR	0.219	111	20	$6.92x10^{-1}$
						SNVer	0.158	120	51	$8.26x10^{-6}$
						Syzygy	0.159	118	57	$1.87x10^{-4}$
rs16854211 (chr3:169338409), intronic (*MECOM*)	*MECOM*	G	T	$7.89x10^{-5}$	$1.58x10^{-4}$	vipR	0.169	116	50	$4.23x10^{-3}$
						SNVer	0.159	117	53	$5.29x10^{-5}$
						Syzygy	0.155	117	54	$7.92x10^{-5}$
rs1895031 (chr3:169354498), intronic (*MECOM*)	*MECOM*	G	C	$7.89x10^{-5}$	$1.58x10^{-4}$	vipR	0.291	138	173	$9x10^{-4}$
						SNVer	0.289	146	178	$5x10^{-5}$
						Syzygy	0.298	150	178	$1.58x10^{-4}$
rs999741 (chr5:147727048), transcript (*RP11-373N22.3*)	*HTR4*	C	G	$1.99x10^{-4}$	$3.98x10^{-4}$	vipR	0.254	88	144	$2.15x10^{-2}$
						SNVer	0.231	113	146	$6.24x10^{-5}$
						Syzygy	0.235	116	143	$3.53x10^{-4}$
rs193259319 (chr5:147823559), downstream (*FBXO38*)	*HTR4*	T	C	$1.99x10^{-4}$	$3.98x10^{-4}$	vipR	0.043	24	2	$2.56x10^{-5}$
						SNVer	0.039	30	6	$1.26x10^{-4}$
						Syzygy	0.033	30	6	$2.9x10^{-4}$
rs138649528 (chr6:30776469), downstream (*NCRNA00243*)	*NCR3*	AT	A	$7.04x10^{-4}$	$1.41x10^{-3}$	vipR	NA	NA	NA	NA
						SNVer	0.032	9	29	$1.02x10^{-4}$

Rs number (chr.: position), function	GWAS gene	Ref. allele	Alt. allele	Threshold	Threshold support	Calling algorithm	M.A.F.	Alt. a.c. cases	Alt. a.c. controls	P-value
						Syzygy	0.033	10	26	$1.11x10^{-3}$
rs35278224;rs67982043 (chr6:32164665), intronic (*NOTCH4*)	*AGER*	CT	C	$3.85x10^{-3}$	$7.69x10^{-3}$	vipR	NA	NA	NA	NA
						SNVer	0.045	13	39	$1.2x10^{-5}$
						Syzygy	0.046	16	34	$1.24x10^{-3}$
rs146088795 (chr6:142640832), intronic (*GPR126*)	*GPR126*	A	G	$1.99x10^{-4}$	$3.98x10^{-4}$	vipR	0.011	0	7	$1.98x10^{-4}$
						SNVer	0.018	3	16	$4.08x10^{-3}$
						Syzygy	0.018	3	17	$3.73x10^{-4}$
rs7174934 (chr15:71571345), intronic (*THSD4*)	*THSD4*	G	A	$8.91x10^{-5}$	$1.78x10^{-4}$	vipR	0.432	191	248	$3.54x10^{-4}$
						SNVer	0.423	221	251	$1.03x10^{-5}$
						Syzygy	0.434	228	249	$9.18x10^{-5}$
rs75958385 (chr16:75403497), intronic (*CFDP1*)	*CFDP1*	G	A	$2.35x10^{-4}$	$4.7x10^{-4}$	vipR	0.013	0	7	$4.16x10^{-4}$
						SNVer	0.015	2	13	$2.75x10^{-3}$
						Syzygy	0.014	1	14	$2.31x10^{-4}$
rs199588075 (chr21:35679578), transcript (*AP000318.2*)	*KCNE2*	CT	C	$6.25x10^{-3}$	$1.25x10^{-2}$	vipR	0.048	25	0	$1.55x10^{-5}$
						SNVer	0.034	29	7	$1.78x10^{-3}$
						Syzygy	NA	NA	NA	NA

Two variants had nominally significant P-values (P < 0.05) in UK BiLEVE (**Table 4-8**). One of these (rs75958385 in the *CFDP1* region) had opposite direction of effect in UK BiLEVE compared to stage 1 (**Table 4-7** and **Table 4-8**), indicating that it was probably a false positive association. The other variant (rs999741 in the *HTR4* region) had a P-value = 0.002 and MAF of 26% in UK BiLEVE and it was in a long non-coding RNA (lncRNA) region. The alternative allele of this variant and of the sentinel SNP previously reported (2)* rs1985524 in this region, were positively correlated (r = 0.45 and r^2 = 0.2 in UK BiLEVE) and both had a protective effect. When conditioning on the previously reported variant (2)* (rs1985524), the rs999741 signal disappeared (**Table 4-9**) indicating that this association was most likely a consequence of its LD with the more significant previously reported SNP (rs1985524).

Table 4-8 Single variant top hits results in stage 2

The column “GWAS gene” presents the gene reported in the lung function GWAS (2)* for each region. Abbreviations: chr. = chromosome, freq = frequency, OR = odds ratio, SE = standard error.

Rs number (chr: position), function	GWAS gene	Non coded allele	Coded allele	Coded allele freq.	Imputation information	Without pack-years adjustment			With pack-years adjustment			Consistent direction of effect?
						OR	SE	P-value	OR	SE	P-value	
rs11678706 (chr2:239908773), intergenic	*HDAC4*	A	C	0.175	0.986	0.953	0.034	$1.52x10^{-1}$	0.955	0.035	$1.88x10^{-1}$	NO
rs16854211 (chr3:169338409), intronic (*MECOM*)	*MECOM*	G	T	0.156	0.977	1.058	0.035	$1.07x10^{-1}$	1.055	0.036	$1.4x10^{-1}$	YES
rs1895031 (chr3:169354498), intronic (*MECOM*)	*MECOM*	G	C	0.295	1	0.987	0.028	$6.5x10^{-1}$	0.978	0.029	$4.5x10^{-1}$	YES
rs999741 (chr5:147727048), transcript (*RP11-373N22.3*)	*HTR4*	C	G	0.256	0.999	0.915	0.029	$2x10^{-3}$	0.928	0.031	$1.5x10^{-2}$	YES
rs193259319 (chr5:147823559), downstream (*FBXO38*)	*HTR4*	T	C	0.021	1	1.053	0.085	$5.46x10^{-1}$	1.034	0.089	$7.11x10^{-1}$	YES
rs138649528 (chr6:30776469), downstream (*NCRNA00243*)	*NCR3*	AT	A	0.028	0.999	1.062	0.073	$5.7x10^{-2}$	1.057	0.077	$9.8x10^{-2}$	NO
rs35278224;rs67982043 (chr6:32164665), intronic (*NOTCH4*)	*AGER*	CT	C	0.047	0.99	1.081	0.058	$1.8x10^{-1}$	1.106	0.06	$9.4x10^{-2}$	NO
rs146088795 (chr6:142640832), intronic (*GPR126*)	*GPR126*	A	G	0.016	0.987	0.934	0.104	$5.1x10^{-1}$	0.907	0.109	$3.69x10^{-1}$	YES
rs7174934 (chr15:71571345), intronic (*THSD4*)	*THSD4*	G	A	0.414	0.987	1.002	0.026	$9.39x10^{-1}$	0.997	0.027	$9x10^{-1}$	NO
rs75958385 (chr16:75403497), intronic (*CFDP1*)	*CFDP1*	G	A	0.029	0.956	1.185	0.075	$2.4x10^{-2}$	1.182	0.079	$3.3x10^{-2}$	NO
rs199588075 (chr21:35679578), transcript (*AP000318.2*)	*KCNE2*	CT	C	0.031	0.955	0.886	0.079	$1.25x10^{-1}$	0.871	0.083	$9.5x10^{-2}$	NO

Table 4-9 Conditional analysis in *HTR4* region

The column “GWAS gene” presents the gene reported in the lung function GWAS (2)* for each region. Abbreviations: chr. = chromosome, freq = frequency, OR = odds ratio, SE =standard error.

Rs number (chr.: position), function	GWAS gene	Non coded allele	Coded allele	Coded allele freq.	Imputation information	Unconditional			Joint			r
						OR	SE	P-value	OR	SE	P-value	
rs999741 (chr5:147727048), transcript (*RP11-373N22.3*)	*HTR4*	C	G	0.256	0.999	0.915	0.029	$2.35x10^{-3}$	0.969	0.033	$3.3x10^{-1}$	0.45
rs1985524 (chr5:147847788), intronic (*HTR4*)	*HTR4*	G	C	0.445	1	0.882	0.026	$1.08x10^{-6}$	0.893	0.029	$8.54x10^{-5}$	

Results for the 26 sentinel SNPs previously reported (or other SNPs in perfect LD with the sentinels) are presented in **Appendix G**. SNPs in four regions (*MECOM*, *HHIP*, *SPATA9*, *HTR4*) out of the 26 had nominally significant P-values when using allele counts for at least two calling algorithms and their direction of effect agreed with the previously reported (2)* effect on lung function (negative effect on lung function and increased risk of COPD, or positive effect on lung function and reduced risk of COPD) (**Table 4-10**). Association with COPD risk for *HHIP* and *HTR4* had already been reported (24, 95)*, but not for *SPATA9* or *MECOM.* The P-values for the *SPATA9* association were only nominally significant (P = 0.020 for SNVer and P = 0.023 for Syzygy) (**Table 4-10**), but the P-values for *MECOM* (P = 5 x 10^{-4} for SNVer and P = 6 x 10^{-4} for Syzygy) met the Bonferroni corrected threshold for 26 tests (2 x 10^{-3}) (**Table 4-10**). Overall, 13 variants of the 26 had consistent direction of effect when using allele counts for at least two calling algorithms (**Appendix G**).

Table 4-10 Single variant results for known variants

Abbreviations: Ref. = reference, Alt. or alt. = alternative, MAF =minor allele frequency, N = number, a.c. = allele count, OR = odds ratio.

Rs number	GWAS gene	Calling algorithm	Ref. allele	Alt. allele	MAF	N. alt. a.c. case	N. alt. a.c. control	OR	P-value
rs1344555	*MECOM*	SNVer	C	T	0.2	143	77	1.72	$4.93x10^{-4}$
		Syzygy	C	T	0.21	148	81	1.69	$5.93x10^{-4}$
		vipR	C	T	0.23	130	54	1.6	$9.29x10^{-3}$
rs11100860	*HHIP*	SNVer	A	G	0.37	202	205	0.73	$1.44x10^{-2}$
		Syzygy	A	G	0.37	204	202	0.76	$3.28x10^{-2}$
		vipR	A	G	0.4	159	203	0.97	$8.37x10^{-1}$
rs153916	*SPATA9*	SNVer	C	T	0.43	360	265	1.33	$2.03x10^{-2}$
		Syzygy	C	T	0.43	363	268	1.33	$2.35x10^{-2}$
		vipR	C	T	0.45	237	191	1.09	$1.33x10^{-1}$
rs1985524	*HTR4*	SNVer	G	C	0.41	227	224	0.75	$2.27x10^{-2}$
		Syzygy	G	C	0.41	230	223	0.77	$3.67x10^{-2}$
		vipR	G	C	0.44	171	223	0.93	$5.89x10^{-1}$

4.3.3.2 Collapsing method

A total of 59 3kb sliding windows from 18 regions out of the 26 regions sequenced, and 23 genes (21 from gene based tests, 1 from exon based tests and 1 that was selected for both) from 19 regions out of the 26 sequenced met the criteria described in 4.3.2.7 (**Appendix G**)**,** and were taken forward to be followed up in UK BiLEVE. Of these, two sliding windows and 3 genes from the gene based analysis were selected due to their P-values in the burden analysis. All the remaining windows and genes were selected because of their C-alpha test P-values. Four of the five regions selected in the burden test were also selected in the C-alpha test (the exception was *GRP126*). Only 32 sliding windows out of the 59 were tested for association in UK BiLEVE, because only 32 windows had at least two variants that met the conditions specified in 4.3.2.8

(MAF < 1%, imputation quality > 0.8 and HWE P > 10^{-6}) in the UK BiLEVE data. All the 23 genes had at least two variants in UK BiLEVE that met the conditions specified in 4.3.2.8.

None of burden test results were significant (P < 0.05) in UK BiLEVE (**Table 4-11**).

Table 4-11 Burden test results in stage 2

The column "GWAS gene" presents the gene reported in the lung function GWAS (2)* for each region.

Locus	GWAS gene	Number of variants	P-value
chr3:168984786-168987786	*MECOM*	2	$4.15x10^{-1}$
FLJ20184 (chr4:106473776-106552837)	*GSTCD*	46	$3.84x10^{-1}$
chr4:145278600-145281600	*HHIP*	2	$8.83x10^{-1}$
ITK (chr5: 156607906-156682109)	*ADAM19*	23	$7.58x10^{-1}$
GPR126 (chr6:142623055-142767403)	*GPR126*	82	$8.7x10^{-1}$

In the C-alpha test undertaken for sliding windows, one sliding window (chr3:169238286-169241286 in the *MECOM* region) met a Bonferroni corrected threshold (P < 8 x 10^{-3}) for that region, and two sliding windows (chr3:25633833-25636833 in the *RARB* region and chr4:145293600-145296600 in the *HHIP* region) showed suggestive evidence (P = 0.053 and P = 0.042 respectively), but did not meet a Bonferroni corrected threshold (0.0125 for both regions) (**Table 4-12 a)**). In the C-alpha test for gene based analysis, *C10orf11* met the significance threshold (P = 0.04, Bonferroni corrected thresthold for 1 test = 0.05) (**Table 4-12 b)**), and *TNXB*, in the *AGER* region, showed

suggestive evidence (P = 0.061, Bonferroni corrected thresthold for 1 test = 0.05) (**Table 4-12 b)**).In the C-alpha test for exon based analysis another gene (*NPNT*) in the *GSTCD* region showed suggestive evidence of association (P = 0.053, Bonferroni corrected threshold = 0.025) (**Table 4-12 c)**).

Table 4-12 C-alpha test stage 2 results

"GWAS gene" is the gene reported in the lung function GWAS (2)* for each region. P-values that reach a Bonferroni corrected threshold as defined in **4.3.2.8** are highlighted in bold.

a) Sliding window

Locus	GWAS gene	Number of variants	Number of alternative allele counts in cases and controls	P-value
chr1:218531175-218534175	*TGFB2*	3	421	$8.13x10^{-1}$
chr2:218807794-218810794	*TNS1*	2	244	$7.57x10^{-1}$
chr2:239973116-239976116	*HDAC4*	2	147	$7.3x10^{-1}$
chr2:240325616-240328616	*HDAC4*	3	401	$6.57x10^{-1}$
chr3:168984786-168987786	*MECOM*	2	252	$8.25x10^{-1}$
chr3:169238286-169241286	*MECOM*	2	573	$\mathbf{2.94x10^{-3}}$
chr3:169310286-169313286	*MECOM*	4	604	$8.35x10^{-1}$
chr3:169311786-169314786	*MECOM*	2	238	$6.73x10^{-1}$
chr3:169340286-169343286	*MECOM*	3	665	$8.87x10^{-1}$
chr3:169341786-169344786	*MECOM*	3	590	$8.71x10^{-1}$
chr3:169371786-169374786	*MECOM*	5	908	$5.26x10^{-1}$
chr3:169373286-169376286	*MECOM*	4	607	$1.99x10^{-1}$
chr3:25464333-25467333	*RARB*	4	466	$4.48x10^{-1}$
chr3:25510833-25513833	*RARB*	6	1155	$8.04x10^{-1}$
chr3:25512333-25515333	*RARB*	4	878	$7.54x10^{-1}$
chr3:25632333-25635333	*RARB*	4	494	$1.51x10^{-1}$
chr3:25633833-25636833	*RARB*	3	363	$5.32x10^{-2}$
chr4:145269600-145272600	*HHIP*	2	158	$7.54x10^{-1}$
chr4:145278600-145281600	*HHIP*	2	210	$7.82x10^{-1}$
chr4:145293600-145296600	*HHIP*	2	167	$4.16x10^{-2}$
chr4:145341600-145344600	*HHIP*	2	299	$4.18x10^{-1}$
chr5:147829118-147832118	*HTR5*	3	276	$5.75x10^{-1}$
chr5:147830618-147833618	*HTR6*	4	597	$7.83x10^{-1}$
chr5:156912906-156915906	*ADAM19*	2	480	$6.32x10^{-1}$
chr9:98180197-98183197	*PTCH1*	6	565	$4.81x10^{-1}$
chr9:98181697-98184697	*PTCH1*	4	358	$6.61x10^{-1}$
chr10:12207674-12210674	*CDC123*	2	423	$8.36x10^{-1}$
chr12:57529676-57532676	*LRP1*	3	280	$3.34x10^{-1}$
chr12:96157082-96160082	*CCDC38*	4	856	$8.65x10^{-1}$
chr12:96158582-96161582	*CCDC38*	5	1115	$8.85x10^{-1}$
chr15:71704287-71707287	*THSD4*	2	205	$7.5x10^{-1}$
chr21:35646821-35649821	*KCNE2*	2	245	$5.77x10^{-1}$

b) Gene based

Locus	GWAS gene	Number of variants	Number of alternative allele counts in cases and controls	P-value
TGFB2 (chr1:218518675-218617961)	*TGFB2*	65	7090	$6.58x10^{-1}$
TNS1 (chr2:218664511-218808796)	*TNS1*	84	13588	$1.55x10^{-1}$
HDAC4 (chr2:239969863-240322643)	*HDAC4*	247	34906	$9.59x10^{-1}$
RARB (chr3:25469833-25639422)	*RARB*	76	9629	$8.39x10^{-1}$
MECOM (chr3:168801286-169381563)	*MECOM*	450	69543	1
FAM13A (chr4:89647105-89978323)	*FAM13A*	123	13561	$6.55x10^{-1}$
FLJ20184 (chr4:106473776-106552837)	*GSTCD*	46	5725	$6.25x10^{-1}$
HHIP (chr4:145567147-145659881)	*HHIP*	31	3882	$7.94x10^{-1}$
ITK (chr5:156607906-156682109)	*ADAM19*	23	3253	$5.33x10^{-1}$
DDR1 (chr6:30856464-30867933)	*NCR3*	7	953	$7.56x10^{-1}$
TNXB (chr6:32008931-32077151)	*AGER*	37	8086	$6.08x10^{-2}$
ARMC2 (chr6:109169618-109295352)	*ARMC2*	65	8782	$9.21x10^{-1}$
LOC153910 (chr6:142847591-142958973)	*GPR126*	86	12820	$5.46x10^{-1}$
PTCH1 (chr9:98205263-98270831)	*PTCH1*	52	6850	$2.58x10^{-1}$
NUDT5 (chr10:12209572-12238143)	*CDC123*	12	1894	$9.24x10^{-1}$
CDC123 (chr10:12237960-12292589)	*CDC123*	32	4619	$9.4x10^{-1}$
C10orf11 (chr10:77542518-78317126)	*C10orf11*	304	39634	$\mathbf{4.03x10^{-2}}$
NTN4 (chr12:96051582-96184536)	*CCDC38*	88	14103	1
HAL (chr12:96367141-96390071)	*CCDC38*	14	1407	$7.11x10^{-1}$
THSD4 (chr15:71433787-72075722)	*THSD4*	291	38799	$7.86x10^{-1}$
CNGB1 (chr16:57916243-58005020)	*MMP15*	44	5877	$9.61x10^{-1}$
MMP15 (chr16:58059281-58080804)	*MMP15*	20	2098	$3.92x10^{-1}$

c) Exon based

Locus	GWAS gene	Number of variants	Number of alternative allele counts in cases and controls	P-value
HDAC4 (chr 2: 239969863-240322643)	*HDAC4*	7	996	$7.43x10^{-1}$
NPNT (chr4: 106816596-106892828)	*GSTCD*	9	1400	$5.25x10^{-2}$

Sensitivity analyses were undertaken in these six loci repeating the C-alpha association test including only independent variants ($r^2 < 0.2$). The variants present in the 3 sliding windows were already independent, and therefore their results were the same (**Table 4-13**). For the three genes, the number of variants included was reduced from 9 to 3, from 37 to 11 and from 304 to 124 for *NPNT*, *TNXB* and *C10orf11* respectively. When testing only independent variants, *NPNT* was no longer significant (**Table 4-13**) but *TNXB* and *C10orf11* became more significant (P=0.047 and 0.028 respectively) (**Table 4-13**) meeting the Bonferroni corrected threshold for these two regions (0.05).

Table 4-13 Sensitivity analysis of top hits in stage 2

The column “GWAS gene” presents the gene reported in the lung function GWAS (2)* for each region. P-values that meet the threshold are shown in bold.

Locus	GWAS gene	Threshold	All variants			Independent variants			
			Number of variants	Number of alternative allele counts	P-value	Number of variants	Number of alternative allele counts	P-value	P-value after permutations
chr3:25633833-25636833	*RARB*	$1.25x10^{-2}$	3	363	$5.32x10^{-2}$	3	363	$5.32x10^{-2}$	$6.25x10^{-2}$
chr3:169238286-169241286	*MECOM*	$8x10^{-3}$	2	573	**$2.94x10^{-3}$**	2	573	**$2.94x10^{-3}$**	$1.92x10^{-2}$
chr4:145293600-145296600	*HHIP*	$1.25x10^{-2}$	2	167	$4.16x10^{-2}$	2	167	$4.16x10^{-2}$	$6.19x10^{-2}$
NPNT (chr4:106816596-106892828)	*GSTCD*	$2.5x10^{-2}$	9	1400	$5.25x10^{-2}$	3	422	$5.1x10^{-1}$	$4.07x10^{-1}$
TNXB (chr6:32008931-32077151)	*AGER*	$5x10^{-2}$	37	8086	$6.08x10^{-2}$	11	1752	**$4.73x10^{-2}$**	$6.69x10^{-2}$
C10orf11 (chr10:77542518-78317126)	*C10orf11*	$5x10^{-2}$	304	39634	**$4.03x10^{-2}$**	124	15529	**$2.77x10^{-2}$**	$5x10^{-1}$

The C-alpha test authors (53) recommend undertaking permutations for the top loci, especially when only a small number of variants are included in the loci tested; which was the case for most of the top hits. Ten thousand permutations were run for the 6 most significant loci including only independent variants. The P-value for *NPNT*, which was no longer significant when testing only independent variants, was slightly smaller after permutations, but remained non-significant (P = 0.410). All other P-values became less significant (**Table 4-13**) after the permutations. The P-value for *C10orf11* became non-significant (P = 0.500). The most significant locus (chr3:169238286-169241286) was nominally significant (P = 0.019) (**Table 4-13**) and the other three were close to nominal significant with P-values ≤ 0.067 (**Table 4-13**); however none of them met the Bonferroni corrected thresholds (**Table 4-13**).

In summary, 3 sliding windows (chr3:25633833-25636833 in the *RARB* region, chr3:169238286-169241286 in the *MECOM* region and chr4:145293600-145296600 in the *HHIP* region) and three genes (*NPNT* in *GSTCD* region, *TNXB* in the *AGER* region and *C10orf11*) that had P-values close to nominal significance (P < 0.061), two (chr3:169238286-169241286 and *C10orf11*) of which met Bonferroni corrected thresholds, were selected to undertake sensitivity analyses. After running the C-alpha test only on independent variants the *NPNT* signal disappeared (P = 0.510), but the other five regions still showed suggestive evidence of association. After running permutations, *NPNT* remained non-significant, *C10orf11* became non-significant and the other

regions showed suggestive evidence of association, although none of them met the Bonferroni corrected threshold. In order to gain more insights into the 4 regions that showed suggestive evidence of association, single variant results for the variants included in each region were examined, and drop one plots were generated, where the C-alpha test was re-run for each region removing one variant at a time, to assess how much influence each individual variant had on the results.

The C-alpha tests for the sliding window chr3:25633833-25636833 in *RARB* were based on 5 variants only present in stage 1 (2 of them not in UK10K+1000G), one variant only present in stage 2 and 2 variants that were present both in stage 1 and stage 2 (**Figure 4-10 a)**). For one of these variants (chr3:25635243) the direction of effect between stage 1 and stage 2 agreed and for the other (chr3:25633946) it did not. Surprisingly, the strongest association both in stage 1 (P = 0.005 for SNVer and P = 0.009 for Syzygy) and stage 2 (P = 0.020) was for the variant with different direction of effect in stage 1 and stage2 (chr3:25633946). The drop one analysis for this region (**Figure 4-10 a)**) showed that the signal both in stage 1 and in stage 2 seemed to be driven by the variant chr3:25633946, indicating that this is likely to be a false positive association.

For the strongest signal in stage 2 (chr3:169238286-169241286) the C-alpha association tests were based on 4 stage 1 variants (2 not in UK10K+1000G) and two stage 2 variants, with no overlap between stage 1 and stage 2 variants (**Figure 4-10 b)**).Two variants in *MECOM*, one in stage 1 (chr3:169238973) and one in stage 2 (chr3:169240816), had nominally significant P-values (P = 0.049 for SNVer and P =0.019 for Syzygy, and P = 0.029 respectively) and they seemed to drive the associations in stage 1 and stage 2 respectively according to the drop one plots (**Figure 4-10 b)**). Therefore in each stage these signals appear to be single variant signals rather than the multiple variant signals the C-alpha test was designed to detect.

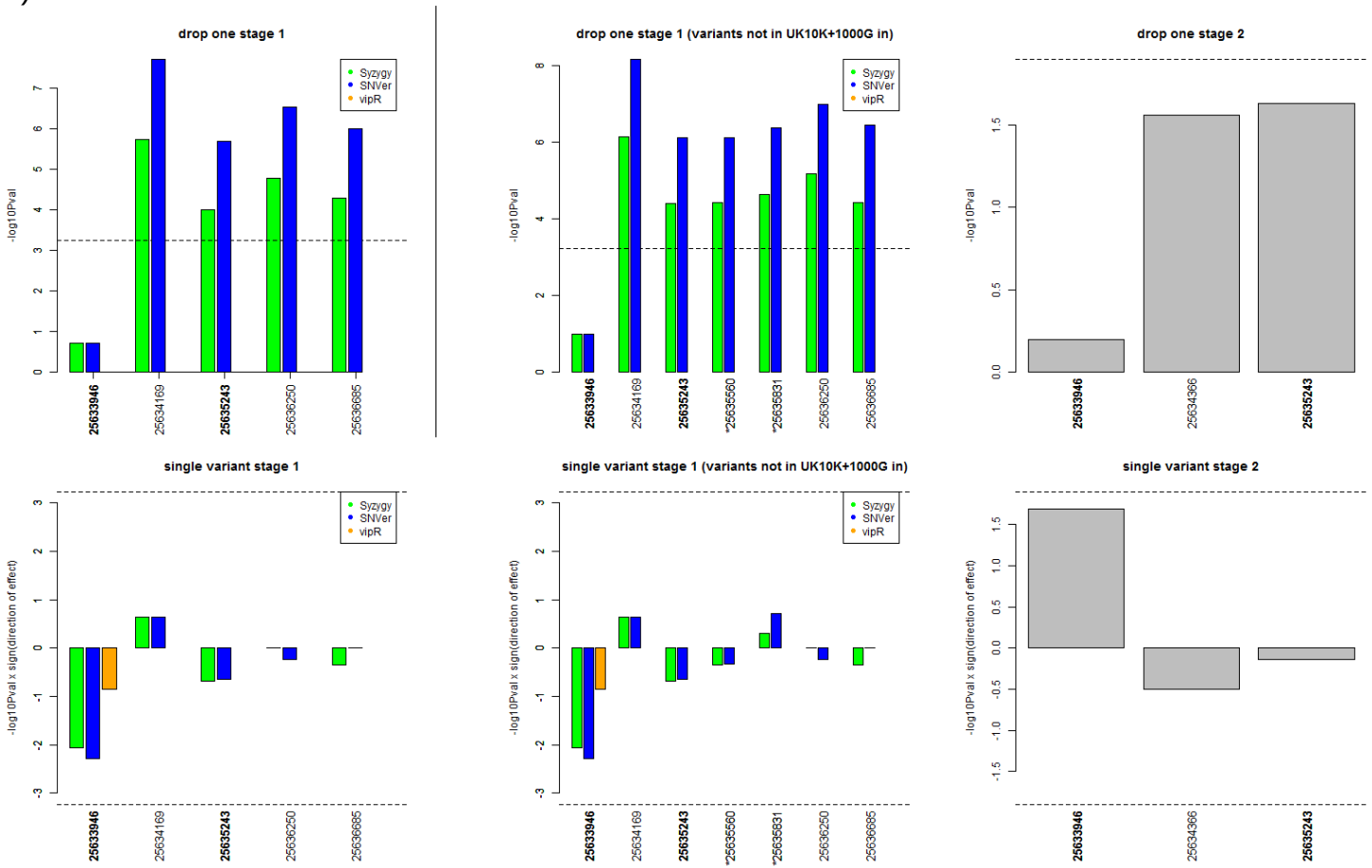
The C-alpha association tests for chr4:145293600-145296600 were based on 3 stage 1 variants (1 not in UK10K+1000G) and 2 stage 2 variants, again with no overlap between stage 1 and 2 (**Figure 4-10 c)**). Two variants upstream of *HHIP,* one in stage 1 (chr4:145295641) and one in stage 2 (chr4:145296265), had nominally significant P-values (P = 0.068 for SNVer and P = 0.018 for Syzygy, and P = 0.020 respectively) and the drop plots show that they have the strongest effect on the results (**Figure 4-10 c)**). The drop one plots (**Figure 4-10 c)**) also show that when removing variant chr4:145296136 the test is no longer significant, indicating that this variant could also be relevant.

The *TNXB* signal was based on 14 stage 1 variants (6 not in UK10K+1000G), 8 stage 2 variants and 3 that were present in both stages (**Figure 4-10 d)**). One of the variants (chr6:32056907) that was present in both stages had consistent direction of effect between stage 1 and stage 2, but the other two (chr6:32041621 and chr6:32061339) did not. **Figure 4-10 d)** shows that chr6:32036678 is the variant with the strongest association in stage 1 (P = 0.063 for SNVer and P = 0.019 for Syzygy); and chr6:32073912 the variant with strongest association in stage 2 (P = 0.004), although chr6:32046050 was also nominally significant in stage 2 (P = 0.010).

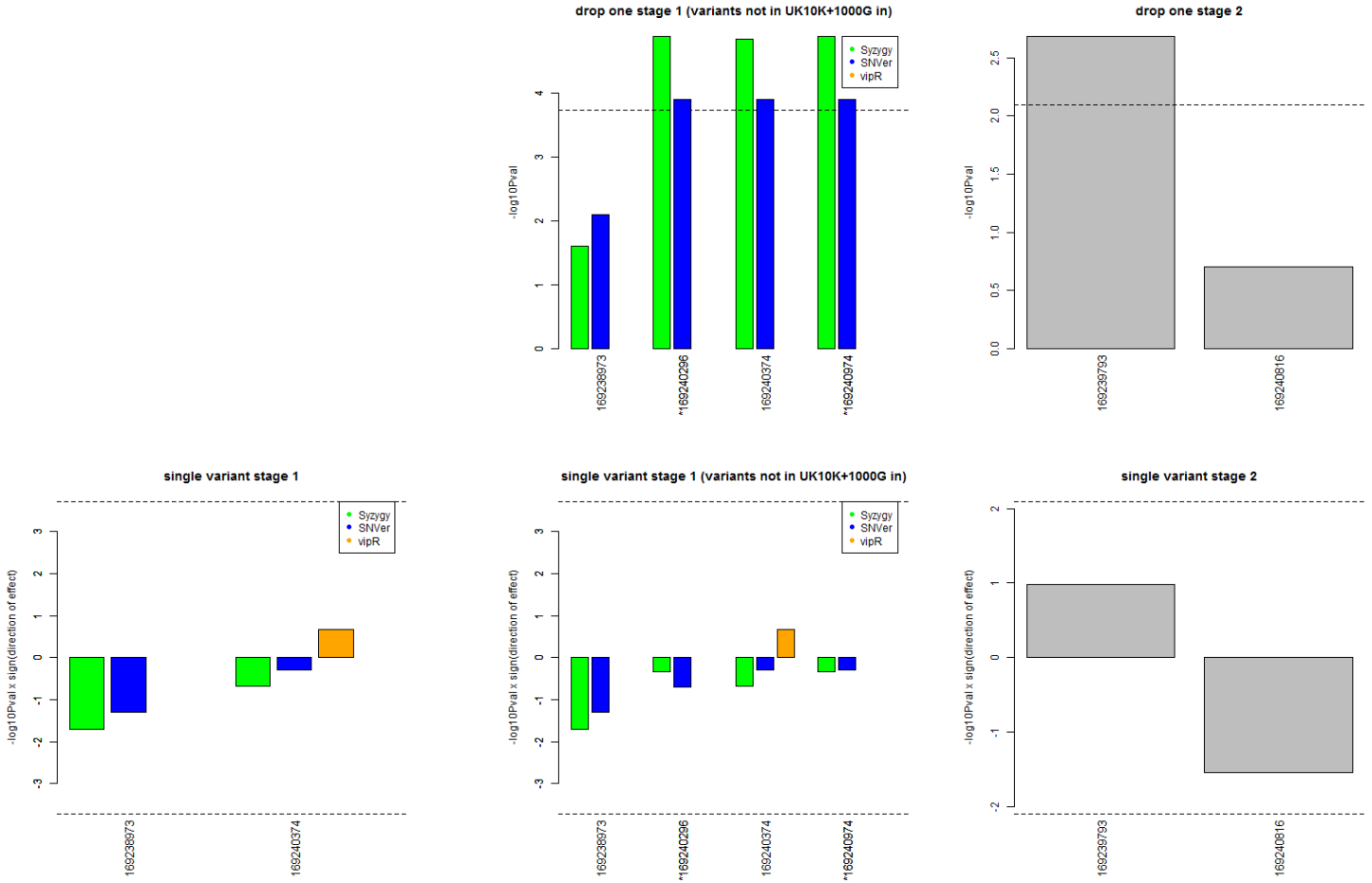
Figure 4-10 Drop one (top) and single variant association results (bottom) plots

A drop one plot for a locus is obtained by undertaking the C-alpha test removing one variant at a time; the P-value plotted for each variant represented on the x-axis is the P-value obtained after removing that variant. Results obtained using calls from different calling algorithms are represented in different colors according to the legend in each figure. Not all calling algorithms called the same set of variants, when there is no visible P-value represented for a variant for a given calling algorithm, it is because that variant was not called by that calling algorithm. For example for region chr3:25633833-25636833, vipR only called one variant (chr3:25633946) and therefore no drop one results were plotted for vipR in this region. If a region only includes two variants, no drop one plot is produced, since no C-alpha test can be undertaken with only one variant. Asymptotic P-values are presented here for the C-alpha test. The single variant plots show the P-values obtained for each variant, and they also show the direction of effect by plotting on the y-axis the -log10(P-value)x direction of effect, with the direction of effect=1 if OR > 1 and =-1 if OR < 1. For each region the first row shows drop one plots and second row shows single variant plots for the same variants; the first column presents results from stage 1 excluding variants not in UK10K+1000G, the second column presents results from stage 1 including variants not in UK10K+1000G (marked with * on the x-axis) and the third column presents results for stage 2. Variants that are included both in stage 1 and stage 2 are highlighted in bold. The horizontal dashed line represents the collapsing method significance threshold for each region (note that different thresholds were used for stage 1 and stage 2).

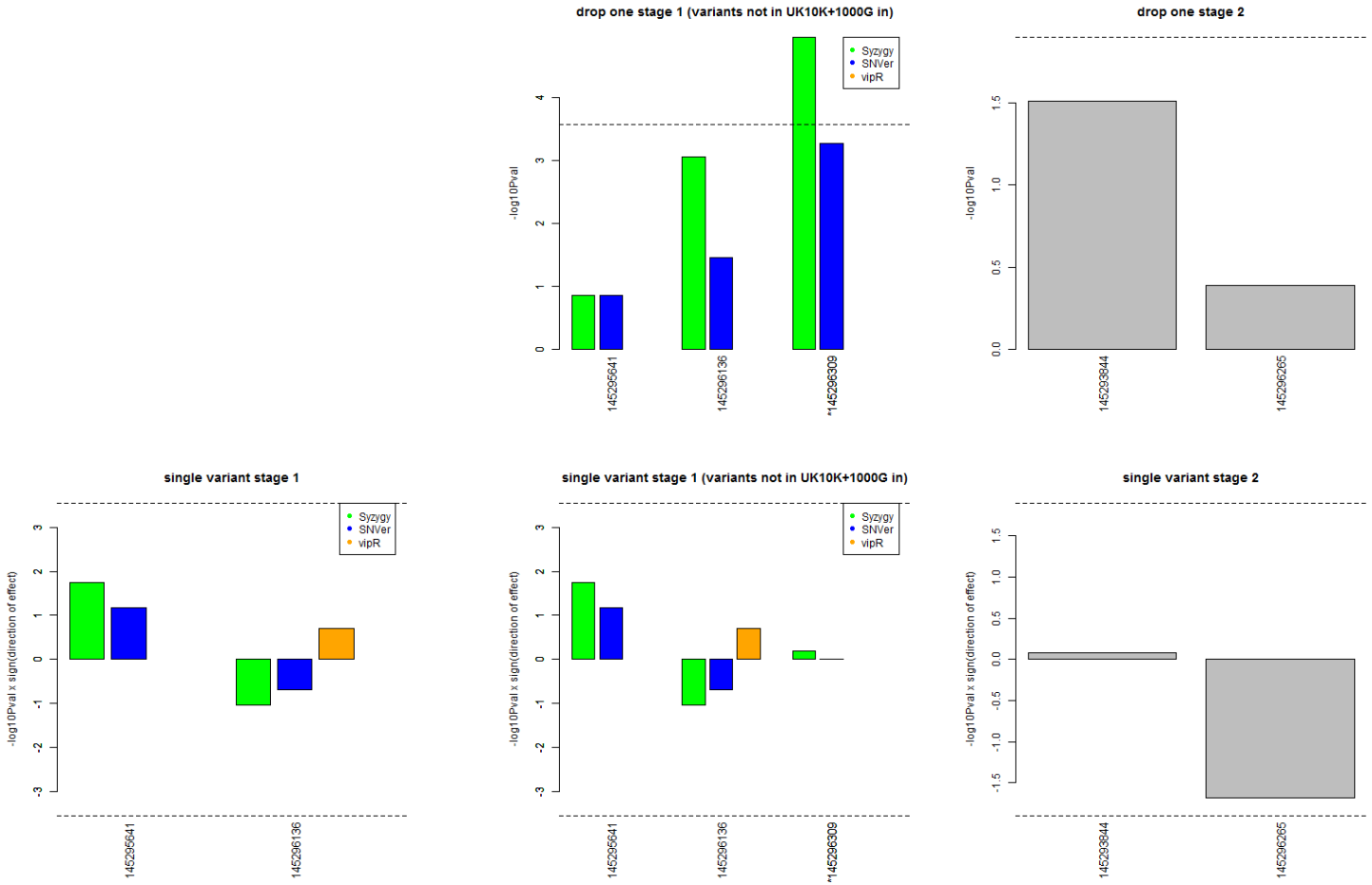
a) chr3:25633833-25636833



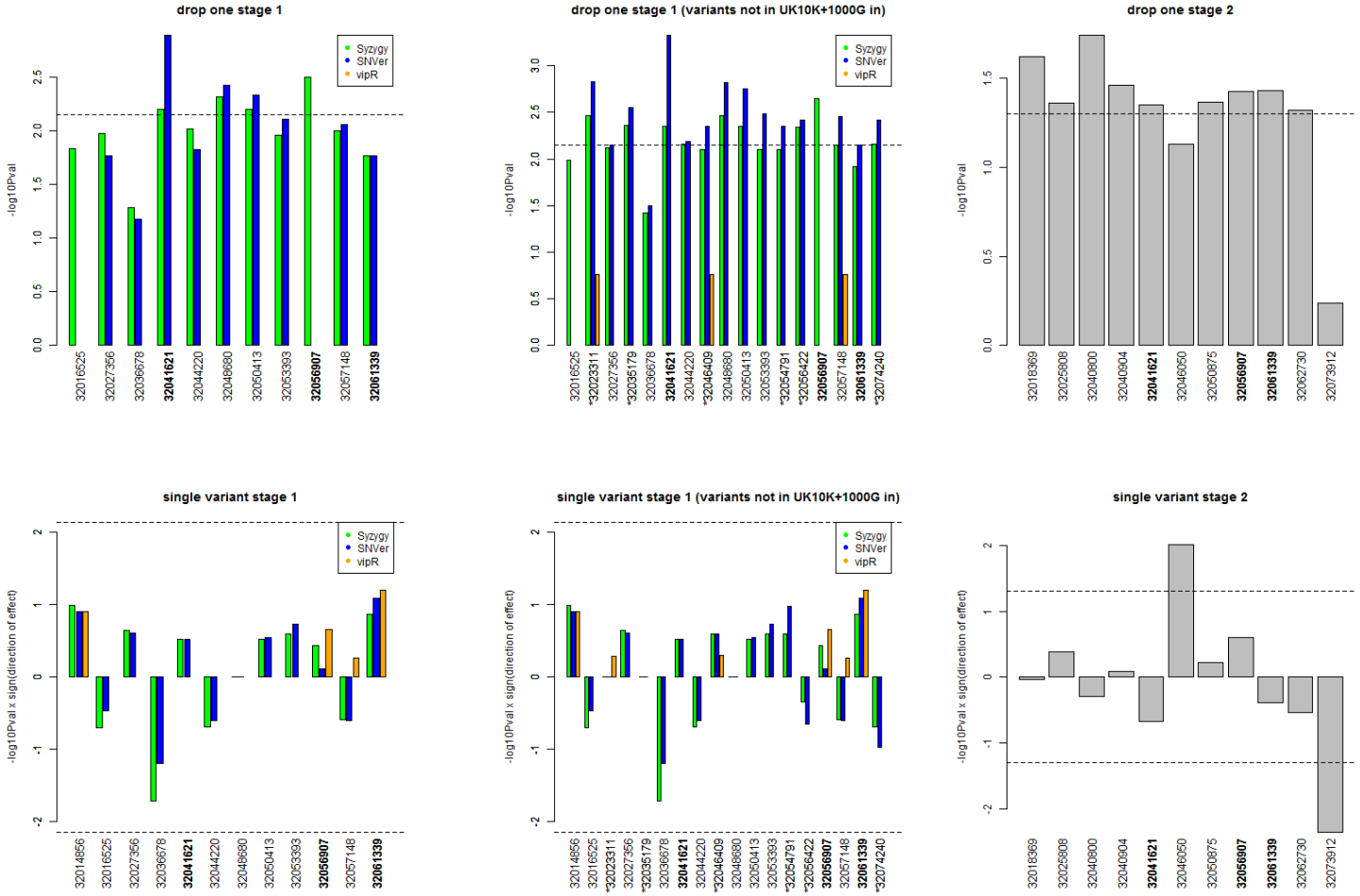
b) chr3:169238286-169241286



c) chr4:145293600-145296600



d) TNXB


drop one stage 1
drop one stage 1 (variants not in UK10K+1000G in)
drop one stage 2
single variant stage 1
single variant stage 1 (variants not in UK10K+1000G in)
single variant stage 2
Syzygy
SNVer
vipR
-log10Pval
-log10Pval x sign(direction of effect)
32014856
32016525
*32023311
32027356
*32035179
32036678
32041621
32044220
*32046409
32048680
32050413
32053393
*32054791
*32056422
32056907
32057148
32061339
*32074240
32018369
32025808
32040800
32040904
32046050
32050875
32062730
32073912

4.3.4 Discussion

The aim of this study was to find low frequency and rare variants in genetic regions known to be associated with lung function in order to gain insights into the biological pathways that link these regions with COPD risk. To do this, 26 regions associated with lung function (2, 23, 94-96)* were sequenced in 300 COPD cases and 300 controls using a cost-effective pooled design. Single variant analyses and collapsing methods were undertaken and top hits were followed up in 4,249 COPD cases and 11,916 controls from the UK BiLEVE study. None of the top hits met the significance threshold defined in stage 2, but suggestive evidence of association was shown for a window in *RARB*, a window in *MECOM*, one intergenic window upstream of *HHIP* and for the *TNXB* gene in the *AGER* region. In addition, the previously reported sentinel SNP for lung function in each region was tested for association with COPD. The strongest association ($P < 6 \times 10^{-4}$ for Syzygy and SNVer) meeting the Bonferroni corrected threshold for 26 tests (2×10^{-3}) was for the *MECOM* sentinel SNP. This is the first report of an association with COPD for this region.

The strongest collapsing signal in stage 2 was for a sliding window in an intronic region of *MECOM*. This gene encodes a transcriptional regulator protein and oncoprotein that might be involved in hematopoiesis, apoptosis, development, and cell differentiation and proliferation. *MECOM* has been associated with osteoporosis (170) and renal function-related traits (171) in East Asians and with blood pressure (172) and ageing (173) in Europeans. The sliding window

~270kb upstream of *HHIP* that showed suggestive evidence of association with COPD is located in a region that contains a DNase hypersensitivity site and transcription factor binding sites found in blood cells, renal epithelium cells and embryonic stem cells (174). This region does not overlap with another region ~85kb upstream of *HHIP* known to interact with the *HHIP* promoter and to function as an *HHIP* enhancer (175). *TNXB* within the *AGER* region showed suggestive evidence of association with COPD risk. This gene encodes a member of the tenascin family of extracellular matrix glycoproteins and it is thought to function in matrix maturation during wound healing. SNPs in *TNXB* have been associated with age-related macular degeneration (176), phospholipid levels in plasma (177), systemic lupus erythematosus (178) and HIV-1 control (179) in Europeans. The sliding window in *RARB* is located in a region that contains two DNase hypersensitivity sites and transcription factor binding sites found in lung fibroblasts and epithelial cells derived from lung carcinoma tissue among other cells (174). However, the association signal for this window seems to be driven by the same intronic variant (chr3:25633946) in stage 1 and stage 2, and the direction of effect is not consistent between stages suggesting that this might be a false positive association.

Single variant association analyses for sentinel variants previously associated with lung function (2)*, confirmed the previously reported associations with COPD risk for *HHIP* (95) and *HTR4* (24)*, showed for the first time the

association of the *MECOM* sentinel variant with COPD risk and provided suggestive evidence of a novel association for the sentinel SNP in *SPATA9*.

The main limitation of this study was power, due to small sample sizes. Assuming a COPD prevalence of 30% among smokers, a study with 300 cases and 300 controls would need an OR of 5 in order to detect a variant with MAF~1% and an OR of 2 for a variant with MAF~5% with 80% power at a nominal level of significance (0.05). Since this study only included 250 controls in the final analysis due to low DNA quality in two pools and as the significance threshold used per region was lower than 0.05, larger OR would have been required in order to detect associations for low frequency rare variants in this study. In addition, despite the large sample size of the stage 2 resource, this study was not ideal since most of the variants followed up were imputed, rather than genotyped, and a considerable proportion (46%) of the top regions did not have enough variants in the stage 2 study to be followed up.

The key strength of this study was the ability to identify novel low frequency or rare variants through sequencing. A total of 1,429 new SNPs and 216 new INDELs passed the quality control checks. However, the pooled design made the variant calling step especially challenging, and validation of a subset of these variants through direct genotyping would have been valuable. Unfortunately, not enough DNA was available for the participants of the study

after having to repeat the sequencing experiment three times due to issues with the enrichment kit. Alternatively, a design like the one presented in (43), where the targeted sequencing is used to identify new variants and then these variants are genotyped in large populations where their association with the trait of interest is tested, would have had greater power, although very costly.

Three variant calling algorithms were used in this study in order to minimize the occurrence of false positive calls. The three algorithms used different statistical methods to call variants and they also performed differently. vipR was less sensitive than the other two algorithms and it called a much smaller number of variants (39,211 SNPs and 459 deletions, compared to 62,506 SNPs and 5,811 INDELs by SNVer and 55,886 SNPs and 5,331 INDELs by Syzygy), most of which had MAF > 1% (85% of SNPs and 99% of INDELs). SNVer also called mainly common (MAF > 1%) variants (81% of SNPs and all INDELs), whereas Syzygy called the largest number of rare variants (43% SNPs and 35% of INDELs). In terms of specificity, Syzygy and vipR show the best specificity when assessing the proportion of SNPs with MAF > 1% included in dbSNP137 (166) (97.08% for vipR and 98.2% for Syzygy), compared to a 72% for SNVer. When calling INDELs all algorithms had lower specificity than when calling SNPs, assessed as the proportion of INDELs with MAF > 1% included in the 1000 Genomes Project (36), (67.55% for vipR, 35% for SNVer and 50% for Syzygy). The Syzygy calculation of allele frequencies was problematic for a subset of INDELs. Despite this issue, Syzygy seems to be the algorithm that performs

best in terms of sensitivity and specificity and also presents additional features that the other algorithms do not such as taking into account uncertainty in estimating allele counts by providing allele dosages. However, in terms of implementation Syzygy was more challenging than the other two algorithms, both in terms of installation due to its many dependencies (other programs are required for the software to work) and in terms of execution. vipR was easy to implement, although a bug in the source code was detected, fixed and reported, and SNVer was the easiest and quickest.

There was a notable difference in the quality of the data for SNPs and INDELs. Twenty-three percent of the SNPs called by at least one algorithm remained after all quality control checks, whereas only 7% of the INDELs passed the quality control checks. In addition, out of the new variants (SNPs not in dbSNP137 (166) or UK10K+1000G and INDELs not in 1000 Genomes Project Phase 1 data (36) or Mills *et al.* (163)) that passed the quality control checks, 93% of the SNPs had MAF < 1%, whereas only 4% of the INDELs had MAF < 1%; and since most common variation is expected to have been already identified, new variants would be expected to be rare.

The potential for false positive calls when working with pooled data led me to apply some strict filters. In some instances these filters might have been over conservative, for instance the repeat mask filter removed 18,622 SNPs and

1,161 INDELs. This filter could have excluded some real variants, however, it would have been very challenging to distinguish a real variant from a sequencing error or a misalignment in a repeat mask region.

Study design impacts on the power of the study and the potential for false positives in the association testing. A flaw in our study design meant that the sequencing lanes included case pools only or control pools only, making a possible lane effect a clear issue for the association testing. The lane test that was implemented to deal with this issue may have been over conservative by removing any true signal that happened to be driven by pools within the same lane. In addition, both the lane test and pool test removed variants whose association was driven either by a single lane or by a single pool with discrepant allele counts in comparison with the rest (treating cases and controls separately). However in some instances they might have also removed variants where the discrepancy for a lane or a pool would have only led to shrinkage to the null of a true association, instead of to a false association. Sensitivity analyses, performing the association testing again after removing the discrepant lane or pool could have been undertaken in order to identify those variants for which the discrepant lane or pool only led to shrinkage to the null. In addition, since cases were ordered by severity of COPD then assigned to pools sequentially and then pools were assigned to lanes sequentially, an association signal driven only by the lane that contained the most extreme COPD cases would have been indistinguishable from a lane artefact.

There are some practical lessons that I have learned while undertaking this study. Descriptive analyses, quality control, and investigation of quality control anomalies remain important in the context of outsourced sequencing. Until the data presented in this chapter was obtained, two previous attempts failed, and it was thanks to thorough quality control checks that I discovered what the issue was and it was possible to negotiate the next experiment with better conditions and free of charge. In addition, when running publicly available software it is very important to have a clear understanding of the methods being implemented in the software and critically assess the output obtained. It also helps to familiarise yourself with the source code being used. In two instances I found bugs in publicly available software, where the software did not implement the methods as intended.

Overall, I have presented here the challenges of working with pooled sequencing data and the strategies I have used to overcome these issues. Suggestive evidence for the association of rare variants in 3 sliding windows and one gene was provided, but no strong findings arose from this study. Additional studies which are better powered will be required to identify rare variants associated with COPD risk. This study has shown for the first time that *MECOM* is associated with COPD risk.

4.4 Conclusion

This chapter has presented two different approaches to study low allele frequency and rare variants and the challenges that they present. The first approach was implemented genome-wide in a large number of individuals (20,941), however these individuals were genotyped using older genotyping arrays, which do not provide good coverage for rare variants; the second approach was a targeted sequencing study (sequencing currently being the optimal approach for measurement of rare variation) but was only undertaken in a limited number of individuals (300 COPD cases and 300 controls). Neither of these studies have been successful in confidently identifying rare variants with an effect on lung function or COPD risk. An ideal study aimed at detecting associations with rare variants would bring together the advantages from both studies, the large sample size and the good coverage for identifying rare variants obtained by sequencing, ideally deep whole genome sequencing. However the cost of these kinds of studies is still very high, and alternative study designs are still required. Alternative study designs include: low depth whole genome sequencing in large numbers of individuals, exome sequencing, the use of genotyping arrays with better coverage for low frequency or rare variants such as, the exome chip, imputation to denser imputation reference panels, such as 1000 Genomes Project (36) or UK10K+1000G, etc. In addition, studies including different ancestries may be beneficial for fine mapping signals (180, 181), especially populations with shorter haplotypes than those typically seen in European populations.

Chapter 5: Conclusion

Chronic obstructive pulmonary disease is a heritable disease (78, 83) predicted to be the third cause of death worldwide in 2030 (182). The biological mechanisms involved in the development and progression of this disease are still poorly understood. The aim of the work presented here was to identify new genomic regions associated with lung function and COPD, but also explore existing regions and risk prediction. These genetic discoveries may improve the knowledge of the biological mechanisms underlying the disease and lead to the development of new preventive and treatment strategies. This chapter gives a brief description of the findings presented in this thesis, discusses some of the analytic approaches and limitations of the studies undertaken, gives an update on other studies that have been undertaken in the field, and discusses potential applications of the findings.

When I started this work, genome-wide association studies had reported 11 loci (*TNS1*, *FAM13A*, *GSTCD*, *HHIP*, *HTR4*, *ADAM19*, *AGER*, *GPR126*, *PTCH1*, *THSD4 and PID1*) associated with lung function (23, 94, 96) and 3 loci (*FAM13*, *HHIP* and *CHRNA 3/5*) associated with COPD (95, 103). Chapter 2 showed that 3 (*GSTCD*, *TNS1* and *HTR4*) out of 5 variants (*TNS1*, *GSTCD*, *HTR4*, *AGER* and *THSD4*) associated with lung function were also associated with COPD; and that individuals with 10 to 12 risk alleles in *TNS1*, *GSTCD*, *HHIP*, *HTR4*, *AGER* and *THSD4* (5% of our population) had 1.6 fold increase in their risk of developing COPD in comparison to individuals with 7 risk alleles in the same

loci (28% of our population). Chapter 3 aimed to identify new common genetic variants associated with lung function, and presented the largest meta-analysis of lung function GWAS at the time. It confirmed the association of 10 (*TNS1*, *FAM13A*, *GSTCD-NPNT*, *HHIP*, *HTR4*, *ADAM19*, *AGER*, *GPR126*, *PTCH1* and *THSD4*) out of the 11 previously reported loci (23, 94, 96) and identified 16 new lung function loci (*MFAP2, TGFB2, HDAC4, RARB, MECOM, SPATA9, ZKSCAN3, NCR3, ARMC2, CDC123, C10orf11, LRP1, CCDC38, MMP15, CFDP1* and *KCNE2*). This brought the total number of loci associated with lung function to 26. Chapter 4 presented two different approaches for identifying low allele frequency and rare genetic variants associated with lung function and COPD. One approach was a burden test applied to a large number individuals (N = 20,941) in a meta-analysis of studies in the SpiroMeta consortium. No convincing novel findings arose from this study, probably due to low coverage for rare variants in the genome-wide arrays used by the studies. The second approach was a targeted sequencing study for the 26 loci associated with lung function. Although this study was strongly limited in power as it included 300 COPD cases and 300 controls, it reported suggestive evidence for the combined effect of rare variants in three sliding windows (chr3:25633833-25636833 in the *RARB* region, chr3:169238286-169241286 in the *MECOM* region and chr4:145293600-145296600 in the *HHIP* region) and one gene (*TNXB* in the *AGER* region). In addition, this study was the first to report the association with COPD of a common variant in *MECOM* known to be associated with lung function.

Throughout the thesis I have worked with different kinds of data. I have mainly used summary data for the analyses undertaken in Chapters 2, 3 and in the first study of Chapter 4. The motivation to use summary data was to enable a simple form of data sharing between studies to achieve the large sample sizes required to detect modest genetic effect sizes. However, having only access to summary data had its challenges. This required me to work closely with a large number of researchers undertaking study-level analyses. There was potential for heterogeneity in the approaches undertaken by study-level analysts despite the standard analysis plan that I developed, and for programming errors to be undetected. I tried to minimize heterogeneity across studies by providing sufficiently detailed analysis plans. In addition, I undertook thorough quality control checks to detect any discrepancy across studies. I liaised with analysts to understand the source of these discrepancies and once they were understood I took the necessary measures. Several issues were found and solved in the various analyses undertaken, highlighting the relevance of a careful quality control pipeline. Another challenge was the inability to undertake additional analyses, such as sensitivity analyses, without having to coordinate large numbers of analysts. I was able to access individual level data for a subset of studies. This allowed me to pilot some of the analyses (section 4.2 Chapter 4) before finalising the analysis plan, to undertake sensitivity analyses (section 2.3 Chapter 2), and to carry out analyses for a subset of studies included in the follow-up stage of the study presented in Chapter 3. In the second study presented in Chapter 4, the sequencing study I designed with my supervisors generated pooled sequencing data and I had access to the raw

data. This gave me the opportunity to process the data myself, and allowed me to try different approaches to analyse the data before choosing the final strategy.

Overall, throughout the thesis I have taken a strict approach in order to avoid false positive associations when aiming to discover associations for lung function or COPD. In some instances this approach may have been over conservative and I might have missed some true associations. In the meta-analysis of lung function GWAS undertaken in Chapter 3 I undertook genomic control (26) at study level twice, before and after meta-analysing smoking strata within studies and then again at the meta-analysis level. Genomic inflation factors are known to increase with sample size for highly polygenic traits, due to the increase in power; and a recent GWAS undertaken for height (183) suggested that single genomic control correction might suffice. Additional signals might have been detected in the meta-analysis of GWAS undertaken in Chapter 3 if a less strict approach was undertaken for genomic control correction, although this would have increased the risk for false positive associations. Data analysis for the pooled targeted sequencing presented in Chapter 4 was particularly challenging due to the similar magnitude of sequencing error rate and minor allele frequencies for rare variants. For this reason, I also applied very strict filters, removing a large number of variants, which may again have removed some true association signals.

Associations with COPD presented in Chapter 2 and Chapter 4 used the GOLD (104) spirometric definition of COPD as a reference. Individuals classified as GOLD stage 2 and above (percent predicted FEV_1 < 80% and FEV_1/FVC < 0.7) (104) were classified as cases and individuals with percent predicted FEV_1 > 80% and FEV_1/FVC > 0.7 as controls, whereas individuals with GOLD stage 1 (percent predicted FEV_1 > 80% and FEV_1/FVC < 0.7) (104) or with percent predicted FEV_1 < 80% and FEV_1/FVC > 0.7 were excluded to avoid misclassification. The GOLD guidelines (55, 104) recommend using post-bronchodilator spirometry for this diagnosis; however, this would have reduced the sample size considerably and therefore the power to detect associations with COPD. For this reason pre-bronchodilator spirometry was used in Chapter 2 and Chapter 4. I showed in a sensitivity analysis undertaken in Chapter 2 (section 2.3) using data from a study with both pre and post-bronchodilator measures that using the criteria described above, excluding individuals classified as GOLD stage 1, only a small number of individuals were misclassified when using pre-bronchidolator spirometry. The most recent GOLD guidelines (55) include assessment of risk of exacerbations and symptoms in order to make a diagnosis, and recommend also an assessment of co-morbidities. The advantage of using only spirometry is that is a well measured and objective criterion; however a more complete diagnosis with information on exacerbations, symptoms and co-morbidities would have the potential to enable genetic studies that could provide insights into more specific aspects of the disease.

Alongside the work presented here, additional analyses have been undertaken which have discovered a number of new loci associated with lung function and with COPD. I took part in a large meta-analysis of GWAS of forced vital capacity undertaken in SpiroMeta and CHARGE (3)*, which discovered 6 additional loci (*EFEMP1*, *BMP6*, *MIR129-2–HSD17B12*, *PRDM11*, *WWOX* and *KCNJ2*). Overall, work undertaken here and by others have now shown association with airflow obstruction or COPD for 13 lung function loci (*TGFB2*, *TNS1, RARB*, *MECOM*, *FAM13, GSTCD*, *HHIP, HTR4, ADAM19, AGER*, *GPR126*, *C10orf11* and *THSD4*) (24, 95, 103, 117, 154, 155)* (Chapter 4, section 4.3), illustrating that studying the genetics of lung function is a powerful approach for understanding the genetics of COPD. Additional studies (155, 184) have identified other new loci (*MMP12*, *RIN3*, *RAB4B-EGLN2-CYP2A6*) for COPD.

The fact that smoking is a major risk factor for COPD but not all smokers develop COPD, seems to point to a gene by smoking interaction, however none of the associations presented here have shown a gene by smoking interaction. I have contributed to a genome-wide gene by smoking interaction study (185)*, which identified three additional signals (*DNER*, *HLA-DQ*, and *KCNJ2/SOX9*) that became genome-wide significant when testing both the main effect and the interaction term together. However, none of these regions showed a very strong interaction, and to my knowledge no other gene by smoking interactions have been reported for lung function. Detecting gene by environment interaction for complex traits has been challenging (186). Particularly, in the case of gene by

smoking interaction, the environment is not easy to measure, given that we often rely on self-reported information. Improvement in the quality of smoking behavior information, as well as increased sample sizes will be required in order to improve the power of these studies.

Analyses presented here have focused on analyzing cross-sectional lung function, and have not covered genetic effects on longitudinal lung function. I have also contributed to a large meta-analysis of longitudinal lung function GWAS (187)*. However, study heterogeneity in number of time points measured, length of follow-up, method used to measure spirometry in each time point, baseline age, cigarette smoking, and the increased power required to detect an effect on slope rather than on baseline for longitudinal lung function, have made these analyses very challenging. Currently little is known about the genetics of decline in lung function in adults, and future studies with more homogeneous data, very large sample sizes and long enough follow-up will be required in order to shade some light into this field. Possibly, data being collected in a homogenous manner by large biobanks will enable this kind of study.

The work undertaken in this thesis, and by others, has identified a number of genetic associations for lung function and COPD. However, these loci tend to have moderate effect sizes, and there is still a large proportion of the variance

of these traits that remains unexplained. Analyses undertaken for other complex traits (32, 43, 188) indicate that it is likely that both common and rare variants play an important role in explaining the remaining proportion of the variance. Additional analyses would give insights into the genetic architecture of lung function. If I had been able to access individual level data genome-wide, I could have estimated the proportion of the variance of the lung function measures and COPD risk that is explained collectively by all common variants measured by a DNA microarray. This kind of analysis has shown that a substantial proportion of the heritability for height, Crohn's disease, bipolar disorder, and type 1 diabetes is explained by common variation (150, 189). Rare variants on the other hand, are expected to have large effect sizes and to also play a role in explaining the missing heritability; in addition they are likely to have a more immediate clinical application due to their potential to be deleterious (34). For these reasons the study of both common and rare variants is relevant. No strong associations with rare variants have been presented here, or have been published for lung function at present, with the exception of those causing alpha1 antitrypsin deficiency (85, 190). This illustrates the challenges of identifying rare variants. A key message that comes across in the thesis is the relevance of sample size to enable genetic discoveries. As sample sizes increase, sequencing costs drop and new analytic techniques are developed to deal with the issues related to analyzing rare variation, the number of reported rare variant associations is likely to rise. I am currently working on a meta-analysis of GWAS imputed to the 1000 Genomes Project Phase 1 data

reference panel (36) which has the potential to discover additional associations for common and low allele frequency variants.

The mechanisms through which discovered loci affect lung function are not yet well understood. Follow-up analyses undertaken for lung function loci presented in Chapter 3 provided some insights into these mechanisms. None of the 26 lung function loci associated with lung function in Chapter 3 showed association with smoking behavior or seemed to be caused by gene by smoking interaction. The direction of effect in children (7-9 years of age) and adults was consistent for 20 out of 26 sentinel variants. This suggests that overall these loci may affect lung function through lung development. An additional study (191) has shown evidence of association with infant lung function for variants in 4 (*HHIP*, *HDAC4*, *NCR3*, *RARB*) of the 26 genes, providing some support for this hypothesis. Some lung function loci have also shown associations with other traits, such as height (127) or lung cancer (128); understanding the mechanisms of these pleiotropic effects will also provide insights into the biological process involved in lung disease. Functional studies have been undertaken for some of the loci associated with lung function, and have given some insights into their function (175, 192, 193)*. However, further studies will be required to fully understand how these loci affect lung function and COPD, and to translate this knowledge into possible treatments.

A more immediate application of the findings could be in the prevention of smoking and in smoking cessation campaigns, by the use of genetic risk scores. In using genetic risk scores it is not necessary to understand the biological processes through which the loci exert an effect on lung function. It is enough to have a number of independent variants robustly associated with lung function and accurate estimates of their effect size. The accuracy of risk scores will increase as the number of variants associated with lung function increases and as we get closer to the causal variants. Loci discovered to date have small effect sizes, but individuals who smoke have a high baseline risk for developing COPD (118), and the small variation in risk provided by the genetic information may become relevant in this context. Presenting personalized genetic risk profiles for COPD could aid in the development of smoking prevention and cessation campaigns. This concept inspired the development of "The Risky Gene Machine", an activity that was part of an exhibition at the Royal Society Summer Science Exhibition in London 2012 ("Breathless genes: the lung and the short of it", http://sse.royalsociety.org/2012/exhibits/breathless-genes/). The Risky Gene Machine was like a "fruit machine", but instead of a random combination of fruits, it provided a random combination of risk and non risk alleles with their associated risk of developing COPD both for smokers and nonsmokers. Results presented in Chapter 2 were used to estimate genetic risk and baseline risk for smokers and nonsmokers was extracted from (118).

The studies that form this thesis have identified 16 new loci associated with lung function, showed association with COPD risk for 4 lung function loci and presented suggestive evidence for the combined effect of rare variants in 3 windows and one gene, within in known loci. These findings have pointed to genomic regions not previously related to lung function and have the potential to lead to the discovery of new molecular pathways involved in lung health and disease, and to the development of new preventive and treatment strategies.

References:

1. Loth DW. Epidemiology of Lung Function and Chronic Obstructive Pulmonary Disease: Erasmus MC; 2013.
2.* Soler Artigas M, Loth DW, Wain LV, Gharib SA, Obeidat M, Tang W, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet. 2011;43(11):1082-90. Epub 2011/09/29.
3.* Loth DW, Soler Artigas M, Gharib SA, Wain LV, Franceschini N, Koch B, et al. Genome-wide association analysis identifies six new loci associated with forced vital capacity. Nat Genet. 2014;46(7):669-77. Epub 2014/06/16.
4. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. Epub 2012/09/08.
5. Dempster ER, Lerner IM. Heritability of Threshold Characters. Genetics. 1950;35(2):212-36. Epub 1950/03/01.
6. Benchek PH, Morris NJ. How meaningful are heritability estimates of liability? Human genetics. 2013;132(12):1351-60. Epub 2013/07/23.
7. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. Nature reviews Genetics. 2013;14(2):139-49. Epub 2013/01/19.
8. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nature reviews Genetics. 2008;9(4):255-66. Epub 2008/03/06.
9. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33 Suppl:228-37. Epub 2003/03/01.
10. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet. 2001;29(3):306-9. Epub 2001/10/16.
11. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews Genetics. 2008;9(5):356-69. Epub 2008/04/10.
12. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7-24. Epub 2012/01/17.
13. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53. Epub 2009/10/09.
14. Gibson G. Rare and common variants: twenty arguments. Nature reviews Genetics. 2011;13(2):135-45. Epub 2012/01/19.
15. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nature protocols. 2010;5(9):1564-73. Epub 2010/11/19.
16. International HapMap Consortium. The International HapMap Project. Nature. 2003;426(6968):789-96. Epub 2003/12/20.

17. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39(7):906-13. Epub 2007/06/19.
18. Li Y, Abecasis GR. Mach 1.0: Rapid haplotype construction and missing genotype inference. Am J Hum Genet. 2006;S79:2290.
19. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44(8):955-9. Epub 2012/07/24.
20. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature reviews Genetics. 2010;11(7):499-511. Epub 2010/06/03.
21. Salanti G, Southam L, Altshuler D, Ardlie K, Barroso I, Boehnke M, et al. Underlying genetic models of inheritance in established type 2 diabetes associations. American journal of epidemiology. 2009;170(5):537-45. Epub 2009/07/16.
22.* Soler Artigas M, Wain LV, Tobin MD. Genome-wide association studies in lung disease. Thorax. 2012;67(3):271-3, 80. Epub 2011/08/23.
23. Repapi E, Sayers I, Wain LV, Burton PR, Johnson T, Obeidat M, et al. Genome-wide association study identifies five loci associated with lung function. Nat Genet. 2010;42(1):36-44. Epub 2009/12/17.
24.* Soler Artigas M, Wain LV, Repapi E, Obeidat M, Sayers I, Burton PR, et al. Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function. American journal of respiratory and critical care medicine. 2011;184(7):786-95. Epub 2011/10/04.
25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9. Epub 2006/07/25.
26. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997-1004. Epub 2001/04/21.
27. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet. 2003;33(2):177-82. Epub 2003/01/14.
28. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Optimal designs for two-stage genome-wide association studies. Genetic epidemiology. 2007;31(7):776-88. Epub 2007/06/06.
29. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature. 2001;411(6837):599-603. Epub 2001/06/01.
30. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet. 2008;40(8):955-62. Epub 2008/07/01.
31. Anderson CA, Soranzo N, Zeggini E, Barrett JC. Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS biology. 2011;9(1):e1000580. Epub 2011/01/27.

32. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009;324(5925):387-9. Epub 2009/03/07.
33. Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. Cell. 2011;147(1):57-69. Epub 2011/10/04.
34. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5-23. Epub 2014/07/06.
35. Neale BM, Purcell S. The positives, protocols, and perils of genome-wide association. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics. 2008;147B(7):1288-94. Epub 2008/05/27.
36. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
37. Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, et al. Sequencing studies in human genetics: design and interpretation. Nature reviews Genetics. 2013;14(7):460-70. Epub 2013/06/12.
38. Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008;26(10):1135-45. Epub 2008/10/11.
39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303. Epub 2010/07/21.
40. Van der Auwera GA CM, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K, Altshuler D, Gabriel S, DePristo M. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. CURRENT PROTOCOLS IN BIOINFORMATICS. 2013;43:11.10.1-11.10.33.
41. Altmann A, Weber P, Quast C, Rex-Haffner M, Binder EB, Muller-Myhsok B. vipR: variant identification in pooled DNA using R. Bioinformatics. 2011;27(13):i77-84. Epub 2011/06/21.
42. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic acids research. 2011;39(19):e132. Epub 2011/08/05.
43. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43(11):1066-73. Epub 2011/10/11.
44. Bansal V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics. 2010;26(12):i318-24. Epub 2010/06/10.
45. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25(17):2283-5. Epub 2009/06/23.

46. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research. 2008;18(11):1851-8. Epub 2008/08/21.
47. Druley TE, Vallania FL, Wegner DJ, Varley KE, Knowles OL, Bonds JA, et al. Quantification of rare allelic variants from pooled genomic DNA. Nature methods. 2009;6(4):263-5. Epub 2009/03/03.
48. Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and rare variants affecting complex traits. Human molecular genetics. 2013;22(R1):R16-21. Epub 2013/08/08.
49. 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). Mutation research. 2007;615(1-2):28-56. Epub 2006/11/15.
50. 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;83(3):311-21. Epub 2008/08/12.
51. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology. 2010;34(2):188-93. Epub 2009/10/08.
52. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5(2):e1000384. Epub 2009/02/14.
53. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genet. 2011;7(3):e1001322.
54. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82-93. Epub 2011/07/09.
55. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management and Prevention of COPD. 2014; Available from: http://www.goldcopd.org/.
56. Halbert RJ, Natoli JL, Gano A, Badamgarav E, Buist AS, Mannino DM. Global burden of COPD: systematic review and meta-analysis. Eur Respir J. 2006;28(3):523-32. Epub 2006/04/14.
57. Fukuchi Y, Nishimura M, Ichinose M, Adachi M, Nagai A, Kuriyama T, et al. COPD in Japan: the Nippon COPD Epidemiology study. Respirology. 2004;9(4):458-65. Epub 2004/12/23.
58. Buist AS, McBurnie MA, Vollmer WM, Gillespie S, Burney P, Mannino DM, et al. International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. Lancet. 2007;370(9589):741-50. Epub 2007/09/04.
59. Trupin L, Earnest G, San Pedro M, Balmes JR, Eisner MD, Yelin E, et al. The occupational burden of chronic obstructive pulmonary disease. Eur Respir J. 2003;22(3):462-9. Epub 2003/10/01.
60. Matheson MC, Benke G, Raven J, Sim MR, Kromhout H, Vermeulen R, et al. Biological dust exposure in the workplace is a risk factor for chronic obstructive pulmonary disease. Thorax. 2005;60(8):645-51. Epub 2005/08/03.

61. Abbey DE, Burchette RJ, Knutsen SF, McDonnell WF, Lebowitz MD, Enright PL. Long-term particulate and other air pollutants and lung function in nonsmokers. American journal of respiratory and critical care medicine. 1998;158(1):289-98. Epub 1998/07/09.
62. Boman C, Forsberg B, Sandstrom T. Shedding new light on wood smoke: a risk factor for respiratory health. Eur Respir J. 2006;27(3):446-7. Epub 2006/03/02.
63. Ezzati M. Indoor air pollution and health in developing countries. Lancet. 2005;366(9480):104-6. Epub 2005/07/12.
64. Gall ET, Carter EM, Earnest CM, Stephens B. Indoor air pollution in developing countries: research and implementation needs for improvements in global public health. American journal of public health. 2013;103(4):e67-72. Epub 2013/02/16.
65. Mannino DM, Homa DM, Akinbami LJ, Ford ES, Redd SC. Chronic obstructive pulmonary disease surveillance--United States, 1971-2000. Morbidity and mortality weekly report Surveillance summaries. 2002;51(6):1-16. Epub 2002/08/30.
66. Burrows B, Cline MG, Knudson RJ, Taussig LM, Lebowitz MD. A descriptive analysis of the growth and decline of the FVC and FEV1. Chest. 1983;83(5):717-24. Epub 1983/05/01.
67. Todisco T, de Benedictis FM, Iannacci L, Baglioni S, Eslami A, Todisco E, et al. Mild prematurity and respiratory functions. European journal of pediatrics. 1993;152(1):55-8. Epub 1993/01/01.
68. Barker DJ, Godfrey KM, Fall C, Osmond C, Winter PD, Shaheen SO. Relation of birth weight and childhood respiratory infection to adult lung function and death from chronic obstructive airways disease. BMJ. 1991;303(6804):671-5. Epub 1991/09/21.
69. Svanes C, Sunyer J, Plana E, Dharmage S, Heinrich J, Jarvis D, et al. Early life origins of chronic obstructive pulmonary disease. Thorax. 2010;65(1):14-20. Epub 2009/09/05.
70. Wilk JB, Djousse L, Arnett DK, Rich SS, Province MA, Hunt SC, et al. Evidence for major genes influencing pulmonary function in the NHLBI family heart study. Genetic epidemiology. 2000;19(1):81-94. Epub 2000/06/22.
71. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, et al. Standardisation of spirometry. Eur Respir J. 2005;26(2):319-38. Epub 2005/08/02.
72. Hankinson JL, Crapo RO, Jensen RL. Spirometric reference values for the 6-s FVC maneuver. Chest. 2003;124(5):1805-11. Epub 2003/11/08.
73. Hankinson JL, Kawut SM, Shahar E, Smith LJ, Stukovsky KH, Barr RG. Performance of American Thoracic Society-recommended spirometry reference values in a multiethnic sample of adults: the multi-ethnic study of atherosclerosis (MESA) lung study. Chest. 2010;137(1):138-45. Epub 2009/09/11.
74. Barnes PJ. Cellular and molecular mechanisms of chronic obstructive pulmonary disease. Clinics in chest medicine. 2014;35(1):71-86. Epub 2014/02/11.

75. Rahman I. Oxidative stress in pathogenesis of chronic obstructive pulmonary disease: cellular and molecular mechanisms. Cell biochemistry and biophysics. 2005;43(1):167-88. Epub 2005/07/27.
76. Hogg JC. Pathophysiology of airflow limitation in chronic obstructive pulmonary disease. Lancet. 2004;364(9435):709-21. Epub 2004/08/25.
77. Soler-Cataluna JJ, Martinez-Garcia MA, Roman Sanchez P, Salcedo E, Navarro M, Ochando R. Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease. Thorax. 2005;60(11):925-31. Epub 2005/08/02.
78. Larson RK, Barman ML. The familial occurrence of chronic obstructive pulmonary disease. Annals of internal medicine. 1965;63(6):1001-8. Epub 1965/12/01.
79. Tager I, Tishler PV, Rosner B, Speizer FE, Litt M. Studies of the familial aggregation of chronic bronchitis and obstructive airways disease. Int J Epidemiol. 1978;7(1):55-62. Epub 1978/03/01.
80. Lebowitz MD, Knudson RJ, Burrows B. Family aggregation of pulmonary function measurements. Am Rev Respir Dis. 1984;129(1):8-11. Epub 1984/01/01.
81. Lewitter FI, Tager IB, McGue M, Tishler PV, Speizer FE. Genetic and environmental determinants of level of pulmonary function. American journal of epidemiology. 1984;120(4):518-30. Epub 1984/10/01.
82. Palmer LJ, Knuiman MW, Divitini ML, Burton PR, James AL, Bartholomew HC, et al. Familial aggregation and heritability of adult lung function: results from the Busselton Health Study. Eur Respir J. 2001;17(4):696-702. Epub 2001/06/13.
83. Ingebrigtsen T, Thomsen SF, Vestbo J, van der Sluis S, Kyvik KO, Silverman EK, et al. Genetic influences on Chronic Obstructive Pulmonary Disease - a twin study. Respiratory medicine. 2010;104(12):1890-5. Epub 2010/06/15.
84. Stoller JK, Aboussouan LS. Alpha1-antitrypsin deficiency. Lancet. 2005;365(9478):2225-36. Epub 2005/06/28.
85. DeMeo DL, Silverman EK. Alpha1-antitrypsin deficiency. 2: genetic aspects of alpha(1)-antitrypsin deficiency: phenotypes and genetic modifiers of emphysema risk. Thorax. 2004;59(3):259-64. Epub 2004/02/27.
86. Silverman EK, Palmer LJ, Mosley JD, Barth M, Senter JM, Brown A, et al. Genomewide linkage analysis of quantitative spirometric phenotypes in severe early-onset chronic obstructive pulmonary disease. Am J Hum Genet. 2002;70(5):1229-39. Epub 2002/03/27.
87. Silverman EK, Chapman HA, Drazen JM, Weiss ST, Rosner B, Campbell EJ, et al. Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. American journal of respiratory and critical care medicine. 1998;157(6 Pt 1):1770-8. Epub 1998/06/25.
88. Palmer LJ, Celedon JC, Chapman HA, Speizer FE, Weiss ST, Silverman EK. Genome-wide linkage analysis of bronchodilator responsiveness and post-bronchodilator spirometric phenotypes in chronic obstructive pulmonary disease. Human molecular genetics. 2003;12(10):1199-210. Epub 2003/04/30.

89. Hall IP, Lomas DA. The genetics of obstructive lung disease: big is beautiful. Thorax. 2010;65(9):760-1. Epub 2010/09/02.
90. Smolonska J, Wijmenga C, Postma DS, Boezen HM. Meta-analyses on suspected chronic obstructive pulmonary disease genes: a summary of 20 years' research. American journal of respiratory and critical care medicine. 2009;180(7):618-31. Epub 2009/07/18.
91. Castaldi PJ, Cho MH, Cohn M, Langerman F, Moran S, Tarragona N, et al. The COPD genetic association compendium: a comprehensive online database of COPD genetic associations. Human molecular genetics. 2010;19(3):526-34. Epub 2009/11/26.
92.* Obeidat M, Wain LV, Shrine N, Kalsheker N, Soler Artigas M, Repapi E, et al. A comprehensive evaluation of potential lung function associated genes in the SpiroMeta general population sample. PLoS One. 2011;6(5):e19382. Epub 2011/06/01.
93.* Wain LV, Soler Artigas M, Tobin MD. What can genetics tell us about the cause of fixed airflow obstruction? Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology. 2012;42(8):1176-82. Epub 2012/07/19.
94. Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. PLoS Genet. 2009;5(3):e1000429. Epub 2009/03/21.
95. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, et al. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. PLoS Genet. 2009;5(3):e1000421. Epub 2009/03/21.
96. Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marciante KD, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nat Genet. 2010;42(1):45-52. Epub 2009/12/17.
97. Hayes JD, Flanagan JU, Jowsey IR. Glutathione transferases. Annual review of pharmacology and toxicology. 2005;45:51-88. Epub 2005/04/12.
98. Dupont LJ, Pype JL, Demedts MG, De Leyn P, Deneffe G, Verleden GM. The effects of 5-HT on cholinergic contraction in human airways in vitro. Eur Respir J. 1999;14(3):642-9. Epub 1999/10/30.
99. Fehrenbach H, Kasper M, Tschernig T, Shearman MS, Schuh D, Muller M. Receptor for advanced glycation endproducts (RAGE) exhibits highly differential cellular and subcellular localisation in rat and human lung. Cellular and molecular biology. 1998;44(7):1147-57. Epub 1998/12/10.
100. Konishi K, Gibson KF, Lindell KO, Richards TJ, Zhang Y, Dhir R, et al. Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. American journal of respiratory and critical care medicine. 2009;180(2):167-75. Epub 2009/04/14.
101. Englert JM, Hanford LE, Kaminski N, Tobolewski JM, Tan RJ, Fattman CL, et al. A role for the receptor for advanced glycation end products in idiopathic pulmonary fibrosis. The American journal of pathology. 2008;172(3):583-91. Epub 2008/02/05.

102. Chen H, Herndon ME, Lawler J. The cell biology of thrombospondin-1. Matrix biology : journal of the International Society for Matrix Biology. 2000;19(7):597-614. Epub 2000/12/05.
103. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. Nat Genet. 2010;42(3):200-2.
104. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management and Prevention of COPD. 2011; Available from: http://www.goldcopd.org.
105. Silverman EK, Vestbo J, Agusti A, Anderson W, Bakke PS, Barnes KC, et al. Opportunities and challenges in the genetics of COPD 2010: an International COPD Genetics Conference report. COPD. 2011;8(2):121-35. Epub 2011/04/19.
106. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, et al. Performance of common genetic variants in breast-cancer risk models. The New England journal of medicine. 2010;362(11):986-93. Epub 2010/03/20.
107. Talmud PJ, Hingorani AD, Cooper JA, Marmot MG, Brunner EJ, Kumari M, et al. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. BMJ. 2010;340:b4838. Epub 2010/01/16.
108. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. American journal of respiratory and critical care medicine. 1999;159(1):179-87. Epub 1999/01/05.
109. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. 2010;42(5):436-40. Epub 2010/04/27.
110. Johannessen A, Omenaas ER, Bakke PS, Gulsvik A. Implications of reversibility testing on prevalence and risk factors for chronic obstructive pulmonary disease: a community study. Thorax. 2005;60(10):842-7. Epub 2005/08/09.
111. Hoenig JM, Heisey DM. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. The American Statistician. 2001;55:19-24.
112. Van Durme YM, Eijgelsheim M, Joos GF, Hofman A, Uitterlinden AG, Brusselle GG, et al. Hedgehog-interacting protein is a COPD susceptibility gene: the Rotterdam Study. Eur Respir J. 2010;36(1):89-95. Epub 2009/12/10.
113. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478(7367):103-9. Epub 2011/09/13.
114. Wain LV, Verwoert GC, O'Reilly PF, Shi G, Johnson T, Johnson AD, et al. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nat Genet. 2011;43(10):1005-11. Epub 2011/09/13.
115. Kohansal R, Martinez-Camblor P, Agusti A, Buist AS, Mannino DM, Soriano JB. The natural history of chronic airflow obstruction revisited: an

analysis of the Framingham offspring cohort. American journal of respiratory and critical care medicine. 2009;180(1):3-10. Epub 2009/04/04.
116. Goddard GH, Lewis CM. Risk categorization for complex disorders according to genotype relative risk and precision in parameter estimates. Genetic epidemiology. 2010;34(6):624-32. Epub 2010/06/23.
117. Castaldi PJ, Cho MH, Litonjua AA, Bakke P, Gulsvik A, Lomas DA, et al. The Association of Genome-Wide Significant Spirometric Loci with COPD Susceptibility. Am J Respir Cell Mol Biol. 2011. Epub 2011/06/11.
118. Lokke A, Lange P, Scharling H, Fabricius P, Vestbo J. Developing COPD: a 25 year follow up study of the general population. Thorax. 2006;61(11):935-9. Epub 2006/10/31.
119. Hole DJ, Watt GC, Davey-Smith G, Hart CL, Gillis CR, Hawthorne VM. Impaired lung function and mortality risk in men and women: findings from the Renfrew and Paisley prospective population study. BMJ. 1996;313(7059):711-5; discussion 5-6. Epub 1996/09/21.
120. Strachan DP. Ventilatory function, height, and mortality among lifelong non-smokers. Journal of epidemiology and community health. 1992;46(1):66-70. Epub 1992/02/01.
121. Young RP, Hopkins R, Eaton TE. Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes. Eur Respir J. 2007;30(4):616-22. Epub 2007/10/02.
122. Lin DY, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genetic epidemiology. 2010;34(1):60-6. Epub 2009/10/23.
123. Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet. 2007;3(7):e114.
124. Guan Y, Stephens M. Practical issues in imputation-based association mapping. PLoS Genet. 2008;4(12):e1000279. Epub 2008/12/06.
125. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529. Epub 2009/06/23.
126. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annual review of genomics and human genetics. 2009;10:387-406. Epub 2009/09/01.
127. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832-8. Epub 2010/10/01.
128. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet. 2009;85(5):679-91. Epub 2009/10/20.
129. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010;42(7):570-5. Epub 2010/06/22.
130. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. European journal of human genetics : EJHG. 2011;19(7):807-12. Epub 2011/03/17.

131. Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, Qi L, et al. Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. PLoS Genet. 2009;5(6):e1000508. Epub 2009/06/27.
132. Elks CE, Perry JR, Sulem P, Chasman DI, Franceschini N, He C, et al. Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. Nat Genet. 2010;42(12):1077-85. Epub 2010/11/26.
133. Gibson MA, Hughes JL, Fanning JC, Cleary EG. The major antigen of elastin-associated microfibrils is a 31-kDa glycoprotein. The Journal of biological chemistry. 1986;261(24):11429-36. Epub 1986/08/25.
134. Faraco J, Bashir M, Rosenbloom J, Francke U. Characterization of the human gene for microfibril-associated glycoprotein (MFAP2), assignment to chromosome 1p36.1-p35, and linkage to D1S170. Genomics. 1995;25(3):630-7. Epub 1995/02/10.
135. Yang J, Liu X, Yue G, Adamian M, Bulgakov O, Li T. Rootletin, a novel coiled-coil protein, is a structural component of the ciliary rootlet. The Journal of cell biology. 2002;159(3):431-40. Epub 2002/11/13.
136. Yang J, Gao J, Adamian M, Wen XH, Pawlyk B, Zhang L, et al. The ciliary rootlet maintains long-term stability of sensory cilia. Molecular and cellular biology. 2005;25(10):4129-37. Epub 2005/05/05.
137. Chazaud C, Dolle P, Rossant J, Mollard R. Retinoic acid signaling regulates murine bronchial tubule formation. Mechanisms of development. 2003;120(6):691-700. Epub 2003/07/02.
138. Feng Q, Hawes SE, Stern JE, Wiens L, Lu H, Dong ZM, et al. DNA methylation in tumor and matched normal tissues from non-small cell lung cancer patients. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008;17(3):645-54. Epub 2008/03/20.
139. Bieganowski P, Shilinski K, Tsichlis PN, Brenner C. Cdc123 and checkpoint forkhead associated with RING proteins control the cell cycle by controlling eIF2gamma abundance. The Journal of biological chemistry. 2004;279(43):44656-66. Epub 2004/08/21.
140. Ito K, Ito M, Elliott WM, Cosio B, Caramori G, Kon OM, et al. Decreased histone deacetylase activity in chronic obstructive pulmonary disease. The New England journal of medicine. 2005;352(19):1967-76. Epub 2005/05/13.
141. Golding J, Pembrey M, Jones R, Team AS. ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. Paediatric and perinatal epidemiology. 2001;15(1):74-87. Epub 2001/03/10.
142. Newnham JP, Evans SF, Michael CA, Stanley FJ, Landau LI. Effects of frequent ultrasound during pregnancy: a randomised controlled trial. Lancet. 1993;342(8876):887-91. Epub 1993/10/09.
143. Williams LA, Evans SF, Newnham JP. Prospective cohort study of factors influencing the relative weights of the placenta and the newborn infant. BMJ. 1997;314(7098):1864-8. Epub 1997/06/28.

144. Evans S, Newnham J, MacDonald W, Hall C. Characterisation of the possible effect on birthweight following frequent prenatal ultrasound examinations. Early human development. 1996;45(3):203-14. Epub 1996/07/19.
145. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet. 2008;40(5):575-83. Epub 2008/04/09.
146. Berrettini W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H, et al. Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. Molecular psychiatry. 2008;13(4):368-73. Epub 2008/01/30.
147. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature. 2008;452(7187):638-42. Epub 2008/04/04.
148. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet. 2008;40(5):616-22. Epub 2008/04/04.
149. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature. 2008;452(7187):633-7. Epub 2008/04/04.
150. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9. Epub 2010/06/22.
151. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369-75, S1-3. Epub 2012/03/20.
152. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nature reviews Genetics. 2013;14(6):379-89. Epub 2013/05/10.
153. Lawrence R, Day-Williams AG, Elliott KS, Morris AP, Zeggini E. CCRaVAT and QuTie-enabling analysis of rare variants in large-scale case control and quantitative trait association studies. BMC bioinformatics. 2010;11:527.
154. Wilk JB, Shrine NR, Loehr LR, Zhao JH, Manichaikul A, Lopez LM, et al. Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. American journal of respiratory and critical care medicine. 2012;186(7):622-32.
155. 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. The Lancet Respiratory medicine. 2014;2(3):214-25. Epub 2014/03/14.
156. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nature reviews Genetics. 2010;11(11):773-85. Epub 2010/10/14.
157. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature reviews Genetics. 2012;13(1):36-46. Epub 2011/11/30.

158. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research. 2012;40(10):e72. Epub 2012/02/11.
159. Stewart CE, Hall IP, Parker SG, Moffat MF, Wardlaw AJ, Connolly MJ, et al. PLAUR polymorphisms and lung function in UK smokers. BMC medical genetics. 2009;10:112.
160. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
161. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
162. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
163. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome research. 2006;16(9):1182-90.
164. Benjamini Y, Heller R. Screening for partial conjunction hypotheses. Biometrics. 2008;64(4):1215-22. Epub 2008/02/12.
165. Watterson GA. On the number of segregating sites in genetical models without recombination. Theoretical population biology. 1975;7(2):256-76.
166. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001;29(1):308-11.
167. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic acids research. 2004;32(Database issue):D493-6.
168. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity. 2005;95(3):221-7.
169. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature methods. 2012;9(2):179-81. Epub 2011/12/06.
170. Hwang JY, Lee SH, Go MJ, Kim BJ, Kou I, Ikegawa S, et al. Meta-analysis identifies a MECOM gene as a novel predisposing factor of osteoporotic fracture. Journal of medical genetics. 2013;50(4):212-9.
171. Okada Y, Sim X, Go MJ, Wu JY, Gu D, Takeuchi F, et al. Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. Nat Genet. 2012;44(8):904-9.
172. Wain LV, Verwoert GC, O'Reilly PF, Shi G, Johnson T, Johnson AD, et al. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nat Genet. 2011;43(10):1005-11.
173. Walter S, Atzmon G, Demerath EW, Garcia ME, Kaplan RC, Kumari M, et al. A genome-wide association study of aging. Neurobiology of aging. 2011;32(11):2109 e15-28.
174. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, et al. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic acids research. 2013;41(Database issue):D56-63.

175. Zhou X, Baron RM, Hardin M, Cho MH, Zielinski J, Hawrylkiewicz I, et al. Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. Human molecular genetics. 2012;21(6):1325-35.
176. Cipriani V, Leung HT, Plagnol V, Bunce C, Khan JC, Shahid H, et al. Genome-wide association study of age-related macular degeneration identifies associated variants in the TNXB-FKBPL-NOTCH4 region of chromosome 6p21.3. Human molecular genetics. 2012;21(18):4138-50.
177. Lemaitre RN, Tanaka T, Tang W, Manichaikul A, Foy M, Kabagambe EK, et al. Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. PLoS Genet. 2011;7(7):e1002193.
178. Chung SA, Taylor KE, Graham RR, Nititham J, Lee AT, Ortmann WA, et al. Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. PLoS Genet. 2011;7(3):e1001323.
179. Fellay J, Ge D, Shianna KV, Colombo S, Ledergerber B, Cirulli ET, et al. Common genetic variation and the control of HIV-1 in humans. PLoS Genet. 2009;5(12):e1000791.
180. Replication DIG, Meta-analysis C, Asian Genetic Epidemiology Network Type 2 Diabetes C, South Asian Type 2 Diabetes C, Mexican American Type 2 Diabetes C, Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples C, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet. 2014;46(3):234-44. Epub 2014/02/11.
181. Franceschini N, van Rooij FJ, Prins BP, Feitosa MF, Karakas M, Eckfeldt JH, et al. Discovery and fine mapping of serum protein loci through transethnic meta-analysis. Am J Hum Genet. 2012;91(4):744-53. Epub 2012/10/02.
182. World Health Oraganization. World Health Statistics. Available from URL: http://www.who.int/gho/publications/world_health_statistics/2008/en/. 2008.
183. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46(11):1173-86. Epub 2014/10/06.
184. Cho MH, Castaldi PJ, Wan ES, Siedlinski M, Hersh CP, Demeo DL, et al. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. Human molecular genetics. 2012;21(4):947-57. Epub 2011/11/15.
185.* Hancock DB, Soler Artigas M, Gharib SA, Henry A, Manichaikul A, Ramasamy A, et al. Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. PLoS Genet. 2012;8(12):e1003098.
186. Aschard H, Lutz S, Maus B, Duell EJ, Fingerlin TE, Chatterjee N, et al. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Human genetics. 2012;131(10):1591-613. Epub 2012/07/05.
187.* Tang W, Kowgier M, Loth DW, Soler Artigas M, Joubert BR, Hodge E, et al. Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. PLoS One. 2014;9(7):e100776. Epub 2014/07/02.

188. Bonnefond A, Clement N, Fawcett K, Yengo L, Vaillant E, Guillaume JL, et al. Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nat Genet. 2012;44(3):297-301. Epub 2012/01/31.
189. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011;88(3):294-305. Epub 2011/03/08.
190. Thun GA, Imboden M, Ferrarotti I, Kumar A, Obeidat M, Zorzetto M, et al. Causal and Synthetic Associations of Variants in the SERPINA Gene Cluster with Alpha1-antitrypsin Serum Levels. PLoS Genet. 2013;9(8):e1003585. Epub 2013/08/31.
191. Collins SA, Lucas JS, Inskip HM, Godfrey KM, Roberts G, Holloway JW, et al. HHIP, HDAC4, NCR3 and RARB polymorphisms affect fetal, childhood and adult lung function. Eur Respir J. 2013;41(3):756-7. Epub 2013/03/05.
192.* Obeidat M, Miller S, Probert K, Billington CK, Henry AP, Hodge E, et al. GSTCD and INTS12 Regulation and Expression in the Human Lung. PLoS One. 2013;8(9):e74630. Epub 2013/09/24.
193. Hodge E, Nelson CP, Miller S, Billington CK, Stewart CE, Swan C, et al. HTR4 gene structure and altered expression in the developing lung. Respiratory research. 2013;14:77. Epub 2013/07/31.

Appendices

A. Articles directly related to the thesis

Soler Artigas, 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. 184(7): p. 786-95.

Effect of Five Genetic Variants Associated with Lung Function on the Risk of Chronic Obstructive Lung Disease, and Their Joint Effects on Lung Function

María Soler Artigas[1], Louise V. Wain[1], Emmanouela Repapi[1,2], Ma'en Obeidat[3], Ian Sayers[3], Paul R. Burton[1], Toby Johnson[4], Jing Hua Zhao[5], Eva Albrecht[6], Anna F. Dominiczak[7], Shona M. Kerr[8], Blair H. Smith[9], Gemma Cadby[10,11], Jennie Hui[12–14], Lyle J. Palmer[10,11], Aroon D. Hingorani[15], S. Goya Wannamethee[16], Peter H. Whincup[17], Shah Ebrahim[18], George Davey Smith[19], Inês Barroso[20,21], Ruth J. F. Loos[5], Nicholas J. Wareham[5], Cyrus Cooper[22], Elaine Dennison[22], Seif O. Shaheen[23], Jason Z. Liu[24], Jonathan Marchini[24], Medical Research Council National Survey of Health and Development (NSHD) Respiratory Study Team[25,26*], Santosh Dahgam[27], Åsa Torinsson Naluai[28], Anna-Carin Olin[27], Stefan Karrasch[29], Joachim Heinrich[30], Holger Schulz[30], Tricia M. McKeever[31,32], Ian D. Pavord[33], Markku Heliövaara[34], Samuli Ripatti[34,35], Ida Surakka[34,35], John D. Blakey[3], Mika Kähönen[36], John R. Britton[31,32], Fredrik Nyberg[27,37], John W. Holloway[38,39], Debbie A. Lawlor[19], Richard W. Morris[16], Alan L. James[13,40], Cathy M. Jackson[41], Ian P. Hall[3], Martin D. Tobin[1], and the SpiroMeta Consortium[†]

[1]Departments of Health Sciences and Genetics, University of Leicester, Leicester; [2]Ludwig Institute for Cancer Research, University of Oxford, Oxford, United Kingdom; [3]Division of Therapeutics and Molecular Medicine, Nottingham Respiratory Biomedical Research Unit, University Hospital of Nottingham, Nottingham, United Kingdom; [4]Clinical Pharmacology, William Harvey Research Institute, Barts and London School of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom; [5]Medical Research Council (MRC) Epidemiology Unit, Institute of Metabolic Science, Cambridge, United Kingdom; [6]Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [7]College of Medical, Veterinary, and Life Sciences, University of Glasgow, Glasgow, Scotland; [8]Medical Genetics, University of Edinburgh, Molecular Medicine Centre, Edinburgh, Scotland; [9]Centre of Academic Primary Care, University of Aberdeen, Scotland, United Kingdom; [10]Ontario Institute for Cancer Research and [11]Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada; [12]Molecular Genetics, PathWest Laboratory Medicine, Nedlands, Western Australia; [13]Busselton Population Medical Research Foundation, Sir Charles Gairdner Hospital, Nedlands, Western Australia; [14]Schools of Population Health and Pathology and Laboratory Medicine, University of Western Australia, Perth, Western Australia, Australia; [15]Department of Epidemiology and Public Health, University College London, London, United Kingdom; [16]Department of Primary Care and Population Health, University College London, London, United Kingdom; [17]Division of Community Health Sciences, St. George's University of London, London, United Kingdom; [18]Non-communicable Diseases Epidemiology Unit, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom; [19]MRC Centre for Causal Analyses in Translational Epidemiology, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom; [20]Wellcome Trust Sanger Institute, Cambridge, United Kingdom; [21]University of Cambridge Metabolic Research Laboratories, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, United Kingdom; [22]MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, United Kingdom; [23]Centre for Health Sciences, Barts and London School of Medicine, London, United Kingdom; [24]Department of Statistics, University of Oxford, Oxford, United Kingdom; [25]MRC National Survey of Health and Development, MRC Unit for Lifelong Health and Ageing, London, United Kingdom; [26]MRC-HPA (Health Protection Agency) Centre for Environment and Health, Imperial College London, St. Mary's Campus, London, United Kingdom; [27]Occupational and Environmental Medicine and [28]Department of Microbiology and Immunology, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; [29]Institute and Outpatient Clinic for Occupational, Social, and Environmental Medicine, Ludwig Maximilian University, Munich, Germany; [30]Institute of Epidemiology I, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; [31]Division of Epidemiology and Public Health, School of Community Health Sciences, University of Nottingham, City Hospital and [32]Nottingham Respiratory Biomedical Research Unit, University of Nottingham, Nottingham, United Kingdom; [33]Institute for Lung Health, Glenfield Hospital, University Hospitals of Leicester National Health Service Trust, Leicester, United Kingdom; [34]National Institute for Health and Welfare, Helsinki, Finland; [35]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland; and; [36]Department of Clinical Physiology, University of Tampere and Tampere University Hospital, Tampere, Finland; [37]AstraZeneca Research and Development, Mölndal, Sweden; [38]Human Genetics Division and [39]Infection, Inflammation, and Immunity Division, School of Medicine, University of Southampton, Southampton General Hospital, Southampton, United Kingdom; [40]Department of Pulmonary Physiology/West Australian Sleep Disorders Research Institute, Sir Charles Gairdner Hospital, Nedlands, Western Australia, Australia; and [41]University of St. Andrews, St. Andrews, Scotland, United Kingdom

(*Received in original form February 1, 2011; accepted in final form May 26, 2011*)

Supported/partially supported through the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE project, grant agreement HEALTH-F4-2007-201413. Complete information concerning funding may be found before the beginning of the REFERENCES.

* See the online supplement for MRC National Survey of Health and Development (NSHD) Respiratory Study Team membership list.

† A list of the SpiroMeta Consortium membership is available in the online supplement.

Am J Respir Crit Care Med Vol 184. pp 786–795, 2011
Originally Published in Press as DOI: 10.1164/rccm.201102-0192OC on June 16, 2011
Internet address: www.atsjournals.org

Complete information concerning author contributions may be found before the beginning of the REFERENCES.

Correspondence and requests for reprints should be addressed to Martin D. Tobin, M.B.Ch.B., MRC Senior Clinical Fellow, Genetic Epidemiology Group, Departments of Health Sciences and Genetics, 2nd Floor, Adrian Building, University of Leicester, University Road, Leicester LE1 7RH, UK. E-mail: mt47@le.ac.uk

This article has an online supplement, which is available from this issue's table of contents at www.atsjournals.org

Rationale: Genomic loci are associated with FEV_1 or the ratio of FEV_1 to FVC in population samples, but their association with chronic obstructive pulmonary disease (COPD) has not yet been proven, nor have their combined effects on lung function and COPD been studied.

Objectives: To test association with COPD of variants at five loci (*TNS1*, *GSTCD*, *HTR4*, *AGER*, and *THSD4*) and to evaluate joint effects on lung function and COPD of these single-nucleotide polymorphisms (SNPs), and variants at the previously reported locus near *HHIP*.

Methods: By sampling from 12 population-based studies (n = 31,422), we obtained genotype data on 3,284 COPD case subjects and 17,538 control subjects for sentinel SNPs in *TNS1*, *GSTCD*, *HTR4*, *AGER*, and *THSD4*. In 24,648 individuals (including 2,890 COPD case subjects and 13,862 control subjects), we additionally obtained genotypes for rs12504628 near *HHIP*. Each allele associated with lung function decline at these six SNPs contributed to a risk score. We studied the association of the risk score to lung function and COPD.

Measurements and Main Results: Association with COPD was significant for three loci (*TNS1*, *GSTCD*, and *HTR4*) and the previously reported *HHIP* locus, and suggestive and directionally consistent for *AGER* and *TSHD4*. Compared with the baseline group (7 risk alleles), carrying 10–12 risk alleles was associated with a reduction in FEV_1 ($\beta = -72.21$ ml, $P = 3.90 \times 10^{-4}$) and FEV_1/FVC ($\beta = -1.53\%$, $P = 6.35 \times 10^{-6}$), and with COPD (odds ratio = 1.63, $P = 1.46 \times 10^{-5}$).

Conclusions: Variants in *TNS1*, *GSTCD*, and *HTR4* are associated with COPD. Our highest risk score category was associated with a 1.6-fold higher COPD risk than the population average score.

Keywords: FEV_1; FVC; genome-wide association study; modeling risk

Chronic obstructive pulmonary disease (COPD), characterized by airflow limitation that is not fully reversible, affects approximately 210 million people worldwide (1) and is among the leading causes of death in developed and developing countries (2, 3). Tobacco smoking is a potent cause of COPD, but not all smokers develop COPD and genetic determinants are also important (4). Identification of the genetic determinants of COPD could provide insight into molecular pathways that may be amenable to improved preventive and treatment strategies. A further potential utility of newly identified genetic associations is to predict disease risk. Current evidence available from common complex diseases where family history may be used, such as type 2 diabetes, suggests that tens of genetic variants with individually modest effects may provide similar but not necessarily substantially improved disease risk prediction over existing scores that incorporate family history (5, 6)

Genetic variants in *SERPINA1* that cause α_1-antitrypsin deficiency have long been known to affect COPD risk. Although there had been limited success in identifying additional susceptibility loci for COPD until 2009 (7), more recent genome-wide association (GWA) studies have shown associations between genetic loci and lung function measures that underpin the diagnosis of COPD (8–10). Two of these studies have investigated genome-wide association with lung function in large sample sizes (>20,000 subjects), focusing exclusively on quantitative lung function measures (8, 9). Since the advent of dense GWA genotyping platforms, the only loci convincingly associated with COPD have been *HHIP* (10, 11), which to date remains the strongest signal, *CHRNA3/5* (11), and *FAM13A* (12).

We hypothesized that genetic variants associated with FEV_1 and FEV_1/FVC would be associated with COPD. In a study of 20,288 individuals with GWA data and follow-up of top signals in a further 54,276 individuals (SpiroMeta Consortium), we previously identified five novel loci showing association ($P < 5 \times 10^{-8}$) with FEV_1 or FEV_1/FVC: in *TNS1* at 2q35, in *GSTCD* at 4q24, in *HTR4* at 5q33, in *AGER* at 6p21, and in *THSD4* at 15q23 (9).

In this article, based on the five loci reported by the SpiroMeta Consortium (9), we further investigate the clinical relevance of these loci. First, we test association of the sentinel single-nucleotide polymorphism (SNP) at each locus with COPD. Second, we investigate the combined effect of the risk alleles at all five novel loci described previously (*TNS1*, *GSTCD*, *HTR4*, *AGER*, and *THSD4*) and the previous association at 4q31 (near *HHIP*) (10) on lung function and COPD risk.

Some of the results of these studies have been previously reported in the form of abstracts (13, 14).

AT A GLANCE COMMENTARY

Scientific Knowledge on the Subject

Genome-wide association studies have reported novel loci for lung function, but the association of these loci with chronic obstructive pulmonary disease (COPD) has not yet been tested, and their effects in combination have not yet been documented.

What This Study Adds to the Field

We show associations between COPD and polymorphisms in *HTR4*, *GSTCD*, and *TNS1*. Using a six-SNP (single-nucleotide polymorphism) risk score, incorporating variants in *HTR4*, *GSTCD*, *TNS1*, *AGER*, *THSD4*, and near *HHIP*, the highest risk category (~5% of Europeans) is associated with a 1.6- fold risk of COPD compared with a common baseline group.

METHODS

Study 1: Single SNP Analysis of *TNS1*, *GSTCD*, *HTR4*, *AGER*, *THSD4* with COPD

Populations, phenotyping, and genotyping. Figure 1 shows the study populations and loci included in the study design. The study population consisted of 31,422 individuals over the age of 40 years from 12 population-based studies. These studies included the European Prospective Investigation into Cancer and Nutrition obese cases cohort (EPIC-obese case subjects) and population cohort (EPIC population-based), Generation Scotland: Scottish Family Health Study (GS: SFHS), Cooperative Health Research in the Region of Augsburg (KORA F4), Adult-onset Asthma and Nitric Oxide (ADONIX) Study, Busselton Health Study (BHS), British Regional Heart Study (BRHS), British Women's Heart and Health Study (BWHHS), Gedling Study (Gedling), Hertfordshire Cohort Study (HCS), Finnish Health 2000 Survey (Health 2000), Nottingham Smokers Study (Nottingham Smokers), and Medical Research Council National Survey of Health and Development (NSHD, or British 1946 Birth Cohort).

FEV_1 and the ratio of FEV_1 to FVC were measured in each study, using the spirometry methods detailed in the online supplement. The percent predicted FEV_1 was calculated according to previously described prediction equations (15, 16). Individuals with percent predicted FEV_1 less than 80% and FEV_1/FVC less than 0.7 (Global Initiative for Chronic Obstructive Lung Disease [GOLD] stages 2-4) were classified as COPD case subjects (17). Individuals with FEV_1 greater than 80% predicted and FEV_1/FVC greater than 0.7 were classified as control subjects. To minimize potential misclassification of COPD case subjects and control subjects, individuals not falling in either category (GOLD stage 1) were excluded from the analyses of COPD risk. We also excluded related individuals from the case-control analyses.

Genotyping was undertaken for a sentinel SNP at each of the following loci: *TNS1* (rs2571445), *GSTCD* (rs10516526), *HTR4* (rs3995090),

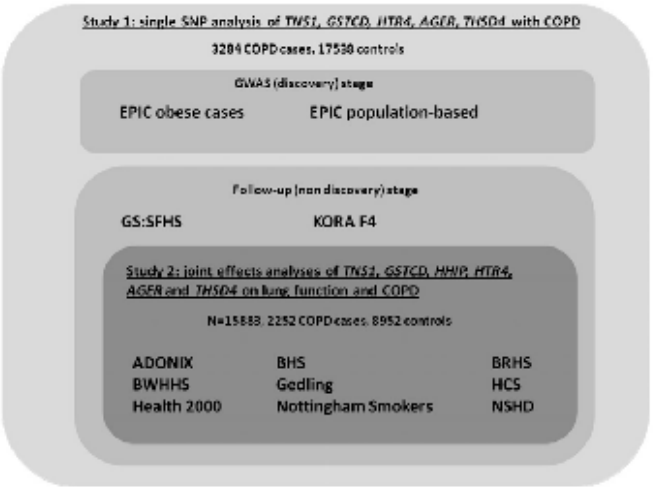


Figure 1. Study design. Single-nucleotide polymorphisms (SNPs) genotyped at each of the loci listed were rs2571445 (*TNS1*), rs10516526 (*GSTCD*), rs12504628 (near *HHIP*), rs3995090 (*HTR4*), rs2070600 (*AGER*), and rs12899618 (*THSD4*). Study 1 included all participating studies. Study 2 included the subset of studies that were genotyped for all six SNPs and that were not included in the discovery set of genome-wide association study (GWAS) data that led to the discovery of our five loci (9). Individuals excluded from study 2 chronic obstructive pulmonary disease (COPD) analyses were as follows: (*1*) individuals under the age of 40 years, (*2*) individuals with stage 1 COPD ($FEV_1/FVC < 0.7$ and percent predicted $FEV_1 > 80\%$), and (*3*) individuals with FEV_1/FVC greater than 0.7 but percent predicted FEV_1 less than 80%. Study abbreviations are as follows: EPIC obese case subjects (European Prospective Investigation into Cancer and Nutrition-obese case subjects) and EPIC population-based (European Prospective Investigation into Cancer and Nutrition cohort), GS:SFHS (Generation Scotland: Scottish Family Health Study), KORA F4 (Cooperative Health Research in the Region of Augsburg), ADONIX (Adult-onset Asthma and Nitric Oxide), BHS (Busselton Health Study), BRHS (British Regional Heart Study), BWHHS (British Women's Heart and Health Study), Gedling (Gedling Study), HCS (Hertfordshire Cohort Study), Health 2000 (Finnish Health 2000 Survey), Nottingham Smokers (Nottingham Smokers Study), and NSHD (Medical Research Council National Survey of Health and Development, also known as the British 1946 Birth Cohort).

AGER (rs2070600), and *THSD4* (rs12899618). Standard quality control approaches were used. To incorporate the previously reported locus near *HHIP* into the analyses, rs12504628 was genotyped in studies that employed KASPar genotyping (KBioscience, Hoddesdon, Herts, UK) and available *in silico* data for rs12504628 were used for EPIC, Health 2000, and a subset of BHS (footnote † in Table 1).

Statistical analysis. Each SNP genotype was coded 0, 1, or 2, corresponding to the number of copies of the coded allele. The effect estimates were oriented to the forward strand of the National Center for Biotechnology Information (NCBI) build 36 reference sequence of the human genome, using the alphabetically higher allele as the coded allele. For each of the five SNPs reported by the SpiroMeta Consortium (9), logistic regression was used within each study population to test association of the SNP with COPD. Adjustments for additional covariates were not used for this analysis given that the percent predicted FEV_1 (used to define COPD) includes adjustments for age, sex, and height (18). We defined Bonferroni-corrected statistical significance as P less than 0.01, to account for the testing of the five independent SNPs. After quality checks of the study population level data, pooled effect size estimates and their standard errors were computed across all studies, using an inverse variance weighting. Although previously shown to be associated with COPD and therefore not the focus of our analysis, for completeness we present the associations with rs12504628 (near *HHIP*).

Study 2: Joint Effects of *TNS1, GSTCD, HHIP, HTR4, AGER,* and *THSD4* Loci on Lung Function and COPD

Populations. The studies and loci included in study 2 are shown in Figure 1. We included the subset of studies that were genotyped for all six SNPs and that were not in the set of GWA data that led to the discovery of these loci (9). This restriction was to avoid cumulative biases resulting from individual SNP estimates of association that tend to be biased away from the null in "hits" from genome-wide discovery sets (winner's curse bias) (19). The phenotyping and genotyping of these studies were undertaken as described for study 1.

Statistical analysis: derivation of the risk score. To derive an unweighted risk allele score theoretically ranging from 0 to 12, each allele previously associated with reduced FEV_1 or FEV_1/FVC (9) contributed one to the risk score. The risk alleles for the six loci were as follows: A for rs2571445 (*TNS1*), A for rs10516526 (*GSTCD*), T for rs12504628 (*HHIP*), A for rs3995090 (*HTR4*), C for rs2070600 (*AGER*), and A for rs12899618 (*THSD4*). We categorized the risk scores into five groups; 0–4, 5–6, 7 (median number of risk alleles, baseline group), 8–9, and 10–12 risk alleles.

Statistical analysis: testing association between the risk score and lung function. We performed linear regressions of FEV_1 and FEV_1/FVC onto age, age squared, sex, and height to obtain residual phenotypes. Linear regressions were undertaken with each residual phenotype as the outcome variable, and an intercept and four indicator variables (for the four nonbaseline risk allele groups) as the explanatory variables to test for association. After quality checks of study level data, we computed pooled estimates using an inverse variance weighting.

Statistical analysis: testing association between the risk score and COPD. Individuals were classified as COPD case subjects (GOLD stages 2–4) or control subjects on the basis of the criteria described previously for study 1. To test the association of the unweighted risk allele score with COPD we used logistic regression with COPD as the outcome, an intercept, and the four indicator variables. Study level estimates were pooled after quality checks using an inverse variance weighting.

RESULTS

Study 1: Single SNP Analysis of *TNS1, GSTCD, HTR4, AGER,* and *THSD4* with COPD

The characteristics of the study participants are shown in Table 1. For the individual SNP associations with COPD a total of 3,284 individuals were classified as COPD case subjects (percent predicted $FEV_1 < 80\%$ and $FEV_1/FVC < 0.7$) and 17,538 individuals as control subjects ($FEV_1 > 80\%$ predicted and $FEV_1/FVC > 0.7$). Of the variants at the five loci shown to be associated with lung function (9), variants at three loci showed significant association with COPD: rs2571445 in *TNS1* (odds ratio [OR] per A allele, 1.10; 95% confidence interval [CI], 1.03–1.16; $P = 1.89 \times 10^{-3}$) (Figures 2 and 3), rs10516526 in *GSTCD* (OR per A allele, 1.24; 95% CI, 1.10–1.40; $P = 3.75 \times 10^{-4}$) (Figures 2 and 3), and rs3995090 in *HTR4* (OR per A allele, 1.12; 95%

TABLE 1. STUDY CHARACTERISTICS

Study Name	Subset	No. in Study 1	No. in Study 2	Male (*n*); Female (*n*)	Age: Mean (SD) (yr)	FEV_1: Mean (SD) (*L*)	FEV_1 Predicted: Mean (SD) (*L*)	FVC: Mean (SD) (*L*)	FEV_1/FVC Ratio: Mean (SD)	Never-smokers (*n*); Ever-smokers (*n*)	COPD Severity (% of COPD Case Subjects in GOLD Stage 3 or 4)	Genotyping
Study 1 Single SNP Analysis: GWAS (Discovery) Stage												
EPIC obese case subjects	All	1,104		476; 628	59.1 (8.8)	2.35 (0.69)	2.91 (0.62)	2.84 (0.87)	0.82(0.17)	489; 615		Affymetrix 500K
	Case subjects	75		47; 28	60.64 (8.45)	1.82 (0.67)	3.09 (0.63)	3.01 (1.05)	0.61 (0.09)	22; 53	30.14	
	Control subjects	599		252; 347	58.30 (8.76)	2.67 (0.62)	2.88 (0.61)	3.13 (0.79)	0.86 (0.07)	281; 318		
EPIC population-based	All	2,336		1,100; 1,236	59.2 (9.0)	2.50 (0.72)	2.95 (0.62)	3.04 (0.90)	0.85 (0.16)	1,061; 1,275		Affymetrix 500K
	Case subjects	190		105; 85	62.31 (8.54)	1.81 (0.62)	2.95 (0.63)	3.00 (0.96)	0.60 (0.09)	72; 118	20.11	
	Control subjects	1,442		677; 765	58.76 (8.81)	2.78 (0.64)	2.94 (0.62)	3.29 (0.82)	0.85 (0.08)	709; 733		
Study 1 Single SNP Analysis: Follow-up (Nondiscovery) Stage												
GSS/HS	All	5,474		2,254; 3,220	46.0 (14.3)	3.15 (0.87)	3.32 (0.75)	4.11 (1.03)	0.77 (0.10)	3,005; 2,469		TaqMan
	Case subjects	335		118; 217	58.4 (9.2)	1.89 (0.54)	2.89 (0.59)	3.32 (0.85)	0.58 (0.10)	123; 212	11.94	
	Control subjects	2,567		1,053;1,514	53.2 (8.5)	3.12 (0.72)	3.11 (0.63)	3.99 (0.91)	0.78 (0.07)	1,457; 1,110		
KORAF4	All	1,305		610; 695	51.6 (5.7)	3.32 (0.81)	3.29 (0.63)	4.28 (1.00)	0.78 (0.06)	499; 806		TaqMan
	Case subjects	59		30; 29	53.4 (6.0)	2.14 (0.66)	3.24 (0.64)	3.47 (0.96)	0.61 (0.06)	12; 47	10.17	
	Control subjects	1,109		512; 597	51.5 (5.7)	3.45 (0.76)	3.28 (0.62)	4.36 (0.96)	0.79 (0.04)	456; 653		
Study 1 Single SNP Analysis and Study 2 Joint Effects Analysis: Follow-up (Nondiscovery) Stage												
ADONIX	All	1,423	1,282	669; 754	49.1 (13.5)	3.34 (0.86)	3.23 (0.66)	4.24 (1.02)	0.79 (0.07)	798; 625		KASPar*
	Case subjects	46	41	27; 19	55.7 (9.3)	2.02 (0.57)	3.23 (0.66)	3.35 (0.87)	0.60 (0.07)	12; 34	13.04	
	Control subjects	783	711	361; 422	61.4 (8.4)	3.23 (0.73)	3.23 (0.67)	4.08 (0.91)	0.79 (0.04)	448; 335		
BHS	All†	4,350	787	1,793; 2,557	50.1 (17.0)	3.02 (0.97)	3.18 (0.82)	3.89 (1.16)	0.77 (0.08)	2,459; 1,891		TaqMan
	Case subjects†	200	92	132; 68	66.9 (11.6)	1.60 (0.60)	2.85 (0.66)	2.73 (0.91)	0.58 (0.09)	67; 133	19.5	
	Control subjects†	2,307	386	944; 1,363	57.9 (12.3)	2.87 (0.83)	2.93 (0.73)	3.66 (1.05)	0.78 (0.05)	1,387; 920		
BRHS	All	3,877	3,415	3,877; 0	68.7 (5.5)	2.57 (0.69)	3.03 (0.4)	3.37 (0.84)	0.77 (0.12)	1,125; 2,752		KASPar*
	Case subjects	641	572	641; 0	69.7 (5.4)	1.76 (0.51)	3 (0.4)	3.01 (0.8)	0.59 (0.09)	111; 530	28.39	
	Control subjects	2,168	1,905	2,168; 0	68.3 (5.5)	2.96 (0.48)	3.03 (0.4)	3.65 (0.65)	0.82 (0.07)	760; 1,408		
BWHHS	All	3,644	3,319	0; 3,644	68.8 (5.5)	1.98 (0.52)	2.16 (0.31)	2.82 (0.76)	0.71 (0.09)	2,060; 1,584		KASPar*
	Case subjects	659	600	0; 659	69.8 (5.4)	1.36 (0.35)	2.14 (0.3)	2.32 (0.54)	0.59 (0.08)	253; 406	15.63	
	Control subjects	1,808	1,653	0; 1,808	68.2 (5.4)	2.23 (0.41)	2.18 (0.3)	2.93 (0.56)	0.76 (0.05)	1,153; 655		
Gesling	All	1,263	1,188	632; 631	56.2 (12.3)	2.85 (0.85)	3.07 (0.69)	3.68 (1.02)	0.77 (0.07)	633; 630		KASPar*
	Case subjects	103	98	67; 36	66.2 (9.1)	1.73 (0.61)	2.88 (0.66)	2.82 (0.83)	0.61 (0.09)	21; 82	24.27	
	Control subjects	840	789	417; 423	57.3 (9.8)	3 (0.73)	3.03 (0.65)	3.8 (0.9)	0.79 (0.04)	431; 409		
HCS	All	2,850	2,343	1,511; 1,339	66.1 (2.8)	2.44 (0.68)	2.80 (0.55)	3.42 (0.92)	0.72 (0.09)	1,319; 1,531		KASPar*
	Case subjects	536	441	308; 228	66.3 (2.8)	1.84 (0.55)	2.87 (0.56)	2.09 (0.85)	0.60 (0.09)	159; 377	15.1	
	Control subjects	1,519	1,264	758; 761	66.0 (2.9)	2.67 (0.60)	2.75 (0.54)	3.51 (0.82)	0.76 (0.04)	837; 682		
Health 2000	All	888	882	427; 456	50.2 (11.0)	3.32 (0.91)	3.14 (0.67)	4.19 (1.08)	0.79 (0.07))	266; 617		Illumina 610K
	Case subjects	32	32	20; 12	60.91 (8.83)	1.78 (0.68)	3.05 (0.66)	3.05 (0.95)	0.58 (0.10)	5; 27	37.5	
	Control subjects	580	580	256; 324	53.19 (8.31)	3.28 (0.77)	3.15 (0.67)	4.09 (0.97)	0.80 (0.05)	192; 388		

(Continued)

TABLE 1. (CONTINUED)

Study Name	Subset	No. in Study 1	No. in Study 2	Male (*n*); Female (*n*)	Age: Mean (SD) (yr)	FEV_1: Mean (SD) (*L*)	FEV_1 Predicted: Mean (SD) (*L*)	FVC: Mean (SD) (*L*)	FEV_1/FVC Ratio: Mean (SD)	Never-smokers (*n*); Ever-smokers (*n*)	COPD Severity (% of COPD Case Subjects in GOLD Stage 3 or 4)	Genotyping
Nottingham Smokers	All	509	466	280; 229	59.5 (10.4)	2.00 (0.95)	2.98 (0.61)	3.02 (1.06)	0.64 (0.16)	0; 509		KASPar*
	Case subjects	242	227	145; 97	63.2 (9.5)	1.28 (0.57)	2.87 (0.59)	2.5 (0.87)	0.51 (0.12)	0; 242	64.46	
	Control subjects	153	138	70; 83	54.8 (8.9)	2.89 (0.61)	3.08 (0.62)	3.69 (0.81)	0.79 (0.05)	0; 153		
NSHD	All	2,404	2,201	1,206; 1,198	53 (0)	2.80 (0.70)	3.20 (0.56)	3.51 (0.89)	0.80 (0.09)	1,003; 1,401		KASPar*
	Case subjects	166	149	102; 64	53 (0)	2.11 (0.58)	3.35 (0.54)	3.46 (0.89)	0.61 (0.08)	49; 117	15.06	
	Control subjects	1,663	1,526	848; 815	53 (0)	3.03 (0.62)	3.20 (0.56)	3.69 (0.81)	0.83 (0.06)	765; 898		
Total	All	31,422	15,883									
	Case subjects	3,284	2,252									
	Control subjects	17,538	8,952									

Definition of abbreviations: ADONIX = Adult-onset Asthma and Nitric Oxide Study; BHS = Busselton Health Study; BRHS = British Regional Heart Study; BWHHS = British Women's Heart and Health Study; COPD = chronic obstructive pulmonary disease; EPIC = European Prospective Investigation into Cancer and Nutrition; GOLD = Global Initiative for Chronic Obstructive Lung Disease; GS:SHS = Generation Scotland: Scottish Family Health Study; GWAS = genome-wide association study; HCS = Hertfordshire Cohort Study; KORA F4 = Cooperative Health Research in the Region of Augsburg; NSHD = Medical Research Council National Survey of Health and Development.

* KASPar genotyping (KBiosciences, Hoddesdon, Herts, UK; http://www.kbioscience.co.uk/).

† The BHS had genotype data for *HHIP* only for a subset of individuals (*n* = 1,168, 131 COPD case subjects and 565 control subjects); this is therefore the subset included in study 2.

CI, 1.05–1.18; $P = 1.79 \times 10^{-4}$) (Figures 2 and 3). The associations with COPD were in the direction expected, given the reported direction of these allelic effects on lung function (Figure 2) (9). The point estimates of the effects on COPD for the sentinel SNPs in *AGER* (rs2070600) and in *THSD4* (rs12899618) were of the magnitude and direction expected, but did not reach statistical significance. Although it was not a primary aim of our study to investigate the association between COPD and the locus near *HHIP* at 4q31 (an association reported by several studies [10, 11, 20]), we were also able to test this in a subset of our data (2,890 case subjects and 13,862 control subjects) for which the rs12504628 genotype was available. We confirmed the association between the 4q31 locus and COPD: OR per rs12504628 T allele, 1.19 (95% CI, 1.12–1.27) and $P = 4.55 \times$

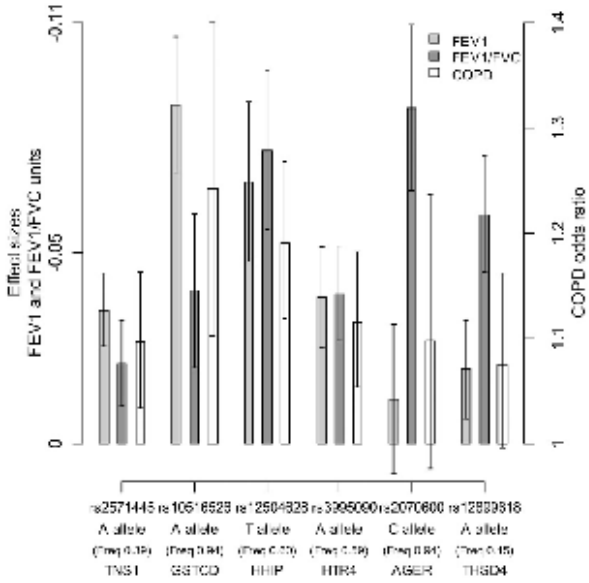


Figure 2. Association of sentinel single-nucleotide polymorphisms (SNPs) at five novel lung function loci and a previously reported *HHIP* SNP with chronic obstructive pulmonary disease (COPD), and comparison with reported associations with FEV_1 and FEV_1/FVC. Shown are the results for COPD after testing for association in 3,284 COPD case subjects and 17,538 control subjects, and a comparison with the associations with FEV_1 and FEV_1/FVC in the combined discovery and follow-up data reported by Repapi and colleagues (9). *Boxes* indicate the point estimates of the effect sizes and *whiskers* the 95% confidence intervals.

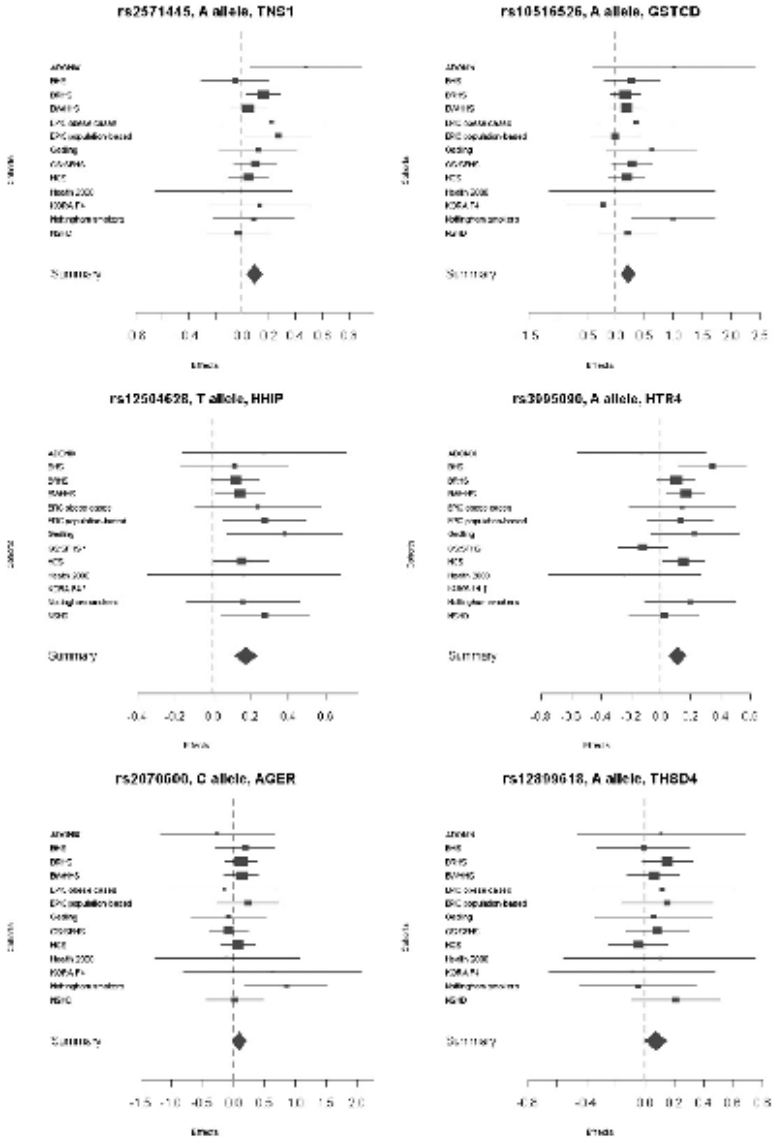


Figure 3. Forest plots of the meta-analysis of association tests with chronic obstructive pulmonary disease (COPD) for the six loci (*TNS1*, *GSTCD*, *HHIP*, *HTR4*, *AGER*, and *THSD4*). *rs12504628 (*HHIP*) data were not available in GS:SFHS or KORA F4. †KORA F4 failed to genotype rs3995090 (*HTR4*). ADONIX = Adult-onset Asthma and Nitric Oxide Study; BHS = Busselton Health Study; BRHS = British Regional Heart Study; BWHHS = British Women's Heart and Health Study; COPD = chronic obstructive pulmonary disease; EPIC = European Prospective Investigation into Cancer and Nutrition; GS:SFHS = Generation Scotland: Scottish Family Health Study; HCS = Hertfordshire Cohort Study; KORA F4 = Cooperative Health Research in the Region of Augsburg; NSHD = Medical Research Council National Survey of Health and Development.

10^{-8}. We tested for heterogeneity of effect sizes across the studies for each SNP; a chi-squared heterogeneity test was not statistically significant ($P > 0.10$) for any of the six SNPs.

As expected, a much higher proportion were ever-smokers among COPD case subjects (72%) than among control subjects (49%). This raises the possibility that these variants could influence COPD risk via an effect on smoking behavior. Therefore we explored whether the effects on COPD risk could be mediated via smoking. First, we examined the effects of an additional pack-years adjustment in smokers in a subset of our data with pack-years available and obtained similar results to those without adjustment for pack-years (*see* Figure E1 in the online supplement). Second, we assessed the association between these SNPs and two smoking behavior phenotypes in the Oxford-GlaxoSmithKline (Ox-GSK) consortium data set (21). As shown in Table E1, none of the SNPs showed even nominal

association ($P < 0.05$) in 18,598 ever-smokers versus 15,041 never-smokers and none of the SNPs showed nominal association with the number of cigarettes smoked per day (n = 15,574). This evidence strongly suggests that the effects of these variants on lung function and COPD risk are not mediated via tobacco addiction.

Study 2: Joint Effects of *HHIP, TNS1, GSTCD, HTR4, AGER,* and *THSD4* Loci on Lung Function and COPD

The characteristics of the study participants are shown in Table 1. In all, 15,883 individuals were included in the analyses of FEV_1 and FEV_1/FVC (Table 1). A trend in lung function (FEV_1 and FEV_1/FVC) was shown across risk allele categories (Figure 4). Notably, compared with the baseline group of seven risk alleles, carrying 10–12 risk alleles was associated with a reduction in FEV_1 (coefficient −72.21 ml [95% CI, −112.12 to −32.30 ml]; $P = 3.90 \times 10^{-4}$) and a reduction in FEV_1/FVC (coefficient −1.53% [95% CI, −2.20% to −0.87%]; $P = 6.35 \times 10^{-6}$). Approximately 5% of the study population carried 10–12 risk alleles and 28% of the study population carried seven risk alleles. The magnitude of decline in FEV_1 compared with the baseline group is equivalent to the physiological average ageing decline in lung function over approximately four years in a non-smoking population (22).

To assess the joint effects of the risk alleles on COPD, data from nondiscovery cohorts (Figure 1) with genotype data available on all six SNPs were used. This included 2,252 COPD case subjects (percent predicted $FEV_1 < 80\%$ and $FEV_1/FVC < 0.7$) and 8,952 control subjects ($FEV_1 > 80\%$ predicted and $FEV_1/FVC > 0.7$) as shown in Figure 1 and Table 1. Again, a clear trend in COPD risk across the categories was observed (Figure 4). Compared with a baseline of seven risk alleles, carrying 10 to 12 risk alleles was associated with an increased risk of COPD (OR, 1.63; 95% CI, 1.31–2.03; $P = 1.46 \times 10^{-5}$) (Figure 4).

DISCUSSION

In a study of the set of novel variants we reported to be associated with lung function (9), we show significant association with COPD for three of the five loci—in *HTR4*, *GSTCD*, and *TNS1*—emphasizing the clinical relevance of these loci to respiratory disease. We also confirm the association between the 4q31 locus near *HHIP* and COPD, and show expected direction and magnitude of effect (although nonsignificant) for the two remaining of the five lung function loci studied. In addition, we provide an estimate of the combined effect sizes of these loci on lung function and COPD in studies independent of the data used to discover these associations. We show that the highest number of risk alleles (10–12 risk alleles, 5% of our population) is associated with a 1.6-fold elevation of COPD risk, compared with a common baseline group of individuals with 7 risk alleles (28% of our population).

The loci associated with COPD may provide important insights into the pathways underlying the development of COPD. The sentinel SNP at the 4q24 locus (rs10516526) is intronic in *GSTCD*, which encodes a glutathione *S*-transferase, C-terminal domain-containing protein. This protein may be involved in cellular detoxification, catalyzing conjugation of glutathione to products of oxidative stress (23), and in regulating the synthesis of prostaglandins and leukotrienes (23). *GSTCD* also shows homology with chloride intracellular channels and could therefore influence lung function via other molecular pathways (9). A second locus associated with COPD was localized to *TNS1* at 2q35, where the sentinel variant was a nonsynonymous coding SNP. Tensin-1 is an actin-binding protein with SH2 (Src homology-2) domains, possibly involved in signal transduction (24) and cell migration (25). At 5q33, the sentinel variant was an intronic SNP (rs3995090) in *HTR4* encoding the 5-hydroxytryptamine receptor-4 (HTR4), which is expressed in neurons and airway epithelial type II cells, where it may

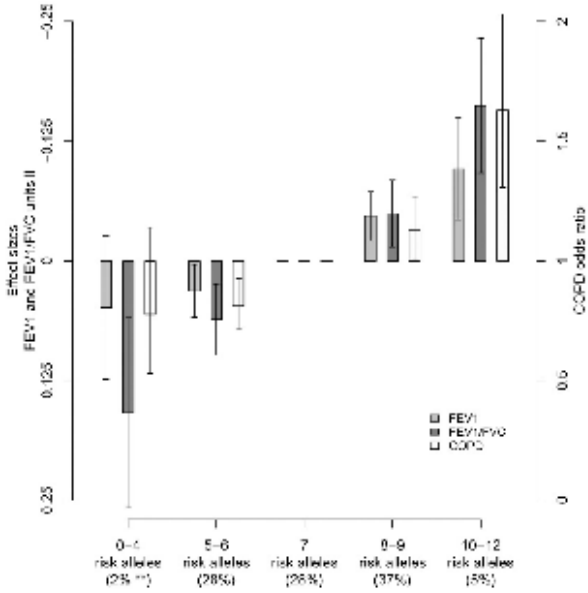


Figure 4. Association of risk scores with lung function and chronic obstructive pulmonary disease (COPD). The risk score theoretically ranges from 0 to 12 across the six loci (*TNS1*, *GSTCD*, *HHIP*, *HTR4*, *AGER*, and *THSD4*). The risk allele category was in each case compared with a baseline of seven risk alleles. *Boxes* indicate the point estimates of the effect sizes and *whiskers* the 95% confidence intervals. ‖To facilitate the plotting of the effect size estimates for FEV_1 and FEV_1/FVC on the same axes, effect sizes are given in terms of the proportion of a standard deviation of FEV_1 and FEV_1/FVC; we used a standard deviation of 752 ml for FEV_1 and 9.45% for FEV_1/FVC (obtained as weighted averages across studies). **Proportion of individuals within each risk score category.

regulate cytokine responses (26). In contrast, the sentinel SNP at 4q31 reported in our study (rs1250468) and that reported in previous studies (rs13147758, $r^2 = 0.97$ with rs1250468) (10, 11), lies in an intergenic region upstream of *HHIP*. The hedgehog (Hh) gene family encodes signaling molecules involved in regulating lung morphogenesis (27), although this locus has been previously associated with height (28), implicating a role in skeletal growth and development.

Although we did not demonstrate a statistically significant association between COPD and two loci, the estimated effects were in the expected direction. These included a nonsynonymous coding SNP (rs2070600) in *AGER*, within a gene-rich region of the major histocompatibility complex (6q21), and an intronic SNP (rs12899618) in *THSD4* (15q23). *AGER* is a strong candidate as it is highly expressed in the lung (29) and altered *AGER* expression has been noted in COPD lung tissue (30) and in subjects with idiopathic pulmonary fibrosis (31). Similarly the gene product of *THSD4* shows homology with the thrombospondin family of extracellular calcium-binding proteins implicated in wound healing, inflammation, and angiogenesis (32) and could also be a good candidate for COPD development. Of the five new loci (9) we examined, the loci showing association with COPD in our data showed strongest association with FEV_1 (*HTR4*, *GSTCD*, and *TNS1*), whereas the two loci that did not show a significant association with COPD (*AGER* and *THSD4*) showed association with FEV_1/FVC, but without a strong association with FEV_1. A simpler explanation for these different findings may be limited statistical power to detect the modest effects of these common genetic variants on a binary outcome.

The effect sizes of the SNPs at the three loci associated with COPD (*HTR4*, *GSTCD*, and *TNS1*) and the locus previously reported (*HHIP*) are modest, in the range 1.10–1.24 per copy of the risk allele at each locus. For the previously reported 4q31 locus near *HHIP*, we estimate an odds ratio of 1.19 (95% CI, 1.12–1.27) for COPD. Although larger effect sizes have been described, for example, an odds ratio for COPD equivalent to approximately 1.4 for the risk allele A at rs13118928 in the U.S. National Emphysema Treatment Trial/Normative Ageing Study and Bergen populations (11); the Rotterdam Study (20) and Framingham Heart Study (10) described odds ratios of 1.25 and 1.10 for this allele. Although the differences in the effect size estimates for the 4q31 locus could be attributable to differences in study characteristics, such as age, these differences could be explained in part by winner's curse bias (19), in which the effect size is generally overestimated in the study that first detects the association at the stringent levels of significance required in GWA studies. Additional evidence from independent populations, such as that provided from our study, can be important in establishing effect size estimates that are likely to be unaffected by such biases.

Most of the constituent studies did not measure postbronchodilation spirometry, and therefore our analysis was limited to prebronchodilation spirometry measures. In the Nottingham Smokers Study, in which both pre- and postbronchodilation spirometry were measured, the positive predictive value of prebronchodilation-defined COPD for diagnosis of postbroncholidation-defined COPD was 98% (Table E2). In all association tests reported in this article, we excluded individuals with GOLD stage 1 COPD from both case subjects and control subjects. As has been shown previously (33), the use of prebroncholidation spirometry to diagnose COPD when GOLD stage 1 COPD case subjects are included would lead to substantial misclassification (e.g., in the Nottingham Smokers Study the positive predictive value would decline to 89%; Table E3). Demonstration of airflow obstruction with postbronchodilator spirometry is required to make a formal diagnosis of COPD. Patients with partly or fully reversible airflow obstruction may have fundamentally different pathological processes contributing to airflow obstruction. One potential bias is that there may be some cases with misclassification of either an asthma diagnosis or the COPD diagnosis. In addition, both asthma and COPD are common diagnoses and may coexist in the same individuals. The associations with lung function that we show were found in the general population and if the SNP effects are specific to asthma and not present in COPD, then the inclusion of patients with asthma could overestimate the contribution of the SNPs to COPD. However, we examined the effects of exclusion of patients with known asthma from the case subjects in a subset of the data with asthma diagnosis available and this did not alter the findings substantially (Figure E2). In addition, within the Nottingham Smokers Study, for which we have in-depth phenotype data and know that individuals with asthma and never-smokers with airflow obstruction have been specifically excluded, we found similar effect estimates. The effect size estimates in Nottingham Smokers were also consistent with effects on COPD for our sentinel SNPs in *HTR4*, *GSTCD*, and *TNS1* (Figure 3).

Various criteria have been proposed for the classification of COPD. Our main analysis was based on GOLD criteria (17). As misclassification is particularly likely to occur when there is poor separation of criteria for case subjects and control subjects, we excluded from our analyses altogether any individuals with mild COPD (stage 1 GOLD criteria: FEV_1/FVC $< 70\%$ but $FEV_1 \geq 80\%$ predicted).

Our conclusions were not materially altered by classifying cases instead on the basis of lower limit of normal equations (18) (Figure E3). Misclassification could also occur as a consequence of spirometry measurement error, and this too would tend to lead to underestimation of the effects of the SNPs studied. An important limitation of our study is that the misclassification and smoking exposure cannot be fully addressed, using these cohorts, because postbronchodilator spirometry and quantification of tobacco smoke exposure is not captured in several of our cohorts. Further studies of patients with COPD with more detailed smoking exposure measurements and more extensive respiratory phenotyping should provide further insight into the precise effects of these SNPs on COPD risk and COPD progression.

We focused on the study of SNPs in six loci: *TNS1*, *GSTCD*, *HHIP*, *HTR4*, *AGER*, and *THSD4*. Further studies will be required to investigate the potential association with COPD of SNPs in the regions of *GPR126*, *ADAM19*, *PTCH*, and *PID1* (additional regions described by the Cohorts for Heart and Aging Research in Genomic Epidemiology [CHARGE] Consortium) (8). The addition of these SNPs, the sentinel SNP from *FAM13A* (associated with COPD and lung function [8, 12]), *CHRNA3/5* (11), and other new loci from ongoing powerful genome-wide association studies to the risk score we constructed would be expected to improve the discrimination of such a score for prediction of COPD. With some exceptions, such as the major histocompatibility region, genome-wide association studies have been successful in localizing association signals (34, 35), but further research is likely to lead to improved resolution of association signals and possibly to detection of multiple independent causal variants at each of these loci. The simple risk score we present here would need to be adapted to incorporate an appropriate weighting as the score incorporates more SNPs, with a greater range of effect sizes. However, for the limited number of SNPs we examined, the inclusion of a weighting (based on the allelic effect size estimate) did not materially alter our findings. We also constructed risk scores in a subset of the data, measuring the effect on COPD *per risk allele* in all individuals and in ever-smokers only; the odds ratios were 1.14 and 1.13, respectively, suggesting that

the effect of these loci on COPD is similar in the general population and in ever-smokers only. As more complete risk scores are constructed, it will be important to investigate their predictive potential in subgroups stratified by smoking status, and to examine whether such a score can improve on risk prediction from conventional risk factors alone (including age, sex, smoking status, history of asthma, and family history, where available).

Our study demonstrates that some loci underlying lung function are associated with COPD, and provides a proof of concept that investigation of genetic determinants of lung function can be a successful strategy to discover molecular pathways underlying COPD. Pathways involving *HTR4*, *GSTCD*, and *TNS1* are strong candidates for potential interventions to prevent or alleviate COPD. We show, in studies unaffected by winner's curse bias, that the highest risk category of a six-SNP risk score is associated with a 1.6-fold risk of COPD compared with a common baseline group of seven risk alleles. It is not yet known whether these SNPs could contribute to a clinically useful strategy for prediction and intervention to prevent COPD, although the high absolute risk of COPD in smokers would suggest that the clinical and public health impact of such an approach is worthy of investigation as more loci are discovered and incorporated into risk scores.

Author Disclosure: I.S. received a sponsored grant from Pfizer. I.P.H. is a board member of the MRC Clinical Training Panel and has received lecture fees from Pfizer. M.A.S., L.W., E.R., M.O., P.B., T.J., J.Z., E.A., A.D., S.K., B.S., G.C., J.H., L.J.P., A.H., S.W., P.W., S.E., G.S., I.B., R.L., N.W., C.C., E.D., S.S., J.L., J.M., S.D., Å.N., A. C.O., S.K., J.H., H.S., T.M.M., I.P., M.H., S.R., I.S., J.B., M.K., J.B., F.N., J.H., D.L., R. M., A.L.J., C.M.J., and M.T. do not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript.

Acknowledgment: The SpiroMeta Consortium acknowledges ENGAGE (Engaging ICT cooperation between Europe and the Asia-Pacific region).

Author Contributions

ADONIX: *Investigators*—Åsa Torinsson Naluai, Fredrik Nyberg, Anna-Carin Olin, Santosh Dahgam; *contributions*—project conception, design, and management: A.-C.O.; phenotype collection and data management: A.-C.O.; genotyping: A.T. N., F.N.; data analysis: A.T.N., F.N., A.-C.O., S.D. **BHS**: *Investigators*—Gemma Cadby, Jennie Hui, Alan L. James, Lyle J. Palmer; *contributions*—project conception, design, and management: A.L.J., L.J.P.; phenotype collection and data management: G.C., A.L.J., L.J.P.; data analysis: G.C., J. Hui, L.J.P.; **BRHS**: *Investigators*—Aroon D. Hingorani, Richard W. Morris, S. Goya Wannamethee, Peter H. Whincup; *contributions*—phenotype collection and data management: R.W.M., S.G.W., P.H.W.; genotyping: A.D.H., R.W.M., P.H.W.; data analysis: R.W.M. **BWHHS**: *Investigators*—George Davey Smith, Shah Ebrahim, Debbie A. Lawlor, Peter H. Whincup; *contributions*—phenotype collection and data management: G.D.S., S. E., D.A.L., P.H.W.; data analysis: D.A.L. **EPIC**: *investigators*—Inês Barroso, Ruth J. F. Loos, Nicholas J. Wareham, Jing Hua Zhao. *contributions*—project conception, design, and management: I.B., R.J.F.L., N.J.W., J.H.Z.; phenotype collection and data management: N.J.W.; genotyping: I.B., R.J.F.L., N.J.W., J.H.Z.; data analysis: R.J.F.L., J.H.Z. **Gedling**: *Investigators*—John R. Britton, Tricia M. McKeever, Ian D. Pavord; *contributions*—phenotype collection and data management: J.R.B., T.M.M., I.D.P.; data analysis: M.O., M.D.T. **GS:SFHS**: *Investigators*—Cathy Jackson, Shona Kerr, Anna Dominiczak, Blair Smith; *contributions*—project conception, design, and management: C.J., A.D., S.K., B.S.; phenotype collection and data management: C.J., B. S., A.D.; genotyping: S.K. **HCS**: *Investigators*—Cyrus Cooper, Elaine Dennison, John W. Holloway, Seif Shaheen; *contributions*—project conception, design, and management: C.C., E.D., J.W.H.; phenotype collection and data management: C.C., E.D., S.S.; genotyping: C.C., E.D., J.W.H.; data analysis: J.W.H., S.S. **Health 2000**: *Investigators*—Markku Heliövaara, Mika Kähönen, Samuli Ripatti, Ida Surakka; *contributions*—project conception, design, and management: M.H., M.K.; phenotype collection and data management: M.H., M.K.; genotyping: S. R., I. Surakka; data analysis: M.K., S.R., I. Surakka. **KORA F4**: *investigators*—Eva Albrecht, Stefan Kaiaasch, Joachim Heinrich, Holger Schulz; *contributions*—project conception, design, and management: S.K., J. Heinrich, H.S.; phenotype collection and data management: S.K., J. Heinrich, H.S.; data analysis: E.A. **Nottingham Smokers**: *investigators*—John D. Blakey, Ian P. Hall, Ma'en Obeidat, Ian Sayers; *contributions*—project conception, design, and management: I.P.H., M.O., I. Sayers; phenotype collection and data management: J.D.B., I.P.H., M.O., I. Sayers; genotyping: I.P.H.; data analysis: M.O., I. Sayers, M.D.T. **NSHD**: *Investigators*—The "NSHD Respiratory Study Team" team members involved were as follows: Zaina Al-Kanaani, Anna Hansell, Rebecca Hardy, Diana Kuh, Andrew Wong; *contributions*—phenotype collection and data management: R.H., D.K., A.W.; genotyping: D.K., A.W.; data analysis: Z.A.-K., A.H., R.H., A.W. **Ox-GSK**: *investigators*—Jason Z. Liu, Jonathan Marchini; *contributions*—project conception, design, and management: J.M.; data analysis: J.Z.L. ***Manuscript Conception, Design, and Management***: I.P.H., M.S.A., M.D.T. ***Data Analysis and Bioinformatics***: P.R.B., I.P.H., T.J., M.O., E.R. I.S., M.S.A., M.D.T., L.V.W. ***Writing the Manuscript***: I.P.H., M.S.A., M.D.T. ***Cohort Funding*** **ADONIX**: The ADONIX Study was funded by the Swedish Research Council for Worklife and Social Research (FAS), grants 2001-0263, 2008-0139, Swedish Heart and Lung Foundation grant 20050561 and a collaborative research grant from AstraZeneca. **BHS**: The Busselton Health Study acknowledges the generous support for the 1994/5 follow-up study from Healthway, Western Australia. The Busselton Health Study is supported by The Great Wine Estates of the Margaret River region of Western Australia. The BHS gratefully acknowledges the assistance of the Western Australian DNA Bank (NHMRC Enabling Facility) with DNA samples and the support provided by the Western Australian Genetic Epidemiology Resource (NHMRC Enabling Facility) for this study. **BRHS**: Blood collection and spirometric measures were performed in 1998–2000 as part of a field study funded by the British Heart Foundation (grant PG/97012). DNA was extracted by Geneservice Limited. A.D.H. has a British Heart Foundation Senior Research Fellowship (FS05/125). **BWHHS**: The British Women's Heart and Health Study (BWHHS) is funded by the U.K. Department of Health Policy Research Program and the British Heart Foundation. A separate British Heart Foundation project grant (PG/06/154/22043) funded cotinine assays. **EPIC**: The EPIC Norfolk Study is funded by Cancer Research UK and the Medical Research Council. I.B. acknowledges funding from the Wellcome Trust (077016/Z/05/Z) and from the U.K. NIHR Cambridge Biomedical Research Centre. **Gedling**: The Nottingham Gedling cohort collection was funded by Asthma UK and the British Lung Foundation. **GS:SFHS**: The Generation Scotland: Scottish Family Health Study is funded by the Chief Scientist Office, part of the Scottish Government Health Directorate (http://www.sehd.scot.nhs.uk/cso), grant reference number CZD/16/6. Genotyping was conducted by the Genetics Core at the Wellcome Trust Clinical Research Facility, University of Edinburgh, Western General Hospital, Edinburgh, UK. **HCS**: The Hertfordshire Cohort Study DNA collection was funded by the Medical Research Council, Arthritis Research Campaign, and International Osteoporosis Foundation. **Health 2000**: This study was financially supported by the Medical Research Fund of the Tampere University Hospital. **KORA F4**: The KORA Augsburg studies were financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany and supported by grants from the German Federal Ministry of Education and Research (BMBF) in the context of the German National Genome Research Network (NGFN-2 and NGFN-plus). Our research was supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ. **Nottingham Smokers**: The Nottingham Smokers cohort collection was funded by the University of Nottingham. **NSHD**: The Medical Research Council National Survey of Health and Development (NSHD, also known as the British 1946 Birth Cohort) is funded by the Medical Research Council. The MRC Unit for Lifelong Health and Ageing is responsible for NSHD. Z.A.-K., R.H., D.K., and A.W. are all affiliated with the MRC National Survey of Health and Development, MRC Unit for Lifelong Health and Ageing, London, UK except A.H., who is affiliated with the MRC-HPA Centre for Environment and Health, Imperial College London. Z.A.-K. has a studentship funded by Department of Health Air Pollution PRP Grant Ref. No. 0020029. **Ox-GSK**: GlaxoSmithKline (GSK), a pharmaceuticals company that is interested in developing therapies for lung disease and new cessation therapies for smoking, funded a postdoctoral fellowship for J.Z.L. at Oxford University. GSK also funded the collection, characterization, and, in some cases, the genotyping and genotype data preparation for several of the cohorts used in this study. A. Roses and P. Matthews played crucial roles in establishing and funding the Medical Genetics activities at GSK.

References

1. World Health Organization. Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach, 2007. Available from: http://www.who.int/gard/publications/GARD_Manual/en/index.html
2. Lopez AD, Shibuya K, Rao C, Mathers CD, Hansell AL, Held LS, Schmid V, Buist S. Chronic obstructive pulmonary disease: current burden and future projections. *Eur Respir J* 2006;27:397–412.
3. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med* 2006;3:e442.
4. Weiss ST. Lung function and airway diseases. *Nat Genet* 2010;42:14–16.
5. McCarthy MI. Exploring the unknown: assumptions about allelic architecture and strategies for susceptibility variant discovery. *Genome Med* 2009;1:66.
6. Talmud PJ, Hingorani AD, Cooper JA, Marmot MG, Brunner EJ, Kumari M, Kivimaki M, Humphries SE. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *BMJ* 2010;340:b4838.
7. Cosio BG, Agusti A. Update in chronic obstructive pulmonary disease 2009. *Am J Respir Crit Care Med* 2010;181:655–660.
8. Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marciante KD, Franceschini N, van Durme YM, Chen TH, Barr RG, *et al.* Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet* 2010;42:45–52.
9. Repapi E, Sayers I, Wain LV, Burton PR, Johnson T, Obeidat M, Zhao JH, Ramasamy A, Zhai G, Vitart V, *et al.* Genome-wide association

study identifies five loci associated with lung function. *Nat Genet* 2010; 42:36–44.
10. Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, Myers RH, Borecki IB, Silverman EK, Weiss ST, *et al.* A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet* 2009;5:e1000429.
11. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, Feng S, Hersh CP, Bakke P, Gulsvik A, *et al.* A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 2009;5:e1000421.
12. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, DeMeo DL, Hunninghake GM, Litonjua AA, Sparrow D, *et al.* Variants in *FAM13A* are associated with chronic obstructive pulmonary disease. *Nat Genet* 2010;42:200–202.
13. Soler Artigas M, Wain L.V, Repapi E, Obeidat M, Sayers I, Hall I.P, Tobin M.D; SpiroMeta Consortium. Five new loci associated with lung function and their joint effect on lung function and COPD risk [abstract]. Presented at the European Mathematical Genetics Meeting, Oxford, 2010
14. Soler Artigas M, Wain L.V, Obeidat M, Repapi E, Sayers I, Hall I.P, Tobin M.D; SpiroMeta Consortium. New loci associated with lung function and chronic obstructive pulmonary disease [abstract]. Presented at the International Genetic Epidemiology Society Meeting, Boston, 2010
15. Hankinson JL, Crapo RO, Jensen RL. Spirometric reference values for the 6-s FVC maneuver. *Chest* 2003;124:1805–1811.
16. Hankinson JL, Kawut SM, Shahar E, Smith LJ, Stukovsky KH, Barr RG. Performance of American Thoracic Society–recommended spirometry reference values in a multiethnic sample of adults: the Multi-ethnic Study of Atherosclerosis (MESA) Lung Study. *Chest* 2010;137:138–145.
17. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for the diagnosis, management and prevention of COPD, 2006. Available from http://www.goldCOPD.org
18. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med* 1999;159:179–187.
19. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003; 33:177–182.
20. Van Durme YM, Eijgelsheim M, Joos GF, Hofman A, Uitterlinden AG, Brusselle GG, Stricker BH. Hedgehog-interacting protein is a COPD susceptibility gene: the Rotterdam Study. *Eur Respir J* 2010;36:89–95.
21. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, Berrettini W, Knouff CW, Yuan X, Waeber G, *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010;42:436–440.
22. Kohansal R, Martinez-Camblor P, Agusti A, Buist AS, Mannino DM, Soriano JB. The natural history of chronic airflow obstruction revisited: an analysis of the Framingham offspring cohort. *Am J Respir Crit Care Med* 2009;180:3–10.
23. Hayes JD, Flanagan JU, Jowsey IR. Glutathione transferases. *Annu Rev Pharmacol Toxicol* 2005;45:51–88.
24. Weigt C, Gaertner A, Wegner A, Korte H, Meyer HE. Occurrence of an actin-inserting domain in tensin. *J Mol Biol* 1992;227:593–595.
25. Chen H, Duncan IC, Bozorgchami H, Lo SH. Tensin1 and a previously undocumented family member, tensin2, positively regulate cell migration. *Proc Natl Acad Sci USA* 2002;99:733–738.
26. Bayer H, Muller T, Myrtek D, Sorichter S, Ziegenhagen M, Norgauer J, Zissel G, Idzko M. Serotoninergic receptors on human airway epithelial cells. *Am J Respir Cell Mol Biol* 2007;36:85–93.
27. Miller L-AD, Wert SE, Clark JC, Xu Y, Perl A-KT, Whitsett JA. Role of Sonic hedgehog in patterning of tracheal–bronchial cartilage and the peripheral lung. *Dev Dyn* 2004;231:57–71.
28. Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, Inouye M, Freathy RM, Attwood AP, Beckmann JS, *et al.* Common variants near *MC4R* are associated with fat mass, weight and risk of obesity. *Nat Genet* 2008;40:768–775.
29. Fehrenbach H, Kasper M, Tschernig T, Shearman MS, Schuh D, Muller M. Receptor for advanced glycation endproducts (RAGE) exhibits highly differential cellular and subcellular localisation in rat and human lung. *Cell Mol Biol* 1998;44:1147–1157.
30. Gaens KH, Ferreira I, van der Kallen CJ, van Greevenbroek MM, Blaak EE, Feskens EJ, Dekker JM, Nijpels G, Heine RJ, 't Hart LM, *et al.* Association of polymorphism in the receptor for advanced glycation end products (RAGE) gene with circulating RAGE levels. *J Clin Endocrinol Metab* 2009;94:5174–5180.
31. Konishi K, Gibson KF, Lindell KO, Richards TJ, Zhang Y, Dhir R, Bisceglia M, Gilbert S, Yousem SA, Song JW, *et al.* Gene expression profiles of acute exacerbations of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2009;180:167–175.
32. Chen H, Herndon ME, Lawler J. The cell biology of thrombospondin-1. *Matrix biology: journal of the International Society for Matrix Biology* 2000;19:597–614.
33. Johannessen A, Omenaas ER, Bakke PS, Gulsvik A. Implications of reversibility testing on prevalence and risk factors for chronic obstructive pulmonary disease: a community study. *Thorax* 2005;60:842–847.
34. Anderson CA, Soranzo N, Zeggini E, Barrett JC. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol* 2011;9:e1000580.
35. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010;467:832–838.

Soler Artigas, M., et al., Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet, 2011. 43(11): p. 1082-90.

Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function

Pulmonary function measures reflect respiratory health and are used in the diagnosis of chronic obstructive pulmonary disease. We tested genome-wide association with forced expiratory volume in 1 second and the ratio of forced expiratory volume in 1 second to forced vital capacity in 48,201 individuals of European ancestry with follow up of the top associations in up to an additional 46,411 individuals. We identified new regions showing association (combined $P < 5 \times 10^{-8}$) with pulmonary function in or near *MFAP2, TGFB2, HDAC4, RARB, MECOM* (also known as *EVI1*), *SPATA9, ARMC2, NCR3, ZKSCAN3, CDC123, C10orf11, LRP1, CCDC38, MMP15, CFDP1* and *KCNE2*. Identification of these 16 new loci may provide insight into the molecular mechanisms regulating pulmonary function and into molecular targets for future therapy to alleviate reduced lung function.

© 2011 Nature America, Inc. All rights reserved.

Pulmonary function, reliably measurable by spirometry, is a heritable trait reflecting the physiological state of the airways and lungs[1]. Pulmonary function measures are important predictors of population morbidity and mortality[2–4] and are used in the diagnosis of chronic obstructive pulmonary disease (COPD), which ranks among the leading causes of death in developed and developing countries[5,6]. A reduced ratio of forced expiratory volume in 1 second (FEV_1) to forced vital capacity (FVC) is used to define airway obstruction, and a reduced FEV_1 is used to grade the severity of airway obstruction[7].

Recently, two large genome-wide association studies (GWAS), each comprising discovery sets of more than 20,000 individuals of European ancestry, identified new loci for lung function[8,9]. Recognizing the need for larger data sets to increase the power to detect loci of individually modest effect size, we conducted a meta-analysis of 23 lung function GWAS comprising a total of 48,201 individuals of European ancestry (stage 1) and followed up potentially new loci in 17 further studies comprising up to 46,411 individuals (stage 2). We identified 16 additional new loci for lung function and provided evidence corroborating the association of loci previously associated with lung function[8–11]. Our findings implicate a number of different mechanisms underlying regulation of lung function and highlight loci shared with complex traits and diseases, including height, lung cancer and myocardial infarction.

RESULTS

Genome-wide analysis (stage 1)

We undertook meta-analyses for cross-sectional lung function measures for approximately 2.5 million genotyped or imputed SNPs across 23 studies with a combined sample size of 48,201 adult individuals of European ancestry. Characteristics of the cohort participants and the genotyping are shown in Supplementary Table 1a and b. We adjusted FEV_1 and FEV_1/FVC measures for ancestry principal components, age, age^2, sex and height as covariates. Our association testing of the inverse-normal–transformed residuals for FEV_1 and FEV_1/FVC assumed an additive genetic model and was stratified by ever-smoking versus never-smoking status. We performed the meta-analyses of the smoking strata within each study and of the study-specific results using inverse-variance weighting (and used the inverse of the standard error squared as the weight). We applied genomic control twice at the study level (to each smoking stratum separately and to the study-level pooled estimates) and also at the meta-analysis level to avoid inflation of the test statistics caused by cryptic population structure or relatedness (see Supplementary Table 1a for study-level estimates). Our application of genomic control at the three stages is likely to be overly conservative because it has recently been shown that in large meta-analyses, test statistics are expected to be elevated under polygenic inheritance even when there is no population structure[12]. The test statistic inflation (λ_{GC}) before applying genomic control at the meta-analysis level was 1.12 for FEV_1 and 1.09 for FEV_1/FVC. Genomic inflation estimates increase with sample size, as has been shown for other traits[13–15]; the standardized estimates to a sample of 1,000 individuals ($\lambda_{GC\_1,000}$) were 1.002 for FEV_1 and 1.002 for FEV_1/FVC. Plots of the meta-analysis P values for FEV_1 and FEV_1/FVC against a uniform distribution of P values expected under the null hypothesis showed deviations which were attenuated, but which persisted, after removal of SNPs in loci reported previously, consistent with additional loci being associated with lung function (Supplementary Fig. 1a).

Follow-up analysis (stage 2)

Twenty-nine new loci showing evidence of association with lung function ($P < 3 \times 10^{-6}$) in stage 1 were followed up in stage 2 by using *in silico* data from seven studies and by undertaking additional genotyping in ten studies for the ten highest ranked SNPs (Fig. 1). Full details of the SNP selection are given in the Online Methods. We performed an inverse-variance–weighting meta-analysis across stages 1 and 2 and obtained two-sided P values for the pooled estimates. Sixteen new loci reached genome-wide significance ($P < 5 \times 10^{-8}$) and showed consistent direction of effects in both stages, comprising 12 new loci for FEV_1/FVC, 3 new loci for FEV_1 and 1 new locus reaching

A full list of author affiliations appears at the end of the paper.

Received 20 April; accepted 19 August; published online 25 September 2011; doi:10.1038/ng.941

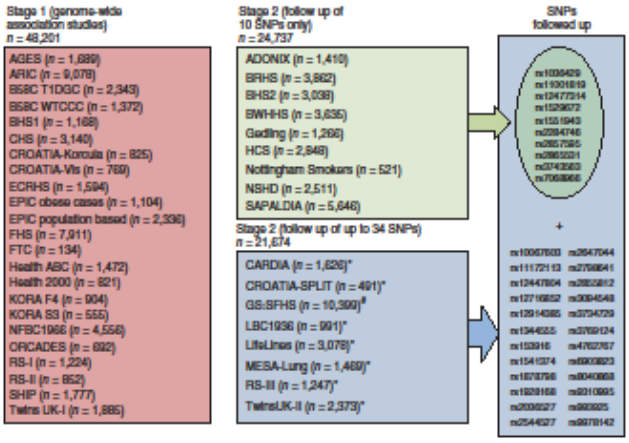


Figure 1 Study design. We followed up in stage 2 a total of 34 SNPs showing new evidence of association ($P < 3 \times 10^{-6}$) with FEV_1 and/or FEV_1/FVC in a meta-analysis of the stage 1 studies. Studies with a combined total of 24,737 individuals undertook genotyping and association testing of the top ten SNPs. Seven studies (marked with an asterisk) with a combined total of 11,275 individuals had genome-wide association data and provided results for up to 34 SNPs. Researchers from GS: SFHS (marked with #) undertook genotyping on a 32-SNP multiplex genotyping platform and so included the 32 top ranking SNPs (including proxies and both SNPs from regions that showed association with both FEV_1 and FEV_1/FVC). This assay failed for one SNP (rs3769124), which was subsequently replaced with the thirty-third SNP (rs4762767). We excluded rs2284746 because of poor clustering. Although rs3743563 was chosen as proxy for rs12447804, which had an effective $N < 80\%$ in the stage 1 meta-analysis, researchers from BHS2 were unable to genotype rs3743563 and so undertook genotyping for rs12447804 instead. See **Table 1** for definitions of all study abbreviations.

genome-wide significance for both traits (**Fig. 2** and **Table 1**). To assess the heterogeneity across the studies included in stage 1 and 2, we performed χ^2 tests for all 16 SNPs, and none of these SNPs was statistically significant after applying a Bonferroni correction for 16 tests. The sentinel SNPs at these loci were in or near *MFAP2* (1p36.13), *TGFB2-LYPLAL1* (1q41), *HDAC4-FLJ43879* (2q37.3), *RARB* (3p24.2), *MECOM* (also known as *EVI1*) (3q26.2), *SPATA9-RHOBTB3* (5q15), *ARMC2* (6q21), *NCR3-AIF1* (6p21.33), *ZKSCAN3* (6p22.1), *CDC123* (10p13), *C10orf11* (10q22.3), *LRP1* (12q13.3), *CCDC38* (12q22), *MMP15* (16q13), *CFDP1* (16q23.1) and *KCNE2-LINC00310* (also known as *C21orf82*) (21q22.11) (**Supplementary Fig. 1b,c**). The strongest signals in *AGER* (rs2070600)[8,9] and two of the new signals (rs6903823 in *ZKSCAN3* and rs2857595, upstream of *NCR3*) lie within a ~3.8-Mb interval at 6p21.32-22.1 that is characterized by long-range linkage disequilibrium (LD). Nevertheless, the leading SNPs in these regions, which are within the major histocompatibility complex (MHC), were statistically independent (**Supplementary Note**).

Gene expression

We investigated mRNA expression of the nearest gene for each of the 16 new loci in human lung tissue and a range of human primary cells including lung, brain, airway smooth muscle cells and bronchial epithelial cells. We detected transcripts for all the selected genes in lung tissue except *CCDC38*, and we also detected transcripts for most genes in airway smooth muscle cells and in bronchial epithelial cells (**Table 2**). As we were unable to detect expression of *CCDC38* in any tissue, we also examined expression of *SNPRF*, which is the gene adjacent to *CCDC38* (**Table 2**), and found its expression in all four cell types. *TGFB2*, *MFAP2*, *EVI1* and *MMP15* were expressed in one or more lung cell types but not in peripheral blood mononuclear cells, providing evidence that these genes may show tissue-specific expression.

We assessed whether SNPs in these new regions or their proxies ($r^2 > 0.6$) were associated with gene expression using a database of expression-associated SNPs in lymphoblastoid cell lines[16]. Four loci showed regional (*cis*) effects on expression ($P < 1 \times 10^{-7}$; **Supplementary Note**). A proxy for our sentinel SNP in *CFDP1*, rs2865531, coincided with the peak of the expression signal for *CFDP1*, and the strongest proxy for rs6903823 in *ZKSCAN3* coincided with the peak of expression for *ZSCAN12*.

Plausible pathways for lung function involving new loci

The putative function of the genes within, or closest to, the association peaks identify a range of plausible mechanisms for affecting lung function. The most statistically significant new signal for FEV_1/FVC ($P = 7.5 \times 10^{-16}$) was in the gene encoding MFAP2, an antigen of elastin-associated microfibrils[17], although correlated SNPs in the region potentially implicate other genes that could plausibly influence lung function, such as *CROCC*, which encodes rootletin, a component of cilia[18]. Our second strongest new signal, also for FEV_1/FVC, was in *RARB*, the gene encoding the retinoic acid receptor β. *Rarb*-null knockout mice have premature alveolar septation[19]. The third most statistically significant new signal for FEV_1/FVC, and the most statistically significant new signal for FEV_1, was in *CDC123*.

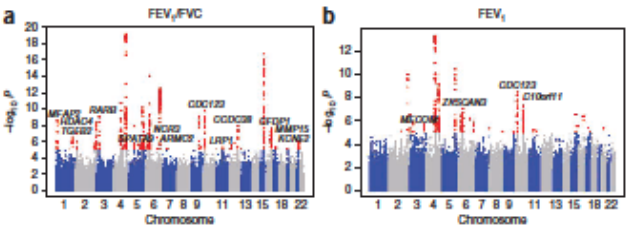


Figure 2 Manhattan plots of association results for FEV_1/FVC and FEV_1 (analysis stage 1). The Manhattan plots for FEV_1/FVC (**a**) and FEV_1 (**b**) are ordered by chromosome position. SNPs for which $-\log_{10} P > 5$ are indicated in red. Newly associated regions that reached genome-wide significance after meta-analysis of stages 1 and 2 are labeled.

© 2011 Nature America, Inc. All rights reserved.

Table 1 Loci associated with lung function

						Stage 1				Stage 2				Joint meta-analysis of all stages	
SNP ID	Chr.	NCBI36 position	Nearest gene	Coded allele	Measure	β (s.e.m.)	P	Coded allele freq.	N	β (s.e.m.)	P	Coded allele freq.	N	β (s.e.m.)	P
rs2284746	1	17,179,262	*MFAP2* (intron)	G	FEV_1/FVC	−0.042 (0.007)	2.47×10^{-9}	0.516	45,944	−0.038 (0.007)	**2.64×10^{-7}**	0.522	35,371	−0.04 (0.005)	**7.50×10^{-16}**
					FEV_1	0.008 (0.007)	2.78×10^{-1}			0.006 (0.007)	3.70×10^{-1}			0.007 (0.005)	1.48×10^{-1}
rs993925	1	216,926,691	*TGFB2* (downstream)	T	FEV_1/FVC	0.040 (0.007)	2.54×10^{-7}	0.308	42,402	0.023 (0.01)	1.76×10^{-2}	0.348	21,414	0.034 (0.006)	**1.16×10^{-8}**
					FEV_1	0.025 (0.007)	1.51×10^{-3}			0.003 (0.007)	7.29×10^{-1}			0.014 (0.005)	8.71×10^{-3}
rs12477314	2	239,542,085	*HDAC4* (downstream)	T	FEV_1/FVC	0.052 (0.008)	4.48×10^{-9}	0.202	45,585	0.031 (0.008)	**8.41×10^{-5}**	0.206	45,821	0.041 (0.006)	**1.68×10^{-12}**
					FEV_1	0.032 (0.008)	2.77×10^{-4}			0.025 (0.007)	**1.82×10^{-4}**			0.028 (0.005)	1.02×10^{-7}
rs1529672	3	25,495,586	*RARB* (intron)	C	FEV_1/FVC	−0.060 (0.009)	7.75×10^{-10}	0.829	40,624	−0.038 (0.009)	**1.16×10^{-5}**	0.831	45,466	−0.048 (0.006)	**3.97×10^{-14}**
					FEV_1	−0.037 (0.009)	1.78×10^{-4}			−0.011 (0.007)	9.33×10^{-2}			−0.020 (0.006)	2.16×10^{-4}
rs1344555	3	170,782,913	*MECOM* (intron)	T	FEV_1/FVC	−0.019 (0.008)	2.61×10^{-2}	0.205	46,067	−0.017 (0.012)	1.55×10^{-1}	0.209	21,313	−0.018 (0.007)	6.65×10^{-3}
					FEV_1	−0.042 (0.008)	1.91×10^{-6}			−0.025 (0.009)	6.44×10^{-3}			−0.034 (0.006)	**2.65×10^{-8}**
rs153916	5	95,062,456	*SPATA9* (upstream)	T	FEV_1/FVC	−0.033 (0.007)	2.06×10^{-6}	0.552	47,530	−0.025 (0.009)	6.67×10^{-3}	0.535	21,647	−0.031 (0.005)	**2.12×10^{-8}**
					FEV_1	−0.001 (0.007)	8.91×10^{-1}			0.004 (0.007)	6.22×10^{-1}			0.001 (0.005)	8.20×10^{-1}
rs6903823	6	28,430,275	*ZKSCAN3* (intron)/ *ZNF323* (intron)	G	FEV_1/FVC	−0.027 (0.008)	2.28×10^{-3}	0.209	47,057	−0.013 (0.011)	2.34×10^{-1}	0.246	21,489	−0.021 (0.007)	1.19×10^{-3}
					FEV_1	−0.046 (0.008)	2.00×10^{-7}			−0.029 (0.008)	**4.75×10^{-4}**			−0.037 (0.006)	**2.18×10^{-10}**
rs2857595	6	31,676,448	*NCR3* (upstream)	G	FEV_1/FVC	0.049 (0.009)	7.86×10^{-8}	0.809	45,540	0.028 (0.008)	**5.36×10^{-4}**	0.796	46,107	0.037 (0.006)	**2.28×10^{-10}**
					FEV_1	0.040 (0.009)	1.46×10^{-5}			0.017 (0.007)	9.41×10^{-3}			0.025 (0.005)	1.30×10^{-6}
rs2798641	6	109,374,743	*ARMC2* (intron)	T	FEV_1/FVC	−0.047 (0.009)	2.81×10^{-7}	0.183	46,369	−0.030 (0.012)	1.57×10^{-2}	0.179	21,173	−0.041 (0.007)	**8.35×10^{-9}**
					FEV_1	−0.046 (0.009)	5.39×10^{-7}			−0.009 (0.01)	3.35×10^{-1}			−0.030 (0.006)	4.69×10^{-6}
rs7068966	10	12,317,998	*CDC123* (intron)	T	FEV_1/FVC	0.045 (0.007)	1.28×10^{-10}	0.519	47,085	0.023 (0.006)	**3.86×10^{-4}**	0.518	46,067	0.033 (0.005)	**6.13×10^{-13}**
					FEV_1	0.040 (0.007)	1.19×10^{-8}			0.022 (0.005)	**3.56×10^{-5}**			0.029 (0.004)	**2.82×10^{-12}**
rs11001819	10	77,985,230	*C10orf11* (intron)	G	FEV_1/FVC	−0.019 (0.007)	6.50×10^{-3}	0.522	45,546	−0.006 (0.006)	3.17×10^{-1}	0.506	45,932	−0.012 (0.005)	7.58×10^{-3}
					FEV_1	−0.041 (0.007)	1.42×10^{-8}			−0.022 (0.005)	**3.10×10^{-5}**			−0.029 (0.004)	**2.98×10^{-12}**
rs11172113	12	55,813,550	*LRP1* (intron)	T	FEV_1/FVC	−0.035 (0.007)	1.36×10^{-6}	0.607	45,387	−0.026 (0.01)	5.83×10^{-3}	0.590	20,509	−0.032 (0.006)	**1.24×10^{-8}**
					FEV_1	−0.021 (0.007)	3.55×10^{-3}			−0.003 (0.007)	6.94×10^{-1}			−0.013 (0.005)	1.19×10^{-2}
rs1036429	12	94,795,559	*CCDC38* (intron)	T	FEV_1/FVC	0.049 (0.008)	1.24×10^{-8}	0.200	47,814	0.028 (0.008)	**3.35×10^{-4}**	0.214	46,311	0.038 (0.006)	**2.30×10^{-11}**
					FEV_1	0.010 (0.008)	2.67×10^{-1}			0.004 (0.006)	5.38×10^{-1}			0.006 (0.005)	2.26×10^{-1}
rs12447804	16	56,632,783	*MMP15* (intron)	T	FEV_1/FVC	−0.053 (0.009)	7.12×10^{-8}	0.208	35,123	−0.021 (0.01)	4.20×10^{-2}	0.222	24,398	−0.038 (0.007)	**3.59×10^{-8}**
					FEV_1	−0.017 (0.009)	8.02×10^{-2}			0.004 (0.007)	5.71×10^{-1}			−0.004 (0.006)	4.73×10^{-1}
rs2865531	16	73,947,817	*CFDP1* (intron)	T	FEV_1/FVC	0.039 (0.007)	2.30×10^{-8}	0.418	47,594	0.024 (0.006)	**1.94×10^{-4}**	0.409	46,304	0.031 (0.005)	**1.77×10^{-11}**
					FEV_1	0.024 (0.007)	6.30×10^{-4}			0.011 (0.005)	3.89×10^{-2}			0.016 (0.004)	1.09×10^{-4}
rs9978142	21	34,574,109	*KCNE2* (upstream)	T	FEV_1/FVC	−0.048 (0.009)	8.23×10^{-7}	0.156	44,577	−0.031 (0.013)	1.75×10^{-2}	0.149	20,944	−0.043 (0.008)	**2.65×10^{-8}**
					FEV_1	−0.012 (0.009)	2.47×10^{-1}			−0.015 (0.01)	1.35×10^{-1}			−0.013 (0.007)	5.57×10^{-2}

Shown are FEV_1 and FEV_1/FVC results for the leading SNPs, ordered by chromosome and position for each independent locus associated ($P < 5 \times 10^{-8}$) with FEV_1 or FEV_1/FVC in a joint analysis of up to 94,612 individuals of European ancestry from the SpiroMeta-CHARGE GWAS (stage 1) and follow up (stage 2). Two-sided *P* values are given for stage 1, stage 2 and the joint meta-analysis of all stages. *P* values reaching genome-wide significance ($P < 5 \times 10^{-8}$) in the joint meta-analysis of all stages are indicated in bold. SNPs reaching independent replication in stage 2 ($P = 0.05/34 = 1.47 \times 10^{-3}$) are indicated with their stage 2 *P* value in bold. The sample sizes (*N*) shown are the effective sample sizes. The effective sample size within each study is the product of sample size and the imputation quality metric. The joint meta-analysis includes data from stage 1 and stage 2. β values reflect effect-size estimates on an inverse-normal transformed scale after adjustments for age, age^2, sex, height and ancestry principal components. The estimated proportion of the variance explained by each SNP can be found in **Supplementary Table 6**. Chr., chromosome; freq., frequency.

© 2011 Nature America, Inc. All rights reserved.

This was the only new region to show genome-wide association with both traits. *CDC123* encodes a homolog of a yeast cell-division–cycle protein that plays a critical role in modulating eukaryotic initiation factor 2 in times of cell stress[20]. The fourth signal for FEV_1/FVC is downstream of *HDAC4*, which encodes a histone deacetylase; reductions in the expression of other histone deacetylases (specifically HDAC2, HDAC5 and HDAC8) have been noted in COPD[21]. The regions we observed in the MHC are much more difficult to localize, with multiple genes being tagged by the top SNP, including non-synonymous SNPs in *ZKSCAN3*, *PGBD1*, *ZSCAN12*, *ZNF323*, *TCF19*, *LTA*, *C6orf15* and *GPANK1* (also known as *BAT4*) (**Supplementary Table 2**). At 6p21.33, we observed the strongest association with lung function for rs2857595, which is in LD ($r^2 = 0.47$) with a non-synonymous SNP in *LTA* (encoding lymphotoxin α) and with a SNP in the upstream promoter region of *TNFA* (encoding tumor necrosis factor α) ($r^2 = 0.86$), both of which are plausible candidates[22,23]. Our top SNP in *MMP15* is in strong LD ($r^2 = 1$) with a non-synonymous SNP (rs3743563, which has an association with FEV_1/FVC at $P = 1.8 \times 10^{-7}$) within the same gene. Among the plausible mechanisms implicated by the other new signals of association with lung function reported here is TGF-β signaling; *TGFB2* expression is upregulated in bronchial epithelial cells in asthma[24]. The putative function of key genes (as defined by LD with the leading SNP) in each of the 16 loci, and relevant findings from animal models, are summarized in **Table 2** and are detailed in **Supplementary Table 2**.

Associations with lung function in children

Alleles representing 11 of the 16 new loci showed directionally consistent effects on lung function in 6,281 children (7–9 years of age) (**Supplementary Table 3a**), suggesting that genetic determination of lung function in adults may in part act through effects on lung development, or alternatively, that some genetic determinants of lung growth and lung function decline are shared.

Association of lung function loci with other traits

Although we stratified for ever smoking versus never smoking, we did not adjust for the amount smoked. In order to investigate the possibility that the associations at any of our 16 new regions were driven by an effect of the SNP on smoking behavior, we evaluated *in silico* data for associations with smoking amount from the Oxford-GlaxoSmithKline (Ox-GSK) consortium[25] for the leading SNPs in these 16 regions. None of these 16 SNPs showed statistically significant association with the number of cigarettes smoked per day (**Supplementary Table 3b**).

Table 2 Expression profiling of candidate genes in the lung and periphery

Sentinel SNP (relationship to gene)	Chr.	Gene	Putative function of encoded protein	Tissue			
				Lung	HASM	HBEC	PBMC
rs993925 (intron)	1	*TGFB2*	Cytokine with roles in pro-fibrotic cytokine modulating epithelial repair mechanisms and extracellular matrix homeostasis including collagen deposition[40].	+	+	–	–
rs2284746 (intron)	1	*MFAP2*	Major antigen of elastin-associated microfibrils[17] and a candidate for involvement in the etiology of inherited connective tissue diseases.	+	+	+	–
rs12477314 (downstream)	2	*HDAC4*	Deacetylase of histone surrounding DNA thus influencing transcription factor access to the DNA and possibly repressing gene transcription.	+	+	+	+
rs1344555 (intron)	3	*EVI1*	Zinc finger transcription factor, encoded as part of *MECOM (MDS1-EVI1* complex locus).	+	+	+	–
rs1529672 (intron)	3	*RARB*	Nuclear retinoic acid receptor responsive to retinoic acid, a vitamin A derivative and which also controls cell proliferation and differentiation.	+	+	+	+
rs153916 (intron)	5	*SPATA9*	Initially identified as a mediator of spermatogenesis, other family members may have a role in pancreatic development and β-cell proliferation[41].	+	+	+	+
rs2798641 (intron)	6	*ARMC2*	Function unknown, although other family members have been identified as having roles in cell signaling, protein degradation and cytoskeleton functions[42].	+	+	+	+
rs2857595 (upstream)	6	*NCR3*	Required for efficient cytotoxicity responses by natural killer cells against normal cells and tumors[43].	+	–	–	+
rs6903823 (intron)	6	*ZKSCAN3*	Transcription factor involved in cell growth, cell cycle and signal transduction.	+	+	+	+
rs7068966 (intron)	10	*CDC123*	Homolog in yeast shown to be a critical control protein modulating eukaryotic initiation factor 2 in times of cell stress.	+	+	+	+
rs11001819 (intron)	10	*C10orf11*	Function unknown.	+	+	+	+
rs11172113 (intron)	12	*LRP1*	Potentially diverse roles including cell signaling and migration[44].	+	+	+	+
rs1036429 (intron)	12	*CCDC38*	Function unknown, although other family members involved in a diverse array of functions skeletal and motor function[45].	–	–	–	–
rs1036429 ($r^2 = 0.96$ with rs4762633 in *SNRPF*)	12	*SNRPF*	Small nuclear ribonucleoprotein F.	+	+	+	+
rs12447804 (intron)	16	*MMP15*	Member of a large protease family with diverse functional roles via protease activity and specificity including tissue remodeling, wound healing, angiogenesis and tumor invasion.	+	+	+	–
rs2865531 (intron)	16	*CFDP1*	Craniofacial development protein 1.	+	+	+	+
rs9978142 (upstream)	21	*KCNE2*	KCNQ1-KCNE2 K+ channels may modulate transepithelial anion secretion in Calu3 airway epithelial cells[46].	+	–	–	+
Reference gene	12	*GAPDH*		+	+	+	+

+ indicates the gene is expressed in the cell type used, and – indicates that we did not detect the gene expression at the mRNA level following 40 cycles of PCR. PCR profiling of gene transcripts in the human lung showed expression of all candidates except *CCDC38*, for which two sets of primers were designed and tested under different optimization conditions. None of these assays detected expression of *CCDC38* in the cell types analyzed. We instead assayed *SNRPF*, which neighbors *CCDC38* and harbors SNPs in strong LD with *CCDC38*'s sentinel SNP. All PCR products were sequence verified. We used *GAPDH* (encoding glyceraldehyde-3-phosphate dehydrogenase) as a positive control for the complementary DNA, and this gene was expressed in all tissues. Chr., chromosome; HASM, human airway smooth muscle; HBEC, human bronchial epithelial cells; PBMC, peripheral blood mononuclear cells.

© 2011 Nature America, Inc. All rights reserved.

In addition, in our stage 1 and 2 datasets combined, we assessed whether the estimated effect sizes of the variants on lung function phenotypes differed substantially between ever smokers and never smokers (Supplementary Table 4) across the 16 loci. For the most strongly associated trait at each locus, we tested the SNP interaction with ever smoking versus never smoking. None of the 16 new loci showed a significant interaction (Bonferroni-corrected threshold for 16 independent SNPs $P = 0.003125$). These analyses suggest that the genetic effects we have identified underlie lung function variability irrespective of smoking exposure.

We adjusted our lung function associations for height, but there are some overlaps between loci associated with height and those associated with lung function. Therefore, we evaluated *in silico* data for height associations of our new regions in the GIANT consortium[14] dataset. The G allele of rs2284746 (in an intron of *MFAP2*), which was associated with decreased FEV_1/FVC, was associated with increased height (Supplementary Table 3c).

Given reported associations between lung cancer and either COPD or lung function decline, we also assessed *in silico* data for sentinel or proxy SNPs in these 16 regions for associations with lung cancer in the International Lung Cancer Consortium (ILCCO) GWAS meta-analysis[26]. Alleles associated with reduced lung function were associated with risk of lung cancer at the strongest available proxy SNP for rs2857595 (upstream of *NCR3*) at 6p21.33 (rs3099844, $r^2 = 0.67$) and for the strongest proxy SNP for rs6903823 (a SNP in an intron of *ZKSCAN3* and *ZNF323*) at 6p22.1 (rs209181, $r^2 = 0.69$) (lung cancer associations at $P = 2.2 \times 10^{-7}$ and $P = 3.4 \times 10^{-5}$, respectively; Supplementary Table 3d). We saw no significant associations with lung cancer at the other new loci (proxy SNPs were available for 15 of the 16 loci, Bonferroni-corrected $P < 0.0033$).

In addition to the effects on height, smoking and lung cancer described above, we examined the literature for evidence of associations with other traits for each of the 16 new loci (detailed in Supplementary Table 2). Genome-wide significant associations ($P < 5 \times 10^{-8}$) have been reported in *KCNE2* with myocardial infarction[27] and at 6p21.33 near *NCR3-AIF1* with neonatal lupus[28] and systemic lupus erythematosus[29]. Other significant complex disease associations have also been noted in the regions of *CDC123* (type 2 diabetes[30]), *CFDP1* (type 1 diabetes[31]) and *MECOM* (blood pressure[32,33]), but with weaker LD ($r^2 < 0.3$) being seen between the reported SNP and the sentinel SNP for lung function in the region (Supplementary Table 2).

Proportion of variance explained by loci discovered to date

Associations in ten loci previously reported for lung function[8,9] reached genome-wide significance ($P < 5 \times 10^{-8}$) in our stage 1 data, namely loci in or near *TNS1*, *FAM13A*, *GSTCD-NPNT*, *HHIP*, *HTR4*, *ADAM19*, *AGER*, *GPR126*, *PTCH1* and *TSHD4* (Supplementary Table 5a). Thus, a total of 26 regions showed genome-wide significant association with lung function in our study. In aggregate, variants at these 26 regions explain approximately 3.2% of the additive polygenic variance for FEV_1/FVC and 1.5% of the variance for FEV_1 (Supplementary Note). Following the approach previously described[34], we estimated that there are a total of 102 (95% confidence interval 57–155) independent variants with similar effect sizes to the 26 variants we report here. In combination, these 102 variants, comprising 26 discovered variants and 76 putative undiscovered variants, collectively explain around 7.5% of the additive polygenic variance for FEV_1/FVC and 3.4% of the variance for FEV_1 (Online Methods, Supplementary Table 6 and Supplementary Note).

DISCUSSION

In meta-analysis of 23 studies comprising 48,201 individuals of European ancestry and follow up in 17 studies comprising up to 46,411 individuals, we report genome-wide significant associations with an additional 12 regions for FEV_1/FVC, an additional 3 regions for FEV_1 and 1 additional region associated with both FEV_1 and FEV_1/FVC. We also confirmed genome-wide association with ten regions previously associated with lung function, bringing to 26 the total number of loci associated with lung function from analyses of these datasets. Most of the new loci are in regions not previously suspected to have been involved in lung development, the control of pulmonary function or the risk of developing COPD. Elucidating the mechanisms through which these regions influence lung function should lead to a more complete understanding of lung function regulation and the pathogenesis of COPD. Four of the new loci (*MFAP2*, *ZKSCAN3*, near *NCR3* and near *KCNE2*) that we showed to be associated with lung function are also associated with other complex traits and diseases (with $P < 5 \times 10^{-8}$ for the other trait at a SNP having $r^2 > 0.3$ with the top lung function SNP in the region). Understanding the intermediates underlying these pleiotropic effects could also lead to crucial insights into the pathophysiology of lung disease. One potential explanation is that these loci underlie control of the mechanisms regulating the development and resolution of inflammation and subsequent tissue remodeling in a range of tissues.

The effect sizes of the variants in the 26 loci associated with lung function collectively explain a modest proportion of the additive genetic variance in FEV_1/FVC and in FEV_1, even after accounting for putative undetected variants with a similar distribution of effect sizes[34]. Our findings are consistent with those from other common complex traits, where it is thought that many as yet unidentified common and rare sequence variants, and potentially structural variants, could explain the remaining heritability[35]. That our study more than doubled the number of loci known to be associated with lung function underlines the utility of large sample sizes to achieve the power to detect common variants associated with complex traits. Nevertheless, it is likely that additional variants with similar effect sizes remain undiscovered[14]. In addition, our study was not designed to detect rare variants or structural variants associated with lung function. Identification of rare variants associated with lung function could be helpful in narrowing the scope of ongoing functional work to those genes most likely to be causally related to the association signals we detected.

Our study focused on cross-sectional measures of lung function. Adult lung function at a particular time point is influenced by the peak lung function achieved by 25–35 years of age as well as the rate of decline of lung function after that peak[36]. The 26 loci now confirmed to be associated with lung function could affect either pre- or postnatal lung development and growth or decline in lung function during adulthood, or both. We showed consistent directions of estimated effects on lung function between adults and children 7–9 years of age for SNPs at 11 of the 16 new loci and 8 of the 10 previously reported loci (Supplementary Table 3a). The results we show for lung function in children provide some indication that these loci affect lung function development, although studies in larger populations of children would provide greater clarity for SNPs in the new loci. Further investigations will be required in large populations with longitudinal data to delineate the influence of these variants on the rates of development of, and decline in, lung function and on the risk of developing COPD.

Of the sentinel SNPs at the 16 new loci associated with lung function, only rs2284746 (*MFAP2*) was associated with height in the GIANT consortium[14] dataset. The G allele of rs2284746 was associated with both increased height and reduced lung function. A similar relationship between lung function and height was previously reported for the G allele of rs3817928 in *GPR126* (refs. 8,14), which is associated with decreased height but with increased FEV_1/FVC. A further 3 of the 180 loci found to be associated with height[14] showed association (for the 180 loci, we used a Bonferroni-corrected threshold of $P = 2.8 \times 10^{-4}$) with either FEV_1 (*CLIC4* and *BMP6*) or FEV_1/FVC (*PIP4K2B*) (Supplementary Table 3e). In each case, the allele associated with an increase in height was associated with a decrease in lung function. This is not the case for the association of rs1032296 near *HHIP*, which has shown consistent directions of effects on lung function and height[11,14]. However, the strongest SNP associated with height in the *HHIP* region lies within an intron of *HHIP* but shows no association with FEV_1 or FEV_1/FVC. Furthermore, although height is an important predictor of FEV_1, this is not true for its ratio to FVC[37]. These observations argue against the associations with lung function at these loci being simply caused by incomplete adjustment for height.

We stratified by ever- and never-smoker status in our analyses, and in our investigation of amount smoked in the Ox-GSK consortium[25], none of the sentinel SNPs in the 16 new regions showed association with the number of cigarettes smoked per day. Additionally, none of these regions was associated with ever smoking in the Ox-GSK consortium data (Supplementary Table 3b). Thus, the SNP associations with lung function we observed are unlikely to have arisen simply as a consequence of inadequate adjustment for smoking.

We did not observe any interactions with ever smoking for any of the sentinel SNPs in the 16 new regions that exceeded a Bonferroni-corrected significance level (for 16 SNPs). Thus, the effects on lung function of the newly associated variants we identified are apparent in both ever smokers and in never smokers, and the effects of smoking and of these genetic variants may be independent and additive.

In other common complex diseases, follow-up studies that incorporate common genetic risk variants into models to predict disease have not been shown to add substantially to existing risk models, particularly when such models already include family history[38,39]. The same may also prove to be true for the 26 genetic variants described in this paper, as the effect size of any individual variant is small, but further work is required in this area. The major utility of our findings will be in the knowledge they provide about previously unknown pathways underlying lung function. Elucidating the mechanisms that these genes are involved in will lead to improved understanding of the regulation of lung function and potentially to new therapeutic targets for COPD.

URLs. R, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the many colleagues who contributed to collection and phenotypic characterization of the clinical sampling, genotyping and analysis of the data. We especially thank those who kindly agreed to participate in the studies.
Major funding for this work is from the following sources (alphabetical): Academy of Finland (project grants 104781, 120315, 129269, 1114194, Center of Excellence in Complex Disease Genetics (213506 and 129680) and SALVE); Althingi (Icelandic Parliament); Arthritis Research Campaign; Asthma UK; AstraZeneca; AXA Research Fund; Biotechnology and Biological Sciences Research Council (BBSRC) (BB/F019394/1, G20234); British Heart Foundation (PG/97012, PG/06/154/22043, FS05/125); British Lung Foundation; Canadian Institutes of Health Research (Grant ID MOP-82893); Cancer Research United Kingdom;

© 2011 Nature America, Inc. All rights reserved.

Chief Scientist Office, Scottish Government Health Directorate (CZD/16/6); Croatian Institute for Public Health; UK Department of Health; Dutch Kidney Foundation; Erasmus Medical Center and Erasmus University, Rotterdam; Estonian Genome Center, University of Tartu, Estonia (SF0180142s08); EU funding (GABRIEL GRANT Number: 018996, ECRHS II Coordination Number: QLK4-CT-1999-01237); European Commission (DG XII, EURO-BLCS, FP-5 QLG1-CT-2000-01643, FP-6 LSHB-CT-2006-018996 (GABRIEL), FP-6 LSHG-CT-2006-018947 (EUROSPAN), FP-6 GenomEUtwin project QLG2-CT-2002-01254, FP7/2007-2013: HEALTH-F2-2008-201865, GEFOS, HEALTH-F2-2008-35627, TREAT-OA, HEALTH-F4-2007-201413 (ENGAGE)); Finnish Foundation for Cardiovascular Research; Finland Attendant Medical Research Institute (FAMRI); German Asthma and COPD Network (COSYCONET: BMBF grant 01GI0883); German Bundesministerium fuer Forschung und Technology (01 AK 803 A-H, 01 IG 07015 G); German Federal Ministry of Education and Research (BMBF) (03ZIK012, 01ZZ9603, 01ZZ0103 and 01ZZ0403): German National Genome Research Network (NGFN-2 and NGFN-plus); German Ministry of Cultural Affairs; GlaxoSmithKline; Gyllenberg Foundations; Healthway, Western Australia; Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; Healthcare and Bioscience iNet (funded by the East Midlands Development Agency, partially financed by the European Regional Development Fund, delivered by Medilink East Midlands); Higher Education Funding Council for England (HEFCE); Hjartavernd (Icelandic Heart Association); Innsbruck Medical University; Institute for Anthropological Research in Zagreb; International Osteoporosis Foundation; Intramural Research Program of the NIH, National Institute on Aging and National Institute of Environmental Health Sciences; Jalmari and Rauha Ahokas Foundation; Juvenile Diabetes Research Foundation International (JDRF); Lifelong Health and Wellbeing Initiative (G0700704/84698); Medical Research Council UK (G1000861, G0501942, G0902313, G0000934, G0800582, G0500539, G0600705, PrevMetSyn/SALVE, G9901462); Medical Research Fund of the Tampere University Hospital; Ministry of Science, Education and Sport of the Republic of Croatia (108-1080315-0302); Medical Research Council Human Genetics Unit; Medisearch–The Leicester Medical Research Foundation; Munich Center of Health Sciences (MC Health) as part of LMUInnovativ; National Health and Medical Research Council of Australia (Grant ID 403981 and ID 003209); National Human Genome Research Institute (NHGRI) (U01-HG-004729, U01-HG-004402); National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centres (Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London and Cambridge University Hospitals NHS Foundation Trust in partnership with the University of Cambridge); Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) (050-060-810); Netherlands Organization for the Health Research and Development (ZonMw); Netherlands Organization for Scientific Research NOW (175.010.2007.006, 175.010.2005.011, 911-03-012); Northern Netherlands Collaboration of Provinces (SNN); Norwegian University of Science and Technology; Novo Nordisk; Ontario Institute of Cancer Research and Canadian Cancer Society Research Institute (CCSRI 020214); Republic of Croatia Ministry of Science, Education and Sports research grants (108-1080315-0302); Research Institute for Diseases in the Elderly (RIDE) (014-93-015: RIDE2); Research Into Ageing (251); Siemens Healthcare, Erlangen, Germany and the Federal State of Mecklenburg–West Pomerania; Social Ministry of the Federal State of Mecklenburg–West Pomerania; Structure Enhancing Fund (FES) of the Dutch government; Swedish Heart and Lung Foundation grant 20050561; Swedish Research Council for Worklife and Social research (FAS), grants 2001-0263, 2003-0139; Swiss National Science Foundation (grants no. 4026-28099,3347CO-108796, 3247BO-104283, 3247BO-104288, 3247BO-104284, 32-65896.01,32-59302.99, 32-52720.97, 32-4253.94); The Asthma, Allergy and Inflammation Research Trust; The Great Wine Estates of the Margaret River region of Western Australia; The Netherlands' Ministry of Economic Affairs, Ministry of Education, Culture and Science and Ministry for Health, Welfare and Sports; The Royal Society; The University of Split and Zagreb Medical Schools; Tromso University; U01 DK062418; UBS Wealth Foundation Grant BA29s8Q7-DZZ; UK Department of Health Policy Research Programme; University Hospital Oulu, Biocenter, University of Oulu, Finland (75617); University Medical Center Groningen; University of Bristol; University of Leicester HEFCE CIF award; University of Nottingham; US National Institutes of Health (NIH) (1P50 CA70907, RO1 CA121197, U19 CA148127, CA55769, CA127219, R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071252, R01HL071258, R01HL071259, UL1RR025005, contracts HHSN268200625226C, HHSN268200782096C, R01-HL084099); US NIH National Cancer Institute (RO1CA111703); US NIH National Center for Research Resources (grants M01-RR00425 and SM01 RR03997); US NIH National Eye Institute (NEI); US NIH National Heart, Lung and Blood Institute (contracts N01-HC-85079 through N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, N01-HC-95095, N01-HC-48047, N01-HC-48048, N01-HC-48049, N01-HC-48050, N01-HC-45134, N01-HC-05187, N01-HC-45205, N01-HC-45204, N01 HC-25195, N01-HC-95159 through N01-HC-95169, RR-024156, N02-HL-6-4278, R01 HL-071022, R01 HL-077612, R01 HL-074104, RC1 HL100543, HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C and HHSN268201100012C, grants HL080295, HL087652, HL105756, R01-HL-084099, R01HL087641, R01HL59367, R01HL086694, HL088133, HL075336 5R01HL087679-02 through the STAMPEED program (1RL1MH083268-01), 1K23HL094531-01); US NIH National Institute of Allergy and Infectious Diseases (NIAID); US NIH National Institute of Child Health and Human Development (NICHD); US NIH National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (DK063491); US NIH National Institute of Environmental Health Sciences (NIEHS) (ZO1 ES49019, ES015794); US NIH National Institute of Mental Health (NIMH) (5R01MH63706:02); US NIH National Institute of Neurological Disorders and Stroke (NINDD); US NIH National Institute on Aging (NIA) (R01 AG032098, RC1 AG035835, N01AG12100, N01AG62101, N01AG62103, N01AG62106, 1R01AG032098-01A1, AG-023269, AG-15928, AG-20098, AG-027058); Wellcome Trust (077016/Z/05/Z, GR069224, 068545/Z/02, 076113/B/04/Z, 079895).

AUTHOR CONTRIBUTIONS

Author contributions are listed in alphabetical order. See Supplementary Note for definitions of study acronyms.

Project conception, design and management. *Stage 1 GWAS, AGES:* G.E., M.G., V.G., T.B.H., L.J.L. *ARIC:* S.J.L., N.F., L.R.L., D.J.C., D.B.H., B.R.J., A.C.M., K.E.N. *BS8C-T1DGC:* D.P.S. *BS8C-WTCCC:* D.P.S. *BHS1:* A.L.J., A.W.M., L.J.P. *CHS:* S.A.G., S.R.H., T.L., B.M.P. *CROATIA-Korcula:* H.C., I.G., S.J., I.R., A.F.W., L.Z. *CROATIA-Vis:* H.C., C.H., O.P., I.R., A.F.W. *ECRHS:* D.L.J., E.O., I.P., M.W. *EPIC:* N.J.W. *FHS:* J.B.W., G.T.O. *FTC:* J.K., K.H.P., T. Rantanen. *Health ABC:* M.C.A., P.A.C., T.B.H., S.B.K., Y.L., B.M. *Health 2000:* M.H., M.K. *KORA F4:* J. Heinrich. *KORA S3:* C.G., H.-E.W. *NFBC1966:* P.E., A.-L.H., M.-R.J., A.P. *ORCADES:* H.C., S.H.W., J.F.W., A.F.W. *RS:* A. Hofman. *SHIP:* S.G., G.H., B.K., H.V. *Twins UK:* T.D.S., G.Z. *Stage 2 follow up, ADONIX:* J. Brisman., A.-C.O. *BHS2:* J. Beilby. *BRHS:* R.W.M., S.G.W., P.H.W. *BWHHS:* G.D.S., S.E., D.A.L., P.H.W. *CARDIA:* A.S. *CROATIA-Split:* M.B., I.K., T.Z. *GS: SFHS:* C.M.J., S.M.K., A.D.M., D.J.P. *HCS:* C.C., J.W.H., A.A.S. *LBC1936:* I.J.D., S.E.H., J.M.S. *LifeLines:* H.M.B., D.S.P., J.M.V., C.W. *MESA-Lung:* R.G.B., J.L.H. *Nottingham smokers:* I.P.H. *NSHD:* R.H., D.K. *SAPALDIA:* N.P.-H., T. Rochat. *Look-up studies, ALSPAC:* R.G., J. Henderson. *ILCCO:* ILCCO data. *Ox-GSK:* C.F., J.M.

Phenotype collection and data management. *Stage 1 GWAS, AGES:* T.A. *ARIC:* D.J.C., N.E., L.R.L., A.C.M., K.E.N. *BS8C-T1DGC:* A.R.R., D.P.S. *BS8C-WTCCC:* A.R.R., D.P.S. *BHS1:* A.L.J., A.W.M., L.J.P. *CHS:* S.A.G., S.R.H., T.L., B.M.P. *CROATIA-Korcula:* I.G., S.J., O.P., I.R., L.Z. *CROATIA-Vis:* H.C., C.H., O.P., I.R., A.F.W. *ECRHS:* D.L.J., E.O., I.P., M.W. *EPIC:* N.J.W. *FHS:* J.B.W., G.T.O. *FTC:* J.K., K.H.P., T. Rantanen. *Health ABC:* P.A.C., B.M., W.T. *Health 2000:* M.H., M.K. *KORA F4:* S.K., H.S. *KORA S3:* N.P.-H. *NFBC1966:* P.E., A.-L.H., M.-R.J., A.P. *ORCADES:* H.C., S.H.W., J.F.W. *RS:* G.G.B., M.E., D.W.L., B.H.Ch.S. *SHIP:* S.G., B.K., H.V. *TwinsUK:* C.J.H., P.G. Hysi, M.M., T.D.S., G.Z. *Stage 2 follow up, ADONIX:* J. Brisman, A.-C.O. *BHS2:* J. Beilby, M.L.H. *BRHS:* R.W.M., S.G.W., P.H.W. *BWHHS:* G.D.S., S.E., D.A.L., P.H.W. *CARDIA:* O.D.W. *CROATIA-Split:* M.B., I.K., T.Z. *GS: SFHS:* C.M.J., A.D.M. *HCS:* C.C., K.A.J., A.A.S. *LBC1936:* I.J.D., L.M.L., J.M.S. *LifeLines:* D.S.P., J.M.V. *MESA-Lung:* R.G.B., J.L.H. *Nottingham smokers:* K.A.A.B., J.D.B., I.P.H., A. Henry, M.O., I. Sayers. *NSHD:* R.H., D.K. *SAPALDIA:* N.P.-H. *Look-up studies, ALSPAC:* R.G., J. Henderson. *ILCCO:* ILCCO. *Raine:* W.Q.A., P.G. Holt, C.E.P., P.D.S.

Genotyping. *Stage 1 GWAS, BS8C-T1DGC:* W.L.M. *BS8C-WTCCC:* W.L.M. *BHS1:* A.L.J., A.W.M., L.J.P. *CHS:* S.R.H., B.M.P., J.I.R. *CROATIA-Vis:* C.H., I.R., A.F.W. *BCRHS:* M.W. *EPIC:* I.B., R.J.F.L., J.H.Z. *FTC:* J.K. *Health ABC:* Y.L., K.L. *Health 2000:* S.R., I. Surakka. *KORA F4:* N.K. *KORA S3:* C.G. *NFBC1966:* P.E., A.-L.H., M.-R.J., A.P., A.R. *ORCADES:* H.C., J.F.W. *RS:* F.R., A.G.U. *SHIP:* G.H. *TwinsUK:* C.J.H., S.-Y.S. *Stage 2 follow up, ADONIX:* S.D., F.N., A.-C.O. *BHS2:* J. Beilby, G.C., J.H. *BRHS:* A.D.H., R.W.M. *BWHHS:* S.E., D.A.L. *CARDIA:* M.F., X.G. *CROATIA-Split:* V.B., T.Z. *Gedling:* J.R.B., T.M. *GS: SFHS:* C.M.J., S.M.K., D.J.P. *HCS:* J.W.H. *LBC1936:* I.J.D., S.E.H., L.M.L., J.M.S. *LifeLines:* C.W. *MESA-Lung:* S.S.R. *NSHD:* D.K., A.W. *SAPALDIA:* M.I., F.K. *Look-up studies, ALSPAC:* S.M.R., W.L.M. *ILCCO:* ILCCO. *Raine:* W.Q.A., C.E.P.

Data analysis. *Stage 1 GWAS, AGES:* G.K.G., A.V.S. *ARIC:* N.E., D.B.H., L.R.L. *BS8C-T1DGC:* A.R.R., D.P.S. *BS8C-WTCCC:* A.R.R., D.P.S. *BHS1:* N.M.W. *CHS:* K.D.M., J.I.R. *CROATIA-Korcula:* C.H., J.E.H., V.V. *CROATIA-Vis:* C.H., V.V. *BCRHS:* D.L.J., A.R. *EPIC:* J.H.Z. *FHS:* J.B.W. *FTC:* I. Surakka. *Health ABC:* P.A.C., Y.L., K.L., W.T. *Health 2000:* M.K., S.R., I. Surakka. *KORA S3:* E.A. *NFBC1966:*

© 2011 Nature America, Inc. All rights reserved.

A.R. *ORCADES:* C.H., V.V. *RS:* M.E., D.W.L. *SHIP:* S.G., G.H., B.K., H.V. *TwinsUK:* M.M., G.Z. *Stage 2 follow-up studies, ADONIX:* S.D., F.N. *BHS2:* G.C. *BRHS:* R.W.M. *BWHHS:* D.A.L. *CARDIA:* M.F., X.G. *HCS:* J.W.H., K.A.J. *LBC1936:* L.M.L. *LifeLines:* H.M.B. *MESA-Lung:* A.M., S.S.R. *Nottingham smokers:* I. Sayers, A. Henry. *NSHD:* D.G., R.H. *SAPALDIA:* J.C., M.I. *Look-up studies, ALSPAC:* D.M.E. *ILCCO:* ILCCO. *Ox-GSK:* J.Z.L. *Rainc:* W.Q.A.

Analysis group: *SpiroMeta consortium:* I.P.H., T.J., M.S.A., M.D.T., L.V.W. *CHARGE consortium:* N.E., S.J.L., D.W.L., K.D.M., A.V.S., W.T., J.B.W.

Expression profiling and bioinformatics group: *SpiroMeta consortium:* I.P.H., M.O., I. Sayers, M.S.A., M.D.T., L.V.W. *CHARGE consortium:* S.A.G., D.W.L.

Writing group: *SpiroMeta consortium:* P.E., I.P.H., M.O., M.S.A., D.P.S., M.D.T., L.V.W. *CHARGE consortium:* S.J.L., D.W.L., S.A.G., G.T.O., V.G., B.H.Ch.S., W.T.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

© 2011 Nature America, Inc. All rights reserved.

1. Wilk, J.B. *et al.* Evidence for major genes influencing pulmonary function in the NHLBI family heart study. *Genet. Epidemiol.* 19, 81–94 (2000).
2. Hole, D.J. *et al.* Impaired lung function and mortality risk in men and women: findings from the Renfrew and Paisley prospective population study. *Br. Med. J.* 313, 711–715, discussion 715–716 (1996).
3. Strachan, D.P. Ventilatory function, height, and mortality among lifelong non-smokers. *J. Epidemiol. Community Health* 46, 66–70 (1992).
4. 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, 616–622 (2007).
5. Lopez, A.D. *et al.* Chronic obstructive pulmonary disease: current burden and future projections. *Eur. Respir. J.* 27, 397–412 (2006).
6. Mathers, C.D. & Loncar, D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med.* 3, e442 (2006).
7. Rabe, K.F. *et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am. J. Respir. Crit. Care Med.* 176, 532–555 (2007).
8. 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).
9. Repapi, E. *et al.* Genome-wide association study identifies five loci associated with lung function. *Nat. Genet.* 42, 36–44 (2010).
10. Pillai, S.G. *et al.* A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.* 5, e1000421 (2009).
11. Wilk, J.B. *et al.* A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet.* 5, e1000429 (2009).
12. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* 19, 807–812 (2011).
13. Elks, C.E. *et al.* Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat. Genet.* 42, 1077–1085 (2010).
14. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–838 (2010).
15. Lindgren, C.M. *et al.* Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet.* 5, e1000508 (2009).
16. Dixon, A.L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* 39, 1202–1207 (2007).
17. Gibson, M.A., Hughes, J.L., Fanning, J.C. & Cleary, E.G. The major antigen of elastin-associated microfibrils is a 31-kDa glycoprotein. *J. Biol. Chem.* 261, 11429–11436 (1986).
18. Yang, J. *et al.* Rootletin, a novel coiled-coil protein, is a structural component of the ciliary rootlet. *J. Cell Biol.* 159, 431–440 (2002).
19. Massaro, G.D. *et al.* Retinoic acid receptor-β: an endogenous inhibitor of the perinatal formation of pulmonary alveoli. *Physiol. Genomics* 4, 51–57 (2000).
20. Bieganowski, P., Shilinski, K., Tsichlis, P.N. & Brenner, C. Cdc123 and checkpoint forkhead associated with RING proteins control the cell cycle by controlling eIF2γ abundance. *J. Biol. Chem.* 279, 44656–44666 (2004).
21. Ito, K. *et al.* Decreased histone deacetylase activity in chronic obstructive pulmonary disease. *N. Engl. J. Med.* 352, 1967–1976 (2005).
22. Wu, H. *et al.* Parental smoking modifies the relation between genetic variation in tumor necrosis factor-α (TNF) and childhood asthma. *Environ. Health Perspect.* 115, 616–622 (2007).
23. Rose, C.E. *et al.* Tumour necrosis factor gene complex polymorphisms in chronic obstructive pulmonary disease. *Respir. Med.* 101, 340–344 (2007).
24. Chu, H.W. *et al.* Transforming growth factor-β2 induces bronchial epithelial mucin expression in asthma. *Am. J. Pathol.* 165, 1097–1106 (2004).
25. Liu, J.Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.* 42, 436–440 (2010).
26. Landi, M.T. *et al.* A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Hum. Genet.* 85, 679–691 (2009).
27. Kathiresan, S. *et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.* 41, 334–341 (2009).
28. Clancy, R.M. *et al.* Identification of candidate loci at 6p21 and 21q22 in a genome-wide association study of cardiac manifestations of neonatal lupus. *Arthritis Rheum.* 62, 3415–3424 (2010).
29. Harley, J.B. *et al.* Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM, PXK, KIAA1542* and other loci. *Nat. Genet.* 40, 204–210 (2008).
30. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* 40, 638–645 (2008).
31. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* 41, 703–707 (2009).
32. Levy, D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* 41, 677–687 (2009).
33. Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.* 41, 666–676 (2009).
34. Park, J.H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* 42, 570–575 (2010).
35. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 43, 519–525 (2011).
36. Kohansal, R. *et al.* The natural history of chronic airflow obstruction revisited: an analysis of the framingham offspring cohort. *Am. J. Respir. Crit. Care Med.* 180, 3–10 (2009).
37. Hankinson, J.L., Odencrantz, J.R. & Fedan, K.B. Spirometric reference values from a sample of the general U.S. population. *Am. J. Respir. Crit. Care Med.* 159, 179–187 (1999).
38. Talmud, P.J. *et al.* Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *Br. Med. J.* 340, b4838 (2010).
39. Wacholder, S. *et al.* Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* 362, 986–993 (2010).
40. Thompson, H.G.R., Min, J.D., Krasieva, T.B., Tromberg, B.J. & George, S.C. Epithelial-derived TGF-β2 modulates basal and wound-healing subepithelial matrix homeostasis. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 291, L1277–L1285 (2006).
41. Maran, C., Tassone, E., Masola, V. & Onisto, M. The story of *SPATA2* (spermatogenesis-associated protein 2): from Sertoli cells to pancreatic β-cells. *Curr. Genomics* 10, 361–363 (2009).
42. Tewari, R., Bailes, E., Bunting, K.A. & Coates, J.C. Armadillo-repeat protein functions: questions for little creatures. *Trends Cell Biol.* 20, 470–481 (2010).
43. Pende, D. *et al.* Identification and molecular characterization of NKp30, a novel triggering receptor involved in natural cytotoxicity mediated by human natural killer cells. *J. Exp. Med.* 190, 1505–1516 (1999).
44. Lillis, A.P., Mikhailenko, I. & Strickland, D.K. Beyond endocytosis: LRP function in cell migration, proliferation and vascular permeability. *J. Thromb. haemost.* 3, 1884–1893 (2005).
45. Burkhard, P., Stetefeld, J. & Strelkov, S.V. Coiled coils: a highly versatile protein folding motif. *Trends Cell Biol.* 11, 82–88 (2001).
46. Cowley, E.A. & Linsdell, P. Characterization of basolateral K^+ channels underlying anion secretion in the human airway cell line Calu-3. *J. Physiol. (Lond.)* 538, 747–757 (2002).

B. Analysis plans

COPD associations analysis plan

Analysis plan for replication studies

Please confirm the numbers of individuals included in the analyses and the distribution of the phenotypes after any exclusions.

For the SNPs of interest, replication studies will provide summary statistics about genotype data (genotype counts, mean phenotype values for the three genotypes, statistics testing for Hardy-Weinberg, counts of genotype inconsistencies in duplicate samples and/or relatives), and association statistics (strand, coded allele, beta, standard error) for the quantitative lung function phenotypes of interest.

Association testing will be based on the following:

For each locus **two tests for association with COPD** should be performed.

a) First restrict datasets to individuals with age>40, then COPD cases & controls should be selected under the following criteria:
   - Cases: FEV1<80% predicted [see below for definition] and FEV1/FVC ratio of <70%
     Controls: FEV1>80% predicted and FEV1/FVC ratio of >70%
   - Individuals with FEV1<80% predicted with FEV1/FVC ratio>70%, or vice versa, should be excluded from both groups.

Perform a logistic regression analysis, with COPD (case=1; control=0) status as the outcome and with the SNP (coded 0, 1, 2 for the number of copies of the coded allele) as the only covariate. **The effects should be reported in logit scale.**

b) As above, first restrict datasets to individuals with age>40 and **pack-years>5**then COPD cases & controls should be selected under the following criteria:
   - Cases: FEV1<80% predicted and FEV1/FVC ratio of <70%
     Controls: FEV1>80% predicted and FEV1/FVC ratio of >70%
   - Individuals with FEV1<80% predicted with FEV1/FVC ratio>70%, or vice versa, should be excluded from both groups.

Perform a logistic regression analysis, with COPD (case=1; control=0) status as the outcome and with the SNP (coded 0, 1, 2 for the number of copies of the coded allele) as the only covariate. **The effects should be reported in logit scale.**

For calculating the predicted FEV1:

- if this has already been calculated in the replication cohort using appropriate reference values for the local population, we suggest use of that value (please let us know what this is when you send in your results)
- if this has not yet been calculated, we suggest use of the following formula[1]:

Males: Expected FEV1 = 0.5536 - 0.01303*age - 0.000172*age^2 +0.00014098*height^2;
Females: Expected FEV1 = 0.4333 - 0.00361*age - 0.000194*age^2 + 0.00011496*height^2.

File format

It would be helpful if the results of the analyses will be given in comma separated value files (csv files) following the naming scheme described in the next paragraph.

The following fields will be required for each SNP. It would be appreciated if the fields are named following the **bold** titles as below.

- **Chr** : Chromosome of the SNP (an integer from 1 to 22)
- **Position**: Position of the SNP (an integer)
- **Markerid**: rs number (a character string beginning with "rs")
- **Markerid2**: other ID when the rs number is not available e.g. affy id(a character string or empty when nothing to report)
- **Bas_all**: baseline allele (a single character: "A" "C" "G" "T")
- **Cod_all**: coded allele (effect allele) (a single character: "A" "C" "G" "T")
- **Strand**: the strand of the baseline and the coded alleles (a single character: "+" or "-")
- **Freq**: allele frequency for **coded allele** (numeric data)
- **Beta**: effect size for each copy of the coded allele (numeric data)
- **Se**: standard errors of beta (numeric data)
- **Type**: whether the SNP was genotyped or imputed (a character string: "gen" or "imp")
- **Imp_info**: r^2.hat or .info for imputed SNPs (numeric data)

We would recommend that *at least* four decimal places will be kept for all the statistics.

In addition, the COPD analyses should include the following:

- **Chr** : Chromosome of the SNP (an integer from 1 to 22)
- **Position**: Position of the SNP (an integer)
- **Markerid**: rs number (a character string beginning with "rs")
- **Markerid2**: other ID when the rs number is not available e.g. affy id(a character string or empty when nothing to report)
- **Bas_all**: baseline allele (a single character: "A" "C" "G" "T")
- **Cod_all**: coded allele (effect allele) (a single character: "A" "C" "G" "T")
- **Strand**: the strand of the baseline and the coded alleles (a single character: "+" or "-")
- **Beta**: effect size for each copy of the coded allele (numeric data) **In logit scale.**
- **Se**: standard errors of beta (numeric data)

Naming scheme

Each analysis should be given in a different file named as:

cohortname_repl_phenotype_dataset_ version.csv
or
cohortname_repl_phenotype _ version.csv (for COPD analyses)

where:

cohortname will be an identifier for the specific cohort
phenotype will be one of “FEV1”,”FF” (for the ratio FEV1/FVC), “COPD” or “COPDpy” (for the analysis with the pack-years criteria)
dataset will be one of “all”,” smk”, “smkPY” (for the pack-years adjustment as defined in 4) or ”nonsmk”
version will be the date of the day of the uploading (ddmmyy)

For example a file name from the cohort ILFGC would be:
ILFGC_repl_FEV1_smk_190109.csv
ILFGC_repl_COPD_190109.csv
or ILFGC_repl_COPDpy_190109.csv

Please address any questions regarding the analysis plan to Martin Tobin mt47@leicester.ac.uk or Emmanouela Repapi er82@leicester.ac.uk

Lung function and COPD risk scores analysis plan

Analysis plan for Modelling the joint effect of Risk Alleles

Please confirm the numbers of individuals included in the analyses and the numbers of the individuals in each group (for both a. and b. analyses) after any exclusions.

Coding of Risk alleles

Risk alleles and their weights are defined in the table below. **Exclude all individuals with any missing genotype data.**

Table1

Chrom	SNPID (Position) Gene	Risk Allele for NCBI B36, HapMap Data Rel24	Risk Allele for NCBI B36.3, dbSNP	Risk allele freq	Weights for FEV1	Weights for FF
2	rs2571445 (218,391,399) *TNS1*	A	T	0.41	1.014	0.345
4	rs10516526 (106,908,353) *GSTCD*	A	A	0.94	2.304	0.687
4	rs12504628 (145,655,774) *HHIP*	T	T	0.56	1.152	0.733
5	rs3995090 (147,826,008) *HTR4*	A	A	0.59	0.825	1.724
6	rs2070600 (32,259,421) *AGER*	C	G	0.94	0.325	1.344
15	rs12899618 (69,432,174) *THSD4*	A	A	0.15	0.380	1.167

For SNPs rs2571445 and rs2070600 the databases don't agree on which base is on the + strand. For rs2571445 the risk allele should be A or T (with the non-risk being G or C respectively) and for rs2070600 the risk allele should be C or G (with the non-risk being T or A respectively) depending on the database that you have the genotypes reported in.

Grouping the individuals:

1. Unweighted analyses

Count the total number of risk alleles an individual carries and group the individuals in 5 groups according to the number of risk alleles that they carry.

Create four indicator variables: ui1 (unweighted indicator 1), ui2, ui3 and ui4. The first one will be 1 for individuals with a total of 0-4 risk alleles (and 0 for the rest), the second-one for individuals with a total of 5-6 risk alleles (and 0 for the rest), the third for individuals with a total of 8-9 risk alleles (and 0 for the rest) and the last for individuals with a total of 10-12 risk alleles (and 0 for the rest), as specified in the following table.

No of risk alleles	Group	ui1	ui2	ui3	ui4
0-4	1	1	0	0	0
5-6	2	0	1	0	0
7	3 (Baseline)	0	0	0	0
8-9	4	0	0	1	0
10-12	5	0	0	0	1

2. Weighted analyses

Multiply the number of risk alleles for each SNP by the appropriate weight. Add up the products for each individual to calculate the total score of Risk allele and assign each individual to a group. Create four indicator variables: wi1 (weighted indicator 1), wi2, wi3 and wi4 and code them as follows:

Risk allele	Group	wi1	wi2	wi3	wi4
0< risk score< 5	1	1	0	0	0
5<=risk score< 7	2	0	1	0	0
7<=risk score< 8	3(Baseline)	0	0	0	0
8<=risk score< 10	4	0	0	1	0
10<= risk score< =12	5	0	0	0	1

Example for weighted analyses:

Genotypes for 6 individuals:

SNP ID	indiv 1	indiv 2	indiv 3	indiv 4	indiv 5	indiv 6
rs10516526	A:A	A:A	A:A	A:A	G:A	A:A
rs12899618	G:A	G:A	G:G	G:G	G:A	G:A
rs2070600	G:G	G:G	G:G	G:G	G:G	G:G
rs2571445	C:C	T:C	T:C	T:T	T:C	C:C
rs3995090	C:C	A:A	C:A	C:C	C:A	A:A
rs12504628	T:C	T:C	T:C	T:C	T:C	T:T

Let's assume that we would like to calculate the risk allele score and the group for each of these 6 individuals for the FEV1 analysis. Taking the risk alleles and the weights for FEV1 from table 1 we have:

SNP ID	Risk allele for our version	Weights for FEV1
rs10516526	A	2.304
rs12899618	A	0.380
rs2070600	G	0.325
rs2571445	T	1.014
rs3995090	A	0.825
rs12504628	T	1.152

The numbers of risk alleles for each individual are:

SNP ID	indiv 1	indiv 2	indiv 3	indiv 4	indiv 5	indiv 6	
rs10516526_risk	2	2	2	2	1	2	
rs12899618_risk	1	1	0	0	1	1	2
rs2070600_risk	2	2	2	2	2	0	2
rs2571445_risk	0	1	1	2	1	2	
rs3995090_risk	0	2	1	0	1		
rs12504628_risk	1	1	1	1	1		

By multiplying them with the weights we have the risk score of each individual for each SNP:

SNP ID	indiv 1	indiv 2	indiv 3	indiv 4	indiv 5	indiv 6
rs10516526_risk	4.608	4.608	4.608	4.608	2.304	4.608
rs12899618_risk	0.38	0.38	0	0	0.38	0.38
rs2070600_risk	0.65	0.65	0.65	0.65	0.65	0.65
rs2571445_risk	0	1.014	1.014	2.028	1.014	0
rs3995090_risk	0	1.650	0.825	0	0.825	1.650
rs12504628_risk	1.152	1.152	1.152	1.152	1.152	2.304

The total risk score and the group each individual will be assigned to are:

	indiv 1	indiv 2	indiv 3	indiv 4	indiv 5	indiv 6
Risk score	6.790	9.454	8.249	8.438	6.325	9.592
wi1	0	0	0	0	0	0
wi2	1	0	0	0	1	0
wi3	0	1	0	0	0	1
wi4	0	0	0	0	0	0

Association testing:

Restrict dataset to those individuals with no missing data for the eversmoking/neversmoking variable and to those with complete data on both FEV1 **and** FVC. **Also exclude all individuals with any missing genotype data.**

1. Residual phenotype calculation

   Undertake linear regression of FEV1 onto age, age^2, sex, and height and use residuals for all subsequent association analyses.

   Repeat using FEV1/FVC ratio in place of FEV1.

   You should **not** transform the phenotypes at any point in this analysis.

2. Unweighted analyses

   a) **Continuous phenotypes.** Separately for the FEV1 and FEV1/FVC residual phenotypes calculated in step 1, perform the following analysis:

      i. Fit a normal **linear *multiple* regression model** with the residual phenotype as the outcome variable, and an intercept and the unweighted indicator variables as the explanatory variables. Report the effect size (beta coefficient) and standard error for the intercept and for each of the indicator variables. Report also the sum of squared errors and the sample size.

      ii. Positive control analyses:

      Part 1: For each SNP fit a normal **linear regression model**, with the residual phenotype as the outcome variable and an intercept and the risk dosage (number of risk alleles an individual carries for each SNP, exactly as used to get the risk score - *same risk alleles*) as the explanatory variables. Report the effect size (beta coefficient) and standard error for the intercept and for the risk dosage of each model.

      Part 2: Fit a normal **linear regression model**, with the identical outcome variable and an intercept and the unweighted Risk score (total number of risk alleles an individual carries before grouping) as the explanatory variable. Report the effect size (beta coefficient) and standard error for the intercept and for the Risk score variable. Report also the sum of squared errors and the sample size.

   b) **COPD analysis.** First restrict datasets to individuals with age>40, then COPD cases & controls should be selected under the following criteria (Note: the definition of COPD is identical to previous analyses):
      - Cases: FEV1<80% predicted [see below for definition] and FEV1/FVC ratio of <70%
        Controls: FEV1>80% predicted and FEV1/FVC ratio of >70%
      - Individuals with FEV1<80% predicted with FEV1/FVC ratio>70%, or vice versa, should be excluded from both groups.

Perform a **logistic multiple regression analysis**, with COPD (case=1; control=0) status as the outcome and with an intercept and the indicator variables as the explanatory variables.

Report the effect size (beta coefficient) and standard error for the intercept and for each of the indicator variables. **The effects should be reported in logit scale.**

For calculating the predicted FEV1:

- if this has already been calculated in the replication cohort using appropriate reference values for the local population, we suggest use of that value (please let us know what this is when you send in your results)
- if this has not yet been calculated, we suggest use of the following formula:

Males: Expected FEV1 = 0.5536 - 0.01303*age - 0.000172*age^2 +0.00014098*height^2;
Females: Expected FEV1 = 0.4333 - 0.00361*age - 0.000194*age^2 + 0.00011496*height^2.

1. Weighted analyses

Repeat analysis 2.a.i above, but using the indicator variables calculated from the **weighted risk score**. Note: the weights and therefore the indicator variables will be different for the analyses of FEV1 and of FEV1/FVC.

At this stage we are not requesting weighted COPD analysis because the COPD weights are expected to change with the inclusion of additional data. [We will therefore review the issue if the referees' comments require such analysis.]

SpiroMeta-CHARGE stage 1 analysis plan

INTERNATIONAL LUNG FUNCTION GENOMICS CONSORTIUM (SPIROMETA): ANALYSIS PLAN

Version: 19th January 2009

ANALYSIS STEPS WITHIN EACH COHORT

DESCRIPTIVE STATISTICS

Cohorts will be asked to provide information on QC (see below) and on the distribution (range, mean, sd) of FEV1, FEV1/FVC, FVC, age, sex, height, smoking status, pack-years of smoking, numbers diagnosed with asthma, COPD.

Cohorts will also be asked to provide histograms of FEV1, residuals from linear regression after adjusting for covariates age, age^2, sex, height (described under "Phenotype for association testing" below).
The same will be requested for FEV1/FVC and FVC.

We have asked for copies of the questionnaires used to collect smoking data, and additional information where needed so that we can assess the consistency of approaches used.

QC

Internal QC of initial genotype data will be undertaken by each of the cohorts, such as exclusion of subjects with poor genotype call rate, subjects with evidence of non-Caucasian ancestry, SNPs with low call rate and SNPs out of HWE. To date many cohorts within GWAS consortia have used no minor allele frequency (MAF) filter prior to imputation, or have filtered out SNPs with MAF<0.01. Exact QC thresholds tend to vary between cohorts. This is probably appropriate given the different technologies. Cohorts must provide information on the quality filters used (see spreadsheet for study information). The analysis working group will be happy to discuss quality filters as required.

IMPUTATION

Imputation can potentially be used to (i) infer untyped HapMap SNPs; (ii) fill in missing genotypes for typed SNPs and (iii) change typed SNP calls where these appear inconsistent with Hapmap haplotypes. Individual cohorts will decide precisely how they implement their imputation – and this may vary with their platform used and data quality - (e.g. whether (ii) is also implemented; (iii) above is normally not implemented). Only cohorts which have imputed untyped HapMap SNPs will have their data included in the meta-analysis. Two of the most commonly used programs are IMPUTE (Marchini et al, Oxford) and MACH (Abecasis et al, UMich). SNPs will be excluded by individual cohorts if they are imputed with low confidence/quality i.e. if .info<0.3 (IMPUTE) or r^2.hat <0.3 (MACH). Cohorts employing other imputation approaches are strongly encouraged to discuss these with the analysis group in advance so that we can ensure consistency as far as is possible. We would not encourage use of PLINK for imputation at present.

ASSOCIATION TESTING AND SOFTWARE

Where possible, ancestry principal components will be estimated using EIGENSTRAT or equivalent software. Association testing will be undertaken within cohorts based on ~2.2 million HapMap SNPs (less a proportion missing due to low imputation quality; this will vary across platforms). A variety of association testing software is available. Commonly used software packages include SNPTEST (Marchini et al, designed to utilize output

use of PLINK for MACH-imputed data is fine. Again, cohorts using different packages for association testing are encouraged to liaise with the analysis working group to confirm that consistent approaches are employed and that consistent output is available. Where possible, we would encourage use of programs/options that make use of the posterior probabilities of the genotype calls (e.g. Proper option in SNPTEST, ProbABEL package) rather than a simple threshold approach. Where genotypes have been coded as missing where the most likely genotype falls below a given threshold, cohorts should supply details of the threshold used.

PHENOTYPES FOR ASSOCIATION TESTING

Initial analyses for FEV1, FVC and FEV1/FVC

ADULT COHORTS:

1. Restrict dataset to those individuals with no missing data for the eversmoking/neversmoking variable and to those with complete data on both FEV1 **and** FVC. Undertake linear regression of age, age^2, sex, height, ancestry principal components on FEV1 and use residuals for all subsequent analyses. Transformation would be taken **once** for each trait (FEV1, FVC and FEV1/FVC ratio) and used for all analyses (including subgroups).
    a. Transformed analysis: transform residuals to ranks and then to normally distributed z-scores. These inverse-normal transformed residuals are then used as the phenotype for association testing under an additive genetic model.
    b. Untransformed analysis: use untransformed residuals for association testing under an additive genetic model (units of millilitres for FEV1 please). This will assist interpretation of findings from a. above.

2. Ever-smokers only: a. analysis as for 1a above
    b. analysis as for 1b above
    c. Undertake linear regression of **age, age^2, sex, height, ancestry principal components and pack-years** on FEV1. **Transform residuals** to ranks and then to normally distributed z-scores. These inverse-normal transformed residuals are then used as the phenotype for association testing under an additive genetic model.

    d. Repeat 2c for FEV1/FVC ratio (using same approach as for FEV1).

3. Never-smokers only: a. analysis as for 1a above
    b. analysis as for 1b above

Repeat the above for outcomes: **FVC** and **FEV1/FVC ratio** (using same approach as for FEV1, untransformed analysis to use millilitres for FVC and percentage for FEV1/FVC ratio).

4. Positive control analysis: Association testing under an additive genetic model using **BMI as the outcome** (no transformation of BMI required, no covariates in the model). No transformation of BMI is required for this analysis as it is simply for positive control purposes. We would be very grateful if cohorts could undertake this analysis as a positive control. To be useful as a positive control the analysis would have to be undertaken by the same analyst that is performing analyses 1-3 above. However, if it inconvenient to perform genome-wide analyses, the output related to chromosome 16 will suffice.

CHILDREN'S COHORTS: undertake analyses 1a, 1b, repeat for FVC and **FEV1/FVC ratio and 4 (the control analysis). No smoking-stratified analyses will be expected.**

ASSOCIATION TESTING OUTPUT

The output statistics and file formats required for each SNP are shown in the file:
"SpiroMeta_file_format_19012009.doc"

In addition, for control purposes, we would also like to receive (from analysis 1a) a file with the output for the first 500 SNPs of chromosome 1 from the software package that has been used by each cohort.

A nominated analyst for each cohort will upload these association test statistics to a sFTP site.

CONSORTIUM ANALYSIS WORKING GROUP: CHECKS AND META-ANALYSIS

A small analysis working group will be established to work out analytic issues applicable to the individual GWASs and the meta-analysis. This will include recommendation to the consortium of the datasets to be merged based on the completeness of data and on checks below by timescales to be decided by the consortium members.

Checks of correctness will be performed and reported, including (i) data from positive controls i.e. genome-wide (or chromosome 16) association with BMI by the same analyst undertaking the association testing with FEV1, FVC & FEV1/FVC. This enables checking of (i) consistency of effect size direction and labelling. (ii) concordance of allele frequency of alphabetically larger allele across studies, and (iii) correlation [or lack of correlation] of effect size estimates across studies.

Careful attention will be paid to alignment of data sets. For each SNP (and for each genetic model assumed if any model other than additive is ultimately examined), the pooled effect size estimate and standard error may be computed using inverse variance weighting or alternate weighting schemes as appropriate. *P*-values will be reported under normality assumptions. For SNPs with *p*-values below some agreed threshold, standard meta-analysis statistics may be reported. QQ plots will be shown after addition of each cohort. The meta-analysis will need to be performed using a suitable criterion (and perhaps using several different criteria) for inclusion/exclusion of individual results according to e.g. imputation quality.

SpiroMeta-CHARGE stage 2 analysis plan

Analysis plan for follow-up studies: version 9 June 2010

Please provide the following summary information for the subset of individuals for which this analysis is undertaken:

N total	N males	N females	Age range at measurement	Mean age y (s.d.)	Mean FEV1 (s.d.)	Mean FEV1/FVC (s.d.)	N never smokers	N ever smokers	N ever smokers with pack-years data*

*See analysis (d) below

For the SNPs of interest, please provide summary statistics about genotype data (genotype counts, mean phenotype values for the three genotypes, statistics testing for Hardy-Weinberg, counts of genotype inconsistencies in duplicate samples and/or relatives), and association statistics (strand, coded allele, beta, standard error) for the quantitative lung function phenotypes of interest.

The analysis plan below assumes unrelated individuals. Please contact us to discuss options if you have related individuals.

Association testing will be based on the following:

a) All individuals: Restrict dataset to those individuals with no missing data for the eversmoking/neversmoking variable and to those with complete data on both FEV1 **and** FEV1/FVC. Undertake linear regression of age, age^2, sex and height on FEV1 (studies with GWAS data please also adjust for ancestry principal components if available) and use residuals for all subsequent analyses. Transformation would be taken **once** for each trait (FEV1 and FEV1/FVC ratio) and used for all analyses (including subgroups). Transform residuals to ranks and then to normally distributed z-scores. These inverse-normal transformed residuals for FEV1 are then used as the phenotype for association testing under an additive genetic model (SNP coded 0, 1, 2 for the number of copies of the coded allele).

b) Never-smokers only: Repeat analysis as for a)

c) Ever-smokers only: Repeat analysis as for a)

d) Ever-smokers only with smoking status (current/past) and Pack-years adjustment: Restrict dataset to those individuals with **no missing data for the smoking status and pack-years variable** and to those with complete data on FEV1. Undertake linear regression of age, age^2, sex, height, ***smoking status and pack-years*** on FEV1. Transformation would be taken **once** for each trait (FEV1 and FEV1/FVC ratio) and used for all analyses (including subgroups).Transform residuals to ranks and then to normally distributed z-scores. These inverse-normal transformed residuals are then used as the phenotype for association testing under an additive genetic model.

Repeat the above for **FEV1/FVC ratio** (using same approach as for FEV1)

File format

It would be helpful if the results of the analyses will be given in comma separated value files (csv files) following the naming scheme described in the next paragraph.

The following fields will be required for each SNP. It would be appreciated if the fields are named following the **bold** titles as below.

- **Chr** : Chromosome of the SNP (an integer from 1 to 22)
- **Position**: Position of the SNP (an integer)
- **Markerid**: rs number (a character string beginning with "rs")
- **Markerid2**: other ID when the rs number is not available e.g. affy id(a character string or empty when nothing to report)
- **Bas_all**: baseline allele (a single character: "A" "C" "G" "T")
- **Cod_all**: coded allele (effect allele) (a single character: "A" "C" "G" "T")
- **Strand**: the strand of the baseline and the coded alleles (a single character: "+" or "-")
- **Freq**: allele frequency for **coded allele** (numeric data)
- **Beta**: effect size for each copy of the coded allele (numeric data)
- **Se**: standard errors of beta (numeric data)
- **Type**: whether the SNP was genotyped or imputed (a character string: "gen" or "imp")
- **Imp_info**: r^2.hat or .info for imputed SNPs (numeric data)

We would recommend that *at least* four decimal places will be kept for all the statistics.

Note for studies with GWAS data: if your analysis pipeline includes a GWAS and then a lookup of the relevant SNPs, *do not* apply a genomic control correction, but *please supply the lambda for analyses (a) to (d)* so that this correction can be applied by us later if required.

Naming scheme

Each analysis should be given in a different file named as:
cohortname_repl_phenotype_dataset_ version.csv

where:

cohortname will be an identifier for the specific cohort
phenotype will be one of "FEV1","FF" (for the ratio FEV1/FVC)
dataset will be one of "all" (analysis 1.a),"nonsmk" (analysis 1.b),"smk" (analysis 1.c), or "smkPY" (analysis 1.d)
version will be the date of the day of the uploading (ddmmyy)

For example a file name from the cohort ILFGC would be:

ILFGC_repl_FEV1_smk_09062010.csv

Please address any questions regarding the analysis plan to Maria Soler Artigas msa20@leicester.ac.uk and Martin Tobin mt47@leicester.ac.uk.

SpiroMeta burden test analysis plan

SpiroMeta gene-based rare variant analysis
[version 26th Jan 2010, adapted from QuTie v4]

Phenotype data preparation and exclusions

*Please note that the method is designed for **unrelated individuals**.*
Note for studies with related individuals: please extract a subset of unrelated individuals.

Restrict dataset to those individuals with no missing data for the eversmoking/neversmoking variable and to those with complete data on both FEV1 **and** FVC.

Undertake linear regression of age, age^2, sex, height, eversmoking/neversmoking, ancestry principal components on FEV1 and use residuals for the subsequent analysis.(*N.B. the residuals need to be recalculated for all the cohorts, since the eversmoking/neversmoking variable is now included as a covariate in the linear regression*)

Transformation would be taken **once** for each trait (FEV1 and FEV1/FVC). Transform residuals to ranks and then to normally distributed z-scores. The rare variants analysis must be run twice: (i) using the inverse-normal transformed residuals for FEV1 as the phenotype and (ii) the inverse-normal transformed residuals for FEV1/FVC as the phenotype.

Please provide the following summary information for the subset of individuals for which this analysis is undertaken (*N.B. this will be the same as that provided for the original SpiroMeta meta analysis if no related individuals need to be excluded*). This can be provided by email to Maria Soler Artigas msa20@le.ac.uk

N total	N males	N females	Age range at measurement	Mean age y (s.d.)	Mean FEV1 l (s.d.)	Mean FEV1/FVC (s.d.)	N never smokers	N ever smokers

Any questions please get in touch with Maria Soler Artigas
[msa20@le.ac.uk, tel:0116 229 7208].

The results can be uploaded to the folders already set up for the initial SpiroMeta analyses. When you are ready for upload please contact Maria Soler Artigas.

Rare variant analysis

Please run the analysis on directly typed SNPs passing QC (but with no MAF exclusions) on unrelated individuals. To run the perl script, please create 22 folders named chr01,

chr02, ... to chr22. In each folder place 3 files, one of each of ped, map and gene (see below).

3 types of input file are needed:
Note: Files should be whitespace (space/tab) delimited

1. A ped file (format example below; you can also use 1,2,3,4 coding for alleles). No naming convention is needed, as long as the file has a .ped extension. No header row is needed. Individuals with missing phenotype values should be removed.

FamilyID	IndID	FatherID	MotherID	Sex	Phenotype(numerical-value)	SNP1-allele1	SNP1-allele2
1	1	0	0	1	21.3	G	G
2	2	0	0	2	24.8	G	C

2. A map file (with .map extension; format example below). Please note that the coordinates here should be based on the same human genome build (build 36) as the coordinates in the gene file (see below). No header row is needed.

Chromosome	SNP	Position(bp)build36
1	rs12345	9876543
1	rs54321	9877654

3. A gene file (centrally provided, based on build 36).

The command line is:

perl QuTie_v4.pl –gene –maf=0.05 –nchr –ext=50 –pout=0.01 –ttest –nperm=100000 –pperm=0.00001 –graph –glog=4

The script takes between a few hours and a few days to run, depending on sample size and on the number of permutations. The different options in the command line denote the following:
-gene: gene-centric analyses will be run
-maf=0.05: only SNPs with MAF≤0.05 will be analysed
-nchr: analysis for all chromosomes
-ext=50: gene intervals will be defined as 50kb either side of the gene coordinates
-pout=0.01: SNP lists with p values<0.01 will be produced
-ttest: t test statistic will also be calculated
-nperm=100000: phenotypes will be permuted 100,000
-pperm=0.00001: permutations will be run for genes with p≤0.00001
-graph: a Manhattan plot will be produced
-glog=4: p values≤0.0001 will be highlighted in the plot.

*if the map file has four columns (chromosome, snp identifier, genetic distance and base-pair position), then the –plink option will be required

Individual chromosome gene-based results will be output as a text file within each chromosomal directory along with chromosome-specific SNP lists. The genome-wide summary graphic and text files will be output to the directory where the script is run.

Further information can be found at
http://www.sanger.ac.uk/resources/software/ccravat-qutie/

File structure and naming convention

From each chromosomal directory, the file containing the gene-based results will be needed. If the map and ped filed are called plink.recode.map and plink.recode.ped this file will be called :

plink.recode_QTRVgene_MAF0.05.txt.

No format editing is required but each file must be re-named as follows:

Cohortname_rv_phenotype_all_transformed_chr_date.txt

For example, if the cohort is named ILFGC, the files for the first chromosome will be:

ILFGC_rv_FEV1_all_transformed_1_270110.txt
ILFGC_rv_FF_all_transformed_1_270110.txt

In total there will be 22 .txt files for FEV1 and 22 .txt files for FEV1/FVC.

In addition the following genome-wide summary graphic and text files saved in the directory from where the script is run must also be provided:

chr1-22.WG_summary_plink.recode_QTRVgene_MAF0.05_histogram.png

chr1-22.WG_summary_plink.recode_QTRVgene_MAF0.05.png

chr1-22.GenWide_Signif_Pvals_Perms_MAF0.05gene.txt

They must be **re-named** (using the convention described above):

Cohortname_rv_phenotype_all_transformed_hist_date.png

Cohortname_rv_phenotype_all_transformed_manh_date.png

Cohortname_rv_phenotype_all_transformed_pval_date.txt

A total of 50 files should therefore be uploaded (25 per phenotype).

All other output files produced by the script must be saved, as they might be needed once the meta-analysis is done but do not need to be uploaded at this point.

In addition, it may be necessary for the individual SNP cluster plots to be viewed at a later date in order to identify where genotyping errors may have led to a false signal. We will contact studies in due course if this becomes necessary but please let us know as soon as possible if you anticipate that this will cause any problems.

C. Chapter 3 additional tables

Genotyping platform and quality control criteria for each study in stage1

Genotyping platforms, filters applied to SNPs and individuals (if any) before imputation, imputation software and genotype-phenotype association software are given. Abbreviations: GWAS= Genome-Wide Association Study, imp'n=imputation, HWE= Hardy Weinberg Equilibrium, MAF= minor allele frequency.

Study	GWAS platform	Calling algorithm	Individual call rate filter applied (before imp'n)	SNP call rate filter applied before imp'n	SNP HWE Filter applied (before imp'n)	SNP MAF filter applied (before imp'n)	Other filter	No of SNPs after filtering (before imp'n)	Imp'n software and version	NCBI; HapMap CEU version for imp'n	Genotype-phenotype association software and version
AGES	Illumina Hu370CNV	BeadStudio	0.97	0.90	$1x10^{-6}$	0.01	remove AT/GC SNPs	208340	MACH 1.0.16	36;21a	ProbABEL 0.1
ARIC	Affymetrix 6.0	Birdseed	0.95	0.95	$1x10^{-6}$	0.01	no chromosomal location	669450	MACH 1.0.16	36;22	ProbABEL 0.1-3
B58C T1DGC	Illumina 550K	ILLUMINUS	0.98	No	No	No	No	520010	MACH 1.0.13	35;21	ProbABEL 0.0-5b
B58C WTCCC	Affymetrix 500K	CHIAMO	0.98	No	No	No	No	490033	IMPUTE 0.2.0	35;21	SNPTEST 1.1.3
BHS1	Illumina 610-Quad	BeadStudio	0.97	0.99 for SNPS with MAF<1%, 0.95 for all other SNPs	$5.7x10^{-7}$	0.01	No	549294	MACH 1.0.16	36	Mach2Qtl 1.0.8
CHS	Illumina 370 CNV	BeadStudio	0.95	0.97	$1x10^{-5}$	heterozygote frequency >0	reproducibility errors<2	306655	BimBam 0.99	36	R

Study	GWAS platform	Calling algorithm	Individual call rate filter applied (before imp'n)	SNP call rate filter applied before imp'n	SNP HWE Filter applied (before imp'n)	SNP MAF filter applied (before imp'n)	Other filter	No of SNPs after filtering (before imp'n)	Imp'n software and version	NCBI; HapMap CEU version for imp'n	Genotype-phenotype association software and version
CROATIA-Korcula	Illumina HumanHap 370cnv	Beadstudio	0.98 (for SNP of call rate >=0.98,MAF>=0.02,HWE>=E-10)	0.98	$1x10^{-6}$	0.01	No	307728	MACH 1.0.15	36;22	GenABEL 1.4.2 , ProbABEL
CROATIA-Vis	Illumina HumanHap 300 v1	Beadstudio	0.97 (for SNP of call rate >=0.98,MAF>=0.02,HWE>=E-10)	0.98	$1x10^{-6}$	0.01	No	305068	MACH 1.0.15	36;22	GenABEL 1.4.2, ProbABEL
ECRHS (population based sample from first survey)	Illumina Quad 610k	GenCall	None	None	None	None	None	582892	MACH 1.0	36;22	ProbABEL 0.0-9
EPIC obese cases	Affymetrix 500K	BRLMM	0.94	0.90	$1x10^{-6}$	0.01	No	397438	IMPUTE 0.3.1	35;21	SNPTEST 1.1.5
EPIC population-based	Affymetrix 500K	BRLMM	0.94	0.90	$1x10^{-6}$	0.01	No	397438	IMPUTE 0.3.1	35;21	SNPTEST 1.1.5
FHS	Affy 500K + 50K Gene focused	Bayesian robust linear modelling using Mahalanobis distance (BRLMM)	0.97	0.97	$1x10^{-6}$	0.01	MISHAP p<10-9, mendelian errors>100	378163	MACH 1.0.15	36;22	GWAF
FTC	Illumina 317K	BeadStudio	0.95	0.90	$1x10^{-5}$	0.01	No	315987	MACH 1.0.16	36;22	PLINK 1.06

Study	GWAS platform	Calling algorithm	Individual call rate filter applied (before imp'n)	SNP call rate filter applied before imp'n	SNP HWE Filter applied (before imp'n)	SNP MAF filter applied (before imp'n)	Other filter	No of SNPs after filtering (before imp'n)	Imp'n software and version	NCBI; HapMap CEU version for imp'n	Genotype-phenotype association software and version
Health 2000	Illumina 610K	Illuminus	0.95	0.95	$1x10^{-6}$	0.01	MDS-plot outliers removed (non-European ancestry)	555388	MACH 1.0	36;22	ProbABEL
Health ABC	Illumina Human 1M-Duo	BeadStudio 3.3.7	0.97	0.95	$1x10^{-6}$	0.01	No sex mismatch, and cryptic relatedness	914263	MACH 1.0.16.a	36;22	R version 2.9.2
KORA F4	Affymetrix 6.0	Birdseed2	0.93	No	No	No	No	909622	IMPUTE 0.4.2	36;22	SNPTEST 1.1.5
KORA S3	Affymetrix 500K	BRLMM	0.93	No	No	No	No	490033	MACH 1.0.9	35;21	MACH2QTL 1.0.4
NFBC1966	Illumina HumanCNV370-Duo	Beadstudio	none	0.95	$1x10^{-4}$	0.01	No	328007	IMPUTE v1.0	35; 21	SNPTEST 1.1.5
ORCADES	Illumina HumanHap 300 v2	Beadstudio	0.98 (for SNP of call rate >=0.98,MAF>=0.02,HWE>=E-10)	0.98	$1x10^{-6}$	0.01	No	306207	MACH 1.0.15	36;22	GenABEL 1.4.2, ProbABEL

Study	GWAS platform	Calling algorithm	Individual call rate filter applied (before imp'n)	SNP call rate filter applied before imp'n	SNP HWE Filter applied (before imp'n)	SNP MAF filter applied (before imp'n)	Other filter	No of SNPs after filtering (before imp'n)	Imp'n software and version	NCBI; HapMap CEU version for imp'n	Genotype-phenotype association software and version
RS-I	Illumina HapMap 550K	BeadStudio	0.98	0.98	$1x10^{-6}$	0.01	excess autosomal heterozygosity, sex mismatch or outlying identity-by-state clustering estimates	512349	MACH 1.0.15	36;22	MACH2QTL as implemented in GRIMP
RS-II	Illumina 550K + 610 Quad	GenomeStudio	0.98	0.98	$1x10^{-6}$	0.01	excess autosomal heterozygosity, sex mismatch or outlying identity-by-state clustering estimates	537405	MACH 1.0.16	36;22	MACH2QTL as implemented in GRIMP
SHIP	Affymetrix 6.0	BirdseedV2	0.92	No	No	No	QC callrate > 0.86 each Chip	869224	IMPUTE 0.5.0	36;22	SNPTEST 1.1.5
TwinsUK-I	Illumina 317K	Beadstudio	0.95	0.95 if MAF>0.05; <0.99 if 0.01<=MAF<0.05	$5.7x10^{-7}$	0.01	unexpected relatedness based on pi_hat	296293	IMPUTE 0.5.0	36;22	GenABEL 1.4.2

Tests for association with lung function for all SNPs followed up in stage 2

Results in stage 2 for the 34 SNPs which showed novel evidence of association ($P < 3 \times 10^{-6}$) in stage 1 are shown. Abbreviations: Chr.=chromosome, N = effective sample size as the product of sample size and imputation quality metric summed up across studies, ns =nonsynonymous, s = synonymous.

Chr.	Measure	SNP_ID (NCBI36 position), function	Coded allele	Stage 1			Stage 2			Stage 1 + stage 2 meta-analysis		
				Beta (Se)	P	N	Beta (Se)	P	N	Beta (Se)	P	N
1	FEV_1/FVC	rs2284746 (17179262), *MFAP2*(intron)	G	-0.042 (0.007)	$2.47x10^{-9}$	45944	-0.038 (0.007)	$2.64x10^{-7}$	35310	-0.04 (0.005)	$7.5x10^{-16}$	81254
1	FEV_1	rs2284746 (17179262), *MFAP2*(intron)	G	0.008 (0.007)	$2.78x10^{-1}$	45944	0.006 (0.007)	$3.7x10^{-1}$	35310	0.007 (0.005)	$1.48x10^{-1}$	81254
1	FEV_1/FVC	rs993925 (216926691), *TGFB2*(downstream)	T	0.04 (0.007)	$2.54x10^{-7}$	42402	0.023 (0.01)	$1.76x10^{-2}$	21162	0.034 (0.006)	$1.16x10^{-8}$	63564
1	FEV_1	rs993925 (216926691), *TGFB2*(downstream)	T	0.025 (0.007)	$1.51x10^{-3}$	42402	0.003 (0.007)	$7.29x10^{-1}$	21162	0.014 (0.005)	$8.71x10^{-3}$	63564
2	FEV_1/FVC	rs2544527 (15843619), *DDX1*(downstream)	T	-0.04 (0.007)	$1.08x10^{-7}$	45352	0 (0.01)	$9.75x10^{-1}$	21115	-0.026 (0.006)	$8.73x10^{-6}$	66467
2	FEV_1	rs2544527 (15843619), *DDX1*(downstream)	T	-0.024 (0.007)	$1.55x10^{-3}$	45352	-0.017 (0.007)	$1.95x10^{-2}$	21115	-0.021 (0.005)	$5.53x10^{-5}$	66467
2	FEV_1/FVC	rs3769124 (239014101), *ASB1*(intron)	G	-0.038 (0.01)	$1.95x10^{-4}$	44924	-0.032 (0.02)	$1.11x10^{-1}$	10579	-0.036 (0.009)	$2.83x10^{-5}$	55503
2	FEV_1	rs3769124 (239014101), *ASB1*(intron)	G	-0.053 (0.01)	$2.76x10^{-7}$	44924	-0.023 (0.02)	$2.44x10^{-1}$	10579	-0.047 (0.009)	$6.5x10^{-8}$	55503
2	FEV_1/FVC	rs12477314 (239542085), *HDAC4*(downstream)	T	0.052 (0.008)	$4.48x10^{-9}$	45585	0.031 (0.008)	$8.41x10^{-5}$	45704	0.041 (0.006)	$1.68x10^{-12}$	91289
2	FEV_1	rs12477314 (239542085), *HDAC4*(downstream)	T	0.032 (0.008)	$2.77x10^{-4}$	45585	0.025 (0.007)	$1.82x10^{-4}$	45704	0.028 (0.005)	$1.02x10^{-7}$	91289
3	FEV_1/FVC	rs1529672 (25495586), *RARB*(intron)	C	-0.06 (0.009)	$7.75x10^{-10}$	40624	-0.038 (0.009)	$1.16x10^{-5}$	45386	-0.048 (0.006)	$3.97x10^{-14}$	86010

Chr.	Measure	SNP_ID (NCBI36 position), function	Coded allele	Stage 1			Stage 2			Stage 1 + stage 2 meta-analysis		
				Beta (Se)	P	N	Beta (Se)	P	N	Beta (Se)	P	N
3	FEV_1	rs1529672 (25495586), *RARB*(intron)	C	-0.037 (0.009)	$1.78x10^{-4}$	40624	-0.011 (0.007)	$9.33x10^{-2}$	45386	-0.02 (0.006)	$2.16x10^{-4}$	86010
3	FEV_1/FVC	rs9310995 (32904119), *TRIM71*(intron)	T	0.017 (0.007)	$1.7x10^{-2}$	44835	-0.013 (0.009)	$1.6x10^{-1}$	21070	0.007 (0.006)	$2.36x10^{-1}$	65905
3	FEV_1	rs9310995 (32904119), *TRIM71*(intron)	T	0.035 (0.007)	$1.28x10^{-6}$	44835	0.009 (0.007)	$2x10^{-1}$	21070	0.023 (0.005)	$3.6x10^{-6}$	65905
3	FEV_1/FVC	rs1344555 (170782913), *MECOM*(intron)	T	-0.019 (0.008)	$2.61x10^{-2}$	46067	-0.017 (0.012)	$1.55x10^{-1}$	21104	-0.018 (0.007)	$6.65x10^{-3}$	67171
3	FEV_1	rs1344555 (170782913), *MECOM*(intron)	T	-0.042 (0.008)	$1.91x10^{-6}$	46067	-0.025 (0.009)	$6.44x10^{-3}$	21104	-0.034 (0.006)	$2.65x10^{-8}$	67171
4	FEV_1/FVC	rs1541374 (106267809), *TET2*(upstream)	T	-0.026 (0.007)	$5.56x10^{-4}$	45221	-0.014 (0.01)	$1.72x10^{-1}$	20516	-0.022 (0.006)	$2.05x10^{-4}$	65737
4	FEV_1	rs1541374 (106267809), *TET2*(upstream)	T	-0.036 (0.007)	$2.43x10^{-6}$	45221	-0.015 (0.007)	$4.36x10^{-2}$	20516	-0.026 (0.005)	$5.8x10^{-7}$	65737
5	FEV_1/FVC	rs1551943 (52230790), *ITGA1*(intron)	G	0.048 (0.008)	$1.2x10^{-8}$	43787	0.007 (0.008)	$3.71x10^{-1}$	45914	0.026 (0.006)	$2.43x10^{-6}$	89701
5	FEV_1	rs1551943 (52230790), *ITGA1*(intron)	G	0.022 (0.008)	$9.93x10^{-3}$	43787	-0.006 (0.006)	$3.53x10^{-1}$	45914	0.004 (0.005)	$3.61x10^{-1}$	89701
5	FEV_1/FVC	rs153916 (95062456), *SPATA9*(upstream)	T	-0.033 (0.007)	$2.06x10^{-6}$	47530	-0.025 (0.009)	$6.67x10^{-3}$	21428	-0.031 (0.005)	$2.12x10^{-8}$	68958
5	FEV_1	rs153916 (95062456), *SPATA9*(upstream)	T	-0.001 (0.007)	$8.91x10^{-1}$	47530	0.004 (0.007)	$6.22x10^{-1}$	21428	0.001 (0.005)	$8.2x10^{-1}$	68958
5	FEV_1/FVC	rs10067603 (131831767), *C5orf56*(downstream)	G	-0.04 (0.008)	$1.6x10^{-6}$	44134	-0.006 (0.011)	$6.03x10^{-1}$	21167	-0.028 (0.006)	$1.46x10^{-5}$	65301
5	FEV_1	rs10067603 (131831767), *C5orf56*(downstream)	G	-0.007 (0.008)	$3.83x10^{-1}$	44134	0.013 (0.008)	$1.14x10^{-1}$	21167	0.002 (0.006)	$6.74x10^{-1}$	65301
6	FEV_1/FVC	rs1928168 (22125717), *AK026189*(intron)	T	0.037 (0.007)	$8.99x10^{-8}$	47936	0.011 (0.009)	$2.4x10^{-1}$	21323	0.028 (0.005)	$1.69x10^{-7}$	69259
6	FEV_1	rs1928168 (22125717), *AK026189*(intron)	T	0.025 (0.007)	$2.61x10^{-4}$	47936	0.002 (0.007)	$7.69x10^{-1}$	21323	0.015 (0.005)	$2.25x10^{-3}$	69259

Chr.	Measure	SNP_ID (NCBI36 position), function	Coded allele	Stage 1			Stage 2			Stage 1 + stage 2 meta-analysis		
				Beta (Se)	P	N	Beta (Se)	P	N	Beta (Se)	P	N
6	FEV_1/FVC	rs6903823 (28430275), *ZKSCAN3*(intron)/*ZNF323*(intron)	G	-0.027 (0.008)	$2.28x10^{-3}$	47057	-0.013 (0.011)	$2.34x10^{-1}$	21428	-0.021 (0.007)	$1.19x10^{-3}$	68485
6	FEV_1	rs6903823 (28430275), *ZKSCAN3*(intron)/*ZNF323*(intron)	G	-0.046 (0.008)	$2x10^{-7}$	47057	-0.029 (0.008)	$4.75x10^{-4}$	21428	-0.037 (0.006)	$2.18x10^{-10}$	68485
6	FEV_1/FVC	rs3094548 (29463181), *OR12D2*(upstream)	G	-0.027 (0.008)	$1.15x10^{-3}$	42516	-0.015 (0.01)	$1.37x10^{-1}$	20733	-0.022 (0.006)	$3.39x10^{-4}$	63249
6	FEV_1	rs3094548 (29463181), *OR12D2*(upstream)	G	-0.042 (0.008)	$4.11x10^{-7}$	42516	-0.016 (0.008)	$3.6x10^{-2}$	20733	-0.029 (0.005)	$1.45x10^{-7}$	63249
6	FEV_1/FVC	rs2855812 (31580699), *MICB*(intron)	T	-0.034 (0.008)	$5.11x10^{-5}$	46921	-0.015 (0.011)	$1.59x10^{-1}$	21190	-0.027 (0.006)	$2.45x10^{-5}$	68111
6	FEV_1	rs2855812 (31580699), *MICB*(intron)	T	-0.045 (0.008)	$8.57x10^{-8}$	46921	-0.013 (0.008)	$1.06x10^{-1}$	21190	-0.03 (0.006)	$1.8x10^{-7}$	68111
6	FEV_1/FVC	rs2857595 (31676448), *NCR3*(upstream)	G	0.049 (0.009)	$7.86x10^{-8}$	45540	0.028 (0.008)	$5.36x10^{-4}$	45657	0.037 (0.006)	$2.28x10^{-10}$	91197
6	FEV_1	rs2857595 (31676448), *NCR3*(upstream)	G	0.04 (0.009)	$1.46x10^{-5}$	45540	0.017 (0.007)	$9.41x10^{-3}$	45657	0.025 (0.005)	$1.3x10^{-6}$	91197
6	FEV_1/FVC	rs2647044 (32775888), *HLA-DQB1*(upstream)	G	0.053 (0.011)	$2.71x10^{-6}$	44610	0.007 (0.022)	$7.63x10^{-1}$	8381	0.044 (0.01)	$5.95x10^{-6}$	52991
6	FEV_1	rs2647044 (32775888), *HLA-DQB1*(upstream)	G	0.031 (0.011)	$6.71x10^{-3}$	44610	0.009 (0.022)	$6.71x10^{-1}$	8381	0.027 (0.01)	$5.62x10^{-3}$	52991
6	FEV_1/FVC	rs2798641 (109374743), *ARMC2*(intron)	T	-0.047 (0.009)	$2.81x10^{-7}$	46369	-0.03 (0.012)	$1.57x10^{-2}$	20999	-0.041 (0.007)	$8.35x10^{-9}$	67368
6	FEV_1	rs2798641 (109374743), *ARMC2*(intron)	T	-0.046 (0.009)	$5.39x10^{-7}$	46369	-0.009 (0.01)	$3.35x10^{-1}$	20999	-0.03 (0.006)	$4.69x10^{-6}$	67368
6	FEV_1/FVC	rs3734729 (150612560), *PPP1R14C*(untranslated-3)	G	-0.045 (0.017)	$8.71x10^{-3}$	43680	-0.058 (0.023)	$1x10^{-2}$	20998	-0.05 (0.013)	$1.93x10^{-4}$	64678
6	FEV_1	rs3734729 (150612560), *PPP1R14C*(untranslated-3)	G	-0.085 (0.016)	$1.08x10^{-6}$	43680	-0.021 (0.017)	$2.24x10^{-1}$	20998	-0.055 (0.012)	$4.48x10^{-6}$	64678
10	FEV_1/FVC	rs1878798 (12283489), *CDC123*(intron)	G	0.042 (0.007)	$3.48x10^{-9}$	46164	0.024 (0.009)	$1.15x10^{-2}$	21086	0.035 (0.005)	$9.56x10^{-11}$	67250

Chr.	Measure	SNP_ID (NCBI36 position), function	Coded allele	Stage 1			Stage 2			Stage 1 + stage 2 meta-analysis		
				Beta (Se)	P	N	Beta (Se)	P	N	Beta (Se)	P	N
10	FEV_1	rs1878798 (12283489), *CDC123*(intron)	G	0.042 (0.007)	$3.11x10^{-9}$	46164	0.015 (0.007)	$3.65x10^{-2}$	21086	0.029 (0.005)	$1.84x10^{-9}$	67250
10	FEV_1/FVC	rs7068966 (12317998), *CDC123*(intron)	T	0.045 (0.007)	$1.28x10^{-10}$	47085	0.023 (0.006)	$3.86x10^{-4}$	45892	0.033 (0.005)	$6.13x10^{-13}$	92977
10	FEV_1	rs7068966 (12317998), *CDC123*(intron)	T	0.04 (0.007)	$1.19x10^{-8}$	47085	0.022 (0.005)	$3.56x10^{-5}$	45892	0.029 (0.004)	$2.82x10^{-12}$	92977
10	FEV_1/FVC	rs11001819 (77985230), *C10orf11*(intron)	G	-0.019 (0.007)	$6.5x10^{-3}$	45546	-0.006 (0.006)	$3.17x10^{-1}$	45677	-0.012 (0.005)	$7.58x10^{-3}$	91223
10	FEV_1	rs11001819 (77985230), *C10orf11*(intron)	G	-0.041 (0.007)	$1.42x10^{-8}$	45546	-0.022 (0.005)	$3.1x10^{-5}$	45677	-0.029 (0.004)	$2.98x10^{-12}$	91223
12	FEV_1/FVC	rs4762767 (19757396), *AEBP2*(downstream)	G	-0.036 (0.007)	$2.42x10^{-6}$	48016	-0.008 (0.011)	$4.47x10^{-1}$	21324	-0.027 (0.006)	$8.15x10^{-6}$	69340
12	FEV_1	rs4762767 (19757396), *AEBP2*(downstream)	G	-0.028 (0.007)	$3.85x10^{-4}$	48016	-0.012 (0.008)	$1.34x10^{-1}$	21324	-0.021 (0.005)	$1.52x10^{-4}$	69340
12	FEV_1/FVC	rs11172113 (55813550), *LRP1*(intron)	T	-0.035 (0.007)	$1.36x10^{-6}$	45387	-0.026 (0.01)	$5.83x10^{-3}$	20256	-0.032 (0.006)	$1.24x10^{-8}$	65643
12	FEV_1	rs11172113 (55813550), *LRP1*(intron)	T	-0.021 (0.007)	$3.55x10^{-3}$	45387	-0.003 (0.007)	$6.94x10^{-1}$	20256	-0.013 (0.005)	$1.19x10^{-2}$	65643
12	FEV_1/FVC	rs1036429 (94795559), *CCDC38*(intron)	T	0.049 (0.008)	$1.24x10^{-8}$	47814	0.028 (0.008)	$3.35x10^{-4}$	46183	0.038 (0.006)	$2.3x10^{-11}$	93997
12	FEV_1	rs1036429 (94795559), *CCDC38*(intron)	T	0.01 (0.008)	$2.67x10^{-1}$	47814	0.004 (0.006)	$5.38x10^{-1}$	46183	0.006 (0.005)	$2.26x10^{-1}$	93997
15	FEV_1/FVC	rs2036527 (76638670), *CHRNA5*(uptream)	G	0.032 (0.007)	$1.19x10^{-5}$	45038	0 (0.01)	$9.82x10^{-1}$	20874	0.022 (0.006)	$1.81x10^{-4}$	65912
15	FEV_1	rs2036527 (76638670), *CHRNA5*(uptream)	G	0.036 (0.007)	$2.4x10^{-6}$	45038	0.015 (0.008)	$5.44x10^{-2}$	20874	0.026 (0.005)	$6.9x10^{-7}$	65912
15	FEV_1/FVC	rs12914385 (76685778), *CHRNA3*(intron)	T	-0.03 (0.007)	$2.28x10^{-5}$	47226	0.002 (0.01)	$8.08x10^{-1}$	21327	-0.019 (0.006)	$5.17x10^{-4}$	68553
15	FEV_1	rs12914385 (76685778), *CHRNA3*(intron)	T	-0.034 (0.007)	$2.95x10^{-6}$	47226	-0.015 (0.007)	$4.1x10^{-2}$	21327	-0.025 (0.005)	$4.72x10^{-7}$	68553

Chr.	Measure	SNP_ID (NCBI36 position), function	Coded allele	Stage 1			Stage 2			Stage 1 + stage 2 meta-analysis		
				Beta (Se)	P	N	Beta (Se)	P	N	Beta (Se)	P	N
15	FEV_1/FVC	rs8040868 (76698236), *CHRNA3*(s)	T	0.04 (0.008)	$1.14x10^{-6}$	35121	-0.005 (0.01)	$6.1x10^{-1}$	21131	0.022 (0.006)	$3.01x10^{-4}$	56252
15	FEV_1	rs8040868 (76698236), *CHRNA3*(s)	T	0.039 (0.008)	$2.98x10^{-6}$	35121	0.012 (0.007)	$9.86x10^{-2}$	21131	0.025 (0.005)	$4.06x10^{-6}$	56252
16	FEV_1/FVC	rs12447804 (56632783), *MMP15*(intron)	T	-0.053 (0.009)	$7.12x10^{-8}$	35123	-0.021 (0.01)	$4.2x10^{-2}$	23693	-0.038 (0.007)	$3.59x10^{-8}$	58816
16	FEV_1	rs12447804 (56632783), *MMP15*(intron)	T	-0.017 (0.009)	$8.02x10^{-2}$	35123	0.004 (0.007)	$5.71x10^{-1}$	23693	-0.004 (0.006)	$4.73x10^{-1}$	58816
16	FEV_1/FVC	rs3743563 (56636666), *MMP15*(missense)	G	0.043 (0.008)	$1.8x10^{-7}$	47179	0.013 (0.008)	$1.22x10^{-1}$	43190	0.028 (0.006)	$6.76x10^{-7}$	90369
16	FEV_1	rs3743563 (56636666), *MMP15*(missense)	G	0.015 (0.008)	$8.52x10^{-2}$	47179	-0.001 (0.007)	$8.74x10^{-1}$	43190	0.006 (0.005)	$2.79x10^{-1}$	90369
16	FEV_1/FVC	rs2865531 (73947817), *CFDP1*(intron)	T	0.039 (0.007)	$2.3x10^{-8}$	47594	0.024 (0.006)	$1.94x10^{-4}$	46286	0.031 (0.005)	$1.77x10^{-11}$	93880
16	FEV_1	rs2865531 (73947817), *CFDP1*(intron)	T	0.024 (0.007)	$6.3x10^{-4}$	47594	0.011 (0.005)	$3.89x10^{-2}$	46286	0.016 (0.004)	$1.09x10^{-4}$	93880
16	FEV_1/FVC	rs12716852 (76746239), *WWOX*(intron)	G	0.011 (0.007)	$1.24x10^{-1}$	47510	-0.004 (0.009)	$6.85x10^{-1}$	21228	0.006 (0.005)	$2.81x10^{-1}$	68738
16	FEV_1	rs12716852 (76746239), *WWOX*(intron)	G	0.036 (0.007)	$3.45x10^{-7}$	47510	0.013 (0.007)	$7.11x10^{-2}$	21228	0.025 (0.005)	$1.92x10^{-7}$	68738
21	FEV_1/FVC	rs9978142 (34574109), *KCNE2*(upstream)	T	-0.048 (0.009)	$8.23x10^{-7}$	44577	-0.031 (0.013)	$1.75x10^{-2}$	20693	-0.043 (0.008)	$2.65x10^{-8}$	65270
21	FEV_1	rs9978142 (34574109), *KCNE2*(upstream)	T	-0.012 (0.009)	$2.47x10^{-1}$	44577	-0.015 (0.01)	$1.35x10^{-1}$	20693	-0.013 (0.007)	$5.57x10^{-2}$	65270

Association of loci influencing lung function with FEV_1 and FEV_1/FVC in children

Effects of the 16 novel SNPs, and effects of SNPs in regions previously reported as associated with lung function, on FEV_1 and FEV_1/FVC in children were looked up in ALSPAC and Raine. To enable a comparison of effect sizes between children and adults, effect sizes in the SpiroMeta-CHARGE stage 2 dataset only (to avoid potential winners' curse bias) are given for the novel loci. Effect sizes in the SpiroMeta-CHARGE GWAS stage 1 are provided for the previously reported regions. For each loci the direction of effects were compared using the most significant SNP in the SpiroMeta-CHARGE dataset across both traits. Abbreviations: ns=nonsynonymous, s= synonymous

Chr.	Measure	SNP_ID(NCBI36 position), function		ALSPAC+Raine meta-analysis			SpiroMeta-CHARGE		
			Coded allele	Beta	Se	P	Beta	Se	P
Novel loci									
1	FEV_1/FVC	rs2284746 (17179262), *MFAP2*(intron)	G	0.004	0.023	$8.6x10^{-1}$	-0.038	0.007	$2.64x10^{-7}$
1	FEV_1	rs2284746 (17179262), *MFAP2*(intron)	G	-0.013	0.024	$5.93x10^{-1}$	0.006	0.007	$3.7x10^{-1}$
1	FEV_1/FVC	rs993925 (216926691), *TGFB2*(downstream)	T	0.043	0.025	$8.9x10^{-2}$	0.023	0.01	$1.76x10^{-2}$
1	FEV_1	rs993925 (216926691), *TGFB2*(downstream)	T	0.039	0.026	$1.26x10^{-1}$	0.003	0.007	$7.29x10^{-1}$
2	FEV_1/FVC	rs12477314 (239542085), *HDAC4*(downstream)	T	0.083	0.028	$4x10^{-3}$	0.031	0.008	$8.41x10^{-5}$
2	FEV_1	rs12477314 (239542085), *HDAC4*(downstream)	T	0.037	0.029	$2.03x10^{-1}$	0.025	0.007	$1.82x10^{-4}$
3	FEV_1/FVC	rs1529672 (25495586), *RARB*(intron)	C	-0.064	0.03	$3.1x10^{-2}$	-0.038	0.009	$1.16x10^{-5}$
3	FEV_1	rs1529672 (25495586), *RARB*(intron)	C	0.033	0.03	$2.76x10^{-1}$	-0.011	0.007	$9.33x10^{-2}$
3	FEV_1/FVC	rs1344555 (170782913), *MECOM*(intron)	T	-0.01	0.029	$7.43x10^{-1}$	-0.017	0.012	$1.55x10^{-1}$
3	FEV_1	rs1344555 (170782913), *MECOM*(intron)	T	-0.03	0.03	$3.2x10^{-1}$	-0.025	0.009	$6.44x10^{-3}$
5	FEV_1/FVC	rs153916 (95062456), *SPATA9*(upstream)	T	0.032	0.022	$1.6x10^{-1}$	-0.025	0.009	$6.67x10^{-3}$
5	FEV_1	rs153916 (95062456), *SPATA9*(upstream)	T	0.017	0.023	$4.64x10^{-1}$	0.004	0.007	$6.22x10^{-1}$
6	FEV_1/FVC	rs6903823 (28430275), *ZKSCAN3*(intron)/*ZNF323*(intron)	G	-0.065	0.027	$1.4x10^{-2}$	-0.013	0.011	$2.34x10^{-1}$

Chr.	Measure	SNP_ID(NCBI36 position), function		ALSPAC+Raine meta-analysis			SpiroMeta-CHARGE		
			Coded allele	Beta	Se	P	Beta	Se	P
6	FEV_1	rs6903823 (28430275), *ZKSCAN3*(intron)/*ZNF323*(intron)	G	0.002	0.028	$9.51x10^{-1}$	-0.029	0.008	$4.75x10^{-4}$
6	FEV_1/FVC	rs2857595 (31676448), *NCR3*(upstream)	G	0.055	0.028	$4.9x10^{-2}$	0.028	0.008	$5.36x10^{-4}$
6	FEV_1	rs2857595 (31676448), *NCR3*(upstream)	G	-0.034	0.029	$2.41x10^{-1}$	0.017	0.007	$9.41x10^{-3}$
6	FEV_1/FVC	rs2798641 (109374743), *ARMC2*(intron)	T	-0.053	0.03	$7.3x10^{-2}$	-0.03	0.012	$1.57x10^{-2}$
6	FEV_1	rs2798641 (109374743), *ARMC2*(intron)	T	-0.097	0.03	$1x10^{-3}$	-0.009	0.01	$3.35x10^{-1}$
10	FEV_1/FVC	rs7068966 (12317998), *CDC123*(intron)	T	0.026	0.022	$2.45x10^{-1}$	0.023	0.006	$3.86x10^{-4}$
10	FEV_1	rs7068966 (12317998), *CDC123*(intron)	T	0.042	0.024	$7.4x10^{-2}$	0.022	0.005	$3.56x10^{-5}$
10	FEV_1/FVC	rs11001819 (77985230), *C10orf11*(intron)	G	-0.038	0.023	$9.7x10^{-2}$	-0.006	0.006	$3.17x10^{-1}$
10	FEV_1	rs11001819 (77985230), *C10orf11(*intron)	G	-0.026	0.024	$2.78x10^{-1}$	-0.022	0.005	$3.1x10^{-5}$
12	FEV_1/FVC	rs11172113 (55813550), *LRP1*(intron)	T	-0.025	0.023	$2.85x10^{-1}$	-0.026	0.01	$5.83x10^{-3}$
12	FEV_1	rs11172113 (55813550), *LRP1*(intron)	T	-0.037	0.024	$1.21x10^{-1}$	-0.003	0.007	$6.94x10^{-1}$
12	FEV_1/FVC	rs1036429 (94795559), *CCDC38*(intron)	T	0.05	0.028	$7.6x10^{-2}$	0.028	0.008	$3.35x10^{-4}$
12	FEV_1	rs1036429 (94795559), *CCDC38*(intron)	T	-0.01	0.03	$7.43x10^{-1}$	0.004	0.006	$5.38x10^{-1}$
16	FEV_1/FVC	rs12447804 (56632783), *MMP15*(intron)	T	-0.017	0.028	$5.52x10^{-1}$	-0.021	0.01	$4.2x10^{-2}$
16	FEV_1	rs12447804 (56632783), *MMP15*(intron)	T	-0.053	0.028	$6.1x10^{-2}$	0.004	0.007	$5.71x10^{-1}$
16	FEV_1/FVC	rs2865531 (73947817), *CFDP1*(intron)	T	-0.004	0.023	$8.67x10^{-1}$	0.024	0.006	$1.94x10^{-4}$
16	FEV_1	rs2865531 (73947817), *CFDP1*(intron)	T	-0.046	0.024	$5x10^{-2}$	0.011	0.005	$3.89x10^{-2}$
21	FEV_1/FVC	rs9978142 (34574109), *KCNE2*(upstream)	T	0.018	0.033	$5.9x10^{-1}$	-0.031	0.013	$1.75x10^{-2}$
21	FEV_1	rs9978142 (34574109), *KCNE2*(upstream)	T	0.008	0.034	$8.13x10^{-1}$	-0.015	0.01	$1.35x10^{-1}$
Previously reported regions									
2	FEV_1/FVC	rs2571445(218391399), *TNS1* (ns)	G	0.011	0.023	$6.31x10^{-1}$	0.033	0.007	$4.46x10^{-6}$
2	FEV_1	rs2571445(218391399), *TNS1*(ns)	G	0.048	0.024	$4.25x10^{-2}$	0.047	0.007	$9.83x10^{-11}$

Chr.	Measure	SNP_ID(NCBI36 position), function		ALSPAC+Raine meta-analysis			SpiroMeta-CHARGE		
			Coded allele	Beta	Se	P	Beta	Se	P
2	FEV_1/FVC	rs10498230(229210747), *PID1*(downstream)	T	0.026	0.043	$5.44x10^{-1}$	0.068	0.014	$1.13x10^{-6}$
2	FEV_1	rs10498230(229210747), *PID1*(downstream)	T	0.07	0.044	$1.13x10^{-1}$	0.03	0.014	$3.6x10^{-2}$
4	FEV_1/FVC	rs2045517(90089987), *FAM13A*(intron)	T	-0.007	0.023	$7.78x10^{-1}$	-0.047	0.007	$2x10^{-11}$
4	FEV_1	rs2045517(90089987), *FAM13A*(intron)	T	-0.028	0.024	$2.43x10^{-1}$	-0.012	0.007	$8.93x10^{-2}$
4	FEV_1/FVC	rs7671167(90103002), *FAM13A*(intron)	T	0.009	0.023	$6.85x10^{-1}$	-0.042	0.007	$1.27x10^{-9}$
4	FEV_1	rs7671167(90103002), *FAM13A*(intron)	T	-0.022	0.023	$3.44x10^{-1}$	-0.017	0.007	$1.64x10^{-2}$
4	FEV_1/FVC	rs10516526(106908353), *GSTCD*(intron)	G	0.106	0.045	$1.98x10^{-2}$	0.039	0.014	$6.17x10^{-3}$
4	FEV_1	rs10516526(106908353), *GSTCD*(intron)	G	0.102	0.047	$2.81x10^{-2}$	0.108	0.014	$4.75x10^{-14}$
4	FEV_1/FVC	rs17331332(107027556), *NPNT*(upstream)	G	-0.081	0.045	$7.14x10^{-2}$	-0.057	0.014	$5.3x10^{-5}$
4	FEV_1	rs17331332(107027556), *NPNT*(upstream)	G	-0.108	0.046	$1.79x10^{-2}$	-0.102	0.014	$1.11x10^{-12}$
4	FEV_1/FVC	rs6823809(107048244), *NPNT*(intron)	T	0.112	0.036	$2.18x10^{-3}$	0.056	0.011	$2.2x10^{-7}$
4	FEV_1	rs6823809(107048244), *NPNT*(intron)	T	0.052	0.038	$1.64x10^{-1}$	0.05	0.011	$4.82x10^{-6}$
4	FEV_1/FVC	rs1032296(145654138), *HHIP*(upstream)	T	-0.004	0.024	$8.68x10^{-1}$	-0.05	0.007	$3.42x10^{-12}$
4	FEV_1	rs1032296(145654138), *HHIP*(upstream)	T	-0.004	0.024	$8.57x10^{-1}$	-0.047	0.007	$8.74x10^{-11}$
4	FEV_1/FVC	rs11100860(145698589), *HHIP*(upstream)	G	0.004	0.024	$8.68x10^{-1}$	0.064	0.007	$6.81x10^{-20}$
4	FEV_1	rs11100860(145698589), *HHIP*(upstream)	G	0.01	0.024	$6.82x10^{-1}$	0.041	0.007	$4.27x10^{-9}$
5	FEV_1/FVC	rs11168048(147822546), *HTR4*(intron)	T	-0.044	0.024	$6.08x10^{-2}$	-0.047	0.007	$5.97x10^{-11}$
5	FEV_1	rs11168048(147822546), *HTR4*(intron)	T	0.018	0.024	$4.7x10^{-1}$	-0.046	0.007	$2.43x10^{-10}$
5	FEV_1/FVC	rs3995090(147826008), *HTR4*(intron)	C	0.049	0.023	$3.12x10^{-2}$	0.046	0.007	$1.04x10^{-10}$
5	FEV_1	rs3995090(147826008), *HTR4*(intron)	C	-0.014	0.023	$5.41x10^{-1}$	0.045	0.007	$3.33x10^{-10}$
5	FEV_1/FVC	rs1985524(147827981), *HTR4*(intron)	G	-0.043	0.023	$6.1x10^{-2}$	-0.045	0.007	$2.9x10^{-10}$
5	FEV_1	rs1985524(147827981), *HTR4*(intron)	G	0.015	0.024	$5.35x10^{-1}$	-0.048	0.007	$3.06x10^{-11}$

Chr.	Measure	SNP_ID(NCBI36 position), function		ALSPAC+Raine meta-analysis			SpiroMeta-CHARGE		
			Coded allele	Beta	Se	P	Beta	Se	P
5	FEV_1/FVC	rs11134779(156869344), *ADAM19*(intron)	G	-0.003	0.024	$9.06x10^{-1}$	-0.042	0.007	$6.01x10^{-9}$
5	FEV_1	rs11134779(156869344), *ADAM19*(intron)	G	-0.021	0.024	$3.93x10^{-1}$	-0.027	0.007	$2.4x10^{-4}$
6	FEV_1/FVC	rs2070600(32259421), *AGER*(ns)	T	0.146	0.045	$1.15x10^{-3}$	0.126	0.016	$9.07x10^{-15}$
6	FEV_1	rs2070600(32259421), *AGER*(ns)	T	0.063	0.046	$1.75x10^{-1}$	0.025	0.016	$1.27x10^{-1}$
6	FEV_1/FVC	rs3817928(142792209), *GPR126*(intron)	G	0.086	0.029	$2.75x10^{-3}$	0.059	0.008	$2.27x10^{-12}$
6	FEV_1	rs3817928(142792209), *GPR126*(intron)	G	-0.011	0.029	$7.11x10^{-1}$	0.023	0.009	$8.63x10^{-3}$
6	FEV_1/FVC	rs262129(142894837), *LOC153910*(unknown)	G	0.098	0.025	$8.19x10^{-5}$	0.056	0.008	$2.91x10^{-13}$
6	FEV_1	rs262129(142894837), *LOC153910*(unknown)	G	-0.008	0.026	$7.61x10^{-1}$	0.031	0.008	$5.44x10^{-5}$
9	FEV_1/FVC	rs16909859(97244613), *PTCH1*(downstream)	G	0.049	0.044	$2.66x10^{-1}$	0.08	0.013	$7.45x10^{-10}$
9	FEV_1	rs16909859(97244613), *PTCH1*(downstream)	G	0.013	0.045	$7.8x10^{-1}$	-0.014	0.013	$2.93x10^{-1}$
9	FEV_1/FVC	rs16909898(97270829), *PTCH1*(intron)	G	-0.068	0.042	$1.06x10^{-1}$	-0.072	0.012	$3.94x10^{-9}$
9	FEV_1	rs16909898(97270829), *PTCH1*(intron)	G	0.006	0.043	$8.85x10^{-1}$	0.015	0.012	$2.21x10^{-1}$
15	FEV_1/FVC	rs12899618(69432174), *THSD4*(intron)	G	0.06	0.032	$6.03x10^{-2}$	0.076	0.01	$1.86x10^{-15}$
15	FEV_1	rs12899618(69432174), *THSD4*(intron)	G	-0.026	0.033	$4.39x10^{-1}$	0.036	0.01	$1.57x10^{-4}$
15	FEV_1/FVC	rs8033889(69467134), *THSD4*(intron)	T	-0.067	0.028	$1.7x10^{-2}$	-0.072	0.008	$2.03x10^{-17}$
15	FEV_1	rs8033889(69467134), *THSD4*(intron)	T	0.051	0.029	$7.41x10^{-2}$	-0.044	0.009	$3.01x10^{-7}$
15	FEV_1/FVC	rs2568494(76528019), *IREB2*(intron)	G	0.018	0.024	$4.48x10^{-1}$	0.029	0.007	$5.25x10^{-5}$
15	FEV_1	rs2568494(76528019), *IREB2*(intron)	G	-0.011	0.025	$6.68x10^{-1}$	0.023	0.007	$1.64x10^{-3}$
15	FEV_1/FVC	rs8034191(76593078), *CHRNA3/5*(intron)	T	0.014	0.024	$5.47x10^{-1}$	0.032	0.007	$9.65x10^{-6}$
15	FEV_1	rs8034191(76593078), *CHRNA3/5*(intron)	T	-0.017	0.024	$4.75x10^{-1}$	0.031	0.007	$2.07x10^{-5}$
15	FEV_1/FVC	rs2036527(76638670), *CHRNA5*(uptream)	G	0.023	0.025	$3.66x10^{-1}$	0.032	0.007	$1.19x10^{-5}$
15	FEV_1	rs2036527(76638670), *CHRNA5*(uptream)	G	0.002	0.026	$9.51x10^{-1}$	0.036	0.008	$2.4x10^{-6}$

Chr.	Measure	SNP_ID(NCBI36 position), function		ALSPAC+Raine meta-analysis			SpiroMeta-CHARGE		
			Coded allele	Beta	Se	P	Beta	Se	P
15	FEV_1/FVC	rs8040868(76698236), *CHRNA3*(s)	T	0.019	0.023	$4.25x10^{-1}$	0.04	0.008	$1.14x10^{-6}$
15	FEV_1	rs8040868(76698236), *CHRNA3*(s)	T	0.016	0.024	$5.09x10^{-1}$	0.039	0.008	$2.98x10^{-6}$

Association of loci influencing lung function with height

Effects of the 16 novel SNPs, and previously reported SNPs, associated with lung function on height were looked up in the GIANT dataset. Both effect sizes for lung and for height can be interpreted as proportion of a standard deviation. Results for the SpiroMeta-CHARGE stage 1 and stage 2 meta-analysis for the lung function measure that showed stronger association are reported for the novel loci. Results from the GWAS Stage 1 for the lung function measure that showed stronger association are reported for the previously reported loci. For each loci the direction of effects were compared using the most significant SNP in the SpiroMeta-CHARGE dataset across both traits. Abbreviations: Chr.= chromosome, ns=nonsynonymous, s= synonymous.

Chr.	SNP ID(NCBI36 position), function	Coded allele	Measure	Lung function (Stage 1+ stage 2 meta-analysis)			Height (GIANT consortium)		
				Beta	Se	P	Beta	Se	P
Novel loci									
1	rs2284746 (17179262), *MFAP2*(intron)	G	FEV_1/FVC	-0.04	0.005	$7.5x10^{-16}$	0.0354	0.0045	$5.64x10^{-15}$
1	rs993925 (216926691), *TGFB2*(downstream)	T	FEV_1/FVC	0.034	0.006	$1.16x10^{-8}$	0.0105	0.005	$3.61x10^{-2}$
2	rs12477314 (239542085), *HDAC4*(downstream)	T	FEV_1/FVC	0.041	0.006	$1.68x10^{-12}$	-0.0029	0.0057	$6.12x10^{-1}$
3	rs1529672 (25495586), *RARB*(intron)	C	FEV_1/FVC	-0.048	0.006	$3.97x10^{-14}$	0.0012	0.0063	$8.49x10^{-1}$
3	rs1344555 (170782913), *MECOM*(intron)	T	FEV_1	-0.034	0.006	$2.65x10^{-8}$	-0.0145	0.0056	$9.68x10^{-3}$
5	rs153916 (95062456), *SPATA9*(upstream)	T	FEV_1/FVC	-0.031	0.005	$2.12x10^{-8}$	0.0027	0.0045	$5.51x10^{-1}$
6	rs6903823 (28430275), *ZKSCAN3*(intron)/*ZNF323*(intron)	G	FEV_1	-0.037	0.006	$2.18x10^{-10}$	-0.0017	0.0056	$7.62x10^{-1}$
6	rs2857595 (31676448), *NCR3*(upstream)	G	FEV_1/FVC	0.037	0.006	$2.28x10^{-10}$	-0.0148	0.006	$1.31x10^{-2}$
6	rs2798641 (109374743), *ARMC2*(intron)	T	FEV_1/FVC	-0.041	0.007	$8.35x10^{-9}$	-0.0042	0.0058	$4.72x10^{-1}$
10	rs7068966 (12317998), *CDC123*(intron)	T	FEV_1/FVC	0.033	0.005	$6.13x10^{-13}$	0.0078	0.0045	$8.52x10^{-2}$
10	rs11001819 (77985230), *C10orf11*(intron)	G	FEV_1	-0.029	0.004	$2.98x10^{-12}$	0.0024	0.0045	$5.96x10^{-1}$
12	rs11172113 (55813550), *LRP1*(intron)	T	FEV_1/FVC	-0.032	0.006	$1.24x10^{-8}$	0.003	0.0047	$5.19x10^{-1}$
12	rs1036429 (94795559), *CCDC38*(intron)	T	FEV_1/FVC	0.038	0.006	$2.3x10^{-11}$	-0.0053	0.0056	$3.44x10^{-1}$

Chr.	SNP ID(NCBI36 position), function	Coded allele	Measure	Lung function (Stage 1+ stage 2 meta-analysis)			Height (GIANT consortium)		
				Beta	Se	P	Beta	Se	P
16	rs12447804 (56632783), *MMP15*(intron)	T	FEV_1/FVC	-0.038	0.007	$3.59x10^{-8}$	0.0077	0.0075	$3.05x10^{-1}$
16	rs2865531 (73947817), *CFDP1*(intron)	T	FEV_1/FVC	0.031	0.005	$1.77x10^{-11}$	-0.0129	0.0045	$4.42x10^{-3}$
21	rs9978142 (34574109), *KCNE2*(upstream)	T	FEV_1/FVC	-0.043	0.008	$2.65x10^{-8}$	-0.0122	0.0062	$4.91x10^{-2}$
Previously reported loci									
2	rs2571445 (218391399), *TNS1*(ns)	G	FEV_1	0.047	0.007	$9.83x10^{-11}$	-0.0032	0.0047	$4.91x10^{-1}$
2	rs10498230 (229210747), *PID1*(downstream)	T	FEV_1/FVC	0.068	0.014	$1.13x10^{-6}$	-0.0111	0.0087	$2.02x10^{-1}$
4	rs2045517 (90089987), *FAM13A*(intron)	T	FEV_1/FVC	-0.047	0.007	$2x10^{-11}$	0.0058	0.0045	$2.01x10^{-1}$
4	rs7671167 (90103002), *FAM13A*(intron)	T	FEV_1/FVC	-0.042	0.007	$1.27x10^{-9}$	0.0072	0.0045	$1.12x10^{-1}$
4	rs10516526 (106908353), *GSTCD*(intron)	G	FEV_1	0.108	0.014	$4.75x10^{-14}$	0.0032	0.0092	$7.28x10^{-1}$
4	rs17331332 (107027556), *NPNT*(upstream)	G	FEV_1	-0.102	0.014	$1.11x10^{-12}$	-0.002	0.0093	$8.3x10^{-1}$
4	rs6823809 (107048244), *NPNT*(intron)	T	FEV_1/FVC	0.056	0.011	$2.2x10^{-7}$	-0.0001	0.0072	$9.89x10^{-1}$
4	rs1032296 (145654138), *HHIP*(upstream)	T	FEV_1/FVC	-0.05	0.007	$3.42x10^{-12}$	-0.0152	0.0047	$1.08x10^{-3}$
4	rs11100860 (145698589), *HHIP*(upstream)	G	FEV_1/FVC	0.064	0.007	$6.81x10^{-20}$	0.0151	0.0045	$8.62x10^{-4}$
5	rs11168048 (147822546), *HTR4*(intron)	T	FEV_1/FVC	-0.047	0.007	$5.97x10^{-11}$	-0.0133	0.0049	$6.52x10^{-3}$
5	rs3995090 (147826008), *HTR4*(intron)	C	FEV_1/FVC	0.046	0.007	$1.04x10^{-10}$	0.0143	0.0049	$3.45x10^{-3}$
5	rs1985524 (147827981), *HTR4*(intron)	G	FEV_1	-0.048	0.007	$3.06x10^{-11}$	-0.015	0.0049	$2.16x10^{-3}$
5	rs11134779 (156869344), *ADAM19*(intron)	G	FEV_1/FVC	-0.042	0.007	$6.01x10^{-9}$	-0.0092	0.0047	$4.79x10^{-2}$
6	rs2070600 (32259421), *AGER*(ns)	T	FEV_1/FVC	0.126	0.016	$9.07x10^{-15}$	0.0094	0.0114	$4.12x10^{-1}$
6	rs3817928 (142792209), *GPR126*(intron)	G	FEV_1/FVC	0.059	0.008	$2.27x10^{-12}$	-0.0368	0.0055	$1.97x10^{-11}$
6	rs262129 (142894837), *LOC153910*(unknown)	G	FEV_1/FVC	0.056	0.008	$2.91x10^{-13}$	-0.0443	0.005	$9.17x10^{-19}$
9	rs16909859 (97244613), *PTCH*1(downstream)	G	FEV_1/FVC	0.08	0.013	$7.45x10^{-10}$	-0.0266	0.0082	$1.23x10^{-3}$
9	rs16909898 (97270829), *PTCH1*(intron)	G	FEV_1/FVC	-0.072	0.012	$3.94x10^{-9}$	0.0313	0.0078	$5.39x10^{-5}$

Chr.	SNP ID(NCBI36 position), function	Coded allele	Measure	Lung function (Stage 1+ stage 2 meta-analysis)			Height (GIANT consortium)		
				Beta	Se	P	Beta	Se	P
15	rs12899618 (69432174), *THSD4*(intron)	G	FEV_1/FVC	0.076	0.01	$1.86x10^{-15}$	-0.0075	0.0061	$2.18x10^{-1}$
15	rs8033889 (69467134), *THSD4*(intron)	T	FEV_1/FVC	-0.072	0.008	$2.03x10^{-17}$	0.0015	0.0054	$7.8x10^{-1}$
15	rs2568494 (76528019), *IREB2*(intron)	G	FEV_1/FVC	0.029	0.007	$5.25x10^{-5}$	0.0021	0.0047	$6.52x10^{-1}$
15	rs8034191 (76593078), *CHRNA3/5*(intron)	T	FEV_1/FVC	0.032	0.007	$9.65x10^{-6}$	-0.0002	0.0047	$9.66x10^{-1}$
15	rs2036527 (76638670), *CHRNA5*(uptream)	G	FEV_1	0.036	0.008	$2.4x10^{-6}$	-0.0005	0.0048	$9.17x10^{-1}$
15	rs8040868 (76698236), *CHRNA3*(s)	T	FEV_1/FVC	0.04	0.008	$1.14x10^{-6}$	-0.006	0.0064	$3.52x10^{-1}$

Association of loci influencing lung function with ever smoking status and number of cigarettes per day

Effects of the 16 novel SNPs, and previously reported SNPs, on two smoking phenotypes (ever-smokers vs. never-smokers and number of cigarettes per day) were looked up in the Oxford-GlaxoSmithKline (Ox-GSK) study, a collaborative effort to investigate the genetic basis of smoking-related behavioral traits. Results for the SpiroMeta-CHARGE joint meta-analysis of stage 1 and stage 2 for the lung function measure that showed stronger association are reported for the novel loci. Results for the SpiroMeta-CHARGE GWAS stage (stage 1) for the lung function measure that showed stronger association are reported for the previously reported loci. Abbreviations: Chr.=chromosome, ns=nonsynonymous, s= synonymous

				Lung function (stage 1 and stage 2 meta-analysis)			**Cigarettes per day**			**Ever vs. Never smoking**		
Chr.	**SNP ID(NCBI36 position), function**	**Coded allele**	**Measure**	**Beta**	**Se**	**P**	**Beta**	**Se**	**P**	**Beta**	**Se**	**P**
Novel loci												
1	rs2284746 (17179262), *MFAP2*(intron)	G	FEV_1/FVC	-0.04	0.005	$7.5x10^{-16}$	0.004	0.01	$6.85x10^{-1}$	-0.015	0.018	$3.99x10^{-1}$
1	rs993925 (216926691), *TGFB2*(downstream)	T	FEV_1/FVC	0.034	0.006	$1.16x10^{-8}$	-0.007	0.011	$5.12x10^{-1}$	-0.027	0.019	$1.53x10^{-1}$
2	rs12477314 (239542085), *HDAC4*(downstream)	T	FEV_1/FVC	0.041	0.006	$1.68x10^{-12}$	-0.021	0.012	$8.83x10^{-2}$	0.014	0.022	$5.2x10^{-1}$
3	rs1529672 (25495586), *RARB*(intron)	C	FEV_1/FVC	-0.048	0.006	$3.97x10^{-14}$	-0.013	0.014	$3.65x10^{-1}$	0.024	0.025	$3.38x10^{-1}$
3	rs1344555 (170782913), *MECOM*(intron)	T	FEV_1	-0.034	0.006	$2.65x10^{-8}$	0.011	0.013	$3.86x10^{-1}$	0.01	0.022	$6.51x10^{-1}$
5	rs153916 (95062456), *SPATA9*(upstream)	T	FEV_1/FVC	-0.031	0.005	$2.12x10^{-8}$	-0.021	0.01	$3.7x10^{-2}$	0.02	0.018	$2.72x10^{-1}$
6	rs6903823 (28430275), *ZKSCAN3*(intron)/*ZNF323*(intron)	G	FEV_1	-0.037	0.006	$2.18x10^{-10}$	0.02	0.012	$8.86x10^{-2}$	-0.026	0.021	$2.16x10^{-1}$
6	rs2857595 (31676448), *NCR3*(upstream)	G	FEV_1/FVC	0.037	0.006	$2.28x10^{-10}$	-0.002	0.014	$8.55x10^{-1}$	-0.009	0.025	$7.12x10^{-1}$
6	rs2798641 (109374743), *ARMC2*(intron)	T	FEV_1/FVC	-0.041	0.007	$8.35x10^{-9}$	0.007	0.013	$5.69x10^{-1}$	-0.036	0.025	$1.39x10^{-1}$
10	rs7068966 (12317998), *CDC123*(intron)	T	FEV_1/FVC	0.033	0.005	$6.13x10^{-13}$	0.001	0.01	$9.22x10^{-1}$	0.02	0.018	$2.74x10^{-1}$
10	rs11001819 (77985230), *C10orf11*(intron)	G	FEV_1	-0.029	0.004	$2.98x10^{-12}$	-0.011	0.01	$2.7x10^{-1}$	0.015	0.018	$4.05x10^{-1}$
12	rs11172113 (55813550), *LRP1*(intron)	T	FEV_1/FVC	-0.032	0.006	$1.24x10^{-8}$	-0.006	0.01	$5.41x10^{-1}$	0.022	0.019	$2.44x10^{-1}$

				Lung function (stage 1 and stage 2 meta-analysis)			Cigarettes per day			Ever vs. Never smoking		
Chr.	**SNP ID(NCBI36 position), function**	**Coded allele**	**Measure**	**Beta**	**Se**	**P**	**Beta**	**Se**	**P**	**Beta**	**Se**	**P**
12	rs1036429 (94795559), *CCDC38*(intron)	T	FEV_1/FVC	0.038	0.006	$2.3x10^{-11}$	0.006	0.012	$6.35x10^{-1}$	0.028	0.021	$1.85x10^{-1}$
16	rs12447804 (56632783), *MMP15*(intron)	T	FEV_1/FVC	-0.038	0.007	$3.59x10^{-8}$	-0.005	0.013	$6.69x10^{-1}$	-0.005	0.022	$8.06x10^{-1}$
16	rs2865531 (73947817), *CFDP1*(intron)	T	FEV_1/FVC	0.031	0.005	$1.77x10^{-11}$	0.019	0.01	$5.3x10^{-2}$	0.002	0.018	$9.13x10^{-1}$
21	rs9978142 (34574109), *KCNE2*(upstream)	T	FEV_1/FVC	-0.043	0.008	$2.65x10^{-8}$	0.007	0.013	$5.84x10^{-1}$	0.007	0.024	$7.59x10^{-1}$
Previously reported loci												
2	rs2571445(218391399), *TNS1*(ns)	G	FEV_1	0.047	0.007	$9.83x10^{-11}$	-0.011	0.011	$3.09x10^{-1}$	0.023	0.02	$2.47x10^{-1}$
2	rs10498230(229210747), *PID1*(downstream)	T	FEV_1/FVC	0.068	0.014	$1.13x10^{-6}$	-0.061	0.019	$1.62x10^{-3}$	-0.05	0.033	$1.38x10^{-1}$
4	rs2045517(90089987), *FAM13A* (intron)	T	FEV_1/FVC	-0.047	0.007	$2x10^{-11}$	-0.01	0.01	$3.23x10^{-1}$	0.004	0.018	$8.4x10^{-1}$
4	rs7671167(90103002), *FAM13A* (intron)	T	FEV_1/FVC	-0.042	0.007	$1.27x10^{-9}$	-0.013	0.01	$1.91x10^{-1}$	0.014	0.018	$4.35x10^{-1}$
4	rs10516526(106908353), *GSTCD*(intron)	G	FEV_1	0.108	0.014	$4.75x10^{-14}$	0.03	0.02	$1.23x10^{-1}$	0.018	0.034	$6.08x10^{-1}$
4	rs17331332(107027556), *NPNT* (upstream)	G	FEV_1	-0.102	0.014	$1.11x10^{-12}$	-0.037	0.019	$5.25x10^{-2}$	-0.031	0.033	$3.51x10^{-1}$
4	rs6823809(107048244), *NPNT*(intron)	T	FEV_1/FVC	0.056	0.011	$2.2x10^{-7}$	0.011	0.026	$6.77x10^{-1}$	-0.045	0.066	$5.02x10^{-1}$
4	rs1032296(145654138), *HHIP*(upstream)	T	FEV_1/FVC	-0.05	0.007	$3.42x10^{-12}$	0.007	0.01	$4.97x10^{-1}$	-0.002	0.018	$9.01x10^{-1}$
4	rs11100860(145698589), *HHIP*(upstream)	G	FEV_1/FVC	0.064	0.007	$6.81x10^{-20}$	-0.002	0.01	$8.58x10^{-1}$	0.001	0.017	$9.32x10^{-1}$
5	rs11168048(147822546), *HTR4*(intron)	T	FEV_1/FVC	-0.047	0.007	$5.97x10^{-11}$	0.003	0.01	$7.34x10^{-1}$	-0.012	0.018	$5.04x10^{-1}$
5	rs3995090(147826008), *HTR4*(intron)	C	FEV_1/FVC	0.046	0.007	$1.04x10^{-10}$	-0.002	0.01	$8.69x10^{-1}$	0.014	0.018	$4.25x10^{-1}$
5	rs1985524(147827981), *HTR4*(intron)	G	FEV_1	-0.048	0.007	$3.06x10^{-11}$	0	0.01	$9.97x10^{-1}$	-0.027	0.018	$1.32x10^{-1}$
5	rs11134779(156869344), *ADAM19*(intron)	G	FEV_1/FVC	-0.042	0.007	$6.01x10^{-9}$	-0.015	0.01	$1.43x10^{-1}$	0.012	0.019	$5.27x10^{-1}$
6	rs2070600(32259421), *AGER*(ns)	T	FEV_1/FVC	0.126	0.016	$9.07x10^{-15}$	0.033	0.026	$1.98x10^{-1}$	0.056	0.044	$2.03x10^{-1}$
6	rs3817928(142792209), *GPR126*(intron)	G	FEV_1/FVC	0.059	0.008	$2.27x10^{-12}$	-0.004	0.012	$7.16x10^{-1}$	0.005	0.021	$8.1x10^{-1}$
6	rs262129(142894837), *LOC153910*(unknown)	G	FEV_1/FVC	0.056	0.008	$2.91x10^{-13}$	0.005	0.011	$6.56x10^{-1}$	0.019	0.02	$3.39x10^{-1}$

				Lung function (stage 1 and stage 2 meta-analysis)			Cigarettes per day			Ever vs. Never smoking		
Chr.	**SNP ID(NCBI36 position), function**	**Coded allele**	**Measure**	**Beta**	**Se**	**P**	**Beta**	**Se**	**P**	**Beta**	**Se**	**P**
9	rs16909859(97244613), *PTCH1*(downstream)	G	FEV_1/FVC	0.08	0.013	$7.45x10^{-10}$	-0.005	0.02	$7.81x10^{-1}$	-0.012	0.033	$7.09x10^{-1}$
9	rs16909898(97270829), *PTCH1*(intron)	G	FEV_1/FVC	-0.072	0.012	$3.94x10^{-9}$	0.024	0.018	$1.91x10^{-1}$	0.019	0.031	$5.45x10^{-1}$
15	rs12899618(69432174), *THSD4*(intron)	G	FEV_1/FVC	0.076	0.01	$1.86x10^{-15}$	-0.013	0.014	$3.36x10^{-1}$	-0.012	0.024	$6.31x10^{-1}$
15	rs8033889(69467134), *THSD4*(intron)	T	FEV_1/FVC	-0.072	0.008	$2.03x10^{-17}$	-0.004	0.012	$7.57x10^{-1}$	-0.02	0.022	$3.59x10^{-1}$
15	rs2568494(76528019), *IREB2*(intron)	G	FEV_1/FVC	0.029	0.007	$5.25x10^{-5}$	-0.082	0.01	$2.15x10^{-15}$	-0.016	0.018	$3.62x10^{-1}$
15	rs8034191(76593078), *CHRNA3/5*(intron)	T	FEV_1/FVC	0.032	0.007	$9.65x10^{-6}$	-0.09	0.011	$1.59x10^{-17}$	-0.02	0.018	$2.75x10^{-1}$
15	rs2036527(76638670), *CHRNA5*(uptream)	G	FEV_1	0.036	0.008	$2.4x10^{-6}$	-0.091	0.011	$6.34x10^{-18}$	-0.016	0.018	$3.86x10^{-1}$
15	rs8040868(76698236), *CHRNA3*(coding-synon)	T	FEV_1/FVC	0.04	0.008	$1.14x10^{-6}$	-0.09	0.011	$2.53x10^{-17}$	-0.023	0.019	$2.26x10^{-1}$

Association of loci influencing lung function with lung cancer

Effects of the 16 novel SNPs, and previously reported SNPs, associated with lung function on lung cancer were assessed in the International Lung Cancer Consortium (ILCCO) GWAS meta-analysis. The ILCCO GWAS meta-analysis only had genotyped data, for this reason proxy SNPs were given when the top SNP was not included in their data. Leading SNP, region name and r^2 between leading SNP and proxy SNP are also provided. Results for the SpiroMeta-CHARGE stage 1 and stage 2 meta-analysis are reported for the SNPs that were followed up in stage 2 and were included in the lung cancer dataset (rs1529672, rs2857595, rs2798641, rs11001819, rs11172113, rs1036429) and results from the GWAS Stage 1 only are provided for the other loci, for the lung function measure that showed stronger association. Abbreviations: Chr.=chromosome, ns=nonsynonymous, s= synonymous

Chr.	Proxy SNP ID (NCBI36 position), function	Leading SNP (region name), r^2 with proxy	Coded allele	Lung function				Lung cancer		
				Measure	Beta	Se	P	Beta	Se	P
Novel loci										
1	rs761423 (17174259), *MFAP2* (intron)	rs2284746 (*MFAP2*), 0.63	T	FEV_1/FVC	0.038	0.007	$8.72x10^{-8}$	0.033	0.017	$6.24x10^{-2}$
1	rs2871775 (17218492), *SDHB* (intron)	rs2284746 (*MFAP2*), 0.66	G	FEV_1/FVC	-0.040	0.007	$9.08x10^{-9}$	-0.015	0.017	$3.81x10^{-1}$
2	rs4591362 (239542675), *HDAC4*(downstream)	rs12477314 (*HDAC4*), 0.94	G	FEV_1/FVC	-0.049	0.009	$8.29x10^{-9}$	-0.020	0.021	$3.49x10^{-1}$
3	rs1529672 (25495586), *RARB* (intron)	rs1529672 (*RARB*), 1	C	FEV_1/FVC	-0.048	0.006	$3.97x10^{-14}$	-0.013	0.027	$6.33x10^{-1}$
3	rs2056777 (25515395), *RARB* (intron)	rs1529672 (*RARB*), 0.77	T	FEV_1/FVC	-0.050	0.009	$5.31x10^{-8}$	-0.011	0.027	$6.8x10^{-1}$
3	rs1362772 (170739927), *MECOM*(intron)	rs1344555 (*MECOM*), 1	T	FEV_1	-0.040	0.009	$3.24x10^{-6}$	-0.010	0.022	$6.63x10^{-1}$
3	rs7642776 (170753972), *MECOM*(intron)	rs1344555 (*MECOM*), 0.94	G	FEV_1	0.038	0.008	$5.38x10^{-6}$	0.013	0.021	$5.37x10^{-1}$
5	rs2548125 (95037182), *SPATA9* (intron)	rs153916 (*SPATA9*), 0.61	G	FEV_1/FVC	-0.029	0.007	$3.5x10^{-5}$	-0.012	0.018	$5.08x10^{-1}$
6	rs209181 (28900456), *LOC401242* (downstream)	rs6903823 (*ZKSCAN3/ZNF323*), 0.69	G	FEV_1	0.035	0.012	$2.25x10^{-3}$	-0.106	0.026	$3.41x10^{-5}$
6	rs3099844 (31556955), *HCG26* (downstream)	rs2857595 (*NCR3*), 0.67	C	FEV_1/FVC	0.058	0.011	$1.92x10^{-7}$	-0.141	0.027	$2.21x10^{-7}$
6	rs2857595 (31676448), *NCR3*(upstream)	rs2857595 (*NCR3*), 1	G	FEV_1/FVC	0.037	0.006	$2.28x10^{-10}$	-0.051	0.022	$1.91x10^{-2}$

Chr.	Proxy SNP ID (NCBI36 position), function	Leading SNP (region name), r^2 with proxy	Coded allele	Lung function				Lung cancer		
				Measure	Beta	Se	P	Beta	Se	P
6	rs1475055 (109350925), *ARMC2* (intron)	rs2798641 (*ARMC2*), 0.73	T	FEV_1/FVC	0.027	0.008	$7.67x10^{-4}$	-0.011	0.020	$5.87x10^{-1}$
6	rs2798641 (109374743), *ARMC2* (intron)	rs2798641 (*ARMC2*), 1	T	FEV_1/FVC	-0.041	0.007	$8.35x10^{-9}$	0.006	0.022	$7.72x10^{-1}$
10	rs1317549 (12285320), *CDC123* (intron)	rs7068966 (*CDC123*), 0.68	T	FEV_1/FVC	-0.038	0.007	$8.95x10^{-8}$	0.013	0.018	$4.78x10^{-1}$
10	rs4478891 (12307660), *CDC123* (intron)	rs7068966 (*CDC123*), 0.85	G	FEV_1/FVC	0.043	0.007	$8.71x10^{-10}$	-0.005	0.018	$7.61x10^{-1}$
10	rs2130800 (77944824), *C10orf11* (intron)	rs11001819 (*C10orf11*), 0.73	T	FEV_1	0.038	0.007	$5.45x10^{-8}$	0.031	0.017	$7.8x10^{-2}$
10	rs11001819 (77985230), *C10orf11* (intron)	rs11001819 (*C10orf11*), 1	G	FEV_1	-0.029	0.004	$2.98x10^{-12}$	-0.051	0.020	$1.21x10^{-2}$
10	rs2637260 (77990352), *C10orf11* (downstream)	rs11001819 (*C10orf11*), 0.72	T	FEV_1	0.035	0.007	$7.38x10^{-7}$	0.024	0.017	$1.72x10^{-1}$
12	rs11172113 (55813550), *LRP1* (intron)	rs11172113 (*LRP1*), 1	T	FEV_1/FVC	-0.032	0.006	$1.24x10^{-8}$	-0.010	0.018	$5.84x10^{-1}$
12	rs1466535 (55820737), *LRP1* (intron)	rs11172113 (*LRP1*), 0.72	G	FEV_1/FVC	-0.025	0.007	$5.76x10^{-4}$	0.004	0.018	$8.45x10^{-1}$
12	rs7307510 (94761701), *SNRPF* (upstream)	rs1036429 (*CCDC38*), 0.96	T	FEV_1/FVC	0.049	0.009	$3.49x10^{-8}$	0.009	0.023	$6.82x10^{-1}$
12	rs1036429 (94795559), *CCDC38* (intron)	rs1036429 (*CCDC38*), 1	T	FEV_1/FVC	0.038	0.006	$2.3x10^{-11}$	0.011	0.022	$6.06x10^{-1}$
16	rs2304488 (56631711), *MMP15* (intron)	rs12447804 (*MMP15*), 0.88	G	FEV_1/FVC	-0.040	0.008	$9.45x10^{-7}$	0.019	0.021	$3.62x10^{-1}$
16	rs12597233 (56657709), *MMP15* (downstream)	rs12447804 (*MMP15*), 0.87	G	FEV_1/FVC	0.038	0.008	$4.8x10^{-6}$	-0.012	0.022	$5.85x10^{-1}$
16	rs4243111 (73878328), *CFDP1*(downstream)	rs2865531 (*CFDP1*), 0.93	T	FEV_1/FVC	-0.037	0.007	$1.52x10^{-7}$	-0.036	0.018	$4.22x10^{-2}$
16	rs1424013 (74053487), *TMEM170* (intron)	rs2865531 (*CFDP1*), 0.82	T	FEV_1/FVC	0.033	0.007	$3.92x10^{-6}$	0.046	0.018	$1x10^{-2}$
21	rs973754 (34555400), *C21orf82* (downstream)	rs9978142 (*KCNE2*), 0.81	G	FEV_1/FVC	-0.043	0.010	$2.09x10^{-5}$	-0.010	0.025	$6.78x10^{-1}$
Previously reported regions										
2	rs1035672 (218383444), *TNS1* (intron)	rs2571445 (*TNS1*), 0.96	G	FEV_1	-0.046	0.007	$3.03x10^{-10}$	-0.020	0.018	$2.66x10^{-1}$
2	rs2571445 (218391399), *TNS1* (ns)	rs2571445 (*TNS1*), 1	G	FEV_1	0.047	0.007	$9.83x10^{-11}$	0.016	0.018	$3.72x10^{-1}$
4	rs2869967 (90088355), *FAM13A* (intron)	rs2045517 (*FAM13A*), 1	T	FEV_1/FVC	0.047	0.007	$2.08x10^{-11}$	0.026	0.018	$1.36x10^{-1}$
4	rs6849143 (90147512), *FAM13A* (intron)	rs2045517 (*FAM13A*), 0.75	T	FEV_1/FVC	-0.038	0.007	$5.99x10^{-8}$	-0.020	0.018	$2.51x10^{-1}$

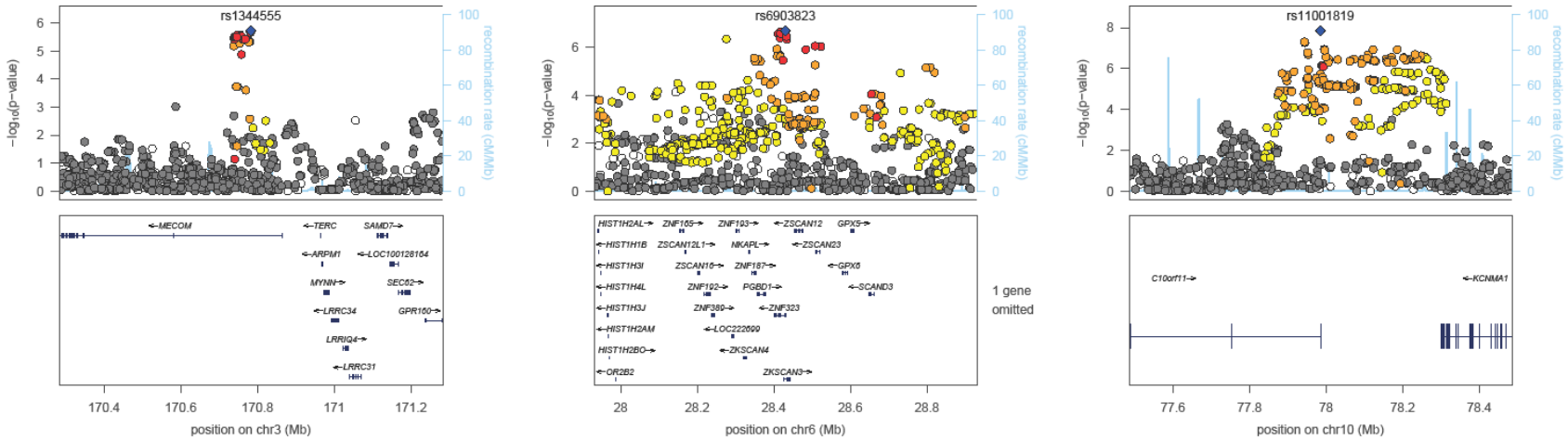
Chr.	Proxy SNP ID (NCBI36 position), function	Leading SNP (region name), r^2 with proxy	Coded allele	Lung function				Lung cancer		
				Measure	Beta	Se	P	Beta	Se	P
4	rs11727735 (106851319), *GSTCD* (intron)	rs10516526 (*GSTCD*), 1	G	FEV_1	0.105	0.014	$1.65x10^{-13}$	-0.029	0.036	$4.19x10^{-1}$
4	rs10516526 (106908353), *GSTCD* (intron)	rs10516526 (*GSTCD*), 1	G	FEV_1	0.108	0.014	$4.75x10^{-14}$	-0.029	0.036	$4.31x10^{-1}$
4	rs1828591 (145700230), *HHIP* (upstream)	rs11100860 (*HHIP*), 1	G	FEV_1/FVC	0.063	0.007	$1.44x10^{-19}$	-0.005	0.018	$7.82x10^{-1}$
4	rs1512288 (145710731), *HHIP* (upstream)	rs11100860 (*HHIP*), 1	G	FEV_1/FVC	-0.062	0.007	$3.46x10^{-19}$	0.000	0.018	$9.89x10^{-1}$
5	rs2277027 (156864954), *ADAM19* (intron)	rs11134779 (*ADAM19*), 1	C	FEV_1/FVC	-0.042	0.007	$6.65x10^{-9}$	0.007	0.018	$6.93x10^{-1}$
5	rs1422795 (156868942), *ADAM19* (ns)	rs11134779 (*ADAM19*), 1	T	FEV_1/FVC	0.041	0.007	$1.05x10^{-8}$	-0.007	0.018	$7.11x10^{-1}$
6	rs2070600 (32259421), *AGER* (ns)	rs2070600 (*AGER*), 1	T	FEV_1/FVC	0.126	0.016	$9.07x10^{-15}$	-0.004	0.044	$9.27x10^{-1}$
6	rs2854050 (32293583), *NOTCH4* (intron)	rs2070600 (*AGER*), 1	G	FEV_1/FVC	-0.083	0.016	$9.34x10^{-8}$	0.022	0.040	$5.8x10^{-1}$
6	rs6570507 (142721265), *GPR126* (intron)	rs262129 (*GPR126*), 0.72	G	FEV_1/FVC	-0.051	0.008	$2.25x10^{-11}$	0.006	0.019	$7.49x10^{-1}$
6	rs11155242 (142733242), *GPR126* (ns)	rs3817928 (*GPR126*), 1	C	FEV_1/FVC	0.055	0.009	$1.88x10^{-10}$	0.003	0.022	$8.99x10^{-1}$
6	rs3748069 (142809326), *GPR126* (downstream)	rs262129 (*GPR126*), 0.84	G	FEV_1/FVC	0.053	0.008	$2.02x10^{-12}$	-0.007	0.019	$7.01x10^{-1}$
6	rs7776356 (142818722), *GPR126* (downstream)	rs3817928 (*GPR126*), 1	G	FEV_1/FVC	0.059	0.008	$4.16x10^{-12}$	-0.001	0.021	$9.69x10^{-1}$
9	rs10512249 (97296130), *PTCH1* (intron)	rs16909859 (*PTCH1*), 0.84	G	FEV_1/FVC	0.066	0.012	$1.54x10^{-8}$	-0.067	0.029	$1.9x10^{-2}$
15	rs1913768 (69436598), *THSD4* (intron)	rs8033889 (*THSD4*), 0.673	G	FEV_1/FVC	0.075	0.009	$2.77x10^{-15}$	-0.014	0.024	$5.49x10^{-1}$
15	rs8033889 (69467134), *THSD4* (intron)	rs8033889 (*THSD4*), 1	T	FEV_1/FVC	-0.072	0.008	$2.03x10^{-17}$	0.014	0.025	$5.85x10^{-1}$
15	rs8034191 (76593078), *AGPHD1* (intron)	rs8040868 (*CHRNA3*), 0.70	T	FEV_1/FVC	0.032	0.007	$9.65x10^{-6}$	-0.258	0.018	$2.19x10^{-46}$
15	rs1051730 (76681394), *CHRNA3* (s)	rs8040868 (*CHRNA3*), 0.76	G	FEV_1/FVC	0.032	0.007	$1.46x10^{-5}$	-0.273	0.018	$1.91x10^{-51}$

D. Chapter 3 additional figures

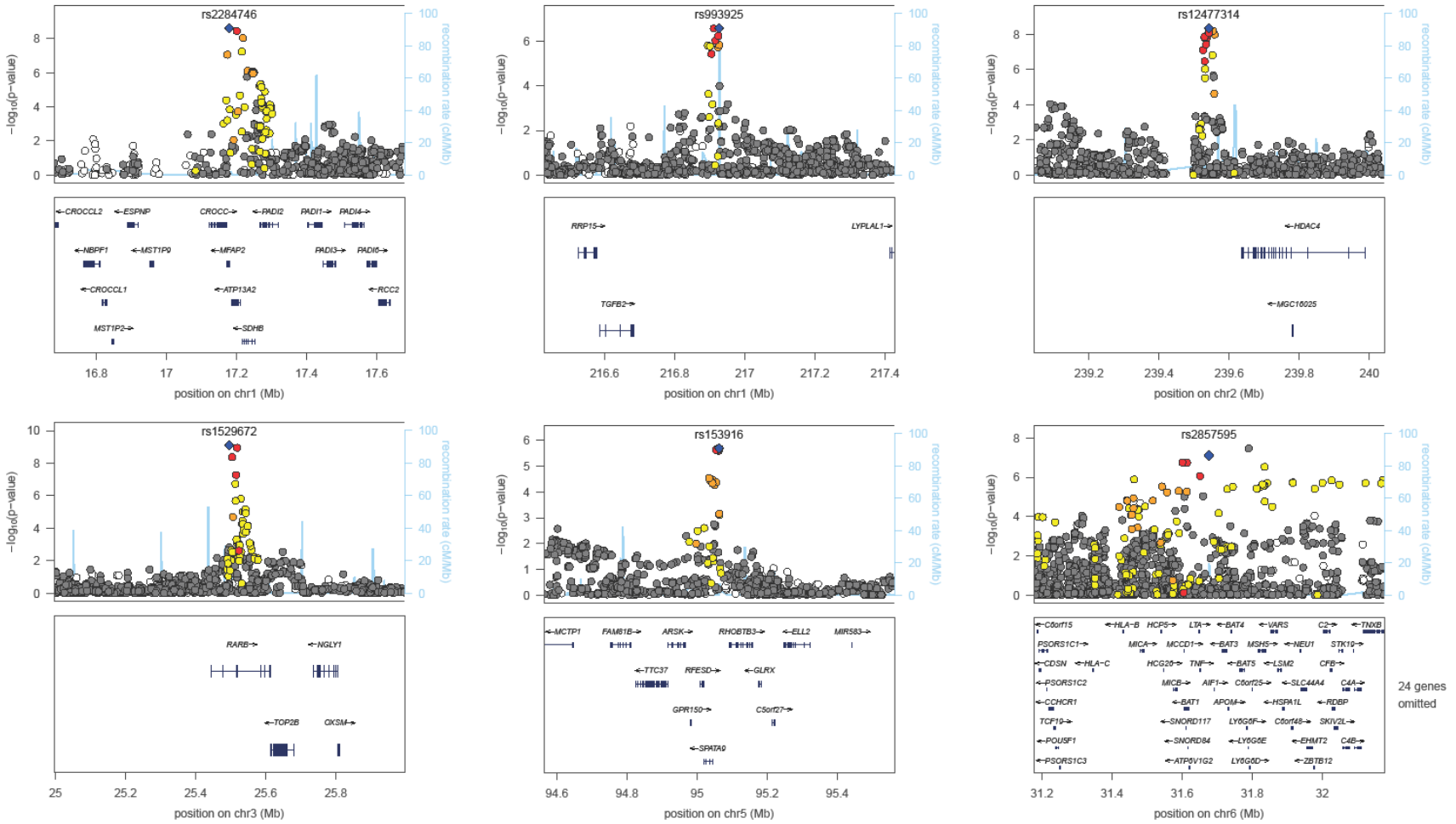
Region plots for the 16 new loci

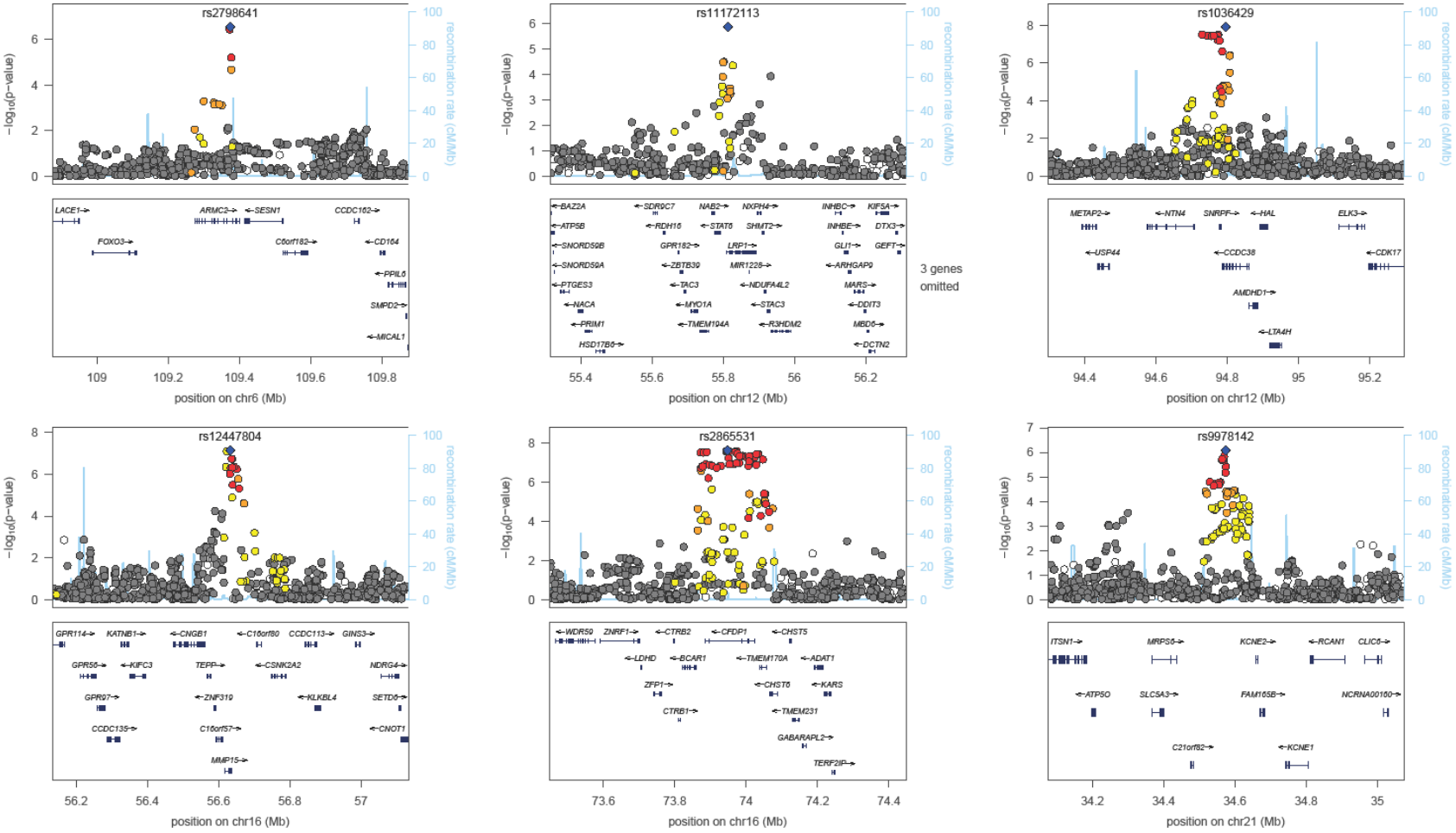
Regional association plots of 16 novel lung function-associated loci. Statistical significance of each SNP on the –log10(P) scale as a function of chromosome position (NCBI Build 36) in the meta-analysis of stage 1 data alone. The sentinel SNP at each locus is shown in blue with the correlations (r^2) of surrounding SNPs to the sentinel indicated by colour (red: r^2>0.8, orange: r^2>0.5, yellow: r^2>0.2, grey: r^2<0.2, white: r^2 unknown). The fine scale recombination rate is shown in blue.

a) FEV_1 only



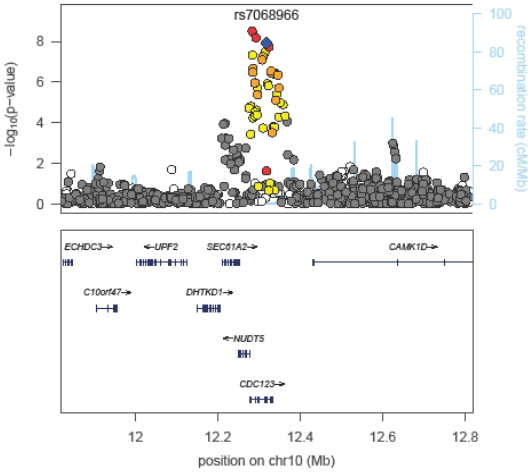
b) FEV_1/FVC only


rs2284746
rs993925
rs12477314
rs1529672
rs153916
rs2857595
−log10(p-value)
recombination rate (cM/Mb)
position on chr1 (Mb)
position on chr1 (Mb)
position on chr2 (Mb)
position on chr3 (Mb)
position on chr5 (Mb)
position on chr6 (Mb)
24 genes omitted

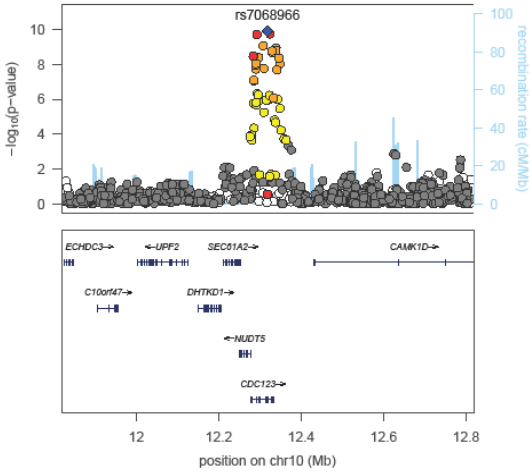

rs2798641
rs11172113
rs1036429
rs12447804
rs2865531
rs9978142
−log10(p−value)
recombination rate (cM/Mb)
3 genes omitted
position on chr6 (Mb)
position on chr12 (Mb)
position on chr16 (Mb)
position on chr21 (Mb)

c) Both FEV_1 and FEV_1/FVC

FEV_1



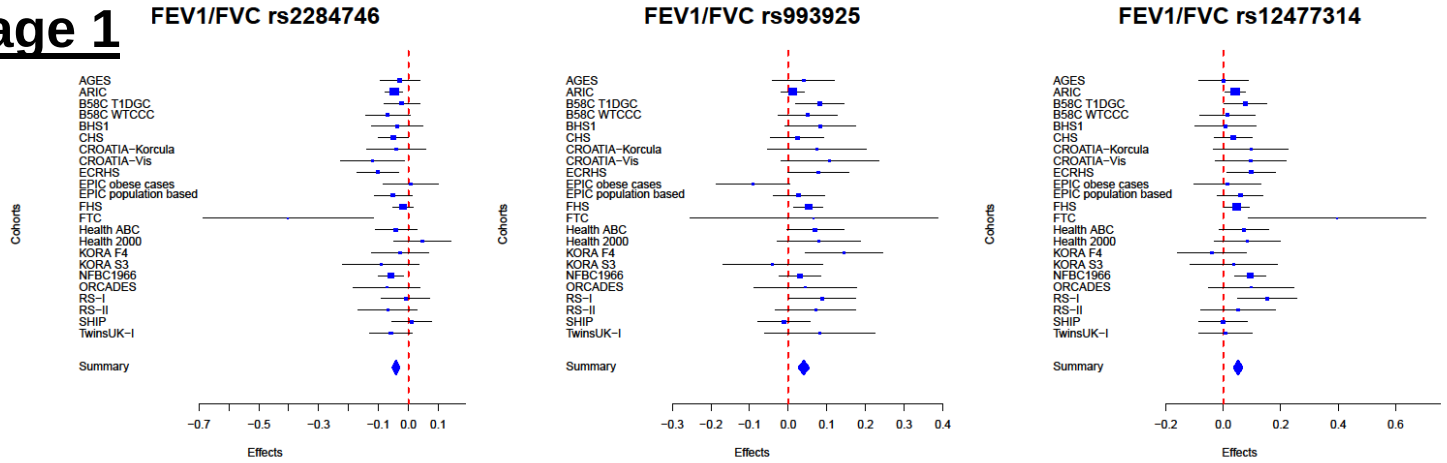
FEV_1/FVC



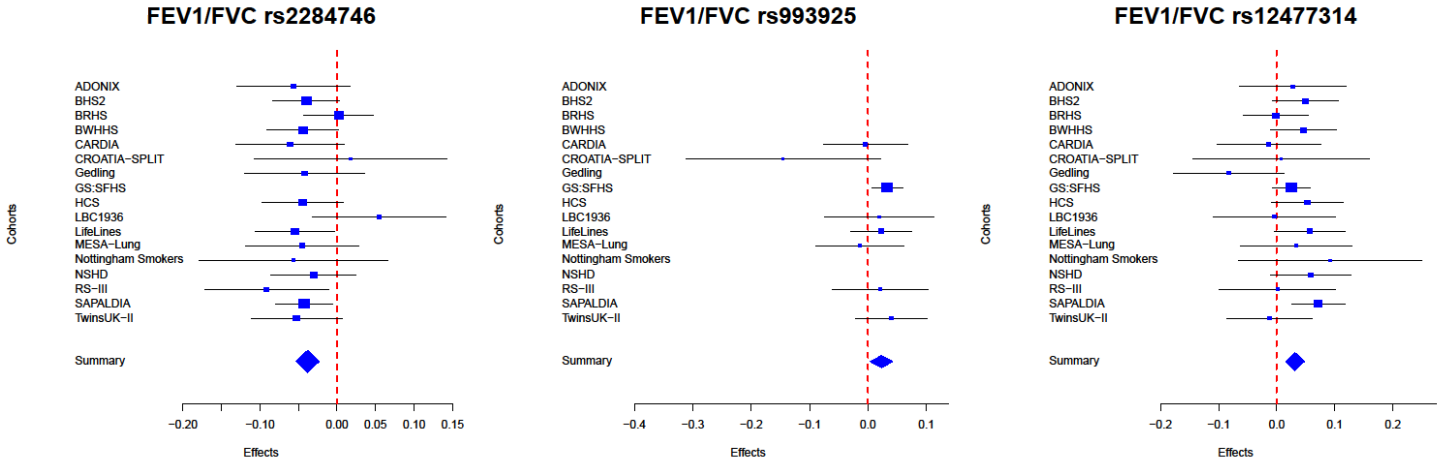
Forest plots for the 16 new loci

Forest plots for the 16 loci associated with lung function for stage 1 and stage 2 separately. Each of the SNPs included in the figure showed genome-wide significant association ($P < 5 \times 10^{-8}$) with either FEV_1 or FEV_1/FVC in the data from stages 1 and 2. For each SNP there is a plot for the meta-analysis of the stage 1 data and another for the meta-analysis of the stage 2 data. The contributing effect (transformed beta) from each study is shown by a square, with confidence intervals indicated by horizontal lines. The contributing weight of each study to the meta-analysis is indicated by the size of the square. The combined meta-analysis estimate is shown at the bottom of each graph.

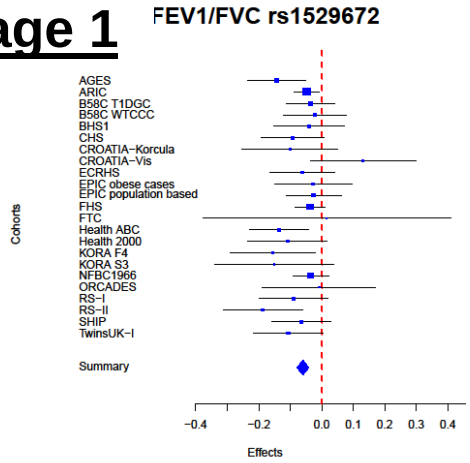
Stage 1

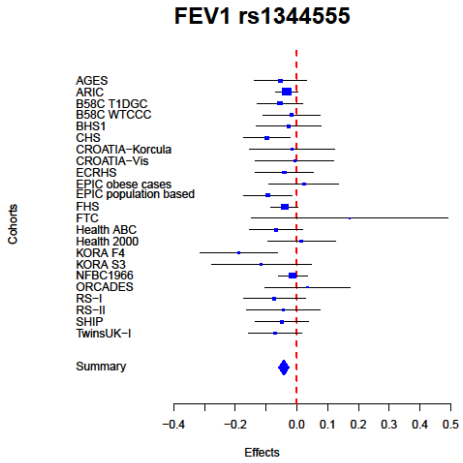


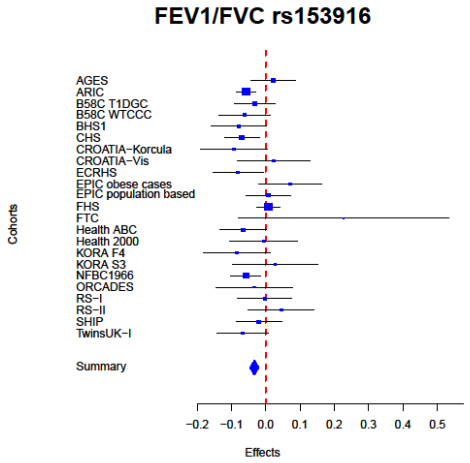
Stage 2



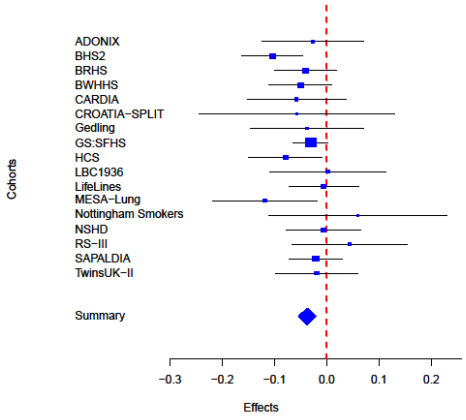
Stage 1

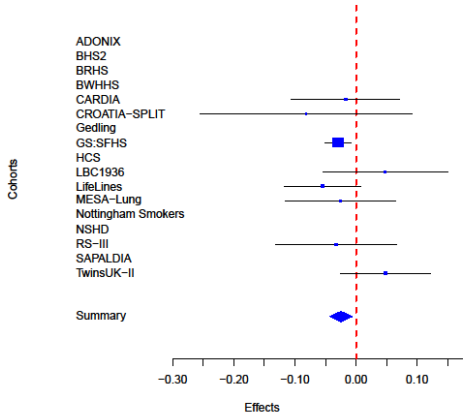


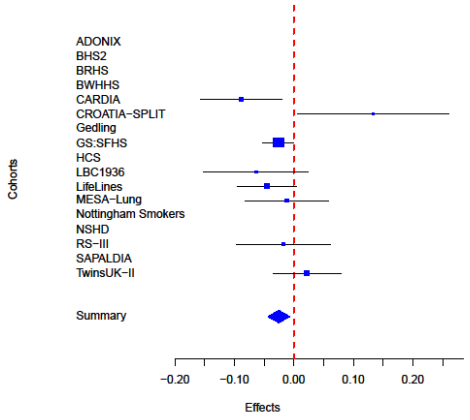




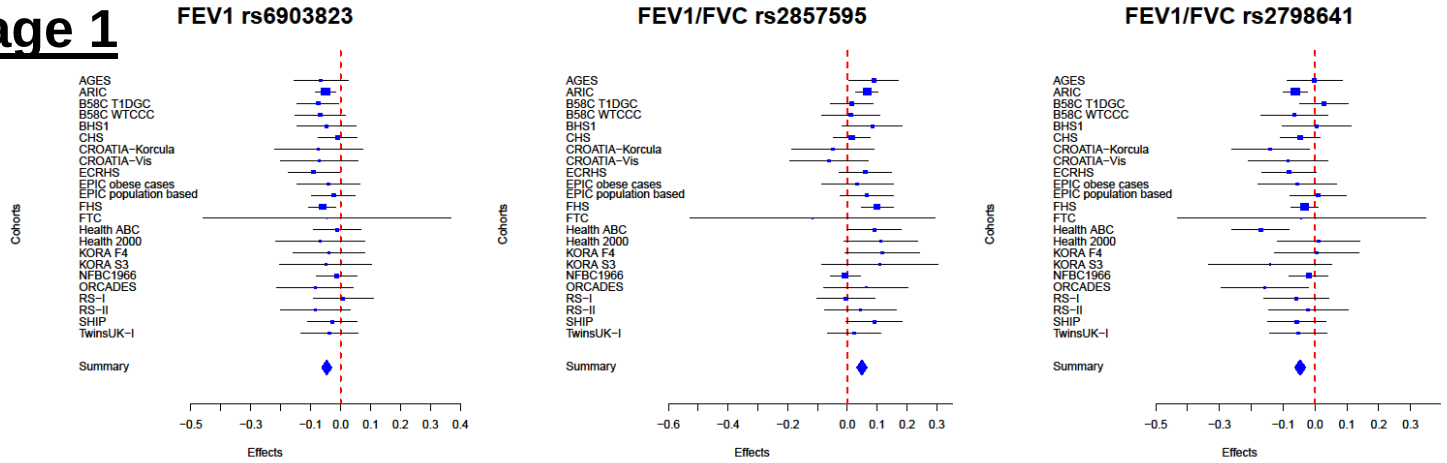
Stage 2



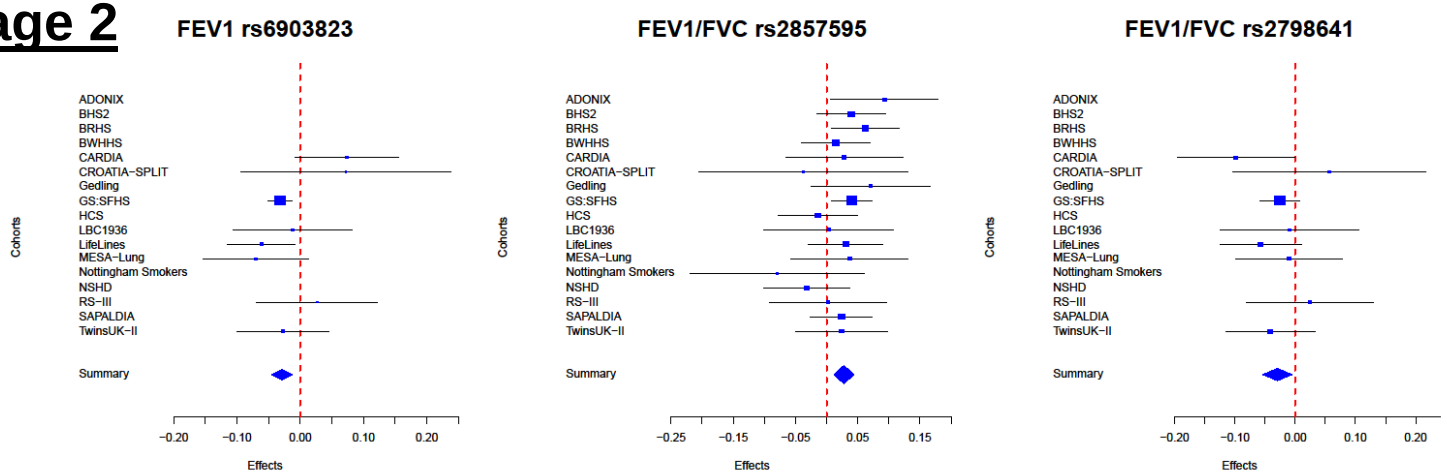




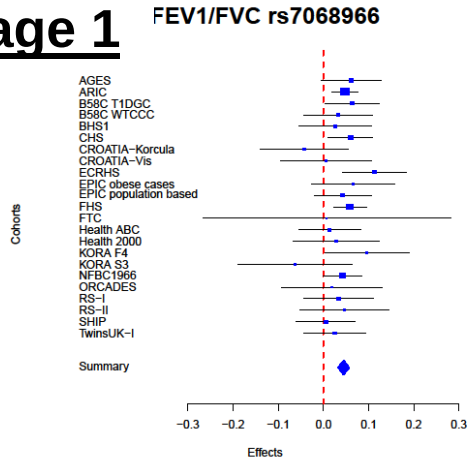
Stage 1

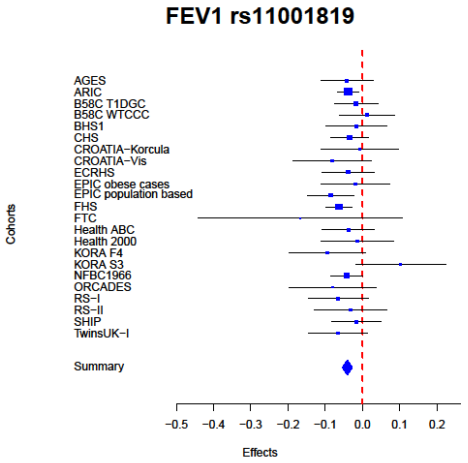


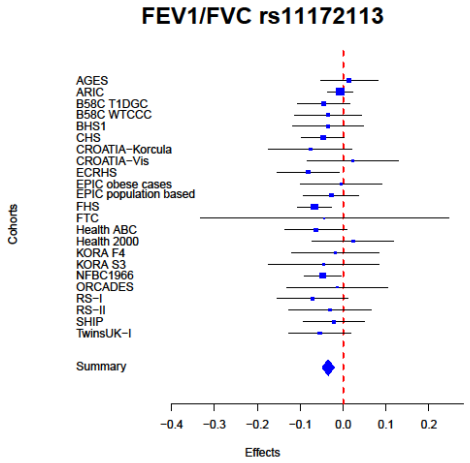
Stage 2



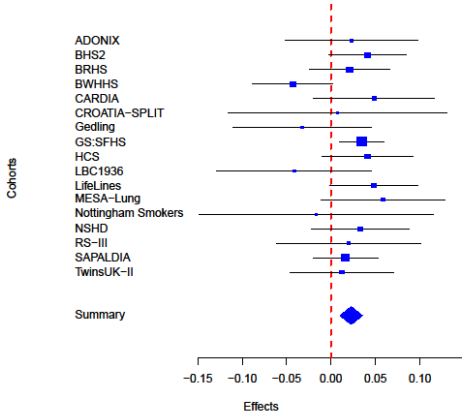
Stage 1

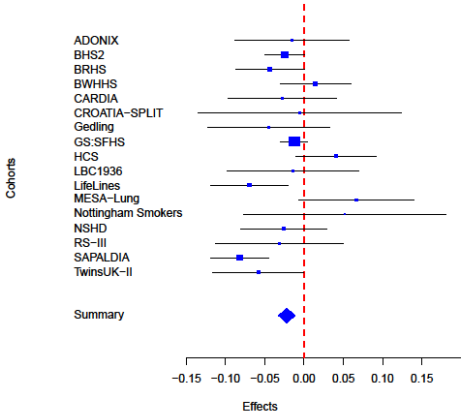


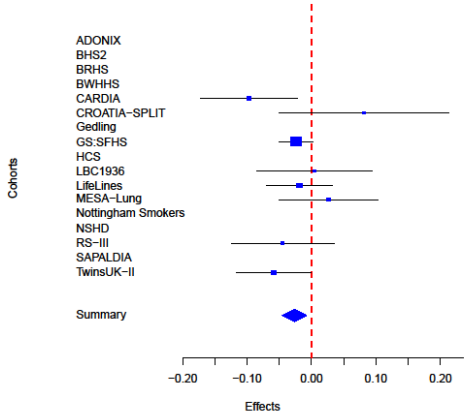




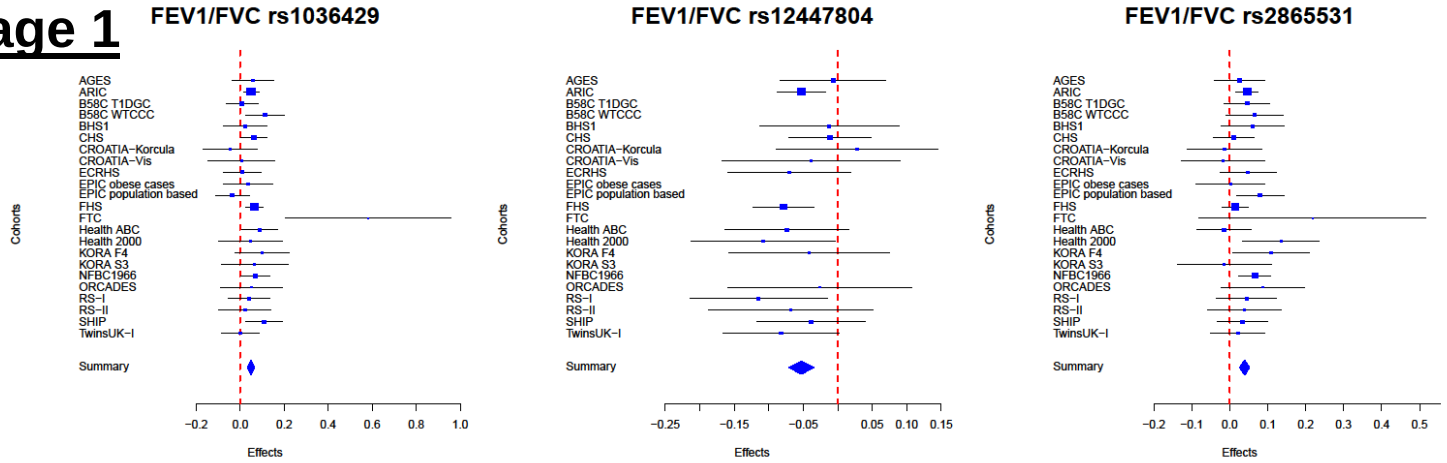
Stage 2



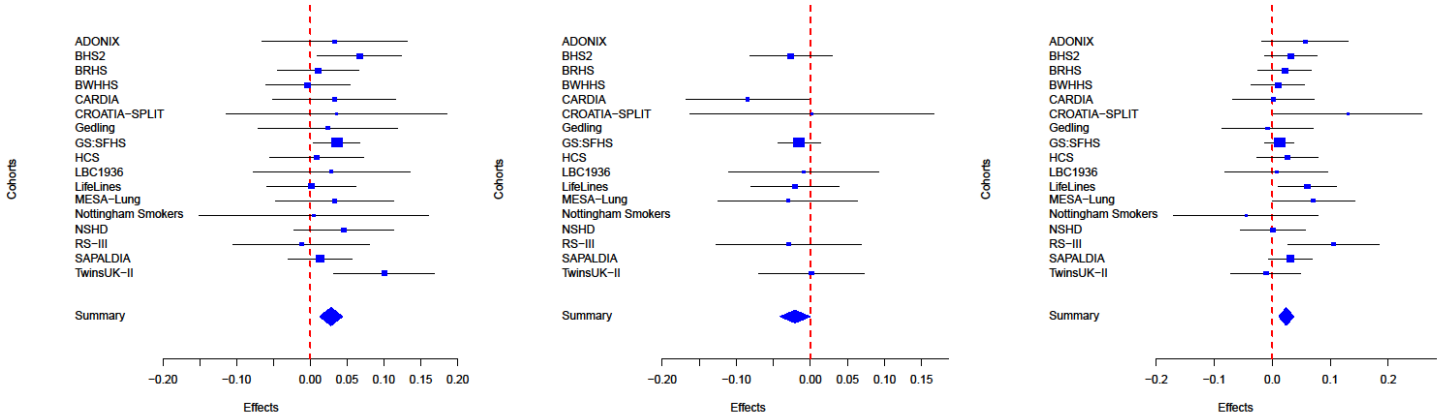




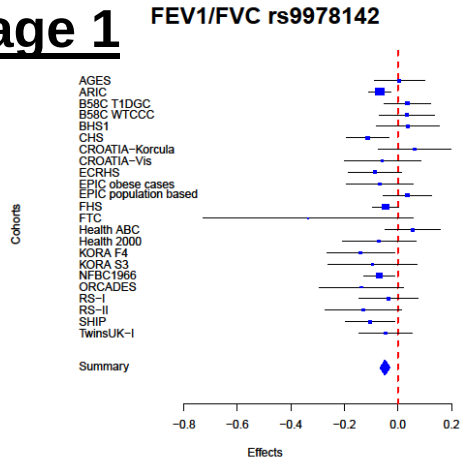
Stage 1



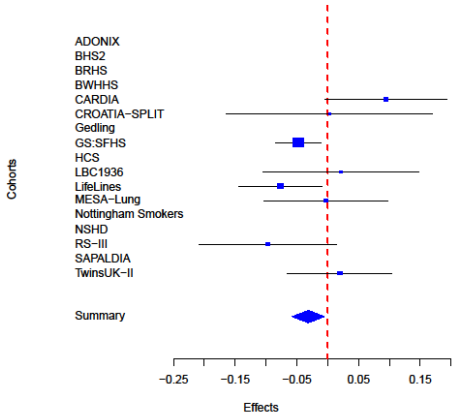
Stage 2



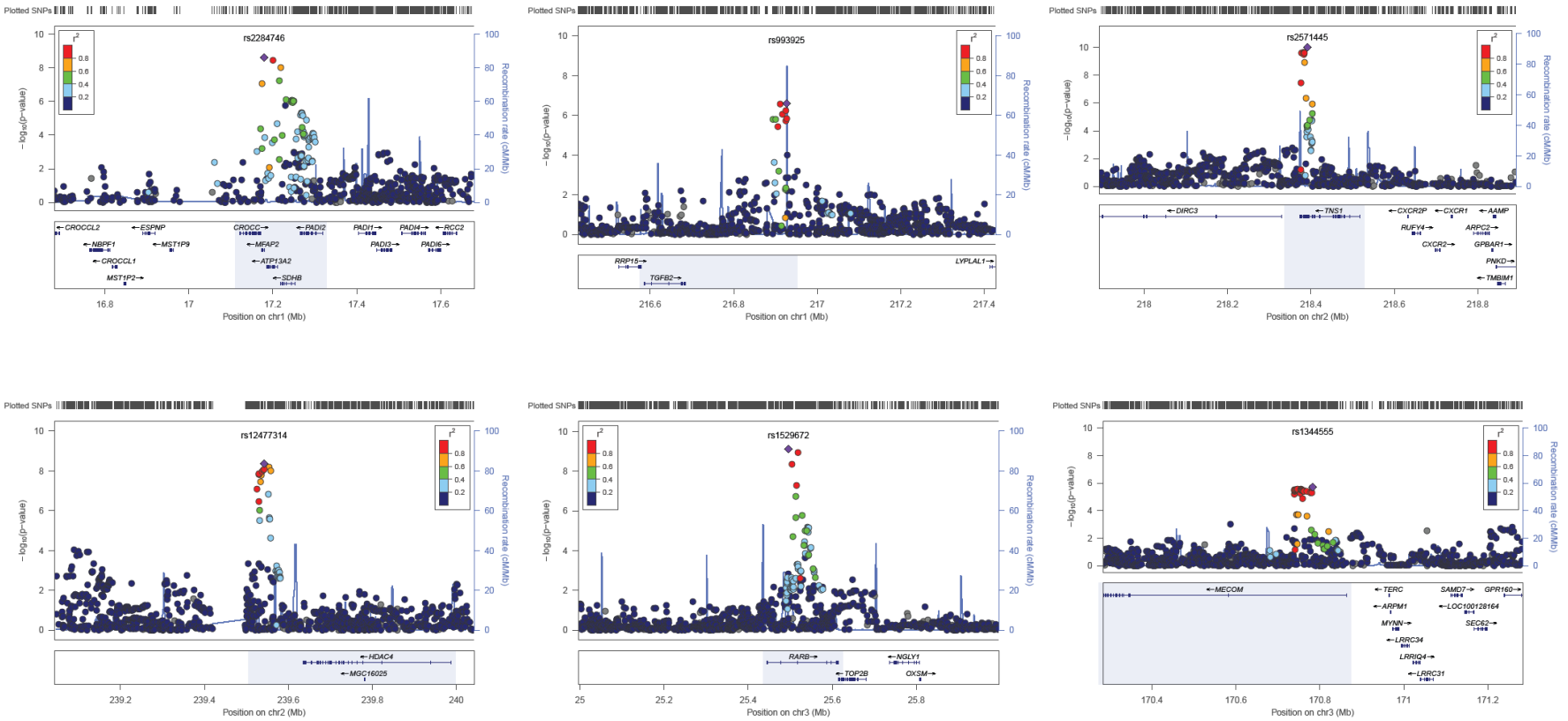
Stage 1

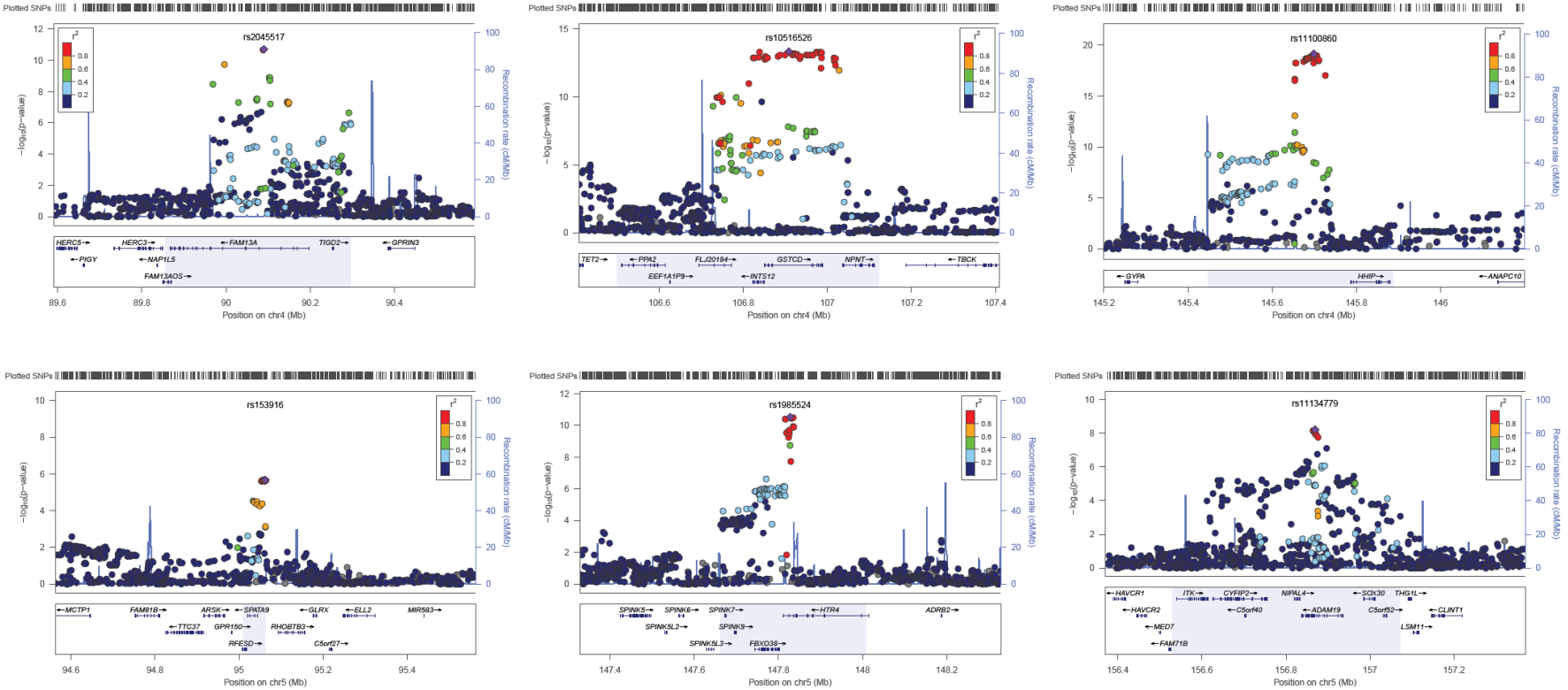


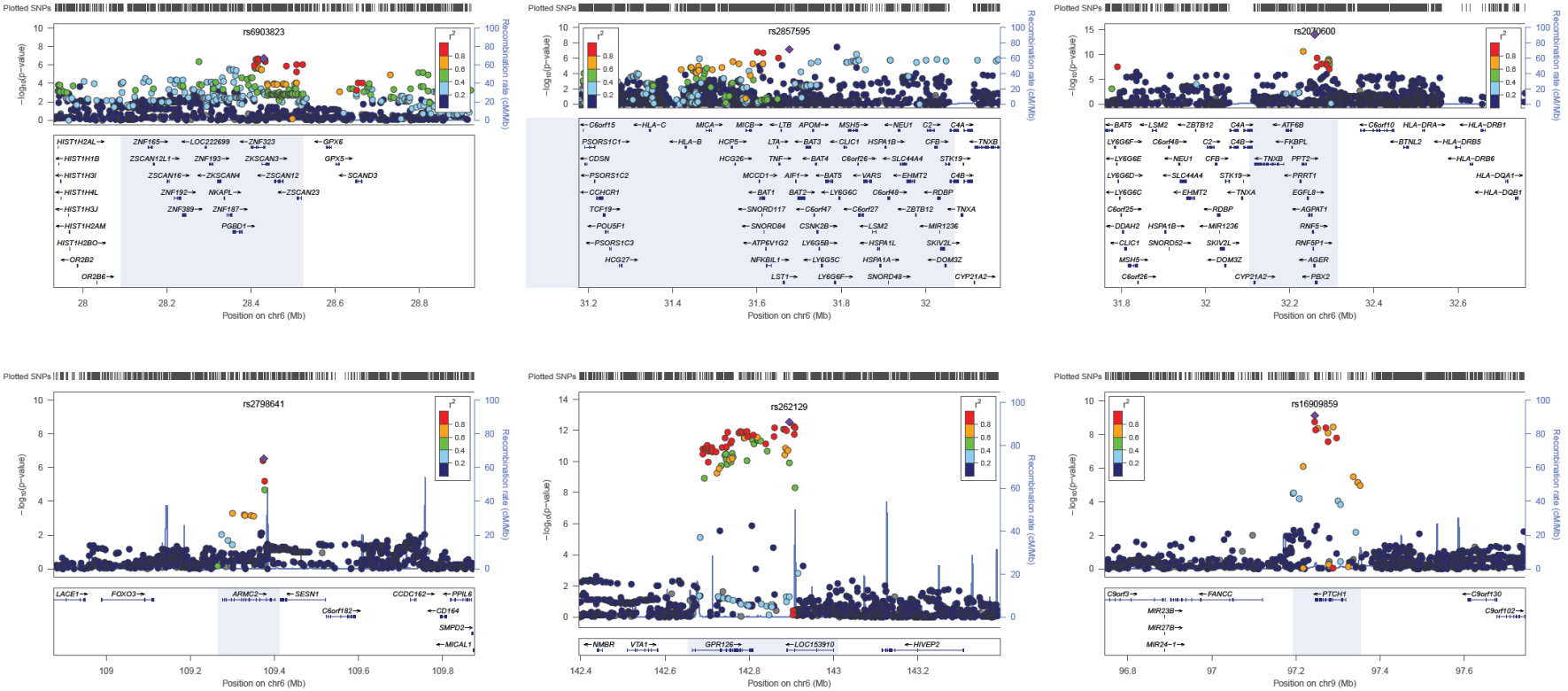
Stage 2

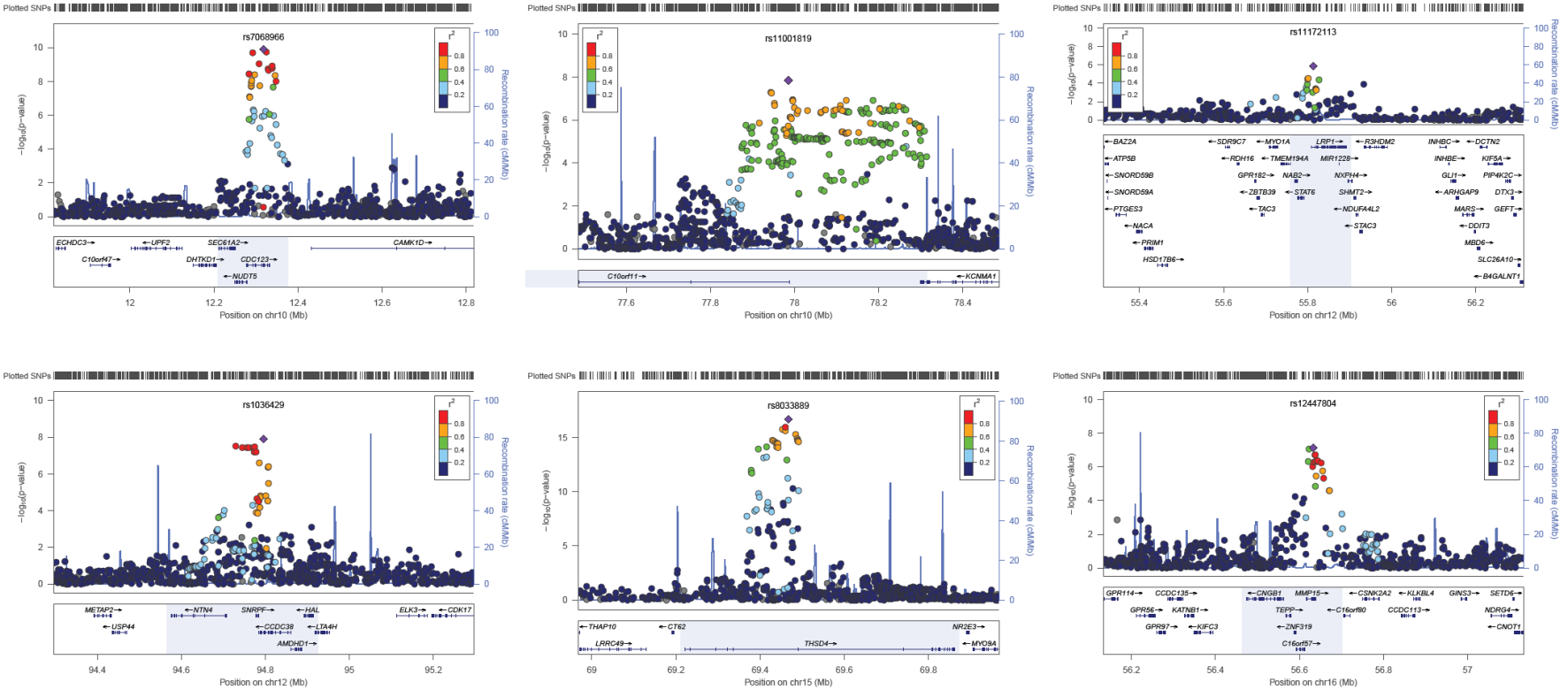


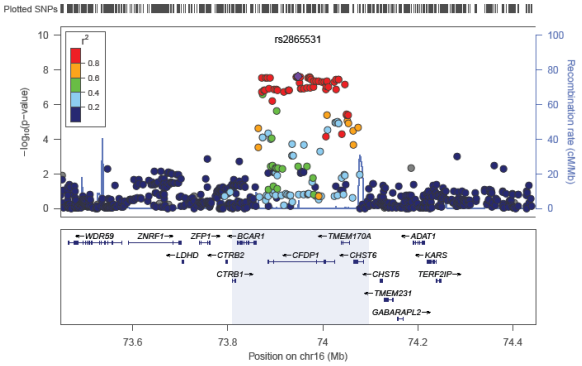
E. Region selection for targeted sequencing

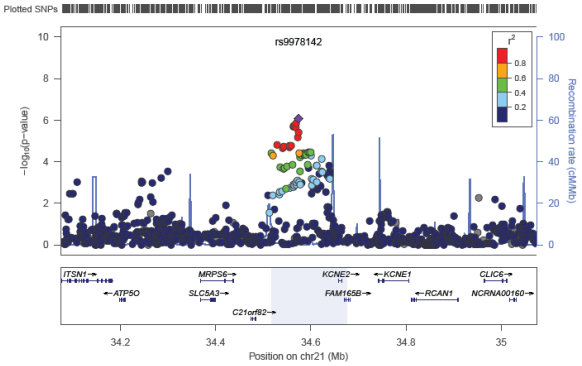



Plotted SNPs
rs2045517
rs10516526
rs11100860
rs153916
rs1985524
rs11134779
r²
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
Position on chr4 (Mb)
Position on chr5 (Mb)
HERC5
HERC3
FAM13A
TIGD2
GPRIN3
PIGY
NAP1L5
FAM13AOS
TET2
PPA2
FLJ20184
GSTCD
NPNT
TBCK
EEF1A1P9
INTS12
GYPA
HHIP
ANAPC10
MCTP1
FAM81B
ARSK
SPATA9
GLRX
ELL2
MIR583
TTC37
GPR150
RHOBTB3
RFESD
C5orf27
SPINK5
SPINK6
SPINK7
HTR4
ADRB2
SPINK5L2
SPINK9
SPINK5L3
FBXO38
HAVCR1
ITK
CYFIP2
NIPAL4
SOX30
THG1L
HAVCR2
C5orf40
ADAM19
C5orf52
CLINT1
MED7
LSM11
FAM71B


Plotted SNPs
rs6903823
rs2857595
rs2030600
rs2798641
rs262129
rs16909859
Recombination rate (cM/Mb)
Position on chr6 (Mb)
Position on chr9 (Mb)


Plotted SNPs
rs7068966
rs11001819
rs11172113
rs1036429
rs8033889
rs12447804
Position on chr10 (Mb)
Position on chr10 (Mb)
Position on chr12 (Mb)
Position on chr12 (Mb)
Position on chr15 (Mb)
Position on chr16 (Mb)
Recombination rate (cM/Mb)


Plotted SNPs
rs2865531
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
WDR59
ZNRF1
ZFP1
BCAR1
TMEM170A
ADAT1
LDHD
CTRB2
CFDP1
CHST6
KARS
CTRB1
CHST5
TERF2IP
TMEM231
GABARAPL2
Position on chr16 (Mb)


Plotted SNPs
rs9978142
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
ITSN1
MRPS6
KCNE2
KCNE1
CLIC6
ATP5O
SLC5A3
FAM165B
RCAN1
NCRNA00160
C21orf82
Position on chr21 (Mb)

F. Additional Syzygy method details

Error rate estimation

Syzygy estimates sequencing error rate by modelling the miscall rate, defined as $(C - REFrc)/C$ for coverage C and. $REFrc$ number of reference allele read counts. To do that it assumes that the factors that explain base to base variation in the miscall rate are: strand, sequence context and coverage around a base. A neighborhood quality score (nqs) is calculated in order to identify bases with lower coverage respect to their neighbors, since this can indicate lower accuracy of the calls for these bases. The neighborhood quality score compares the coverage at a given position with the coverage of the neighboring bases (+/- 10kb)

$$nqs = \frac{C\ at\ a\ base}{median\ (C\ of\ bases\ +and-10\ bp, with\ C > 0)}$$

Syzygy selects bases not included in dbSNP137 with miscall rate < 0.01, neighborhood quality score (nqs) > 0.2 and ≤ 1.5. and coverage > 1. Then, it models the miscall rate for these bases using nqs and the $sequencing\ context$ (a factor of trinucleotides for a base +/- 1bp) as covariates for each strand separately. After that, it uses the estimated effect sizes of nqs and $sequencing\ context$ to estimate the error rate for all bases with $nqs \geq 0.2$ and ≥ 1.7. This is done across pools (for example to estimate the miscall rate, C and $REFac$ are added up across pools) and for each strand separately, so it produces error rate estimates per position and per strand ($ER_{position,strand}$). Error rates specific for each pool, as used in the LOD score calculations, that vary per position, strand and pool ($ER_{pool,position,strand}$) are obtained as follows:

$$ER_{pool,position,strand} = \max(\min ER_{pool,strand}, ER_{position,strand})$$

$$\text{for } \min ER_{pool,strand} = \max(\text{miscall rate}_{pool,strand}, 0.001)$$

$$\text{for miscall rate}_{pool,strand} = \frac{\sum_{p \epsilon P} C_{pool,strand,p} - \sum_{p \epsilon P} REFrc_{pool,strand,p}}{\sum_{p \epsilon P} C_{pool,strand,p}}$$

$$and\ P = all\ positions$$

Strand bias test

Syzygy undertakes an additional test for strand bias, more specifically it tests whether the minor allele frequency estimated for the forward strand (MAF_{fwd}) and the reverse strand (MAF_{rev}) are equal to the overall allele frequency across both strands (MAF) or the allele frequency in one strand is equal to the overall allele frequency and the allele frequency in the other strand is 0.

$$H_0: MAF_{fwd} = MAF_{rev} = MAF$$

$$H_1: MAF_{fwd} = MAF\ and\ MAF_{rev} = 0$$

$$H_2: MAF_{rev} = MAF\ and\ MAF_{fwd} = 0$$

To do this it constructs a strand logarithm of odds (LOD) score ($strand\ LOD\ score$) comparing the maximum of the likelihood (L) of obtaining the data under H_1or H_2vs the likelihood of obtaining the data under H_0.

$$strand\ LOD\ score = \log_{10}(\frac{\max(L(H_1), L(H_2))}{L(H_0)})$$

Likelihoods are computed as explained previously using the Bayes' Rule.

G. Chapter 4 additional tables

Single variant results for known variants

Results of COPD association for variants previously associated with lung function are presented here and ordered by chromosome and position. P-values < 0.05 are highlighted in bold. The column “GWAS gene” presents the gene reported in the lung function GWAS undertaken in Chapter 3 for each region. Abbreviations: MAF = minor allele frequency, N. alt. a.c. = number of alternative allele counts, N. ref. a.c. = number of reference allele counts, O.R. = odds ratio.

Rs number (chr.: position)	GWAS gene	Calling algorithm	Ref. allele	Alt. allele	MAF	N. alt. a.c. case/ N. ref. a.c. case	N. alt. a.c. control / N. ref. a.c. control	O.R.	P-value	Consistent direction of effect?
rs2284746 (chr1:17306675)	*MFAP2*	vipR	G	C	0.48	246/254	207/243	1.14	$3.3x10^{-1}$	NO
rs2284746 (chr1:17306675)	*MFAP2*	syzygy	C	G	0.46	328/272	266/234	1.06	$6.28x10^{-1}$	YES
rs2284746 (chr1:17306675)	*MFAP2*	SNVer	C	G	0.48	318/282	258/242	1.06	$6.71x10^{-1}$	YES
rs993925 (chr1:218860068)	*TGFB2*	SNVer	C	T	0.38	237/363	179/321	1.17	$2.12x10^{-1}$	NO
rs993925 (chr1:218860068)	*TGFB2*	syzygy	C	T	0.38	239/361	182/318	1.16	$2.62x10^{-1}$	NO
rs993925 (chr1:218860068)	*TGFB2*	vipR	C	T	0.39	220/330	149/251	1.12	$4.19x10^{-1}$	NO
rs2571445 (chr2:218683154)	*TNS1*	SNVer	A	G	0.49	300/300	268/232	0.87	$2.5x10^{-1}$	YES
rs2571445 (chr2:218683154)	*TNS1*	vipR	G	A	0.44	118/82	52/48	1.33	$2.68x10^{-1}$	YES
rs2571445 (chr2:218683154)	*TNS1*	syzygy	A	G	0.48	304/296	266/234	0.90	$4.31x10^{-1}$	YES
rs12477314 (chr2:239877148)	*HDAC4*	SNVer	C	T	0.18	114/486	86/414	1.13	$4.8x10^{-1}$	NO
rs12477314 (chr2:239877148)	*HDAC4*	syzygy	C	T	0.18	114/486	88/412	1.10	$5.84x10^{-1}$	NO
rs12477314 (chr2:239877148)	*HDAC4*	vipR	C	T	0.21	98/352	70/280	1.11	$6x10^{-1}$	NO
rs1529672 (chr3:25520582)	*RARB*	SNVer	C	A	0.15	86/514	83/417	0.84	$3.14x10^{-1}$	YES

Rs number (chr.: position)	GWAS gene	Calling algorithm	Ref. allele	Alt. allele	MAF	N. alt. a.c. case/ N. ref. a.c. case	N. alt. a.c. control / N. ref. a.c. control	O.R.	P-value	Consistent direction of effect?
rs1529672 (chr3:25520582)	*RARB*	syzygy	C	A	0.16	89/511	84/416	0.86	$4.06x10^{-1}$	YES
rs1529672 (chr3:25520582)	*RARB*	vipR	C	A	0.16	85/465	82/418	0.93	$7.36x10^{-1}$	YES
rs1344555 (chr3:169300219)	*MECOM*	SNVer	C	T	0.2	143/457	77/423	1.72	$4.93x10^{-4}$	YES
rs1344555 (chr3:169300219)	*MECOM*	syzygy	C	T	0.21	148/452	81/419	1.69	$5.93x10^{-4}$	YES
rs1344555 (chr3:169300219)	*MECOM*	vipR	C	T	0.23	130/370	54/246	1.60	$9.29x10^{-3}$	YES
rs2045517 (chr4:89870964)	*FAM13A*	vipR	C	T	0.45	228/272	200/250	1.05	$7.44x10^{-1}$	YES
rs2045517 (chr4:89870964)	*FAM13A*	SNVer	C	T	0.43	260/340	219/281	0.98	$9.03x10^{-1}$	NO
rs2045517 (chr4:89870964)	*FAM13A*	syzygy	C	T	0.44	264/336	219/281	1.01	$9.51x10^{-1}$	YES
rs10516526 (chr4:106688904)	*GSTCD*	vipR	A	G	0.09	37/313	26/324	1.47	$1.86x10^{-1}$	NO
rs10516526 (chr4:106688904)	*GSTCD*	SNVer	A	G	0.07	42/558	30/470	1.18	$5.42x10^{-1}$	NO
rs10516526 (chr4:106688904)	*GSTCD*	syzygy	A	G	0.07	44/556	32/468	1.16	$5.54x10^{-1}$	NO
rs11100860 (chr4:145479139)	*HHIP*	SNVer	A	G	0.37	202/398	205/295	0.73	$1.44x10^{-2}$	YES
rs11100860 (chr4:145479139)	*HHIP*	syzygy	A	G	0.37	204/396	202/298	0.76	$3.28x10^{-2}$	YES
rs11100860 (chr4:145479139)	*HHIP*	vipR	A	G	0.4	159/241	203/297	0.97	$8.37x10^{-1}$	YES
rs153916 (chr5:95036700)	*SPATA9*	SNVer	C	T	0.43	360/240	265/235	1.33	$2.03x10^{-2}$	YES
rs153916 (chr5:95036700)	*SPATA9*	syzygy	C	T	0.43	363/237	268/232	1.33	$2.35x10^{-2}$	YES
rs153916 (chr5:95036700)	*SPATA9*	vipR	T	C	0.45	191/259	237/263	0.82	$1.33x10^{-1}$	YES
rs1985524 (chr5:147847788)	*HTR4*	SNVer	G	C	0.41	227/373	224/276	0.75	$2.27x10^{-2}$	YES
rs1985524 (chr5:147847788)	*HTR4*	syzygy	G	C	0.41	230/370	223/277	0.77	$3.67x10^{-2}$	YES
rs1985524 (chr5:147847788)	*HTR4*	vipR	G	C	0.44	171/229	223/277	0.93	$5.89x10^{-1}$	YES
rs11134779 (chr5:156936766)	*ADAM19*	vipR	A	G	0.35	166/334	150/250	0.83	$1.83x10^{-1}$	NO

Rs number (chr.: position)	GWAS gene	Calling algorithm	Ref. allele	Alt. allele	MAF	N. alt. a.c. case/ N. ref. a.c. case	N. alt. a.c. control / N. ref. a.c. control	O.R.	P-value	Consistent direction of effect?
rs11134779 (chr5:156936766)	*ADAM19*	SNVer	A	G	0.32	187/413	171/329	0.87	$3.01x10^{-1}$	NO
rs11134779 (chr5:156936766)	*ADAM19*	syzygy	A	G	0.33	190/410	169/331	0.91	$4.78x10^{-1}$	NO
rs6903823 (chr6:28322296)	*ZKSCAN3*	vipR	A	G	0.25	135/465	103/247	0.70	$1.98x10^{-2}$	NO
rs6903823 (chr6:28322296)	*ZKSCAN3*	syzygy	A	G	0.23	132/468	117/383	0.92	$6.13x10^{-1}$	NO
rs6903823 (chr6:28322296)	*ZKSCAN3*	SNVer	A	G	0.24	120/380	122/378	0.98	$9.41x10^{-1}$	NO
rs2857595 (chr6:31568469)	*NCR3*	vipR	G	A	0.24	114/336	102/348	1.16	$3.91x10^{-1}$	YES
rs2857595 (chr6:31568469)	*NCR3*	SNVer	G	A	0.22	133/467	109/391	1.02	$9.42x10^{-1}$	YES
rs2857595 (chr6:31568469)	*NCR3*	syzygy	G	A	0.22	133/467	112/388	0.99	$9.42x10^{-1}$	NO
rs2070600 (chr6:32151443)	*AGER*	SNVer	C	T	0.06	37/463	24/426	1.42	$2.33x10^{-1}$	NO
rs2070600 (chr6:32151443)	*AGER*	syzygy	C	T	0.07	48/552	31/469	1.32	$2.91x10^{-1}$	NO
rs2070600 (chr6:32151443)	*AGER*	vipR	C	T	0.09	31/319	23/227	0.96	$8.86x10^{-1}$	YES
rs2798641 (chr6:109268050)	*ARMC2*	vipR	C	T	0.19	96/454	94/356	0.80	$1.7x10^{-1}$	NO
rs2798641 (chr6:109268050)	*ARMC2*	SNVer	C	T	0.18	100/500	99/401	0.81	$1.82x10^{-1}$	NO
rs2798641 (chr6:109268050)	*ARMC2*	syzygy	C	T	0.18	103/497	96/404	0.87	$3.88x10^{-1}$	NO
rs262129 (chr6:142853144)	*LOC153910*	SNVer	A	G	0.29	177/423	144/356	1.03	$8.42x10^{-1}$	NO
rs262129 (chr6:142853144)	*LOC153910*	vipR	A	G	0.31	169/381	140/310	0.98	$9.45x10^{-1}$	YES
rs262129 (chr6:142853144)	*LOC153910*	syzygy	A	G	0.29	175/425	146/354	1.00	1	YES
rs16909859 (chr9:98204792)	*PTCH1*	SNVer	G	A	0.05	25/575	32/468	0.64	$1.03x10^{-1}$	NO
rs16909859 (chr9:98204792)	*PTCH1*	syzygy	G	A	0.05	28/572	31/469	0.74	$2.84x10^{-1}$	NO
rs16909859 (chr9:98204792)	*PTCH1*	vipR	G	A	0.06	16/284	26/324	0.70	$3.38x10^{-1}$	NO

Rs number (chr.: position)	GWAS gene	Calling algorithm	Ref. allele	Alt. allele	MAF	N. alt. a.c. case/ N. ref. a.c. case	N. alt. a.c. control / N. ref. a.c. control	O.R.	P-value	Consistent direction of effect?
rs7068966 (chr10:12277992)	*CDC123*	vipR	C	T	0.48	259/241	230/220	1.03	$8.46x10^{-1}$	NO
rs7068966 (chr10:12277992)	*CDC123*	SNVer	C	T	0.5	298/302	252/248	0.97	$8.56x10^{-1}$	YES
rs7068966 (chr10:12277992)	*CDC123*	syzygy	C	T	0.5	302/298	250/250	1.01	$9.52x10^{-1}$	NO
rs11001819 (chr10:78315224)	*C10orf11*	vipR	G	A	0.48	272/278	208/242	1.14	$3.4x10^{-1}$	NO
rs11001819 (chr10:78315224)	*C10orf11*	syzygy	G	A	0.47	291/309	228/272	1.12	$3.63x10^{-1}$	NO
rs11001819 (chr10:78315224)	*C10orf11*	SNVer	G	A	0.46	283/317	225/275	1.09	$5.04x10^{-1}$	NO
rs11172113 (chr12:57527283)	*LRP1*	vipR	T	C	0.44	161/189	148/202	1.16	$3.61x10^{-1}$	NO
rs11172113 (chr12:57527283)	*LRP1*	SNVer	T	C	0.4	233/367	204/296	0.92	$5.36x10^{-1}$	YES
rs11172113 (chr12:57527283)	*LRP1*	syzygy	T	C	0.4	234/366	205/295	0.92	$5.36x10^{-1}$	YES
rs1036429 (chr12:96271428)	*CCDC38*	vipR	C	T	0.25	96/304	89/261	0.93	$6.72x10^{-1}$	YES
rs1036429 (chr12:96271428)	*CCDC38*	syzygy	T	C	0.22	467/133	386/114	1.04	$8.28x10^{-1}$	YES
rs1036429 (chr12:96271428)	*CCDC38*	SNVer	T	C	0.23	460/140	383/117	1.00	1	YES
rs8033889 (chr15:71680080)	*THSD4*	SNVer	G	T	0.2	116/484	107/393	0.88	$4.08x10^{-1}$	NO
rs8033889 (chr15:71680080)	*THSD4*	syzygy	G	T	0.21	120/480	109/391	0.90	$5.02x10^{-1}$	NO
rs8033889 (chr15:71680080)	*THSD4*	vipR	G	T	0.21	105/395	105/395	1.00	1	NA
rs12447804 (chr16:58075282)	*MMP15*	vipR	C	T	0.22	100/350	75/275	1.05	$7.97x10^{-1}$	YES
rs12447804 (chr16:58075282)	*MMP15*	SNVer	C	T	0.2	119/481	101/399	0.98	$8.8x10^{-1}$	NO
rs12447804 (chr16:58075282)	*MMP15*	syzygy	C	T	0.2	120/480	100/400	1.00	1	NA
rs35263058 (chr16:75391937)	*CFDP1*	SNVer	T	C	0.41	348/252	304/196	0.89	$3.56x10^{-1}$	NO
rs35263058 (chr16:75391937)	*CFDP1*	syzygy	T	C	0.41	352/248	302/198	0.93	$5.79x10^{-1}$	NO

Rs number (chr.: position)	GWAS gene	Calling algorithm	Ref. allele	Alt. allele	MAF	N. alt. a.c. case/ N. ref. a.c. case	N. alt. a.c. control / N. ref. a.c. control	O.R.	P-value	Consistent direction of effect?
rs35263058 (chr16:75391937)	*CFDP1*	vipR	C	T	0.44	221/279	170/230	1.07	$6.36x10^{-1}$	NO
rs9978142 (chr21:35652239)	*KCNE2*	SNVer	A	T	0.16	100/500	76/424	1.12	$5.63x10^{-1}$	YES
rs9978142 (chr21:35652239)	*KCNE2*	vipR	A	T	0.17	99/501	77/373	0.96	$8.03x10^{-1}$	NO
rs9978142 (chr21:35652239)	*KCNE2*	syzygy	A	T	0.16	98/502	79/421	1.04	$8.69x10^{-1}$	YES

Burden test top hits in stage 1

Locus (sliding windows or genes) that reach the threshold for follow-up after sensitivity analyses either with ("Independent variants and variants not in UK10K+1000G") or without ("Independent variants") including variants not in UK10K+1000G for the burden test. The column "GWAS gene" presents the gene reported in the lung function GWAS undertaken in Chapter 3 for each region. Abbreviations: N.: number of variants

Locus	Threshold	All variants						Independent variants						Independent variants and variants not in UK10K+1000G					
		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy	
		P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N.
chr3: 168984786-168987786	$1.87x10^{-4}$	NA	NA	$2.25x10^{-4}$	5	$7.87x10^{-5}$	5	NA	NA	$7.24x10^{-3}$	2	$9.81x10^{-4}$	2	NA	NA	$2.25x10^{-4}$	5	$7.87x10^{-5}$	5
FLJ20184	$1x10^{-2}$	1	27	$1.28x10^{-2}$	42	$5.12x10^{-4}$	46	1	15	$1.43x10^{-2}$	23	$1.12x10^{-4}$	25	1	22	$4.21x10^{-2}$	37	$1.57x10^{-3}$	41
chr4: 145278600-145281600	$2.76x10^{-4}$	NA	NA	$2.34x10^{-4}$	4	$2.15x10^{-4}$	4	NA	NA	$3.76x10^{-3}$	2	$1.38x10^{-3}$	2	NA	NA	$2.34x10^{-4}$	4	$2.15x10^{-4}$	4
ITK	$8.33x10^{-3}$	1	13	$3.98x10^{-4}$	28	$1.41x10^{-4}$	29	1	10	$5.81x10^{-5}$	20	$5.01x10^{-5}$	21	1	12	$3.98x10^{-4}$	28	$1.64x10^{-4}$	28
GPR126	$2.5x10^{-2}$	$7.18x10^{-1}$	63	$2.04x10^{-3}$	110	$3.21x10^{-2}$	109	$7.98x10^{-1}$	33	$8.11x10^{-2}$	61	$3.91x10^{-1}$	60	$5.15x10^{-1}$	50	$1.51x10^{-3}$	93	$4.65x10^{-2}$	91

C-alpha test top hits in stage 1

Locus (sliding windows: a), genes: b) or exon based genes: c)) that reach the threshold for follow-up after sensitivity analyses either with ("Independent variants and variants not in UK10K+1000G") or without ("Independent variants") including variants not in UK10K+1000G for the burden test. The column "GWAS gene" presents the gene reported in the lung function GWAS for each region. Abbreviations: N.: number.

a) Sliding window

Locus	Threshold	All variants						Independent variants						Independent variants and variants not in UK10K+1000G					
		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy	
		N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P
chr1:218531175-218534175	$4.24x10^{-4}$	1	NA	2	$7.67x10^{-7}$	2	$7.67x10^{-7}$	1	NA	2	$7.67x10^{-7}$	2	$7.67x10^{-7}$	1	NA	2	$7.67x10^{-7}$	2	$7.67x10^{-7}$
chr2:218807794-218810794	$1.09x10^{-3}$	3	$6.33x10^{-5}$	4	$1.91x10^{-5}$	3	$2.92x10^{-3}$	2	$9.25x10^{-4}$	3	$1.19x10^{-4}$	2	$3.5x10^{-2}$	2	$9.25x10^{-4}$	3	$1.19x10^{-4}$	2	$3.5x10^{-2}$
chr2:239890616-239893616	$3.68x10^{-4}$	1	NA	3	$5.76x10^{-6}$	3	$6.33x10^{-5}$	NA	NA	2	$6.05x10^{-6}$	2	$6.65x10^{-5}$	1	NA	3	$5.76x10^{-6}$	3	$6.33x10^{-5}$
chr2:239971616-239974616	$3.68x10^{-4}$	1	NA	3	$8.03x10^{-11}$	3	$1.33x10^{-9}$	1	NA	3	$8.03x10^{-11}$	3	$1.33x10^{-9}$	1	NA	3	$8.03x10^{-11}$	3	$1.33x10^{-9}$
chr2:239973116-239976116	$3.68x10^{-4}$	1	NA	4	$2.71x10^{-13}$	4	$3.4x10^{-12}$	1	NA	3	$8.03x10^{-11}$	3	$1.33x10^{-9}$	1	NA	3	$8.03x10^{-11}$	3	$1.33x10^{-9}$
chr2:240325616-240328616	$3.68x10^{-4}$	2	$2.61x10^{-6}$	2	$3.95x10^{-11}$	2	$7.05x10^{-3}$	2	$2.61x10^{-6}$	2	$3.95x10^{-11}$	2	$7.05x10^{-3}$	2	$2.61x10^{-6}$	2	$3.95x10^{-11}$	2	$7.05x10^{-3}$
chr3:168984786-168987786	$1.87x10^{-4}$	2	$2.48x10^{-2}$	5	$1.5x10^{-5}$	5	$2.74x10^{-6}$	1	NA	2	$1.12x10^{-4}$	2	$1.06x10^{-5}$	2	$2.48x10^{-2}$	5	$1.5x10^{-5}$	5	$2.74x10^{-6}$
chr3:169238286-169241286	$1.87x10^{-4}$	1	NA	4	$4.52x10^{-5}$	4	$5.34x10^{-6}$	1	NA	2	$3.5x10^{-4}$	2	$1.46x10^{-5}$	1	NA	4	$4.52x10^{-5}$	4	$5.34x10^{-6}$
chr3:169310286-169313286	$1.87x10^{-4}$	2	$2.48x10^{-2}$	3	$1.99x10^{-8}$	3	$1.99x10^{-8}$	2	$2.48x10^{-2}$	2	$6.65x10^{-5}$	2	$2.85x10^{-6}$	2	$2.48x10^{-2}$	2	$6.65x10^{-5}$	2	$2.85x10^{-6}$
chr3:169311786-169314786	$1.87x10^{-4}$	2	$2.48x10^{-2}$	3	$1.99x10^{-8}$	3	$1.99x10^{-8}$	2	$2.48x10^{-2}$	2	$6.65x10^{-5}$	2	$2.85x10^{-6}$	2	$2.48x10^{-2}$	2	$6.65x10^{-5}$	2	$2.85x10^{-6}$

Locus	Threshold	All variants						Independent variants						Independent variants and variants not in UK10K+1000G					
		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy	
		N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P
chr3:169340286-169343286	$1.87x10^{-4}$	3	$2.95x10^{-9}$	4	$5.29x10^{-6}$	4	$8.75x10^{-6}$	2	$7.82x10^{-9}$	2	$7.06x10^{-5}$	2	$7.06x10^{-5}$	2	$7.82x10^{-9}$	3	$3.82x10^{-5}$	3	$6.7x10^{-5}$
chr3:169341786-169344786	$1.87x10^{-4}$	6	$2.54x10^{-10}$	5	$3.03x10^{-6}$	6	$1.37x10^{-10}$	4	$1.7x10^{-9}$	2	$7.06x10^{-5}$	3	$9.65x10^{-10}$	5	$6.58x10^{-10}$	4	$2.05x10^{-5}$	5	$4.54x10^{-10}$
chr3:169371786-169374786	$1.87x10^{-4}$	2	$7.36x10^{-2}$	7	$2.69x10^{-5}$	6	$1.36x10^{-5}$	1	NA	4	$2.18x10^{-4}$	3	$6.3x10^{-5}$	2	$7.36x10^{-2}$	7	$2.69x10^{-5}$	6	$1.36x10^{-5}$
chr3:169373286-169376286	$1.87x10^{-4}$	4	$4.59x10^{-2}$	6	$2.64x10^{-7}$	6	$9.73x10^{-7}$	3	$1.26x10^{-1}$	3	$5.39x10^{-6}$	3	$5.39x10^{-6}$	4	$4.59x10^{-2}$	6	$2.64x10^{-7}$	6	$9.73x10^{-7}$
chr3:25464333-25467333	$5.81x10^{-4}$	3	$1.25x10^{-4}$	4	$1.63x10^{-5}$	5	$9.01x10^{-3}$	3	$1.25x10^{-4}$	3	$1.51x10^{-5}$	4	$5.69x10^{-3}$	3	$1.25x10^{-4}$	4	$1.63x10^{-5}$	5	$9.01x10^{-3}$
chr3:25510833-25513833	$5.81x10^{-4}$	4	$1.62x10^{-3}$	3	$5.19x10^{-8}$	2	$5.02x10^{-7}$	4	$1.62x10^{-3}$	3	$5.19x10^{-8}$	2	$5.02x10^{-7}$	4	$1.62x10^{-3}$	3	$5.19x10^{-8}$	2	$5.02x10^{-7}$
chr3:25512333-25515333	$5.81x10^{-4}$	3	$2.03x10^{-3}$	2	$6x10^{-8}$	2	$5.02x10^{-7}$	3	$2.03x10^{-3}$	2	$6x10^{-8}$	2	$5.02x10^{-7}$	3	$2.03x10^{-3}$	2	$6x10^{-8}$	2	$5.02x10^{-7}$
chr3:25527333-25530333	$5.81x10^{-4}$	1	NA	2	$6.65x10^{-5}$	2	$6.65x10^{-5}$	1	NA	2	$6.65x10^{-5}$	2	$6.65x10^{-5}$	1	NA	2	$6.65x10^{-5}$	2	$6.65x10^{-5}$
chr3:25599333-25602333	$5.81x10^{-4}$	2	$3.76x10^{-5}$	2	$1.52x10^{-4}$	2	$1.52x10^{-4}$	2	$3.76x10^{-5}$	2	$1.52x10^{-4}$	2	$1.52x10^{-4}$	2	$3.76x10^{-5}$	2	$1.52x10^{-4}$	2	$1.52x10^{-4}$
chr3:25632333-25635333	$5.81x10^{-4}$	2	$2.48x10^{-2}$	6	$1.14x10^{-5}$	6	$7.35x10^{-4}$	2	$2.48x10^{-2}$	6	$1.14x10^{-5}$	6	$7.35x10^{-4}$	2	$2.48x10^{-2}$	6	$1.14x10^{-5}$	6	$7.35x10^{-4}$
chr3:25633833-25636833	$5.81x10^{-4}$	1	NA	7	$3.36x10^{-7}$	7	$1.87x10^{-5}$	1	NA	5	$9.43x10^{-7}$	5	$4.73x10^{-5}$	1	NA	7	$3.36x10^{-7}$	7	$1.87x10^{-5}$
chr4:106514233-106517233	$2.69x10^{-4}$	1	NA	2	$3.12x10^{-9}$	2	$7.37x10^{-9}$	1	NA	2	$3.12x10^{-9}$	2	$7.37x10^{-9}$	1	NA	2	$3.12x10^{-9}$	2	$7.37x10^{-9}$
chr4:106515733-106518733	$2.69x10^{-4}$	2	$4.19x10^{-3}$	5	$3.15x10^{-9}$	5	$8.26x10^{-9}$	1	NA	3	$2.17x10^{-8}$	3	$5.09x10^{-8}$	2	$4.19x10^{-3}$	5	$3.15x10^{-9}$	5	$8.26x10^{-9}$
chr4:145265100-145268100	$2.76x10^{-4}$	3	$5.88x10^{-1}$	6	$4.45x10^{-7}$	7	$1.26x10^{-11}$	3	$5.88x10^{-1}$	3	$1.71x10^{-4}$	4	$1.76x10^{-8}$	3	$5.88x10^{-1}$	5	$6.4x10^{-5}$	6	$1.66x10^{-9}$
chr4:145266600-145269600	$2.76x10^{-4}$	5	$5.57x10^{-3}$	9	$1.04x10^{-8}$	8	$6.12x10^{-11}$	3	$1.73x10^{-2}$	4	$1.44x10^{-5}$	4	$2.66x10^{-7}$	5	$5.57x10^{-3}$	8	$1.33x10^{-6}$	7	$7.91x10^{-9}$
chr4:145268100-145271100	$2.76x10^{-4}$	8	$1.45x10^{-3}$	9	$9.28x10^{-7}$	8	$5.78x10^{-8}$	5	$5.57x10^{-3}$	5	$6.14x10^{-6}$	5	$2.39x10^{-7}$	8	$1.45x10^{-3}$	9	$9.28x10^{-7}$	8	$5.78x10^{-8}$

Locus	Threshold	All variants						Independent variants						Independent variants and variants not in UK10K+1000G					
		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy	
		N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P
chr4:145269600-145272600	$2.76x10^{-4}$	4	$4.59x10^{-2}$	4	$1.37x10^{-4}$	4	$4.21x10^{-5}$	3	$5.88x10^{-2}$	3	$1.71x10^{-4}$	3	$3.89x10^{-5}$	4	$4.59x10^{-2}$	4	$1.37x10^{-4}$	4	$4.21x10^{-5}$
chr4:145272600-145275600	$2.76x10^{-4}$	3	$1.01x10^{-2}$	4	$8.26x10^{-6}$	4	$3.01x10^{-6}$	3	$1.01x10^{-2}$	4	$8.26x10^{-6}$	4	$3.01x10^{-6}$	3	$1.01x10^{-2}$	4	$8.26x10^{-6}$	4	$3.01x10^{-6}$
chr4:145278600-145281600	$2.76x10^{-4}$	3	$1.26x10^{-1}$	4	$2.55x10^{-4}$	4	$1.46x10^{-5}$	2	$1.75x10^{-1}$	2	$4.02x10^{-4}$	2	$1.69x10^{-5}$	3	$1.26x10^{-1}$	4	$2.55x10^{-4}$	4	$1.46x10^{-5}$
chr4:145289100-145292100	$2.76x10^{-4}$	4	$4.59x10^{-2}$	7	$2.91x10^{-6}$	7	$4.52x10^{-6}$	3	$5.88x10^{-2}$	5	$3.74x10^{-9}$	5	$2.05x10^{-8}$	3	$5.88x10^{-2}$	5	$3.74x10^{-9}$	5	$2.05x10^{-8}$
chr4:145290600-145293600	$2.76x10^{-4}$	2	$1.75x10^{-1}$	5	$2.66x10^{-7}$	5	$1.28x10^{-5}$	1	NA	3	$6.81x10^{-7}$	3	$3.33x10^{-5}$	2	$1.75x10^{-1}$	4	$6.31x10^{-7}$	4	$3.06x10^{-5}$
chr4:145293600-145296600	$2.76x10^{-4}$	3	$8.88x10^{-3}$	6	$1.58x10^{-8}$	5	$3.92x10^{-7}$	1	NA	2	$5.4x10^{-4}$	2	$1.11x10^{-5}$	1	NA	3	$5.31x10^{-3}$	3	$1.19x10^{-4}$
chr4:145332600-145335600	$2.76x10^{-4}$	5	$8.25x10^{-3}$	6	$1.14x10^{-9}$	7	$3.66x10^{-13}$	2	$3.47x10^{-2}$	3	$1.84x10^{-5}$	4	$7.88x10^{-7}$	2	$3.47x10^{-2}$	3	$1.84x10^{-5}$	4	$7.88x10^{-7}$
chr4:145334100-145337100	$2.76x10^{-4}$	7	$2.72x10^{-5}$	9	$1.48x10^{-17}$	10	$2.85x10^{-17}$	2	$1.12x10^{-2}$	2	$9.84x10^{-6}$	4	$3.08x10^{-4}$	3	$8.88x10^{-3}$	3	$9.18x10^{-5}$	4	$3.08x10^{-4}$
chr4:145335600-145338600	$2.76x10^{-4}$	10	$1.72x10^{-7}$	13	$1.25x10^{-27}$	13	$2.12x10^{-26}$	2	$3.61x10^{-3}$	2	$1.09x10^{-7}$	3	$3.62x10^{-5}$	3	$2.85x10^{-3}$	4	$7.58x10^{-7}$	4	$3.77x10^{-5}$
chr4:145341600-145344600	$2.76x10^{-4}$	2	$7.34x10^{-2}$	4	$3.65x10^{-8}$	4	$3.54x10^{-12}$	1	NA	2	$1.42x10^{-4}$	2	$9.27x10^{-8}$	1	NA	2	$1.42x10^{-4}$	2	$9.27x10^{-8}$
chr4:145382100-145385100	$2.76x10^{-4}$	5	$2.78x10^{-3}$	6	$3.33x10^{-9}$	6	$4.3x10^{-10}$	4	$7.04x10^{-3}$	5	$4.57x10^{-7}$	5	$5.79x10^{-8}$	4	$7.04x10^{-3}$	5	$4.57x10^{-7}$	5	$5.79x10^{-8}$
chr4:145383600-145386600	$2.76x10^{-4}$	3	$5.56x10^{-2}$	4	$9.98x10^{-7}$	4	$4.91x10^{-7}$	2	$1.75x10^{-1}$	3	$1.43x10^{-4}$	3	$6.99x10^{-5}$	2	$1.75x10^{-1}$	3	$1.43x10^{-4}$	3	$6.99x10^{-5}$
chr4:89812605-89815605	$5x10^{-4}$	2	$7.36x10^{-2}$	2	$9.25x10^{-4}$	2	$3.76x10^{-5}$	2	$7.36x10^{-2}$	2	$9.25x10^{-4}$	2	$3.76x10^{-5}$	2	$7.36x10^{-2}$	2	$9.25x10^{-4}$	2	$3.76x10^{-5}$
chr4:89814105-89817105	$5x10^{-4}$	2	$7.36x10^{-2}$	3	$5.76x10^{-6}$	3	$2.71x10^{-6}$	2	$7.36x10^{-2}$	3	$5.76x10^{-6}$	3	$2.71x10^{-6}$	2	$7.36x10^{-2}$	3	$5.76x10^{-6}$	3	$2.71x10^{-6}$
chr5:147826118-147829118	$4.9x10^{-4}$	2	$3.76x10^{-5}$	2	$9.87x10^{-14}$	2	$4.7x10^{-6}$	2	$3.76x10^{-5}$	2	$9.87x10^{-14}$	2	$4.7x10^{-6}$	2	$3.76x10^{-5}$	2	$9.87x10^{-14}$	2	$4.7x10^{-6}$
chr5:147829118-147832118	$4.9x10^{-4}$	NA	NA	3	$2.86x10^{-7}$	3	$6.81x10^{-7}$	NA	NA	2	$3.03x10^{-7}$	2	$7.22x10^{-7}$	NA	NA	3	$2.86x10^{-7}$	3	$6.81x10^{-7}$

Locus	Threshold	All variants						Independent variants						Independent variants and variants not in UK10K+1000G					
		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy	
		N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P
chr5:147830618-147833618	$4.9x10^{-4}$	NA	NA	3	$2.86x10^{-7}$	3	$6.81x10^{-7}$	NA	NA	2	$3.03x10^{-7}$	2	$7.22x10^{-7}$	NA	NA	3	$2.86x10^{-7}$	3	$6.81x10^{-7}$
chr5:156912906-156915906	$5.88x10^{-4}$	1	NA	3	$2.33x10^{-8}$	3	$5.11x10^{-5}$	1	NA	2	$6x10^{-8}$	2	$5.53x10^{-5}$	1	NA	3	$2.33x10^{-8}$	3	$5.11x10^{-5}$
chr9:98180197-98183197	$9.8x10^{-4}$	1	NA	3	$1.88x10^{-6}$	3	$1.06x10^{-4}$	1	NA	2	$2.68x10^{-7}$	2	$3.5x10^{-5}$	1	NA	3	$1.88x10^{-6}$	3	$1.06x10^{-4}$
chr9:98181697-98184697	$9.8x10^{-4}$	1	NA	4	$1.02x10^{-5}$	4	$2.84x10^{-4}$	1	NA	3	$1.88x10^{-6}$	3	$1.06x10^{-4}$	1	NA	4	$1.02x10^{-5}$	4	$2.84x10^{-4}$
chr10:12207674-12210674	$1.09x10^{-3}$	2	$2.7x10^{-2}$	4	$9.11x10^{-5}$	4	$1.6x10^{-7}$	2	$2.7x10^{-2}$	3	$4.8x10^{-4}$	3	$1.02x10^{-6}$	2	$2.7x10^{-2}$	4	$9.11x10^{-5}$	4	$1.6x10^{-7}$
chr10:12209174-12212174	$1.09x10^{-3}$	2	$2.7x10^{-2}$	4	$9.11x10^{-5}$	4	$1.6x10^{-7}$	2	$2.7x10^{-2}$	3	$4.8x10^{-4}$	3	$1.02x10^{-6}$	2	$2.7x10^{-2}$	4	$9.11x10^{-5}$	4	$1.6x10^{-7}$
chr10:77609018-77612018	$1.71x10^{-4}$	3	$1.55x10^{-1}$	5	$2.17x10^{-4}$	5	$6.3x10^{-5}$	3	$1.55x10^{-1}$	4	$2.04x10^{-4}$	4	$5.91x10^{-5}$	3	$1.55x10^{-1}$	5	$2.17x10^{-4}$	5	$6.3x10^{-5}$
chr12:57529676-57532676	$1.56x10^{-3}$	1	NA	4	$7.04x10^{-6}$	4	$6.14x10^{-4}$	1	NA	4	$7.04x10^{-6}$	4	$6.14x10^{-4}$	1	NA	4	$7.04x10^{-6}$	4	$6.14x10^{-4}$
chr12:96134582-96137582	$5.05x10^{-4}$	4	$4.66x10^{-3}$	3	$4.44x10^{-6}$	2	$5.33x10^{-4}$	3	$1.02x10^{-2}$	2	$9.28x10^{-6}$	2	$5.33x10^{-4}$	4	$4.66x10^{-3}$	3	$4.44x10^{-6}$	2	$5.33x10^{-4}$
chr12:96136082-96139082	$5.05x10^{-4}$	3	$6.69x10^{-2}$	3	$4.44x10^{-6}$	2	$5.33x10^{-4}$	2	$2.24x10^{-1}$	2	$9.28x10^{-6}$	2	$5.33x10^{-4}$	3	$6.69x10^{-2}$	3	$4.44x10^{-6}$	2	$5.33x10^{-4}$
chr12:96157082-96160082	$5.05x10^{-4}$	1	NA	2	$4.69x10^{-7}$	2	$4.84x10^{-9}$	1	NA	2	$4.69x10^{-7}$	2	$4.84x10^{-9}$	1	NA	2	$4.69x10^{-7}$	2	$4.84x10^{-9}$
chr12:96158582-96161582	$5.05x10^{-4}$	2	$7.36x10^{-2}$	3	$1.22x10^{-4}$	3	$2.78x10^{-6}$	2	$7.36x10^{-2}$	3	$1.22x10^{-4}$	3	$2.78x10^{-6}$	2	$7.36x10^{-2}$	3	$1.22x10^{-4}$	3	$2.78x10^{-6}$
chr12:96335582-96338582	$5.05x10^{-4}$	1	NA	3	$2.56x10^{-9}$	3	$5.15x10^{-5}$	1	NA	3	$2.56x10^{-9}$	3	$5.15x10^{-5}$	1	NA	3	$2.56x10^{-9}$	3	$5.15x10^{-5}$
chr15:71704287-71707287	$2.66x10^{-4}$	3	$1.45x10^{-3}$	3	$8.69x10^{-6}$	3	$2.42x10^{-4}$	2	$4.19x10^{-3}$	2	$2.46x10^{-6}$	2	$7.9x10^{-5}$	3	$1.45x10^{-3}$	3	$8.69x10^{-6}$	3	$2.42x10^{-4}$
chr16:58032243-58035243	$1.61x10^{-3}$	NA	NA	3	$1.67x10^{-7}$	4	$2.68x10^{-6}$	NA	NA	2	$3.99x10^{-9}$	2	$3.25x10^{-4}$	NA	NA	2	$3.99x10^{-9}$	3	$1.81x10^{-5}$
chr21:35645321-35648321	$1.25x10^{-3}$	3	$1.71x10^{-4}$	4	$1.99x10^{-2}$	5	$1.82x10^{-4}$	2	$4.14x10^{-4}$	2	$5.4x10^{-2}$	3	$4.83x10^{-4}$	3	$1.71x10^{-4}$	3	$3.33x10^{-2}$	4	$2.98x10^{-4}$

Locus	Threshold	All variants						Independent variants						Independent variants and variants not in UK10K+1000G					
		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy	
		N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P
chr21:35646821-35649821	$1.25x10^{-3}$	4	$1.59x10^{-6}$	5	$1.65x10^{-2}$	5	$1.82x10^{-4}$	2	$4.14x10^{-4}$	2	$5.4x10^{-2}$	3	$4.83x10^{-4}$	3	$1.71x10^{-4}$	3	$3.33x10^{-2}$	4	$2.98x10^{-4}$

b) Gene based

Locus	Threshold	All variants						Independent variants						Independent variants and variants not in UK10K+1000G					
		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy	
		N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P
TGFB2	$5x10^{-2}$	34	$1.61x10^{-3}$	60	$7.43x10^{-7}$	65	$2.9x10^{-5}$	21	$3.34x10^{-3}$	31	$3.54x10^{-8}$	33	$5.57x10^{-9}$	30	$1.39x10^{-3}$	45	$2.92x10^{-9}$	49	$7.7x10^{-8}$
TNS1	$5x10^{-2}$	22	$5.32x10^{-4}$	69	$2.07x10^{-2}$	72	$3.74x10^{-1}$	12	$1.52x10^{-2}$	41	$3.38x10^{-2}$	42	$3.08x10^{-1}$	19	$5.75x10^{-3}$	61	$5.2x10^{-3}$	62	$1.56x10^{-1}$
HDAC4	$5x10^{-2}$	65	$3.43x10^{-34}$	174	$1.42x10^{-22}$	189	$2.95x10^{-4}$	32	$3.2x10^{-24}$	93	$2.07x10^{-18}$	100	$2.15x10^{-5}$	41	$7.87x10^{-26}$	141	$8.81x10^{-25}$	151	$2.63x10^{-5}$
RARB	$2.5x10^{-2}$	68	$1.48x10^{-17}$	125	$7.24x10^{-12}$	127	$2.44x10^{-7}$	48	$4.33x10^{-13}$	72	$1.17x10^{-7}$	76	$5.33x10^{-4}$	60	$1.02x10^{-15}$	111	$6.04x10^{-11}$	112	$3.14x10^{-8}$
MECOM	$5x10^{-2}$	331	$5.13x10^{-32}$	496	$1.89x10^{-29}$	509	$7.08x10^{-25}$	95	$7.8x10^{-19}$	164	$2.02x10^{-7}$	172	$3.36x10^{-11}$	168	$9.08x10^{-19}$	308	$8.71x10^{-25}$	315	$2.85x10^{-25}$
FAM13A	$5x10^{-2}$	65	$8.86x10^{-3}$	131	$7.67x10^{-4}$	137	$3.48x10^{-2}$	38	$1.75x10^{-2}$	63	$1.1x10^{-2}$	68	$6.18x10^{-2}$	52	$2.28x10^{-3}$	109	$1.58x10^{-4}$	115	$9.38x10^{-3}$
FLJ20184	$1x10^{-2}$	27	$5.52x10^{-11}$	42	$7.52x10^{-7}$	46	$1.93x10^{-3}$	15	$1.82x10^{-8}$	23	$9.67x10^{-6}$	25	$1.24x10^{-3}$	22	$1.65x10^{-11}$	37	$6.27x10^{-7}$	41	$2.37x10^{-3}$
HHIP	$5x10^{-2}$	37	$1.89x10^{-8}$	66	$4.31x10^{-8}$	65	$1.45x10^{-3}$	21	$2.04x10^{-6}$	31	$5.1x10^{-4}$	28	$8.33x10^{-3}$	33	$1.01x10^{-5}$	57	$3.5x10^{-7}$	56	$7.07x10^{-3}$
ITK	$8.33x10^{-3}$	13	$1.05x10^{-1}$	28	$9.31x10^{-6}$	29	$5.94x10^{-3}$	10	$2.31x10^{-1}$	20	$2.66x10^{-5}$	21	$8.36x10^{-3}$	12	$1.75x10^{-1}$	28	$9.31x10^{-6}$	28	$2.41x10^{-3}$
DDR1	$1.22x10^{-3}$	5	$5.2x10^{-3}$	6	$9.8x10^{-8}$	6	$1.66x10^{-6}$	4	$6.62x10^{-3}$	4	$1.11x10^{-6}$	4	$3.87x10^{-6}$	5	$5.2x10^{-3}$	6	$9.8x10^{-8}$	6	$1.66x10^{-6}$
TNXB	$7.14x10^{-3}$	6	$5.54x10^{-3}$	17	$7.03x10^{-3}$	21	$3.81x10^{-2}$	1	$NAx10^{NA}$	9	$6.91x10^{-3}$	11	$7.65x10^{-3}$	3	$1.26x10^{-1}$	15	$2.84x10^{-3}$	17	$5.45x10^{-3}$
ARMC2	$5x10^{-2}$	55	$9.82x10^{-11}$	63	$3.91x10^{-5}$	68	$7.16x10^{-6}$	29	$4.27x10^{-7}$	39	$1.45x10^{-3}$	44	$2.19x10^{-3}$	39	$4.71x10^{-7}$	58	$1.33x10^{-5}$	62	$1.94x10^{-5}$
LOC153910	$2.5x10^{-2}$	28	$3.06x10^{-10}$	44	$3.15x10^{-4}$	45	$2.15x10^{-1}$	19	$2.83x10^{-9}$	28	$8.85x10^{-5}$	30	$1.17x10^{-1}$	22	$3.12x10^{-10}$	41	$1.82x10^{-5}$	43	$1.56x10^{-1}$
PTCH1	$5x10^{-2}$	22	$6.15x10^{-14}$	57	$5.21x10^{-4}$	54	$1.66x10^{-1}$	6	$1.64x10^{-2}$	27	$1.71x10^{-2}$	26	$2.99x10^{-1}$	13	$3.07x10^{-3}$	41	$3.87x10^{-3}$	40	$2.34x10^{-1}$
CDC123	$1.67x10^{-2}$	10	$7.91x10^{-4}$	18	$2.86x10^{-3}$	23	$4.78x10^{-1}$	7	$3.03x10^{-3}$	11	$6.57x10^{-3}$	14	$5.82x10^{-1}$	10	$7.91x10^{-4}$	18	$2.86x10^{-3}$	22	$4.78x10^{-1}$
NUDT5	$1.67x10^{-2}$	12	$5.96x10^{-3}$	22	$1.44x10^{-2}$	26	$3.29x10^{-3}$	7	$8.47x10^{-4}$	11	$1.36x10^{-3}$	13	$2.72x10^{-3}$	11	$9.6x10^{-4}$	18	$2.02x10^{-4}$	20	$6.13x10^{-4}$
C10orf11	$5x10^{-2}$	221	$5.53x10^{-32}$	370	$7.13x10^{-10}$	389	$9.73x10^{-17}$	102	$3.95x10^{-18}$	163	$7.56x10^{-9}$	176	$9.54x10^{-6}$	149	$3.45x10^{-24}$	275	$1.48x10^{-12}$	292	$2.85x10^{-15}$
HAL	$8.3x10^{-3}$	7	$2.84x10^{-2}$	17	$2.01x10^{-3}$	19	$1.08x10^{-2}$	5	$4.8x10^{-2}$	10	$2.89x10^{-3}$	11	$3.39x10^{-3}$	7	$2.84x10^{-2}$	16	$1.62x10^{-3}$	18	$7.76x10^{-3}$
NTN4	$8.3x10^{-3}$	43	$1.44x10^{-6}$	57	$1.39x10^{-2}$	62	$1.36x10^{-6}$	21	$1.73x10^{-8}$	31	$8.06x10^{-3}$	35	$3.06x10^{-9}$	28	$1.6x10^{-7}$	47	$5.11x10^{-3}$	50	$1.64x10^{-8}$

Locus	Threshold	All variants						Independent variants						Independent variants and variants not in UK10K+1000G					
		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy	
		N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P
THSD4	$5x10^{-2}$	150	$2.1x10^{-39}$	293	$3.37x10^{-6}$	340	$4.89x10^{-3}$	87	$3.68x10^{-24}$	144	$6.88x10^{-10}$	157	$2.41x10^{-6}$	111	$9.78x10^{-30}$	224	$4.51x10^{-13}$	243	$1.97x10^{-5}$
CNGB1	$1x10^{-2}$	9	$1.88x10^{-3}$	25	$2.14x10^{-3}$	25	$4.16x10^{-2}$	7	$1.14x10^{-2}$	18	$5.04x10^{-4}$	18	$1.36x10^{-2}$	7	$1.14x10^{-2}$	21	$2.47x10^{-4}$	21	$1.55x10^{-2}$
MMP15	$1x10^{-2}$	5	$2.78x10^{-3}$	5	$7x10^{-8}$	6	$3.89x10^{-6}$	4	$3.5x10^{-3}$	3	$3.25x10^{-7}$	4	$8.63x10^{-6}$	5	$2.78x10^{-3}$	4	$2.63x10^{-7}$	5	$7x10^{-6}$

c) Exon based

Locus	Threshold	All variants						Independent variants						Independent variants and variants not in UK10K+1000G					
		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy		vipR		SNVer		Syzygy	
		N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P	N	P
HDAC4	$5x10^{-2}$	1	NA	4	$3.66x10^{-11}$	4	$2.11x10^{-8}$	1	NA	4	$3.66x10^{-11}$	4	$2.11x10^{-8}$	1	NA	4	$3.66x10^{-11}$	4	$2.11x10^{-8}$
NPNT	$1.25x10^{-2}$	7	$3.38x10^{-6}$	10	$6.76x10^{-5}$	12	$5.36x10^{-4}$	3	$1.81x10^{-3}$	6	$2.88x10^{-3}$	5	$1.75x10^{-1}$	3	$1.81x10^{-3}$	8	$2.84x10^{-3}$	7	$1.96x10^{-1}$