

Manuscript Number:

Title: Surfactant Protein D is a Causal Risk Factor for Chronic Obstructive Pulmonary Disease: Results of Mendelian Randomization

Article Type: Articles (Original Research)

Keywords: SP-D, GWAS, pQTLs, eQTLs, COPD, biomarker, Mendelian randomization

Corresponding Author: Dr. Ma'en Obeidat, PhD

Corresponding Author's Institution: University of British Columbia

First Author: Ma'en Obeidat, PhD

Order of Authors: Ma'en Obeidat, PhD; Xuan Li; Stephen Burgess, PhD; Guohai Zhou, PhD; Nick Fishbane, MSc; Nadia N Hansel; Yohan Bossé; Philippe Joubert; Ke Hao; David C Nickle; Maarten van den van den ; Wim Timens; Michael H Cho; Brian D Hobbs; Kim de Jong; H. Marike Boezen; Rayjean Hung; Nicholas Rafaels; Rasika Mathias; Ingo Ruczinski; Terri H Beaty; Kathleen C Barnes; Peter D Paré; Don D Sin; International COPD Genetics Consortium; Lung eQTL Consortium; Lung Health Study

Manuscript Region of Origin: AFGHANISTAN

Abstract: Background

Chronic obstructive pulmonary disease (COPD) is currently the third leading cause of death worldwide. Surfactant protein D (SP-D) is a protein produced primarily in the lung and is involved in regulating pulmonary surfactants, maintaining lipid homeostasis in the lung and promoting innate immunity to protect the lungs from microbial and chemical insults. COPD is associated with reduced levels of SP-D in bronchial and bronchoalveolar lavage (BAL) fluid, as well as increased serum SP-D levels. It remains uncertain whether SP-D is a biomarker of COPD or is part of the causal pathway leading to COPD.

Methods

The Lung Health Study (LHS) is a longitudinal multicenter study of mild-to-moderate COPD subjects, who were followed over 11 years. In year five, SP-D serum levels were measured in 4,061 individuals who also had whole genome genotypes. To identify genetic variants that act as protein quantitative trait loci (pQTLs) for serum SP-D levels, a genome-wide association study (GWAS) for SP-D levels in 4,041 individuals from LHS was undertaken using 1000 genome imputed genotypes and adjusting for age, sex, body mass index (BMI), and genetic principal components. The identified pQTLs were additionally tested for association with variation in lung tissue expression of the SP-D encoding gene; SFTPD in 1,037 individuals and genes co-expression patterns were also utilized to uncover genes whose expression strongly correlate with SFTPD. Finally, a Mendelian randomization (MR) framework was applied whereby SP-D pQTLs

were tested for association with COPD risk in large GWAS meta-analysis of COPD (11,157 cases and 36,699 controls) from the International COPD Genetic Consortium (ICGC) and with longitudinal decline of FEV1 over 11 years in individuals from LHS.

Results

Three regions on chromosomes 6, 10 and 16 achieved genome-wide significance ($P < 5 \times 10^{-8}$) as pQTLs for serum SP-D levels. The most significantly associated pQTL was rs61860415 in the SFTPD gene on chromosome 10 ($P = 6.34 \times 10^{-92}$). The sentinel SNP in the second associated locus was rs116688442 in the HLA region on chromosome 6 ($P = 2.24 \times 10^{-54}$). The third locus harboured rs62048789 in the ATP2C2 gene on chromosome 16 ($P = 2.61 \times 10^{-17}$). Conditional analysis uncovered an additional three independent SNPs exceeding genome-wide significance on chromosome 10. Variants in the SFTPD locus were also associated with lung tissue expression of SFTPD and the SNPs' effect on serum protein levels was significantly related to their effect on mRNA expression in lung. The MR analysis revealed that SNPs associated with increased serum SP-D levels decreased the risk of COPD (estimate = -0.13, $P = 4.06 \times 10^{-5}$) in the ICGC COPD GWAS results. In the LHS, MR analysis revealed those SNPs that increased serum SP-D levels also slowed the FEV1 decline over 11 years (estimate = 0.0018, $P = 5.38 \times 10^{-3}$).

Interpretation

The integration of genetic variation effect on serum levels, lung tissue expression and disease phenotypes provided novel insights into SP-D biology and established a novel causal link between increased SP-D levels and protection against COPD risk and accelerated FEV1 decline. SP-D represents a very promising biomarker and therapeutic target for COPD which warrants further study.

Surfactant Protein D is a Causal Risk Factor for Chronic Obstructive Pulmonary Disease: Results of Mendelian Randomization

Ma'en Obeidat¹, Xuan Li¹, Stephen Burgess², Guohai Zhou¹, Nick Fishbane¹, Nadia N. Hansel³, Yohan Bossé^{4,5}, Philippe Joubert⁴, Ke Hao⁶, David C. Nickle⁷, Maarten van den Berge⁸, Wim Timens⁹, Michael H. Cho^{10,11}, Brian D. Hobbs^{10,11}, Kim de Jong¹², Marika Boezen¹², Rayjean Hung¹³, Nicholas Rafaels¹⁴, Rasika Mathias¹⁵, Ingo Ruczinski¹⁶, Terri H. Beaty¹⁷, Kathleen C. Barnes¹⁴, Peter D. Paré¹, and Don D. Sin¹, International COPD Genetics Consortium, Lung eQTL Consortium, and Lung Health Study.

¹ The University of British Columbia Center for Heart Lung Innovation, St Paul's Hospital Vancouver, Canada

² Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

³ Pulmonary and Critical Care Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

⁴ Institut universitaire de cardiologie et de pneumologie de Québec, Laval University, Québec, Canada.

⁵ Department of Molecular Medicine, Laval University, Québec, Canada.

⁶ Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA.

⁷ Merck Research Laboratories, Boston, MA, USA

⁸ University of Groningen, University Medical Center Groningen, Department of Pulmonology, GRIAC research institute, Groningen, The Netherlands

⁹ University of Groningen, University Medical Center Groningen, Department of Pathology and Medical Biology, GRIAC research institute, Groningen, The Netherlands.

¹⁰ Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA

¹¹ Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA

¹² University of Groningen, University Medical Center Groningen, Department of Epidemiology, GRIAC Research Institute, Groningen, The Netherlands

¹³ Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada

¹⁴ Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado School of Medicine, Anschutz Medical Campus, Aurora, CO, USA

¹⁵ Division of Genetic Epidemiology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

¹⁶ Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

¹⁷ Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

Corresponding Author: Don D. Sin, MD, UBC Centre for Heart Lung Innovation, St Paul's Hospital, 1081

Burrard Street, Vancouver, BC V6Z 1Y6, Canada. Telephone: 604-806-8346; e-mail: don.sin@hli.ubc.ca

Background

Chronic obstructive pulmonary disease (COPD) is currently the third leading cause of death worldwide. Surfactant protein D (SP-D) is a protein produced primarily in the lung and is involved in regulating pulmonary surfactants, maintaining lipid homeostasis in the lung and promoting innate immunity to protect the lungs from microbial and chemical insults. COPD is associated with reduced levels of SP-D in bronchial and bronchoalveolar lavage (BAL) fluid, as well as increased serum SP-D levels. It remains uncertain whether SP-D is a biomarker of COPD or is part of the causal pathway leading to COPD.

Methods

The Lung Health Study (LHS) is a longitudinal multicenter study of mild-to-moderate COPD subjects, who were followed over 11 years. In year five, SP-D serum levels were measured in 4,061 individuals who also had whole genome genotypes. To identify genetic variants that act as protein quantitative trait loci (pQTLs) for serum SP-D levels, a genome-wide association study (GWAS) for SP-D levels in 4,041 individuals from LHS was undertaken using 1000 genome imputed genotypes and adjusting for age, sex, body mass index (BMI), and genetic principal components. The identified pQTLs were additionally tested for association with variation in lung tissue expression of the SP-D encoding gene; *SFTPD* in 1,037 individuals and genes co-expression patterns were also utilized to uncover genes whose expression strongly correlate with *SFTPD*. Finally, a Mendelian randomization (MR) framework was applied whereby SP-D pQTLs were tested for association with COPD risk in large GWAS meta-analysis of COPD (11,157 cases and 36,699 controls) from the International COPD Genetic Consortium (ICGC) and with longitudinal decline of FEV1 over 11 years in individuals from LHS.

Results

Three regions on chromosomes 6, 10 and 16 achieved genome-wide significance ($P < 5 \times 10^{-8}$) as pQTLs for serum SP-D levels. The most significantly associated pQTL was rs61860415 in the *SFTPD* gene on chromosome 10 ($P = 6.34 \times 10^{-92}$). The sentinel SNP in the second associated locus was rs116688442 in the HLA region on chromosome 6 ($P = 2.24 \times 10^{-54}$). The third locus harboured rs62048789 in the *ATP2C2* gene on chromosome 16 ($P = 2.61 \times 10^{-17}$). Conditional analysis uncovered an additional three independent SNPs exceeding genome-wide significance on chromosome 10. Variants in the *SFTPD* locus were also associated with lung tissue expression of *SFTPD* and the SNPs' effect on serum protein levels was significantly related to their effect on mRNA expression in lung. The MR analysis revealed that SNPs associated with increased serum SP-D levels decreased the risk of COPD (estimate = -0.13, $P = 4.06 \times 10^{-5}$) in the ICGC COPD GWAS results. In the LHS, MR analysis revealed those SNPs that increased serum SP-D levels also slowed the FEV1 decline over 11 years (estimate = 0.0018, $P = 5.38 \times 10^{-3}$).

Interpretation

The integration of genetic variation effect on serum levels, lung tissue expression and disease phenotypes provided novel insights into SP-D biology and established a novel causal link between increased SP-D levels and protection against COPD risk and accelerated FEV1 decline. SP-D represents a very promising biomarker and therapeutic target for COPD which warrants further study.

Funding

Funding for the SP-D measurements was provided by the Canadian Institutes of Health Research (CIHR). Genotyping for the LHS samples was provided by the Gene, Environment Association Studies grant U01HG004738. The gene expression and genotyping for the lung eQTL study was funded by Merck.

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is currently the third leading cause of death worldwide¹. The disease is characterized by airway obstruction that is not fully reversible, and is caused by a complex interaction between genetic and environmental risk factors (e.g. smoking)². Although one third of variability in lung function can be attributed to genetic factors, genes directly controlling risk to COPD remain largely unknown³.

A critical class of molecules involved in lung growth and homeostasis are surfactants. One surfactant that has been associated with COPD, even in the absence of cigarette smoke, is surfactant protein D (SP-D)⁴. SP-D is a large multimeric collagenous glycoprotein weighing ~43 kDa and is part of the collectin family of proteins⁵. SP-D is produced primarily by type II alveolar cells, as well as club cells^{6,7}. SP-D plays an important role in regulating pulmonary surfactants, maintaining lipid homeostasis in the lung⁸ and promoting innate immunity⁹ to protect the lungs from microbial and chemical insults¹⁰. SP-D deficiency in mice leads to alveolar macrophage activation, increased oxidant stress in the airways and emphysematous changes in lung parenchyma by 3 weeks of age.¹¹ In humans, single nucleotide polymorphisms (SNPs) in the surfactant protein D (SFTPD) gene on chromosome 10 have been associated with COPD in several different cohorts¹². The most well studied SFTPD variant (rs721917) is a coding SNP, which results in the substitution of methionine for threonine at amino acid 11 (Met11Thr)¹³. This SNP is associated with SP-D serum levels and emphysema¹⁴. COPD is associated with reduced levels of SP-D in bronchial and bronchoalveolar lavage (BAL) fluid¹⁵, as well as increased serum SP-D levels¹⁶. Despite these associations it remains uncertain whether SP-D is a biomarker of COPD or is part of the causal pathway leading to COPD.

One way of evaluating causality in biomarker-disease associations is to use Mendelian randomization (MR), which utilizes genetic variation to ascertain causal associations¹⁷⁻¹⁹. MR exploits two unique attributes of genotypes enabling robust causal interrogation: 1) the random allocation of parental alleles to zygotes at meiosis, independent of environmental exposures later in life, and 2) the unidirectional flow of biological information from genes to transcript to protein, thus avoiding reverse causation²⁰. Within the MR framework, if the relationship between SP-D and COPD is truly causal, the genetic variant(s) influencing serum SP-D levels should also be significantly associated with COPD risk and/or FEV1 decline over time.

In the present study, we unravelled the genetic architecture of SP-D serum levels in a large cohort of COPD patients to determine SP-D protein and mRNA expression quantitative trait loci (pQTL and eQTL)

and used MR to demonstrate a causal relationship between serum SP-D and two important phenotypes: COPD risk and FEV1 decline over time.

METHODS

Overall study design

The overall study design is depicted in **Figure 1**. Firstly, we identified protein quantitative trait loci (pQTLs) for SP-D using a genome wide association study (GWAS) of serum SP-D levels in COPD patients, who participated in the Lung Health Study (LHS)²¹. Secondly, using data from the Lung eQTL Consortium²², we determined which of these pQTLs were also associated with changes in gene expression (mRNA) and potentially represented expression trait quantitative loci (eQTLs) in lung tissue. We also used gene expression data from lung tissues to identify networks of genes strongly related to SFTPD. Thirdly, we determined which of the SP-D pQTLs were associated with risk to COPD, cross sectional lung function and COPD progression defined by the rate of decline in FEV1 over 11 years. Finally, we applied a MR framework to determine whether there was a causal relationship between serum SP-D levels and COPD risk or progression.

Cohorts

Lung Health Study (LHS)

The details of the LHS have been published previously^{21, 23}. Briefly, LHS was a multicenter clinical study that evaluated the effects of ipratropium bromide, a short acting antimuscarinic agent, and smoking cessation on lung function decline in current smokers with mild to moderate COPD. For the first 5 years, the lung function of participants was measured annually, followed by another lung function measurement at year 11. At each visit the subjects' smoking status was determined using a questionnaire, which was validated by salivary cotinine and exhaled carbon monoxide levels as previously described²¹. Based on these data, the subjects were categorized as "sustained quitters" if they were non-smokers at all of the follow-up visits, "continuous smokers" if they were smoking at all of the follow-up visits and "intermittent quitters" if their smoking status varied during the follow-up period. In year 5 of LHS, venipuncture was carried out on 5,413 LHS participants who were alive and eligible for venipuncture at this visit. The blood samples were taken when participants were stable and free of exacerbations for at least 4 weeks and were separated into buffy coat and serum²⁴.

Genotyping

From the buffy coat samples of 4,251 European Americans in LHS, SNP genotyping was performed. The details of genotyping and quality control have been previously described²⁵. Briefly, samples were

genotyped using the Illumina Human660WQuad v.1_A BeadChip. Overall, 98.4 % of samples ($n = 4,181$) passed initial quality control standards and genotypes were available for 559,766 SNPs. An additional 133 samples were removed because they failed quality control, which resulted in a final sample of $n = 4,048$ for the present analysis. Imputation was undertaken with the software IMPUTE2²⁶ using the 'all ancestries' 1000 Genomes Project (1000G) reference panel, March 2012 release²⁷. Variants were excluded if the imputation r^2 was < 0.3 .

Serum SP-D measurements in LHS

Using one aliquot of serum samples, SP-D levels were measured in duplicate in 4,754 LHS participants using a commercially available solid phase sandwich enzyme-linked immunosorbent assay according to manufacturer's instructions (ELISA; BioVendor Laboratory Medicine, Modrice, Czech Republic) by study personnel, who were blinded to the subject's clinical or genotyping information. The lower limit of detection for the assay was 0.2 ng/mL, and the coefficient of variation was 3.9%²⁸.

Protein quantitative trait loci (pQTL) of serum SP-D: GWAS of SP-D serum levels

We performed a genome-wide association study for log serum SP-D levels using SNPTEST²⁹ assuming an additive genetic model and adjusting for age, sex, body mass index (BMI), smoking status (sustained quitters versus intermittent quitters versus continuous smokers) and the first 5 genetic principal components (PCs). The final GWAS dataset included 4,041 subjects. SP-D pQTLs were defined as the sentinel SNPs meeting genome-wide significance ($P < 5 \times 10^{-8}$).

For each of the genome-wide significant pQTL loci identified, we performed conditional analysis to reveal independent SNPs in these regions associated with SP-D levels. For each locus, we regressed log SP-D levels against the most significant SNP along with age, sex, BMI, and smoking status to obtain regression residuals. With these residuals, a second genetic association was performed, adjusting for the first 5 genetic PCs. The residual genetic association was repeated while conditioning on the sentinel SNP yielding genome-wide significance until all significant SNPs for SP-D levels (defined by $P < 5 \times 10^{-8}$) were exhausted at each locus. Using this approach, we identified 3 additional independent SNPs on the chr10 locus. The other regions (chr6 and 16) did not harbour any additional loci that were independently associated with SP-D levels.

A recent large GWAS meta-analysis evaluating SNPs associated with COPD³⁰ in 15,256 COPD cases and 47,936 controls identified a coding SNP in the SFTPD gene, rs721917 (Met11Thr), which was significantly associated with risk of COPD ($P = 2.5 \times 10^{-8}$ in the overall meta-analysis which included an additional 9,498

COPD cases and 9,748 controls from the UK BiLEVE study³¹. This SNP effect on serum SP-D was investigated in our GWAS results and it was also included in all subsequent analyses.

Association of SP-D pQTLs with lung tissue expression in the Lung eQTL study

Because lung tissue samples were not collected in LHS participants, we used data from the lung eQTL study to identify potential lung eQTLs for SFTPD. The study details and the subjects' characteristics have been previously described²². Briefly, lung eQTLs were derived from a meta-analysis of genotyping and RNA expression data performed in non-tumor lung tissue samples from 1,111 patients who underwent lung resection surgery at three participating sites: University of British Columbia (UBC; n=339), Laval University (n=409) and the University of Groningen (n=363). Gene expression profiling was performed using an Affymetrix custom array (GPL10379) testing 51,627 non-control probe sets where normalization was performed using robust multi-array average (RMA)³². The expression data are available at NCBI Gene Expression Omnibus repository (GEO, <http://www.ncbi.nlm.nih.gov/geo>) through accession number GSE23546.

Genotyping was performed on DNA extracted from blood or lung tissue using the Illumina Human1M-Duo BeadChip array, and imputed with MaCH/Minimac software³³ using the 1000G reference panel, March 2012 release. The eQTL analysis was adjusted for age, sex and smoking status. The resulting eQTLs were categorized into *cis*-acting (if the signal was less than 1Mb away from transcription start site) or *trans* eQTLs (if the signal was further than 1Mb away or on a different chromosome). Following standard microarray and genotyping quality control steps, association testing for each variant with mRNA expression in either *cis* (SNP within 1Mb of transcript) or *trans* (more than 1Mb away or on a different chromosome) position was undertaken separately for each centre and then pooled together using inverse variance weighting meta-analysis.

All genome-wide significant pQTLs for SP-D levels were tested for association with lung tissue mRNA levels (i.e. lung eQTLs). Scatter plots were used to visualize the SNPs effect on both SP-D serum levels and lung mRNA.

Association of SP-D pQTLs with COPD risk and progression

The relationship of SNPs identified as pQTLs for serum SP-D with COPD risk was investigated in the International COPD Genetics Consortium (ICGC) Study. ICGC is the largest COPD GWAS to date, and included a total of 15,256 COPD cases and 47,936 controls³⁰. The look-up in ICGC was performed in

GWAS results for individuals with European ancestry with a sample size of 11,157 cases and 36,699 controls.

SP-D pQTLs were also tested for association with FEV1 decline over the 11 years of follow-up in the LHS using a multiple linear mixed effects model which takes into account multiple measurements of FEV1 over time within the same individual. These pQTLs were additionally tested for association with cross sectional FEV1 and FEV1/FVC in the LHS cohort using linear regression. The analysis of FEV1 decline in LHS used SNP dosage as a predictor and the (time x genotype) interaction as the response variable, while the analysis of cross sectional measures used FEV1 and FEV1/FVC as the response variables. All analyses in LHS were adjusted for age, sex, BMI, and smoking status as noted previously.

Mendelian randomization (MR) analysis

The independent pQTLs for SP-D identified in this study (including the coding SNP) were used as genetic instruments in MR analyses. The instrumental variables linear regression analysis of the SP-D pQTLs effect on outcome (COPD risk and lung function phenotypes) versus the SNPs' allelic effect on observed SP-D serum levels was weighted by the inverse variance of the outcome effect estimate, and constrained (forced to pass through the origin) ³⁴.

Weighted gene co-expression analysis (WGCNA)

The WGCNA R package ³⁵ was used to cluster modules of co-expressed genes into a co-expression network. The expression values from 1,037 lung tissue samples were used in this analysis. A weighted gene co-expression network reconstruction algorithm was used to create a consensus co-expression network among the 18,403 genes that were common to all three centres ³⁶. WGCNA first creates a matrix of Pearson correlations across all genes, which is then transformed into an adjacency matrix through 'soft thresholding' by raising these correlations to some power β . In this study $\beta = 10$ was selected to enable the matrix to approximate scale-free topology, thus creating a topological overlap matrix (TOM) ³⁷. Average linkage hierarchical clustering was used to group genes based on their topological overlap of connectivity, followed by a dynamic cut-tree algorithm to cluster dendrogram branches into distinct gene modules ³⁸. Modules were defined as groups of highly interconnected genes. For each gene, Module Membership (MM) was calculated with values ranging between 0 and 1 by correlating the gene's expression with the module eigengene, which was determined by the first principal component of the gene expression profile in that module. We deemed "hub" genes as those with the highest MM. The

generated TOM values in this WGCNA estimate the strength of connection between the different genes in a network^{39 40}, and were used to identify genes strongly correlated to *SFTPD*.

Statistical analysis software

All analyses were performed with R version 3.2.1

Role of the funding source

The funding sources have no roles in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

RESULTS

Descriptive demographics of LHS Participants

The demographics of LHS participants across the quintiles of serum SP-D levels are shown in **Table 1**. Serum SP-D levels were significantly related to increasing age, decreasing BMI and FEV1, and smoking status (with continuous smokers having higher average serum SP-D levels than sustained quitters).

Genome-wide association study of serum SP-D levels

Variants with effective sample sizes ($N_{\text{effective}}$, computed as the product of sample size and imputation quality summed across studies) $< 70\%$ were filtered out, leaving 9,355,308 variants in this GWAS analysis. Quantile–quantile (QQ) plots are presented in **Supplementary Figure 1** and showed a deviation from the expected distribution only for low p-values indicating strong association signals. The genomic inflation factor (λ) was 1.02, suggesting no systematic deviation in the association statistics due to factors such as population structure.

A total of 1,933 SNPs representing three regions on chromosomes 6, 10 and 16 achieved genome-wide significance ($P < 5 \times 10^{-8}$) as described previously in Kim et al.⁴¹. Manhattan plot is shown in **Figure 2** and the region plots are shown in **Figure 3**. The most significantly associated pQTL was an intronic SNP (rs61860415) in the *SFTPD* gene ($P = 6.34 \times 10^{-92}$). The sentinel SNP in the second associated locus was an intergenic SNP (rs116688442) in the HLA region on chromosome 6 ($P = 2.24 \times 10^{-54}$). The third locus harboured the intronic SNP (rs62048789) in the *ATP2C2* gene on chromosome 16 ($P = 2.61 \times 10^{-17}$).

To identify additional independent signals, association testing was repeated using residuals from the sentinel SNP in the GWAS model. The analysis of these residuals on chromosome 10 identified an additional 3 independent signals exceeding genome-wide significance (**Supplementary Figure 2**). The regions on chromosomes 6 and 16 did not harbour any additional independent loci significant at the $P < 5 \times 10^{-8}$ threshold. All pQTL sentinel SNPs meeting the genome-wide association threshold for serum SP-D in these three regions are presented in **Table 2**. SNP rs721917; recently associated with COPD in the ICGC study³⁰ was a pQTL for serum SP-D ($P = 7.5 \times 10^{-12}$). This SNP had a modest linkage disequilibrium (LD) with the other sentinel SNPs on chr 10 region ranging from $r^2 = 0.16$ with SNP rs80271335 to $r^2 = 0.53$ with SNP rs12358676. The association of this SNP with SP-D serum levels is reported in **Table 2** and it is also included in all the subsequent analyses.

The effect of SP-D pQTLs on lung mRNA expression (eQTLs)

Given that SP-D is predominantly synthesized in the lung, the three loci harbouring pQTLs for serum SP-D were evaluated in lung tissue to determine whether these regions also acted as an expression quantitative trait loci (eQTLs) for lung tissue mRNA²². The top eQTL results (based on the lowest eQTL P-value per gene) for each of the loci are shown in **Table 3**. The pQTL SNPs in the SFTPD locus on chr10 were also lung eQTLs for *SFTPD*, and the SNPs' effect on serum protein levels was significantly related to their effect on mRNA expression in lung tissue (**Figure 4**). Together, these data suggest SNPs in the *SFTPD* gene region affect serum SP-D by altering the levels of gene expression in lung tissue. In contrast, the pQTL SNPs in chr6 and chr16 were not related to *SFTPD* mRNA levels in the lung, indicating they affect serum levels through other mechanisms (**Supplementary Figure 3**).

Association of SP-D pQTLs with COPD and lung function

The 7 pQTL SNPs for serum SP-D levels (including the coding SNP), were tested for association with COPD risk and progression (defined by observed FEV1 decline over 11 years). The association results with COPD risk were extracted from the ICGC meta-analysis. These the LHS. The association results of SP-D pQTLs were additionally tested for association with FEV1 decline over 11 years in participants in the LHS. Additional testing was also performed for cross sectional FEV1 and FEV1/FVC measures in pQTL SNPs with COPD phenotypes and lung function are summarized in **Table 4**. Interestingly, the coding SNP (rs721917) in the SFTPD gene was significantly associated with all phenotypes studied ($P < 0.05$) and showed the lowest P-values of all tested SNPs.

Mendelian randomization (MR) analyses

The combined SNPs allelic effects on serum SP-D levels and risk of COPD and FEV1 decline were evaluated using MR. In the ICGC cohort, MR analysis revealed that SNPs associated with increased serum SP-D levels decreased the risk of COPD in a dose-dependent manner (estimate = -0.13 change in COPD log odds ratio for every unit increase in log SP-D, $P = 4.06 \times 10^{-5}$; **Figure 5A**). In the LHS, MR analysis revealed those SNPs that increased serum SP-D levels also increased FEV1/FVC ratio (estimate = 0.46 increase in FEV1/FVC per unit increase in log SP-D, $P = 0.037$; **Supplementary Figure 4**) and slowed the FEV1 decline over 11 years (estimate = 0.0018 less decline in FEV1 slope per unit increase in log SP-D, $P = 5.38 \times 10^{-3}$; **Figure 5B**), all in a dose-dependent manner (i.e. a greater SNP effect on observed serum SP-D level was associated with a greater protection against COPD and less FEV1 decline over time).

WGCNA and SFTPD co-expression in lung tissue

To further understand the biology of SP-D, we searched for genes that were strongly correlated with SFTPD gene expression in lung tissue. Using the consensus network approach, we identified 38 distinct modules in the expression data on lung tissue, which consisted of strongly co-expressed genes. The module containing the SFTPD gene included 364 different genes which were enriched for pathways related to lipid, cholesterol and fatty acids metabolic and biosynthesis processes (**Supplementary Table 1**). This network construction additionally allowed us to identify 'hub' genes, which, by definition, had the highest module membership value for a given module. The hub gene in the module containing SFTPD was surfactant associated 3 (SFTA3). The top 50 genes with the strongest correlation with SFTPD are shown in **Supplementary Figure 5**. Genes strongly co-expressed with SFPD included CRTAC1, CACNA2D2, and S100.

DISCUSSION

Despite its high prevalence and impact, there are currently no biomarkers or disease modifying therapies for COPD. Establishing causality for selected proteins and pathways is one promising step toward their development as both biomarkers and therapeutic targets.

In this study, we determined the effect of genetic variation on SP-D protein levels in serum, mRNA levels in lung tissue and disease phenotypes to identify a potential causal role for SP-D in the etiology of COPD. We identified 6 independent loci that were significantly associated with serum SP-D levels, including four independent SNPs at the *SFTPD* locus on chr10, plus two distal regions of the genome (in the HLA locus on chr6 and the *ATP2C2* gene region on chr16). Importantly, using a lung tissue-specific eQTL dataset, we showed the effects of the SNPs in the *SFTPD* gene on serum SP-D is likely by modulating *SFTPD* mRNA levels in the lung. The genetic signal on chr16, on the other hand, appears to be related to expression of the *ATP2C2* gene and thus its effect on serum SP-D is likely indirect. Most importantly, using MR, we showed SNPs associated with higher serum SP-D levels seem to impart protection against COPD and FEV1 decline over time. Together these data suggest a causal role of SP-D in the pathogenesis of COPD.

A previous study by Kim et al. identified the three loci that were pQTLs for SP-D on chr6, 10 and 16⁴¹. A more recent study by Sun et al. additionally identified SNP rs2146192 in the *SFTPD* gene region as associated with SP-D blood levels in COPD subjects⁴². This SNP is in modest LD with the top SP-D pQTL in our study ($r^2=0.53$). Using conditional analyses, Sun et al. additionally identified two independent pQTL SNPs; one on chr10 and the other in the HLA region on chr6. In the current study, we confirm and extend the findings of Kim et al. and Sun et al. We replicated the same 3 genes from Kim et al. and identified several additional SNPs in the *SFTPD*, *HLA* and *ATP2C2* regions act as pQTLs for SP-D levels. In total, the 3 genes explained 30% of the variation in serum SP-D levels. The strongest pQTLs for SP-D in this study were in the *SFTPD* gene encoding the SP-D protein, suggesting serum levels are driven in part by SP-D production in lung tissue. Consistent with this notion, we observed SNPs' estimated effect on *SFTPD* mRNA levels in lung tissue appears significantly related to their effect on serum SP-D levels. In addition, there were no distal genetic determinants of *SFTPD* expression in lung tissue suggesting the SP-D serum pQTLs on chr10 captured most of the regulatory effects on SFTPD production in the lung.

The identification of distal pQTLs may provide important insights into SP-D biology. The chr16 locus contains the ATPase secretory pathway Ca²⁺ transporting 2 (*ATP2C2*) gene, which is expressed in many tissues (including the lungs) and functions as a transporter of Ca²⁺ ions⁴³. SNPs in the *ATP2C2* region had no effect on SFTPD mRNA levels in lungs, however, mRNA levels of *ATP2C2* were positively correlated

with *SFTPD* mRNA expression ($r=0.2$, $P= 6.9 \times 10^{-11}$). Together, these data suggest *ATP2C2* may regulate serum SP-D through an indirect mechanism such as post translational modification, transport, and/or clearance, among others or through a shared functional pathway.

In lung tissue, *SFTPD* was mapped to a module of 364 strongly co-expressed genes which were enriched for pathways related to lipid and cholesterol biosynthesis. Surfactant associated 3 (*SFTA3*) was the 'hub' gene and its expression was positively related to *SFTPD*. *SFTA3* is a recently characterised immunoregulatory protein in the lung, and may play an important role in inflammation and immune defence⁴⁴. Other genes that were strongly co-expressed with *SFTPD* included cartilage acidic protein 1 (*CRTAC1*), the calcium channel, voltage-dependent, alpha 2/delta subunit 2 (*CACNA2D2*), and S100 calcium binding protein A14 (*S100A14*) genes. *CRTAC1* is up-regulated (>10 fold) in type 2 epithelial cells as they differentiate in culture⁴⁵. Previous studies have shown binding of SP-D to phosphatidylinositol⁴⁶, TLR2 and TLR4⁴⁷, and carbohydrate structures on the surface of bacteria, viruses, and fungi is calcium dependent⁴⁸⁻⁵³. Calcium may also cause SP-D oligomers to self-aggregate, as has been shown previously with SPA⁵⁴. Further work is needed to unravel the mechanisms underlying calcium related gene associations with SP-D gene expression and serum levels.

In the LHS study, there was a modest negative correlation between serum SP-D levels and FEV1 ($r=-0.063$, $P<1 \times 10^{-4}$), consistent with several other studies^{49, 50}. Superficially these relationships appear to be paradoxical to the main findings of the present study (opposite direction of association). However, cross-sectional relationships between lung function and serum SP-D have to be interpreted cautiously because of the possibility of reverse causation (i.e. pathological processes which contribute to COPD causing increased leakage of SP-D from lungs into the systemic circulation) and confounding (e.g. smoking). For instance, we and others have previously shown serum SP-D is affected not only by lung synthesis of SP-D but also by permeability of the alveolar-capillary interface, which is perturbed in COPD and also increased by active smoking and lung inflammation^{55, 56}. Genetic analyses avoid most of these pitfalls since the causal pathway between genetic variants and protein and disease is unidirectional. In the present study, MR results supported a significant protective role for SP-D on both risk of COPD and FEV1 decline.

One of the important considerations in MR studies is the selection of appropriate genetic variants to infer causality of biomarker-disease associations. We used a hypothesis-free GWAS to identify local and distal pQTLs for serum SP-D levels, and in addition included the coding SNP rs721917 given its association with COPD risk in the ICGC study. The functional consequences of this protein coding variants are not clear, but it was associated with serum SP-D levels in the present study. The SNP may also affect the multimeric

structure of SP-D, thereby disrupting the protein's host defence and immune functions by reducing its the binding affinity to pathogen ligands or the antioxidant properties.

There is a huge unmet clinical need to identify novel biomarkers and therapeutic targets for COPD. SP-D represents an excellent candidate since: 1) it is a lung specific protein that can be reliably assayed in blood, BAL and lung tissue; 2) it tracks well with disease severity and exacerbations⁵⁷; 3) it is responsive to steroid treatment⁵⁷; and 4) it has inherent functional and biological attributes suggesting a role in COPD pathogenesis¹⁰. Moreover, this study is the first to our knowledge to report a novel causal link between genetically elevated SP-D serum levels and both decreased risk for COPD and FEV1 decline.

The current study has a number of limitations. First, LHS study participants were all smokers with mild or moderate COPD at the time of recruitment, and this may have affected the serum measures of SP-D as shown by the discordance between genetic and observational analyses. Second, the serum and lung mRNA expression were measured in different individuals, which clearly prevents direct comparisons between the two sites. However, the use of genetic variation provided a robust method to estimate and extrapolate the findings for mRNA and protein between lung tissue and blood. Third, SP-D was detected in serum using ELISA, which may not differentiate between intact (dodecameric) and monomeric or trimeric forms of SP-D. Further quantitative studies of different multimeric SP-D protein structures are warranted.

In conclusion, the current study investigated the effect of genetic variation on serum SP-D levels, lung tissue gene expression and COPD risk and progression. The findings provided novel insights into SP-D biology and establish a novel causal link between increased SP-D levels and protection against COPD risk and accelerated FEV1 decline. SP-D represents a very promising biomarker and therapeutic target for COPD which warrants further study.

Authors' Contributions

Conceived and designed the study: MO, PDP, DDS

Lung eQTL data collection and analysis: KH, DCN, MVB, WT, YB, PJ

International COPD Genetics Consortium (ICGC) GWAS data collection and analysis: MHC, BDD, KDJ, MB

Lung Health Study (LHS) data analyses: MO, XL, GZ, NF, NH, THB, NR, RM, IR, KCB,

Statistical support and advice on study conduct: RH, SB,

Wrote the manuscript: MO, PDP, DDS

Discussed results and implications and commented on the manuscript at all stages: all co-authors.

Competing financial interests

David C. Nickle is employed by Merck & Co. Inc.

REFERENCES

1. WHO. The top 10 causes of death. World Health Organization; 2014.
2. Jørgen V, Suzanne SH, Alvar GA, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine* 2013; **187**(4): 347-65.
3. Hukkinen M, Kaprio J, Broms U, et al. Heritability of lung function: a twin study among never-smoking elderly women. *Twin research and human genetics : the official journal of the International Society for Twin Studies* 2011; **14**(5): 401-7.
4. Wert SE, Yoshida M, LeVine AM, et al. Increased metalloproteinase activity, oxidant production, and emphysema in surfactant protein D gene-inactivated mice. *Proceedings of the National Academy of Sciences of the United States of America* 2000; **97**(11): 5972-7.
5. Atochina EN, Beers MF, Hawgood S, et al. Surfactant protein-D, a mediator of innate lung immunity, alters the products of nitric oxide metabolism. *American journal of respiratory cell and molecular biology* 2004; **30**(3): 271-9.
6. Voorhout WF, Veenendaal T, Kuroki Y, Ogasawara Y, van Golde LM, Geuze HJ. Immunocytochemical localization of surfactant protein D (SP-D) in type II cells, Clara cells, and alveolar macrophages of rat lung. *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society* 1992; **40**(10): 1589-97.
7. Madsen J, Kliem A, Tornøe I, Skjodt K, Koch C, Holmskov U. Localization of lung surfactant protein D on mucosal surfaces in human tissues. *J Immunol* 2000; **164**(11): 5866-70.
8. Fisher JH, Sheftelyevich V, Ho YS, et al. Pulmonary-specific expression of SP-D corrects pulmonary lipid accumulation in SP-D gene-targeted mice. *American journal of physiology Lung cellular and molecular physiology* 2000; **278**(2): L365-73.
9. Wright JR. Immunoregulatory functions of surfactant proteins. *Nature reviews Immunology* 2005; **5**(1): 58-68.
10. Sin DD, Pahlavan PS, Man SF. Surfactant protein D: a lung specific biomarker in COPD? *Therapeutic advances in respiratory disease* 2008; **2**(2): 65-74.

11. Wert SE, Yoshida M, LeVine AM, et al. Increased metalloproteinase activity, oxidant production, and emphysema in surfactant protein D gene-inactivated mice. *Proceedings of the National Academy of Sciences of the United States of America* 2000; **97**(11): 5972-7.
12. Foreman MG, Kong X, DeMeo DL, et al. Polymorphisms in Surfactant Protein–D Are Associated with Chronic Obstructive Pulmonary Disease. *American journal of respiratory cell and molecular biology* 2011; **44**(3): 316-22.
13. Bowler RP. Surfactant protein D as a biomarker for chronic obstructive pulmonary disease. *Copd* 2012; **9**(6): 651-3.
14. Ishii T, Hagiwara K, Kamio K, et al. Involvement of surfactant protein D in emphysema revealed by genetic association study. *European journal of human genetics : EJHG* 2012; **20**(2): 230-5.
15. Honda Y, Takahashi H, Kuroki Y, Akino T, Abe S. Decreased contents of surfactant proteins A and D in BAL fluids of healthy smokers. *Chest* 1996; **109**(4): 1006-9.
16. Mutti A, Corradi M, Goldoni M, Vettori MV, Bernard A, Apostoli P. Exhaled metallic elements and serum pneumoproteins in asymptomatic smokers and patients with COPD or asthma. *Chest* 2006; **129**(5): 1288-97.
17. White J, Sofat R, Hemani G, et al. Plasma urate concentration and risk of coronary heart disease: a Mendelian randomisation analysis. *The Lancet Diabetes & Endocrinology*; **4**(4): 327-36.
18. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 2003; **32**(1): 1-22.
19. Voight BF, Peloso GM, Orho-Melander M, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*; **380**(9841): 572-80.
20. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics* 2014; **23**(R1): R89-98.
21. Anthonisen NR, Connett JE, Kiley JP, et al. Effects of smoking intervention and the use of an inhaled anticholinergic bronchodilator on the rate of decline of FEV1. The Lung Health Study. *Jama* 1994; **272**(19): 1497-505.
22. Hao K, Bossé Y, Nickle DC, et al. Lung eQTLs to Help Reveal the Molecular Underpinnings of Asthma. *PLoS Genetics* 2012; **8**(11): e1003029.
23. Kanner RE, Connett JE, Williams DE, Buist AS. Effects of randomized assignment to a smoking cessation intervention and changes in smoking habits on respiratory symptoms in smokers with early chronic obstructive pulmonary disease: the Lung Health Study. *The American journal of medicine* 1999; **106**(4): 410-6.
24. Man SF, Connett JE, Anthonisen NR, Wise RA, Tashkin DP, Sin DD. C-reactive protein and mortality in mild to moderate chronic obstructive pulmonary disease. *Thorax* 2006; **61**(10): 849-53.
25. Hansel NN, Ruczinski I, Rafaels N, et al. Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. *Human genetics* 2013; **132**(1): 79-90.
26. 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.
27. Consortium TGP. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**(7422): 56-65.
28. Hill J, Heslop C, Man SFP, et al. Circulating surfactant protein-D and the risk of cardiovascular morbidity and mortality. *European Heart Journal* 2011.
29. 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.
30. Brian D. Hobbs DKdJ, Nick R.G. Shrine, Annah B Wyss, Stephanie J London, Michael Cho. A Genome-Wide Association Study of Chronic Obstructive Pulmonary Disease Identifies 22 Risk Loci. *Nature Genetics* 2017.

31. Wain LV, Shrine N, Miller S, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory Medicine* 2015; **3**(10): 769-81.
32. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003; **4**(2): 249-64.
33. 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.
34. Johnson T. Efficient Calculation for Multi-SNP Genetic Risk Scores. American Society of Human Genetics Annual Meeting. San Francisco, CA, USA; 2012.
35. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* 2008; **9**: 559.
36. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* 2005; **4**: Article17.
37. Yip A, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics* 2007; **8**(1): 22.
38. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008; **24**(5): 719-20.
39. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 2007; **8**: 22.
40. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical Organization of Modularity in Metabolic Networks. *Science* 2002; **297**(5586): 1551-5.
41. Kim DK, Cho MH, Hersh CP, et al. Genome-Wide Association Analysis of Blood Biomarkers in Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine* 2012; **186**(12): 1238-47.
42. Sun W, Kechris K, Jacobson S, et al. Common Genetic Polymorphisms Influence Blood Biomarker Measurements in COPD. *PLoS Genetics* 2016; **12**(8): e1006011.
43. Vanoevelen J, Dode L, Van Baelen K, et al. The Secretory Pathway Ca²⁺/Mn²⁺-ATPase 2 Is a Golgi-localized Pump with High Affinity for Ca²⁺ Ions. *Journal of Biological Chemistry* 2005; **280**(24): 22800-8.
44. Schicht M, Rausch F, Finotto S, et al. SFTA3, a novel protein of the lung: three-dimensional structure, characterisation and immune activation. *European Respiratory Journal* 2014; **44**(2): 447-56.
45. Ballard PL, Lee JW, Fang X, et al. Regulated gene expression in cultured type II cells of adult human lung. *American Journal of Physiology - Lung Cellular and Molecular Physiology* 2010; **299**(1): L36-L50.
46. Ogasawara Y, Kuroki Y, Akino T. Pulmonary surfactant protein D specifically binds to phosphatidylinositol. *The Journal of biological chemistry* 1992; **267**(29): 21244-9.
47. Ohya M, Nishitani C, Sano H, et al. Human Pulmonary Surfactant Protein D Binds the Extracellular Domains of Toll-like Receptors 2 and 4 through the Carbohydrate Recognition Domain by a Mechanism Different from Its Binding to Phosphatidylinositol and Lipopolysaccharide. *Biochemistry* 2006; **45**(28): 8657-64.
48. Meschi J, Crouch EC, Skolnik P, et al. Surfactant protein D binds to human immunodeficiency virus (HIV) envelope protein gp120 and inhibits HIV replication. *The Journal of general virology* 2005; **86**(Pt 11): 3097-107.
49. Ferguson JS, Voelker DR, McCormack FX, Schlesinger LS. Surfactant Protein D Binds to Mycobacterium tuberculosis Bacilli and Lipoarabinomannan via Carbohydrate-Lectin Interactions Resulting in Reduced Phagocytosis of the Bacteria by Macrophages1. *The Journal of Immunology* 1999; **163**(1): 312-21.

50. Murray E, Khamri W, Walker MM, et al. Expression of Surfactant Protein D in the Human Gastric Mucosa and during *Helicobacter pylori* Infection. *Infection and Immunity* 2002; **70**(3): 1481-7.
51. Nadesalingam J, Reid KB, Palaniyar N. Collectin surfactant protein D binds antibodies and interlinks innate and adaptive immune systems. *FEBS letters* 2005; **579**(20): 4449-53.
52. Hickling TP, Bright H, Wing K, et al. A recombinant trimeric surfactant protein D carbohydrate recognition domain inhibits respiratory syncytial virus infection in vitro and in vivo. *Eur J Immunol* 1999; **29**(11): 3478-84.
53. Madan T, Eggleton P, Kishore U, et al. Binding of pulmonary surfactant proteins A and D to *Aspergillus fumigatus* conidia enhances phagocytosis and killing by human neutrophils and alveolar macrophages. *Infection and Immunity* 1997; **65**(8): 3171-9.
54. van Iwaarden JF, Teding van Berkhout F, Whitsett JA, Oosting RS, van Golde LM. A novel procedure for the rapid isolation of surfactant protein A with retention of its alveolar-macrophage-stimulating properties. *Biochemical Journal* 1995; **309**(Pt 2): 551-5.
55. Winkler C, Atochina-Vasserman EN, Holz O, et al. Comprehensive characterisation of pulmonary and serum surfactant protein D in COPD. *Respiratory research* 2011; **12**(1): 1-11.
56. Tkacova R, McWilliams A, Lam S, Sin DD. Integrating lung and plasma expression of pneumo-proteins in developing biomarkers in COPD: a case study of surfactant protein D. *Med Sci Monit* 2010; **16**.
57. Lomas DA, Silverman EK, Edwards LD, et al. Serum surfactant protein D is steroid sensitive and associated with exacerbations of COPD. *European Respiratory Journal* 2009; **34**(1): 95-102.

ACKNOWLEDGEMENTS

Ma'en Obeidat is a Postdoctoral Fellow of the Michael Smith Foundation for Health Research (MSFHR) and the Canadian Institute for Health Research (CIHR) Integrated and Mentored Pulmonary and Cardiovascular Training program (IMPACT). He is also a recipient of British Columbia Lung Association Research Grant.

Ke Hao is partially supported by National Natural Science Foundation of China grant no. 21477087.

Michael Cho is recipient of NIH grant R01 HL113264, and he also received grant support from GlaxoSmithKline (GSK).

Figure 1

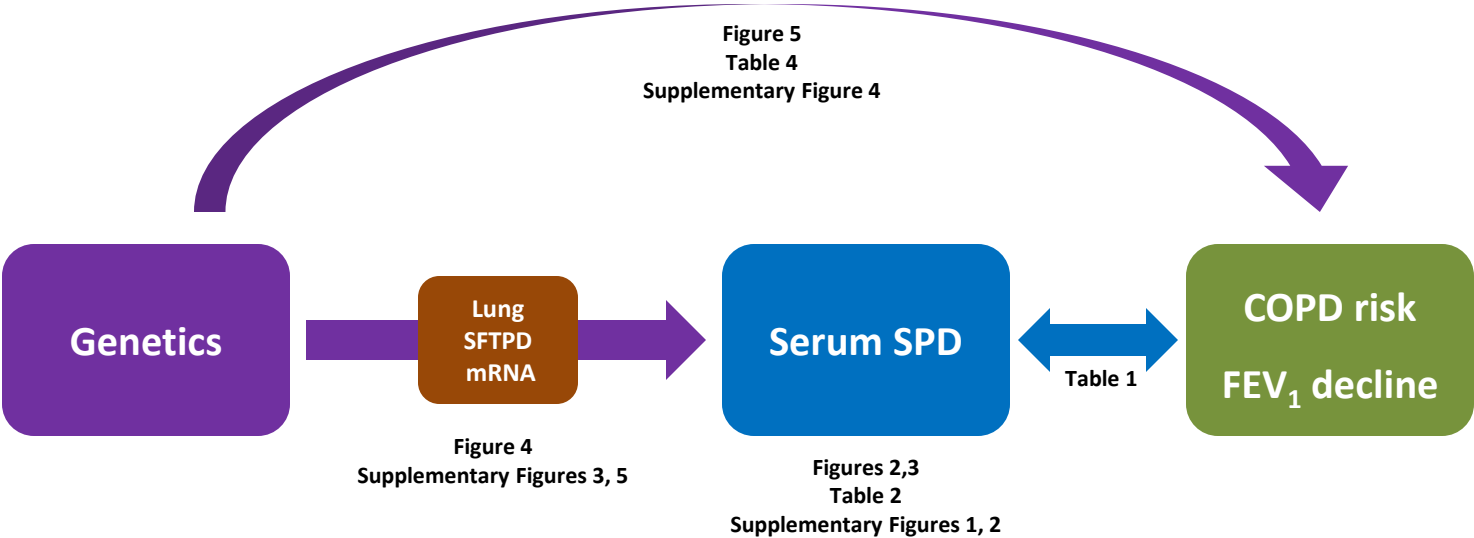


Figure 1: Study design

Figure 2

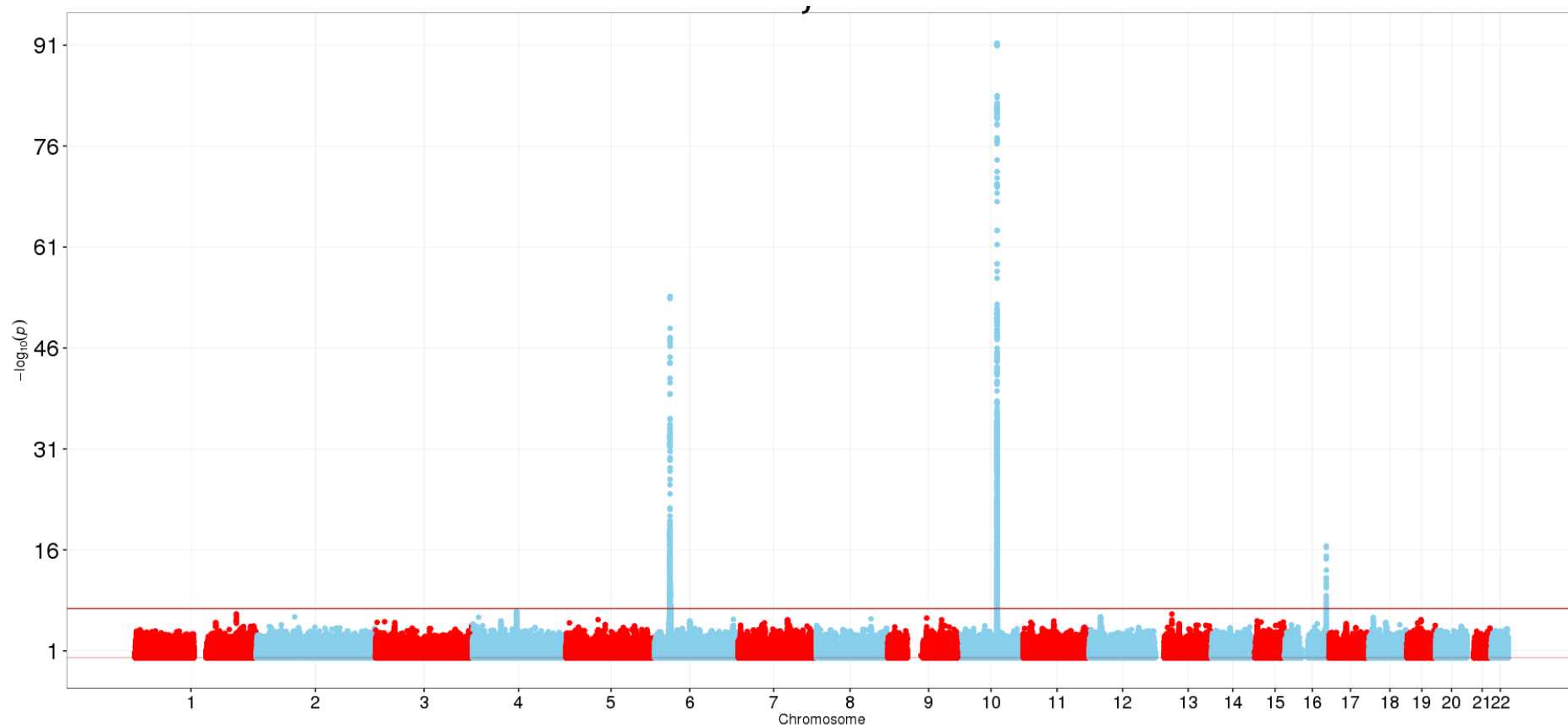


Figure 2. SP-D GWAS Manhattan plot. Manhattan plot shows the P values ($-\log_{10}$ scale) on the Y axis and the SNP positions across 22 autosomal chromosomes on the X axis. The horizontal red line represents the genome-wide cut-off of 5×10^{-8} .

Figure 3

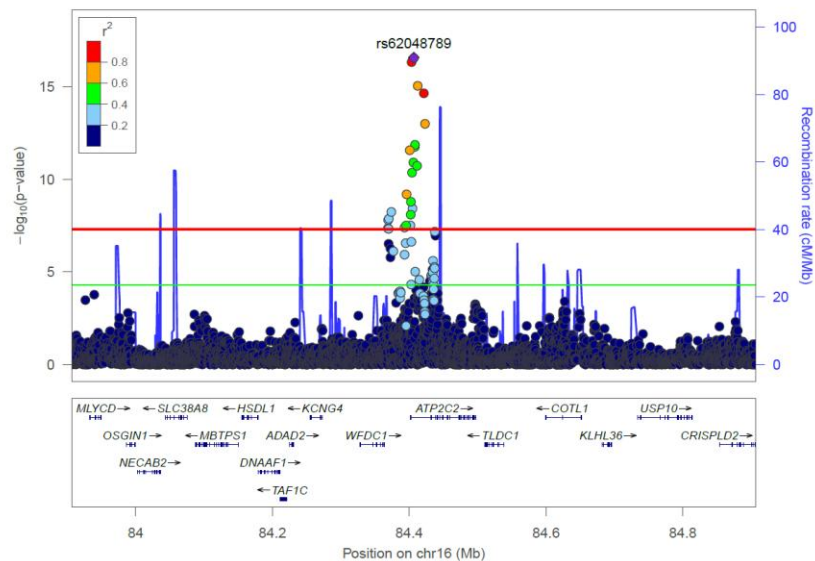
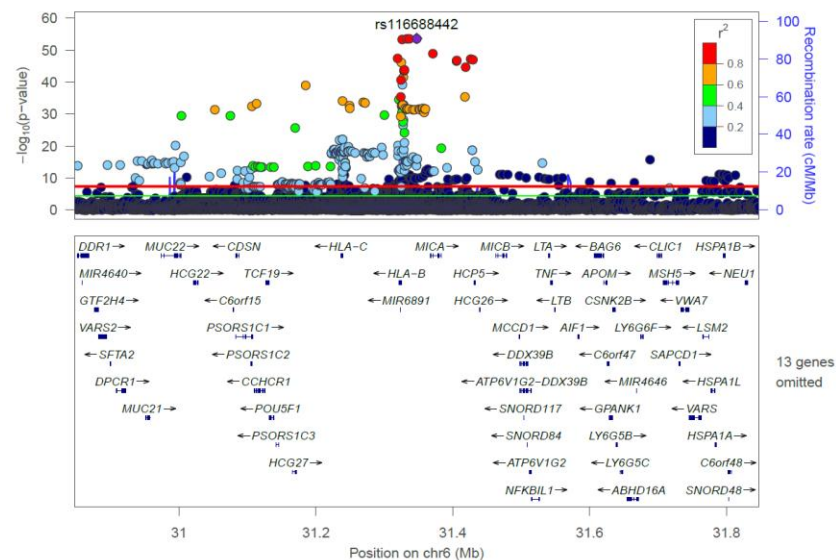
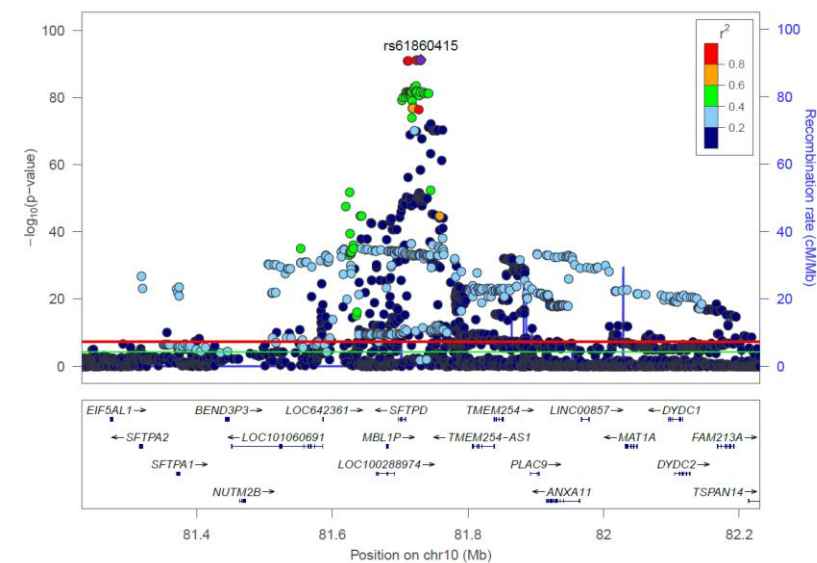


Figure 3. Region plots of the SP-D pQTLs loci. The Y axis represent the P values in the ($-\log_{10}$ scale) and the X axis is the genomic position. Gene names and their corresponding coordinates are shown below. The sentinel SNP is shown as a purple diamond and the color coding of SNPs reflects the degree of linkage disequilibrium (LD) with the sentinel SNP using 1000G reference.

Figure 4

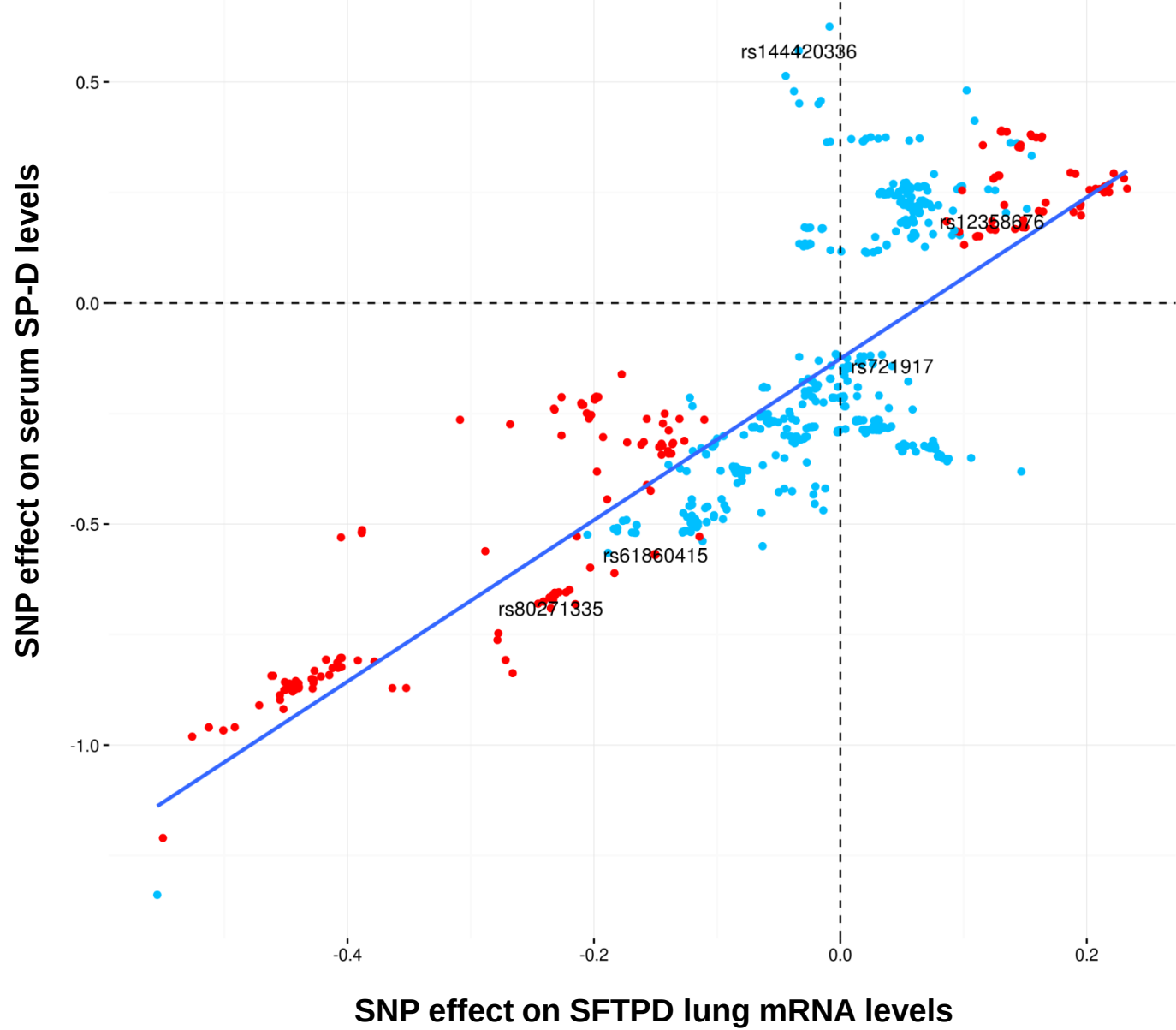


Figure 4. SP-D serum pQTLs on chr10 as lung eQTLs. Plotted are all the genome-wide significant SNPs for serum SP-D levels. The Y axis is the effect estimates of the SNPs on serum SP-D levels. The X axis is their effect estimates on lung tissue mRNA levels of SFTPD. SNPs in red are ones that have a significant effect on mRNA levels using a nominal P value of $P < 0.05$. The graph shows SNPs associated with increased serum levels in the GWAS are also associated with increases the lung mRNA expression and vice versa.

Figure 5

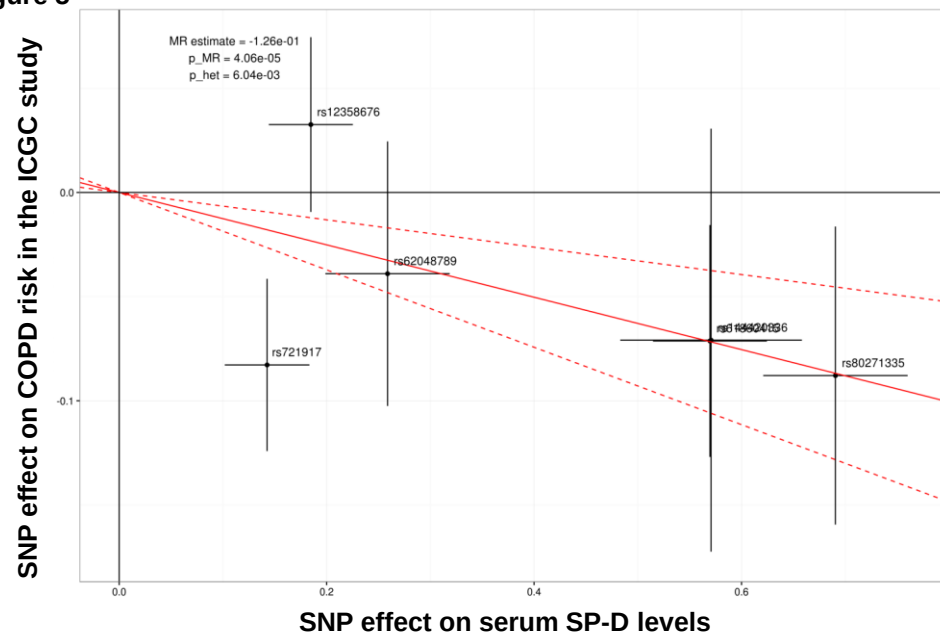


Figure 5: Association of individual SNPs with SP-D serum levels and COPD risk and progression. A)

Estimates of SP-D on COPD risk, and B) estimates of SP-D on FEV1 decline. Estimates are derived from serum GWAS for SP-D and from ICGC and LHS GWAS for COPD risk and FEV1 decline, respectively. Error bars represent 95% CIs. The SNPs rs numbers are shown. The slope of the line is the instrumental variable regression estimate of the effect of SP-D on COPD risk and FEV1 decline.

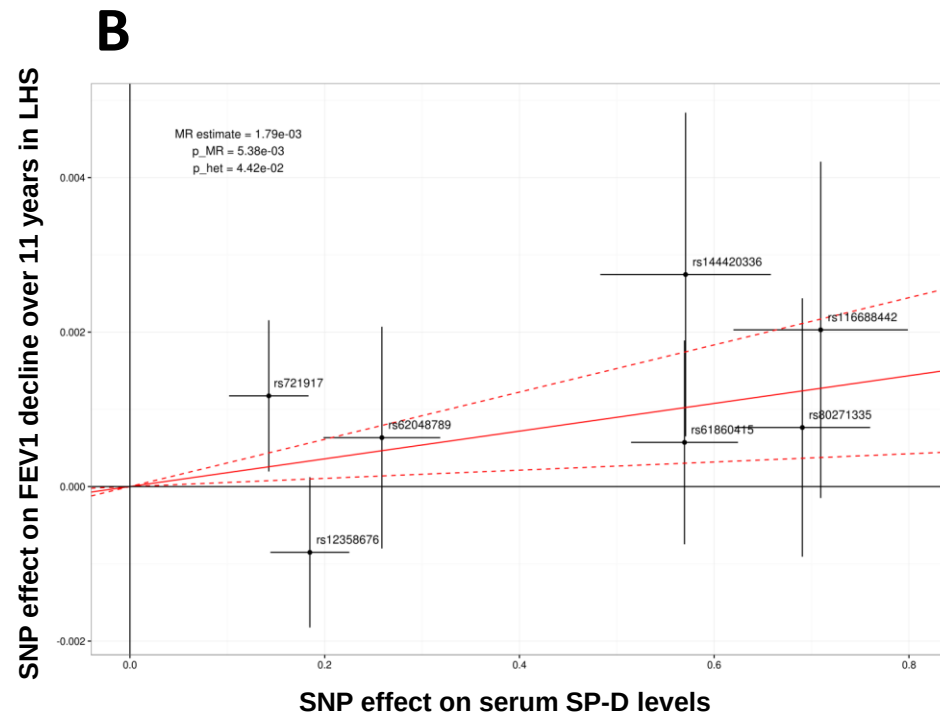


Table 1. Demographic and clinical characteristics of subjects with genotypes across quintiles of serum surfactant protein D (SP-D) in the Lung Health Study

	Quint 1 (n=786)	Quint 2 (n=792)	Quint 3 (n=812)	Quint 4 (n=822)	Quint 5 (n=829)	P value
SP-D, ng/ml	≤59.7	59.7 to 81.3	81.3 to 105.4	105.4 to 139.5	> 139.5	--
Age, years	52.1±6.8	52.8±6.8	54.0±6.6	53.8±6.6	55.1±6.3	<.0001
Men, %	63.6%	65.5%	61.6%	62.2%	63.1%	0.524
BMI, kg/m²	27.9±4.3	27.4±4.4	26.8±4.3	26.4±4.2	26.0±4.2	<.0001
Smokers, %[†]	70.9%	78.7%	82.9%	87.2%	92.5%	<.0001
FEV₁, % Predicted	78.5±11.6	75.7±12.0	74.9±12.0	74.6±11.9	73.3±12.2	<.0001

[†] Either continuous or intermittent smokers, defined as smokers (verified objectively using salivary cotinine or exhaled carbon monoxide) at the time of at least in one annual follow-up visit. All values are for the year 5 values because this was the visit during which blood samples were collected. For age, BMI, FEV₁, the values are mean±SD. P value is generated using the analysis of variance (ANOVA) test for continuous variables and a chi square test for categorical variables.

Table 2. The protein expression trait loci (pQTL) for serum surfactant protein D (SP-D)

SNP	Chr	Position (hg19)	Coded allele/refe rence allele	MAF	Imputati on r2	beta	SE	P value	% Variance
rs116688442	6	31347214	A / G	0.055	0.998	0.709	0.046	2.24E-54	6.013
rs61860415	10	81730579	C / T	0.176	0.992	-0.570	0.028	6.34E-92	9.056
rs80271335	10	81723739	C / T	0.100	0.976	-0.691	0.036	3.26E-84	8.330
rs144420336	10	81716293	A / G	0.063	0.918	0.571	0.045	2.37E-37	3.677
rs12358676	10	81755644	G / T	0.464	0.987	0.185	0.021	3.87E-19	1.787
rs721917	10	81706324	G / A	0.422	1	-0.143	0.021	7.50E-12	1.071
rs62048789	16	84406620	A / G	0.135	0.967	-0.259	0.031	2.61E-17	1.680

Chr: chromosome. Coded allele: the allele used to derive the statistics. MAF: minor allele frequency. Imputation r2: imputation quality control metrics. SE: standard error. % variance is the variance explained by that particular SNP. Beta, SE and P value for Chr10 pQTLs are from the SP-D pQTL GWAS.

Table 3. Lung tissue eQTL results for serum SP-D pQTLs

Chr	SNP	Position (b37)	Alleles	Z.Laval	Z.Groningen	Z.UBC	Lung eQTL value	Gene Symbol	SP-D GWAS P value
6	rs9380221	31150873	T C	-15.8	-17.8	-13.7	3.87E-163	PSORS1C3	1.28E-08
10	rs12267802	81749610	A G	2.9	1.7	3.4	3.75E-06	SFTPD	2.22E-31
16	rs12445532	84404091	G A	-9.3	-4.2	-3.7	4.53E-27	ATP2C2	4.30E-11

Chr: chromosome. Alleles: the SNP alleles, the first letter is the coded/effective allele. Lung eQTL P value: is the eQTL meta-analysis p value from UBC, Groningen, and Laval studies. Z.Laval: the eQTL Z score in the in Laval cohort. Z.UBC: the Z score in the in UBC cohort. Z.Groningen: the Z score in the in Groningen cohort. Gene Symbol: The gene that is regulated by the eQTL. PSORS1C3: psoriasis susceptibility 1 candidate 3, SFTPD: surfactant protein D, ATP2C2: ATPase, Ca++ transporting, type 2C, member 2.

Table 4

Table 4. Association SP-D pQTLs with COPD and related phenotypes.

Phenotype	rs116688442 (chr6) estimate (P value) (alleles)	rs12358676 (chr10) estimate (P value) (alleles)	rs144420336 (chr10) estimate (P value) (alleles)	rs61860415 (chr10) estimate (P value) (alleles)	rs62048789 (chr16) estimate (P value) (alleles)	rs721917 (chr10) estimate (P value) (alleles)	rs80271335 (chr10) estimate (P value) (alleles)
LHS FEV1 decline N= ~3,800	-0.00203 (p = 0.0678) (G_A)	0.000852 (p = 0.0863) (T_G)	-0.00274 (p = 0.0102) (G_A)	0.000572 (p = 0.396) (T_C)	0.000633 (p = 0.387) (G_A)	0.00117 (p = 0.0187) (A_G)	0.000765 (p = 0.37) (T_C)
LHS FEV1 Y5 N= ~3,800	0.00395 (p = 0.836) (G_A)	0.0164 (p = 0.0567) (T_G)	-0.0312 (p = 0.09) (G_A)	0.0114 (p = 0.327) (T_C)	-0.0153 (p = 0.23) (G_A)	0.0216 (p = 0.0127) (A_G)	0.0216 (p = 0.144) (T_C)
LHS FEV1FVC Y5 N= ~3,800	0.331 (p = 0.381) (G_A)	0.331 (p = 0.0528) (T_G)	-0.462 (p = 0.205) (G_A)	0.366 (p = 0.113) (T_C)	-0.00906 (p = 0.971) (G_A)	0.484 (p = 0.00495) (A_G)	0.617 (p = 0.0349) (T_C)
ICGC COPD 11,157 cases / 36,699 controls*	NA	-0.0326 (p = 0.128) (T_G)	-0.0709 (p = 0.172) (A_G)	-0.0713 (p = 0.012) (T_C)	0.039 (p = 0.229) (A_G)	-0.0828 (p = 8.87e-05) (A_G)	-0.0879 (p = 0.0161) (T_C)

LHS: Lung Health Study. ICGC: International COPD Genetics Consortium. FEV1 decline was estimated using Linear mixed effect. For each study, the alleles shown are the coded_noncoded allele. Estimates and P values are computed for the coded allele. P values <0.05 is are shown in bold.
*COPD results are reported for meta-analysis across the 17 ICGC European ancestry cohorts with genome-wide data available.

Supplementary Material

[Click here to download Necessary Additional Data: SP-D supplement TLRM.docx](#)