Title: Genome-wide gene-by-smoking interaction study of Chronic Obstructive Pulmonary Disease

Authors: Woori Kim, Dmitry Prokopenko, Phuwanat Sakornsakolpat, Brian D. Hobbs, Sharon M. Lutz, John E. Hokanson, Louise V. Wain, Carl A. Melbourne, Nick Shrine, Martin D. Tobin, Edwin K. Silverman, Michael H. Cho, Terri H. Beaty

Author Affiliations: Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, Maryland (Woori Kim and Terri H. Beaty); Genetics and Aging Research Unit, Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts (Dmitry Prokopenko); Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand (Phuwanat Sakornsakolpat); Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts (Brian D. Hobbs, Edwin K. Silverman, Michael H. Cho); Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, Massachusetts (Brian D. Hobbs, Edwin K. Silverman, Michael H. Cho); PRecisiOn Medicine Translational Research (PROMoTeR) Center, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care, Boston, Massachusetts (Sharon M. Lutz); Colorado School of Public Health, University of Colorado Denver, Aurora, Colorado (John E. Hokanson); Department of Health Sciences, University of Leicester, Leicester, UK (Louise V. Wain, Carl A. Melbourne, Nick Shrine, Martin D. Tobin); National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK (Louise V. Wain, Martin D. Tobin)

Abbreviations: 1df (1-degree-of-freedom); 2df (2-degree-of-freedom); AAT (Alpha-1 Antitrypsin); COPD (Chronic Obstructive Pulmonary Disease); CPD (Cigarettes Per Day); eQTL (expression Quantitative Trait Locus); FEV_1 (Forced Expiratory Volume in 1 second); FTND (Fagerstrom Test for Nicotine Dependence); FVC (Forced Vital Capacity); GOLD (Global Initiative for Chronic Obstructive Lung Disease); GWAS (Genome-Wide Association Study); ICGC (International COPD Genetics Consortium); MAF (Minor Allele Frequency); NHW (Non-Hispanic White); PC (Principal Component); PRS (Polygenic Risk Score); PY (Pack-Years); QC (Quality Controls); SNP (Single Nucleotide Polymorphism); UKB (UK Biobank)

Correspondence to: Dr. Terri H. Beaty, 615 N. Wolfe Street, Baltimore, MD 21205, Email (tbeaty1@jhu.edu)

Running head: Genome-wide gene-by-smoking interaction study of Chronic Obstructive Pulmonary Disease

Word count: 3500/ 3500

ABSTRACT

Risk for Chronic Obstructive Pulmonary Disease (COPD) is determined by both cigarette smoking and genetic susceptibility, but little is known about gene-by-smoking interactions. We performed a genome-wide association analysis of 179,689 controls and 21,077 COPD cases from UK Biobank subjects of European ancestry, considering genetic main effects and gene-by-smoking interaction effects simultaneously (2-degree-of-freedom (2df) test) as well as interaction effects alone (1-degree-of-freedom (1df) interaction test). We sought to replicate significant results in the COPDGene study and SpiroMeta Consortium. We considered two smoking variables: (1) ever/never and (2) current/non-current. In the 1df interaction test, we identified one genome-wide significant locus on 15q25.1 (*CHRNB4*) and identified PI*Z allele (rs28929474) *SERPINA1* and 3q26.2 (*MECOM*) in an analysis of previously reported COPD loci. In the 2df test, most of the significant signals were also significant for genetic marginal effects, aside from 16q22.1 (*SMPD3*) and 19q13.2 (*EGLN2*). The significant effects at 15q25.1 and 19q13.2 loci, both previously described in prior genome-wide association studies of COPD or smoking, but not 16q22.1 or 3q26.2, were replicated in the COPDGene and SpiroMeta. In our study, we identified interaction effects at previously reported COPD loci, however, we failed to identify novel susceptibility loci.

Words count: 194/200

Keywords: Chronic Obstructive Pulmonary Disease, Gene-Environment Interaction, Gene-by-Smoking Interaction, Genome-Wide Association Study, Smoking

Risk for Chronic Obstructive Pulmonary Disease (COPD) is determined by both cigarette smoking and genetic susceptibility. Adverse effects of smoking on risk of COPD may differ by an individual's genetic susceptibility, which raises the potential for important gene-by-smoking interactions. However, little is known about gene-by-smoking interactions on COPD risk.

A significant interaction between PI*Z allele (rs28929474) of *SERPINA1* and cigarette smoking on spirometric measure of lung function was reported in European-ancestry subjects (1,2). For COPD-related traits, genome-wide gene-by-smoking interaction studies have focused on quantitative measures of lung function based on spirometry (3,4). While spirometric measures of lung function are used to diagnose COPD, no genome-wide studies have investigated gene-by-smoking interaction on risk of COPD itself.

A recent large-scale genome-wide association study (GWAS) identified 82 distinct loci associated with COPD risk (5). However, these identified variants explained less than 10% of the phenotypic variability on a liability scale. To fill the gap of the genetics of COPD explained by common variants, more of the phenotypic variability might be explained by including gene-by-smoking interactions in GWAS model.

A major challenge of studying gene-by-environment interactions is the much larger sample size required compared to conventional GWAS to detect marginal effects of genes (6). The 2-degree-of-freedom (2df) joint test leverages genetic main effects and gene-by-environment interaction effects simultaneously and can provide better power than a standard interaction test, which is a 1-degree-of-freedom (1df) test (7). Using a 2df test, recent large-scale genome-wide gene-by-

environment interaction studies of complex traits have identified new genetic factors as well as possible gene-by-environment interactions (3,8–12). The availability of the large-scale UK Biobank (UKB) study, which collected a wide range of phenotypes as well as genetic data, could provide a promising opportunity to detect possible gene-by-environment interactions.

Here, we performed genome-wide gene-by-smoking interaction analyses of COPD in the UKB study to identify novel genetic variants for risk to COPD while accounting for potential smoking interactions, and assessed the impact of potential gene-by-smoking interactions on risk of COPD at known COPD and lung function GWAS loci.

METHODS

Study populations

The UKB is a population-based cohort of volunteers where over 500,000 individuals were originally recruited (13). We used UKB subjects as our discovery set. We also used two additional datasets, the COPDGene Study and SpiroMeta Consortium, to further investigate significant results from the UKB. COPDGene recruited former- and current-smokers whose smoking history was at least 10 pack-years (14). SpiroMeta is comprised of a total of 79,055 individuals from 22 studies (15). All participants provided written informed consent and studies were approved by local Research Ethics Committees and/or Institutional Review Boards.

Spirometric measures and genetic data

Details of quality controls (QC) of spirometric measures, genetic markers and subjects in the UKB study have been previously described (5,13,15). Briefly, to determine lung function,

measures of forced expiratory volume in 1 second (FEV_1) and forced vital capacity (FVC) were derived from spirometry volume-time series data, subjected to additional quality control based on ATS/ERS criteria (15,16). Genotyping was performed using Axiom UK BiLEVE array and Axiom Biobank array (Affymetrix, Santa Clara, California, USA) and imputed to the Haplotype Reference Consortium panel (version 1.1). We included independent subjects of European ancestry based on a combination of self-reported ethnicity data and principal components (PCs) data provided by UKB.

Measures of smoking exposure

We assigned smoking status to individuals in UKB based on their responses on questionnaires. Never-smokers included non-current-smokers or those who smoked less than 100 cigarettes in their life. Ever-smokers were defined as either current, most days (current or all days in the past) or smoked occasionally.

To test for possible gene-by-smoking interactions, we considered 2 smoking variables: Ever/Never and Current/Non-current smoker. For Ever/Never smokers, former- and current-smokers were included into the ever-smoker group. For Current/Non-current smokers, former- and never-smokers were included into the non-current smoker group. Smoking variables were coded as 0 and 1 for unexposed and exposed groups, respectively. Here, we refer Ever/Never-smoker analysis as "GxEver-smoking analysis" and Current/Non-current smoker analysis as "GxCurrent-smoking analysis".

Outcome

We defined COPD cases based on pre-bronchodilator spirometry following the modified Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria for moderate airflow limitation: FEV_1 less than 80% of predicted value (using reference equations from (17)), and the ratio of FEV_1/FVC less than 0.7.

Genetic analysis

We included markers with minor allele frequency (MAF) ≥ 0.01 and imputation quality score (r^2) ≥ 0.5. We performed a logistic regression analysis considering genetic main effects and gene-by-smoking interaction effects simultaneously (2df joint test) as well as interaction effects alone (1df interaction test), adjusting for age, sex, genotyping array and the first 10 PCs. We used the 2df joint test to search for new genetic variants of COPD, and the 1df interaction test to assess interaction effects alone. If a marker shows a significance in the 2df joint test, it is associated with the outcome across exposure groups. If a marker shows significance in the 1df interaction test, its genetic effect should differ by exposure group. Additionally, marginal GWAS were run stratified by each smoking variable. All genome-wide analyses were performed using the Plink software (version 2.0, www.cog-genomics.org/plink/2.0).

Conditional analysis

We defined distinct 'loci' using a 1-Mb window (+/- 500kb) around the lead variant (i.e. most significant single nucleotide polymorphism (SNP)). As our joint analysis was likely to include substantial overlap with previously described association studies of marginal effects for risk of COPD, we performed conditional analysis of each lead variant to determine whether our signals

were independent of known risk loci for COPD (5) or lung function (15). As the current GCTA (http://cnsgenomics.com/software/gcta) tool does not account for gene-by-environment interactions in their conditional analysis, we took a stratified approach for this analysis. We stratified by smoking exposed and unexposed groups, and conditioned on recognized SNPs from previous GWASs of COPD (5) or lung function (15) within 2-Mb of the lead variant. The conditioned 2df test for genetic main effects and interaction effects was then calculated on the conditioned stratified results using the following equations (9,18). For the 1df test,

$$\mathrm{Z} = \frac{\gamma_G^{(1)} - \gamma_G^{(0)}}{\sqrt{SE(\gamma_G^{(1)})^2 + SE(\gamma_G^{(0)})^2 - 2rSE(\gamma_G^{(1)})SE(\gamma_G^{(0)})}}$$

where $\gamma_G^{(1)}$ and $\gamma_G^{(0)}$ represent stratum-specific genetic effects; $SE(\gamma_G^{(1)})$ and $SE(\gamma_G^{(0)})$ are their respective standard errors (SE); and r is the Spearman rank correlation coefficient between $\gamma_G^{(1)}$ and $\gamma_G^{(0)}$, calculated from the genome-wide results. This Z statistics approximately follows a standard normal distribution under $H_0: \beta_{GE} = 0$. For the 2df test,

$$\mathrm{X} = \left[\frac{\gamma_G^{(1)}}{SE(\gamma_G^{(1)})}\right]^2 + \left[\frac{\gamma_G^{(0)}}{SE(\gamma_G^{(0)})}\right]^2$$

which approximately follows a 2df chi-squared distribution under $H_0: \beta_G = \beta_{GE} = 0$ when the two strata are independent.

Dose response analysis

To further characterize our significant results, we conducted a dose response analysis in all subjects and in ever-smokers as a secondary analysis. We tested gene-by-smoking dose interaction using the standard 1df test. We considered three quantitative measures of smoking

dose: smoking duration, pack-years (PY) and cigarettes per day (CPD). We considered exposures as both a quantitative variable and a categorical variable grouped based on quartiles.

Replication

As the COPDGene cohort is enriched for heavy smokers, we hypothesized SNPs presenting a stronger association among the exposed group in the UKB would also show some marginal effects on COPD risk in COPDGene subjects. For the selected SNPs, we tested for a marginal association between each SNP and COPD risk, adjusting for age, sex, smoking status, pack-years and genetic ancestry PCs in 5,342 Non-Hispanic White (NHW) subjects from COPDGene. We further tested gene-by-smoking dose interaction based on the 1df test. As the Fagerstrom Test for Nicotine Dependence (FTND) measure was collected for current-smokers in COPDGene, we also tested gene-by-FTND interaction. We considered both a quantitative FTND score and a categorical variable grouped into mild (0-3), moderate (4-6) and severe (7-10) (19).

We also attempted to replicate our results by lookup in a genome-wide association analysis of spirometric measures of lung function (FEV_1, FVC and FEV_1/FVC) stratified by ever- and never-smoker groups in SpiroMeta. Using summary statistics from these stratified results, we calculated test statistics for a 1df interaction test and a 2df joint test based on the same approach used in our stratified conditional analysis (9,18). Analyses on SpiroMeta data has been previously published (15). Briefly, each of the 22 studies performed a linear regression adjusting for age, age^2, sex, and height, by using rank-based inverse normal transformation, adjusting for population substructure by including genetic ancestry PCs or as linear mixed models, and

performing separate analyses for ever- and never-smokers. Results were combined under a fixed-effects meta-analysis.

RESULTS

Subject characteristics

We analyzed 200,766 subjects of European ancestry, including 179,689 controls and 21,077 COPD cases in the UKB study (**Table 1**). These UKB subjects included 71,591 ever-smokers (former- and current-smokers combined) and 129,175 never-smokers; and 14,590 current-smokers and 186,176 non-current-smokers (never- and former-smokers combined). NHW COPDGene subjects included 3,361 former-smokers and 1,981 current-smokers. While UKB subjects had a higher proportion of COPD cases among current-smokers (31.5%) compared to former- (13.8%) and never-smokers (6.7%), COPDGene subjects (who were enriched for moderate-to-severe COPD) showed a higher proportion of COPD cases among former-smokers (54.5%) than current-smokers (49.4%).

Genome-wide Results

The analysis workflow is depicted in **Figure 1**.

2df joint test. We identified 48 loci for GxEver- and 55 loci for GxCurrent-smoking analysis (defined using 1-Mb windows) achieving genome-wide significance (P<5.00E-08) (**Supplemental Table 1** and **Supplemental Figure 1**). The lead variants at 15 of these loci for GxEver- and 19 loci for GxCurrent-smoking analysis were the same as previously identified in

GWAS of COPD or lung function (5,15). For the remaining loci, we conducted a conditional analysis to search for new signals (see **Methods** and **Supplemental Table 2**). After adjusting for previously reported variants, 2 loci, 16q22.1 - *SMPD3* (lead variant: rs141322661, P_{2df}=3.92E-09 from GxEver- and P_{2df}=1.45E-08 from the GxCurrent-smoking analysis) and 19q13.2 - *EGLN2* (lead variant: rs2604894, P_{2df} =5.87E-09 from GxCurrent-smoking analysis), maintained genome-wide significance (**Table 2** and **Figure 2**).

In a previous UKB GWAS of COPD examining only marginal genetic effects (5), rs141322661 at 16q22.1 reached genome-wide significance (P=1.88E-09), but not in a meta-analysis of the UKB and the International COPD Genetics Consortium (ICGC) (P=1.90E-08), and rs2604894 at 19q13.2 did not reach genome-wide significance (P=1.17E-04). Other signals were attenuated and did not reach genome-wide significance, indicating that those findings are not novel.

1df interaction test. We identified one locus - 15q25.1 (defined using 1-Mb windows) as achieving genome-wide significance (P <5.00E-08) for both GxEver- and GxCurrent-smoking analyses (**Table 2, Figure 3** and **Supplemental Figure 2**). In the GxEver-smoking analysis, the lead variant (rs12440014 in *CHRNB4*) showed $P_{1df\ interaction}$=8.96E-12, presenting as a significant association among ever-smokers (OR (95% CI)=0.85 (0.82-0.88), P=3.39E-19), but not among never-smokers.

Interaction of reported variants

We examined possible gene-by-smoking interactions on risk to COPD at 82 known COPD-associated loci, 279 known lung function-associated loci and 2 loci previously reporting smoking

interactions for either lung function or COPD (**Supplemental Table 3**). Because results from the 2df test for these known loci predominantly showed genetic main effects, we evaluated results from the 1df interaction test under Bonferroni corrected significance thresholds.

At known loci for risk to COPD, rs55676755 in *CHRNA3* and rs28534575 in *CHRNB4*, significantly interacted with smoking, presenting as significant associations in ever-smokers (rs55676755: OR (95% CI)=1.19 (1.15-1.22), P=8.74E-28 and rs28534575: OR (95% CI)=0.85 (0.82-0.88), P=2.53E-18), but not in never-smokers (**Supplemental Table 4**). SNP rs7642001 at 3q26.2 – *MECOM* showed a significant interaction among current smokers ($P_{1df\ interaction}$=3.65E-04) but not among non-current smokers ($P_{1df\ interaction}$=3.11E-01). At known loci for lung function, there was no evidence of significant interactions with smoking. At loci previously reporting smoking interactions, PI*Z allele (rs28929474) - *SERPINA1* significantly interacted with both ever smoking ($P_{1df\ interaction}$=6.70E-04) and current smoking ($P_{1df\ interaction}$=7.19E-03).

Selected SNPs

To further investigate significant results, we selected SNPs at 5 loci (**Table 3** and **Figure 4**). In the 2df joint test, rs141322661 at 16q22.1 – *SMPD3* and rs2604894 at 19q13.2 – *EGLN2* reached genome-wide significance, independent of previously described loci of COPD and lung function. In the 1df interaction test, we included rs12440014 at 15q25.1 – *CHRNB4*. Among previously reported variants, we selected rs7642001 at 3q26.2 – *MECOM* and rs28929474 in *SERPINA1*, with evidence of interaction.

To examine whether these five selected SNPs were associated with smoking behavior, we checked regions of selected SNPs in the most recent and largest GWAS of smoking itself (20) (**Supplemental Table 5**). Markers at 15q25.1 and 19q13.2 were reported to be associated with CPD and current smoking.

To further characterize these potential gene-by-smoking interactions, we conducted a dose response analysis (**Supplemental Table 6**). SNPs rs7642001 at 3q26.2 – *MECOM,* rs28929474 in *SERPINA1*, rs12440014 at 15q25.1 – *CHRNB4* and rs2604894 at 19q13.2 – *EGLN2* showed nominally significant interactions with smoking duration on COPD risk (P<0.05). In a dose response analysis among ever-smokers, the significance of dose response at these SNPs were attenuated but rs7642001 and rs12440014 were still nominally significant (P<0.05).

Replication

To replicate our findings, we used COPDGene and summary statistics from SpiroMeta (**Table 3**).

COPDGene. As COPDGene is enriched for heavy smokers, we hypothesized any SNPs showing a stronger association among exposed group in the UKB should also show some marginal associations with COPD risk among NHW COPDGene subjects. SNPs rs7642001 at 3q26.2 - *MECOM*, rs28929474 in *SERPINA1,* rs12440014 at 15q25.1 - *CHRNB4* and rs2604894 at 19q13.2 - *EGLN2* were nominally significantly associated with COPD risk (P<0.05). In a dose response analysis, a stronger association between rs7642001 at 3q26.2 – *MECOM* and COPD was observed with longer duration of smoking ($P_{1df\ interaction}$=6.20E-04) (**Supplemental Table 6**).

For the 1,937 current smokers in COPDGene available for FTND scores, rs12440014 at 15q25.1 - *CHRNB4* showed evidence of interaction with higher nicotine dependence ($P_{1df\ interaction}$=4.37E-02).

SpiroMeta. We replicated a significant interaction for rs12440014 at 15q25.1 – *CHRNB4* with ever smoking on FEV_1 ($P_{1df\ interaction}$=7.33E-03), presenting as a stronger association among ever-smokers compared to never-smokers in SpiroMeta (**Table 3**). We observed a significant interaction for rs7642001 at 3q26.2 - *MECOM* on FEV_1 ($P_{1df\ interaction}$=1.97E-03). However, the direction of this apparent interaction effects was opposite between UKB and SpiroMeta. Allele "A" at rs7642001 was more significantly associated with decreased FEV_1 among never-smokers (Beta (95% CI)=-0.04 (-0.05, -0.02), P=3.31E-05) compared to ever-smokers (Beta (95% CI)=-0.01 (-0.03, 0.005), P=1.93E-01). In the UKB, this SNP was more significantly associated with increased risk of COPD among current-smokers (OR (95% CI)=1.20 (1.13-1.27), P=4.54E-10) compared to non-current-smokers (OR (95% CI)=1.07 (1.04-1.09), P=3.09E-07). The stratified analyses for other measures of lung function are listed in **Supplemental Table 7**.

DISCUSSION

We conducted a genome-wide association analysis of COPD accounting for smoking interaction to identify novel susceptibility loci and to assess the potential gene-by-smoking interactions in UK Biobank subjects of European ancestry. Most of the significant signals in the 2df joint test were also significant for genetic marginal effects, aside from 16q22.1 (*SMPD3*) and 19q13.2 (*EGLN2*). In the 1df interaction test, we identified one genome-wide significant locus, 15q25.1 (*CHRNB4*), and identified PI*Z allele at *SERPINA1* and 3q26.2 (*MECOM*) in an analysis of

previously reported COPD risk loci. The estimated effects at the 15q25.1 and 19q13.2 loci, both previously described in prior GWAS of COPD or smoking, but not 16q22.1 or 3q26.2, were replicated in the COPDGene and SpiroMeta.

SNP rs141322661 at 16q22.1 had reached genome-wide significance in previous GWAS of COPD in UKB subjects but not in the meta-analysis of UKB and ICGC. SNP rs141322661, an intronic variant of *SMPD3*, occurs at a very low frequency in European populations (the G allele has a frequency between 0.01 and 0.02), which makes it difficult to replicate this signal. Further investigation of this 16q22.1 region will be required to confirm this finding.

SNP rs2604894 at 19q13.2 could only be observed through genome-wide association analysis accounting for current-smoking interaction in UKB subjects. In the previous UKB GWAS of COPD (which did not incorporate interaction in the model), rs2604894 did not reach genome-wide significance (5). In conventional GWAS of COPD using cohorts enriched for smokers, rs2604894 was reported to be significantly associated with COPD risk (OR (95% CI)=0.74 (0.65-0.84), P=3.41E-08) (21). This study included cohorts such as COPDGene (14) and ECLIPSE (22) designed to identify genetic factors for COPD by recruiting exclusively former- and current-smokers. The different study design between the UKB study, a population-based cohort, and cohorts at high risk to COPD because of their smoking history may have attenuated the statistical significance for rs2604894 association and hindered the replication in previous marginal GWAS in the UKB. Our finding highlights the importance of accounting for heterogeneity in genetic effects across exposure groups in association discovery studies.

SNP rs2604894 at 19q13.2 is an intronic variant of *EGLN2*, a gene known to be involved in regulating hypoxia tolerance and apoptosis in cardiac and skeletal muscle. Markers at 19q13.2 were reported to be associated with CPD and current smoking (23). Significant lung expression quantitative traits locus (eQTLs) (but not including rs2604894) have been detected at 19q13.2 (24). Further functional studies of the 19q13.2 region is clearly warranted to verify the possible contribution to COPD.

We identified genome-wide significant gene-by-smoking interaction effects at 15q25.1 in UKB and replicated these findings in SpiroMeta, revealing associations primarily in ever-smokers. The *CHRNA5/A3/B4* gene cluster on 15q25.1 encodes the nicotinic acetylcholine receptor subunits α5, α3 and β4. Variants in this gene cluster have been robustly associated with several lung-related traits, such as lung cancer (25) and COPD (5) as well as smoking-related phenotypes, such as smoking quantity (20,26–28) and nicotine dependence (26). Because smoking is the most important environmental risk factor for COPD, it is quite likely that the association of variants in 15q25.1 region with COPD mediates through smoking behavior (29). Collectively, we speculate genes in the 15q25.1 region exerts both gene-by-smoking interaction and mediation effects.

We noted a significant dose response for rs12440014 at 15q25.1 in the UKB, but not in COPDGene study. This may be simply due to smaller sample sizes in COPDGene study. However, given COPDGene was limited to heavy smokers and thus enriched for severe COPD cases compared to UKB (a population-based cohort), our results could reflect other possibilities: 1) the genetic susceptibility of the 15q25.1 region on COPD could be substantial at relatively low

levels of smoking exposure, and/or 2) COPD patients may be more likely to quit smoking as their symptoms worsen, diluting any association between markers in 15q25.1 and COPD.

We confirmed a known gene-by-smoking interaction on COPD, *SERPINA1* (PI*Z allele)-by-smoking interaction in our study population (1,2). In a previous study, a PI*Z-by-smoking interaction was identified for FEV_1 (P=0.03) and COPD status (P=0.01) in subjects of European ancestry (2). The *SERPINA1*, which encodes the Alpha-1 Antitrypsin (AAT) protein, influences the risk to COPD (30). Homozygosity for PI*Z allele is the most common cause of AAT deficiency. Although it is a Mendelian syndrome, there is marked variability in the development and severity of COPD among PI*ZZ individuals. Our replication in the UKB study helps to understand variable manifestations of COPD risk among individuals with AAT deficiency.

Our objective was to identify genetic loci associated with COPD risk, which may have been missed when considering only genetic main effects in the conventional GWAS approach used in Sakornsakolpat P. et al., 2019 (5). However, our study incorporating potential smoking interactions did not reveal novel loci. There are several possible explanations for our lack of finding. First, power for detecting gene-by-smoking interactions and discovering novel genetic risk factors may be limited even in this large sample size (31). Second, we only included independent subjects of European ancestry. Investigation of more ethnically diverse populations would allow more robust inferences of gene-by-environment interaction by increasing diversity of not only environmental exposure but also genetic determinants (32). Third, we used self-reported smoking history. Measurement errors of smoking exposure could lead to our lack of findings of gene-by-smoking interactions (6). Fourth, our study included variants with MAF ≥

0.01. Rare variants are traditionally thought to exert large gene effects and could lead to larger gene-by-smoking interaction effects.

Despite our large sample size (n=200,766), there are limitations in our study. First, our use of two smoking measures (ever smoking and current smoking) in genome-wide investigation may have limited interpretation of our results. Smokers with more severe COPD are more likely to reduce or quit smoking and those without symptoms are more likely to continue smoking and thus be current smokers, often described as "healthy smoker effect". Such a phenomenon is highly possible in COPDGene and may also be relevant for UKB as well (33) . Second, a "healthy volunteer" selection bias exists in UKB study. The UKB cohort is not fully representative of the general population; its participants are less likely to smoke and have fewer self-reported health conditions compared with the general population of the UK (34). However, generalizability is not necessary to draw inferences about associations. Its large sample size and heterogeneity of smoking exposures should still make our findings valid. Third, our use of the 2df joint test may be limited to understand gene-by-smoking interaction on COPD risk or more broadly, gene-by-environment interaction. Integration of genetic markers and other "-omics" data (transcriptomic, proteomic or epigenomic data) could be helpful. For example, genetic markers influencing other biomarkers (such as eQTL) may be more likely to interact with smoking (35).

In summary, our genome-wide investigation incorporating smoking interaction did not identify novel susceptibility loci of COPD. However, we identified interaction effects at previously reported COPD loci, 15q25.1 (*CHRNB4*) and PI*Z allele in *SERPINA1* on COPD risk. Cigarette

smoking is the most important environmental risk factor for COPD, but individuals vary in their susceptibility to the damaging effects of cigarette smoke. It raises the possibility of detectable gene-by-smoking interactions, but we identified few significant interactions in our large-scale study. Considering diverse populations and other approaches may better help further elucidate gene-by-environment interactions on COPD risk.

References

1. Silverman EK, Province MA, Campbell EJ, Pierce JA, Rao DC, Boerwinkle E. Family study of α1-antitrypsin deficiency: Effects of cigarette smoking, measured genotype, and their interaction on pulmonary function and biochemical traits. Genet Epidemiol [Internet]. 1992 [cited 2017 Dec 19];9(5):317–31. Available from: http://doi.wiley.com/10.1002/gepi.1370090504

2. Castaldi PJ, Demeo DL, Hersh CP, Lomas DA, Soerheim IC, Gulsvik A, et al. Impact of non-linear smoking effects on the identification of gene-by-smoking interactions in COPD genetics studies. Thorax [Internet]. 2011 Oct [cited 2017 Jul 15];66(10):903–9. Available from: http://thorax.bmj.com/content/thoraxjnl/66/10/903.full.pdf

3. Hancock DB, Soler Artigas MM, 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 [Internet]. 2012;8(12):e1003098. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23284291

4. Park B, Koo S-M, An J, Lee M, Kang HY, Qiao D, et al. Genome-wide assessment of gene-by-smoking interactions in COPD. Sci Rep [Internet]. 2018 Dec 18 [cited 2018 Jul 20];8(1):9319. Available from: www.nature.com/scientificreports

5. Sakornsakolpat P, Prokopenko D, Lamontagne M, Reeve NF, Guyatt AL, Jackson VE, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. Nat Genet [Internet]. 2019 Mar 25 [cited 2018 Jul 31];51(3):494–505. Available from: https://www.biorxiv.org/content/early/2018/06/26/355644

6. Aschard H, Lutz S, Maus B, Duell EJ, Fingerlin TE, Chatterjee N, et al. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Hum Genet [Internet]. 2012 Oct 4 [cited 2017 Jul 18];131(10):1591–613. Available from: https://link.springer.com/content/pdf/10.1007%2Fs00439-012-1192-0.pdf

7. Kraft P, Yen Y-CC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Hum Hered [Internet]. 2007;63(2):111–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17283440

8. Sidney S, de Faire U, Faul JD, Katsuya T, Pedersen NL, de las Fuentes L, et al. Multi-Ancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions. Am J Epidemiol [Internet]. 2019 Jan 29 [cited 2019 Mar 13]; Available from: https://academic.oup.com/aje/advance-article-abstract/doi/10.1093/aje/kwz005/5304469

9. Bentley AR, Sung YJ, Brown MMRM, Winkler TW, Kraja AT, Ntalla I, et al. Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. Nat Genet [Internet]. 2019 Apr 29;51(4):636–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30926973

10. Xu J, Gaddis NC, Bartz TM, Hou R, Manichaikul AW, Pankratz N, et al. Omega-3 Fatty Acids and Genome-wide Interaction Analyses Reveal DPP10 -Pulmonary Function Association. Am J Respir Crit Care Med [Internet]. 2018 [cited 2018 Nov 29];30:rccm.201802-0304OC. Available from: https://www-atsjournals-org.ezp.welch.jhmi.edu/doi/pdf/10.1164/rccm.201802-0304OC

11. Sung YJ, Winkler TW, de Las Fuentes L, Bentley AR, Brown MR, Kraja AT, et al. A Large-Scale Multi-ancestry Genome-wide Study Accounting for Smoking Behavior Identifies Multiple Significant Loci for Blood Pressure. Am J Hum Genet [Internet]. 2018 [cited 2019 Feb 13];102(3):375–400. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5985266/pdf/main.pdf

12. Justice AE, Winkler TW, Feitosa MF, Graff M, Fisher VA, Young K, et al. Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. Nat Commun [Internet]. 2017 Apr 26 [cited 2019 Mar 21];8:14977. Available from: http://www.nature.com/doifinder/10.1038/ncomms14977

13. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature [Internet]. 2018 Oct 10 [cited 2019 Jan 19];562(7726):203–9. Available from: https://doi.org/10.1038/s41586-018-0579-z

14. Regan E a, Hokanson JE, Murphy JR, Make B, Lynch D a, Beaty TH, et al. Genetic Epidemiology of COPD (COPDGene) Study Design. COPD J Chronic Obstr Pulm Dis [Internet]. 2011 Feb 9;7(1):32–43. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2924193&tool=pmcentrez&rendertype=abstract

15. Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat Genet [Internet]. 2019 Mar 25 [cited 2018 Jul 31];51(3):481–93. Available from: https://www.biorxiv.org/content/early/2018/06/12/343293

16. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. Am J Respir Crit Care Med [Internet]. 2017 Mar 1;195(5):557–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28128970

17. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric Reference Values from a Sample of the General U.S. Population [Internet]. Vol. 159, Am J Respir Crit Care Med. 1999 [cited 2019 Sep 3]. Available from: www.atsjournals.org

18. Sung YJ, Winkler TW, Manning AK, Aschard H, Gudnason V, Harris TB, et al. An Empirical Comparison of Joint and Stratified Frameworks for Studying G × E Interactions: Systolic Blood Pressure and Smoking in the CHARGE Gene-Lifestyle Interactions Working Group. Genet Epidemiol [Internet]. 2016;40(5):404–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27230302

19. Hancock DB, Reginsson GW, Gaddis NC, Chen X, Saccone NL, Lutz SM, et al. Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence. Transl Psychiatry [Internet]. 2015 Oct 6 [cited 2019 Oct 26];5:e651. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26440539

20. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet [Internet]. 2019 Feb 14 [cited 2019 Mar 27];51(2):237–44. Available from: https://doi.org/10.1038/s41588-018-0307-5

21. 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. Hum Mol Genet [Internet]. 2012 Feb 15 [cited 2019 Jul 12];21(4):947–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22080838

22. Vestbo J, Anderson W, Coxson HO, Crim C, Dawber F, Edwards L, et al. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). Eur Respir J [Internet]. 2008 Apr 1 [cited 2018 Oct 4];31(4):869–73. Available from: www.erj.ersjournals.com/

23. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2

million individuals yield new insights into the genetic etiology of tobacco and alcohol use [Internet]. Vol. 51, Nature Genetics. 2019 [cited 2019 Apr 30]. p. 237–44. Available from: https://doi.org/10.1038/s41588-018-0307-5

24. Lamontagne M, Couture C, Postma DS, Timens W, Sin DD, Paré PD, et al. Refining Susceptibility Loci of Chronic Obstructive Pulmonary Disease with Lung eqtls. Miao X-P, editor. PLoS One [Internet]. 2013 Jul 30;8(7):e70220. Available from: https://dx.plos.org/10.1371/journal.pone.0070220

25. 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 [Internet]. 2008 Apr 3 [cited 2017 Jan 26];452(7187):633–7. Available from: http://www.nature.com/doifinder/10.1038/nature06885

26. 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 [Internet]. 2008 Apr 3 [cited 2017 Jan 26];452(7187):638–42. Available from: http://www.nature.com/articles/nature06846

27. Furberg H, Kim Y, Dackor J, Boerwinkle E, Franceschini N, Ardissino D, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet [Internet]. 2010 May 25 [cited 2016 Nov 22];42(5):441–7. Available from: http://www.nature.com/doifinder/10.1038/ng.571

28. Lutz SM, Frederiksen B, Begum F, McDonald M-LN, Cho MH, Hobbs BD, et al. Common and Rare Variants Genetic Association Analysis of Cigarettes per Day Among Ever-Smokers in Chronic Obstructive Pulmonary Disease Cases and Controls. Nicotine Tob Res [Internet]. 2019 May 21;21(6):714–22. Available from: https://academic.oup.com/ntr/article/21/6/714/4996129

29. Lutz SM, Hokanson JE. Genetic influences on smoking and clinical disease: Understanding behavioral and biological pathways with mediation analysis. Ann Am Thorac Soc. 2014;11(7):1082–3.

30. DeMeo DL, Silverman EK. α1-Antitrypsin deficiency • 2: Genetic aspects of α1-antitrypsin deficiency: Phenotypes and genetic modifiers of emphysema risk [Internet]. Vol. 59, Thorax. 2004 [cited 2019 Jan 17]. p. 259–64. Available from: www.thoraxjnl.com

31. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, et al. Size matters: Just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol. 2009;38(1):263–73.

32. Ritz BR, Chatterjee N, Garcia-Closas M, Gauderman WJ, Pierce BL, Kraft P, et al. Lessons Learned From Past Gene-Environment Interaction Successes. Am J Epidemiol [Internet]. 2017 Oct 1 [cited 2017 Aug 1];186(7):778–86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28978190

33. Becklake MR, Lalloo U. The 'Healthy Smoker': A Phenomenon of Health Selection? Respiration [Internet]. 1990 [cited 2019 Oct 14];57(3):137–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2274712

34. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. Am J Epidemiol [Internet]. 2017 Nov 1 [cited 2019 Sep 12];186(9):1026–34. Available from: https://academic.oup.com/aje/article/186/9/1026/3883629

35. Favé M-J, Lamaze FC, Soave D, Hodgkinson A, Gauvin H, Bruat V, et al. Gene-by-environment interactions in urban populations modulate risk phenotypes. Nat Commun [Internet]. 2018 Dec 6 [cited 2018 Mar 30];9(1):827. Available from: http://www.nature.com/articles/s41467-018-03202-2

Table 1. Subject characteristics stratified by smoking status

	UK Biobank			COPDGene	
	Never	**Former**	**Current**	**Former**	**Current**
N	129175	57001	14590	3361	1981
Moderate COPD	8631 (6.70)	7857 (13.80)	4589 (31.50)	1831 (54.50)	978 (49.40)
Age	55.55 (8.05)	57.91 (7.64)	54.12 (8.06)	64.95 (8.23)	57.48 (7.80)
Female	50194 (38.90)	29175 (51.20)	6941 (47.60)	1620 (48.20)	907 (45.80)
Pack-years	0.00 (0.00)	19.21 (17.50)	28.40 (18.25)	47.15 (27.30)	48.13 (24.49)
BMI	26.87 (4.56)	27.95 (4.58)	26.73 (4.68)	28.97 (5.89)	27.58 (5.79)
FEV_1 % predicted	96.44 (13.95)	93.52 (16.30)	85.10 (18.51)	70.03 (29.35)	75.29 (25.05)
FEV_1/FVC	0.77 (0.06)	0.76 (0.07)	0.72 (0.09)	0.61 (0.19)	0.65 (0.16)

Mean (SD) for continuous variable; N (%) for categorical variable; FEV_1: Forced expiratory volume in one second; FVC: Forced vital capacity

Table 2. Significant results of 2df join test and 1df interaction test in UK Biobank

rsID	Chr:Position	Nearest Gene	Effect/Ref. Allele	EAF	Smoking Exposure	Genetic Main		Interaction		2DF Joint
						OR (95% CI)	P	OR (95% CI)	P	P
Significant from 2df joint test*										
rs141322661	16:68398875	*SMPD3*	G/A	0.01	Ever Smoking	0.76 (0.65-0.88)	2.09E-04	0.94 (0.78-1.15)	5.72E-01	3.92E-09
rs141322661	16:68398875	*SMPD3*	G/A	0.01	Current Smoking	0.77 (0.69-0.86)	1.47E-06	0.81 (0.62-1.07)	1.45E-01	1.45E-08
rs2604894	19:41292404	*EGLN2*	A/G	0.45	Current Smoking	0.96 (0.94-0.98)	1.28E-03	0.9 (0.84-0.95)	4.11E-04	5.87E-09
Significant from 1df interaction test*										
rs7170068	15:78912943	*CHRNA3*	A/G	0.22	Current Smoking	0.96 (0.93-0.98)	2.38E-03	0.79 (0.74-0.86)	1.82E-09	6.08E-16
rs12440014	15:78926726	*CHRNB4*	G/C	0.24	Ever Smoking	1.02 (0.98-1.06)	3.14E-01	0.83 (0.79-0.88)	8.96E-12	6.96E-18
Significant from 1df interaction test in candidate variants**										
rs7642001	3:168746145	*MECOM*	A/G	0.37	Current Smoking	1.07 (1.04-1.09)	3.20E-07	1.12 (1.05-1.19)	3.65E-04	1.73E-14

rs28929474	14:94844947	*SERPINA1*	T/C	0.02	Ever Smoking	0.95 (0.85-1.07)	4.02E-01	1.3 (1.12-1.52)	6.70E-04	1.10E-04

EAF: Effect Allele Frequency; Interaction: Interaction test with 1 degree of freedom; 2DF Joint: Joint test with 2 degrees of freedom of genetic main and interaction effects; *Genome-wide statistical significance (P<5.00E-08) applied; **Bonferroni corrected statistical significance (P<1.00E-04) applied.

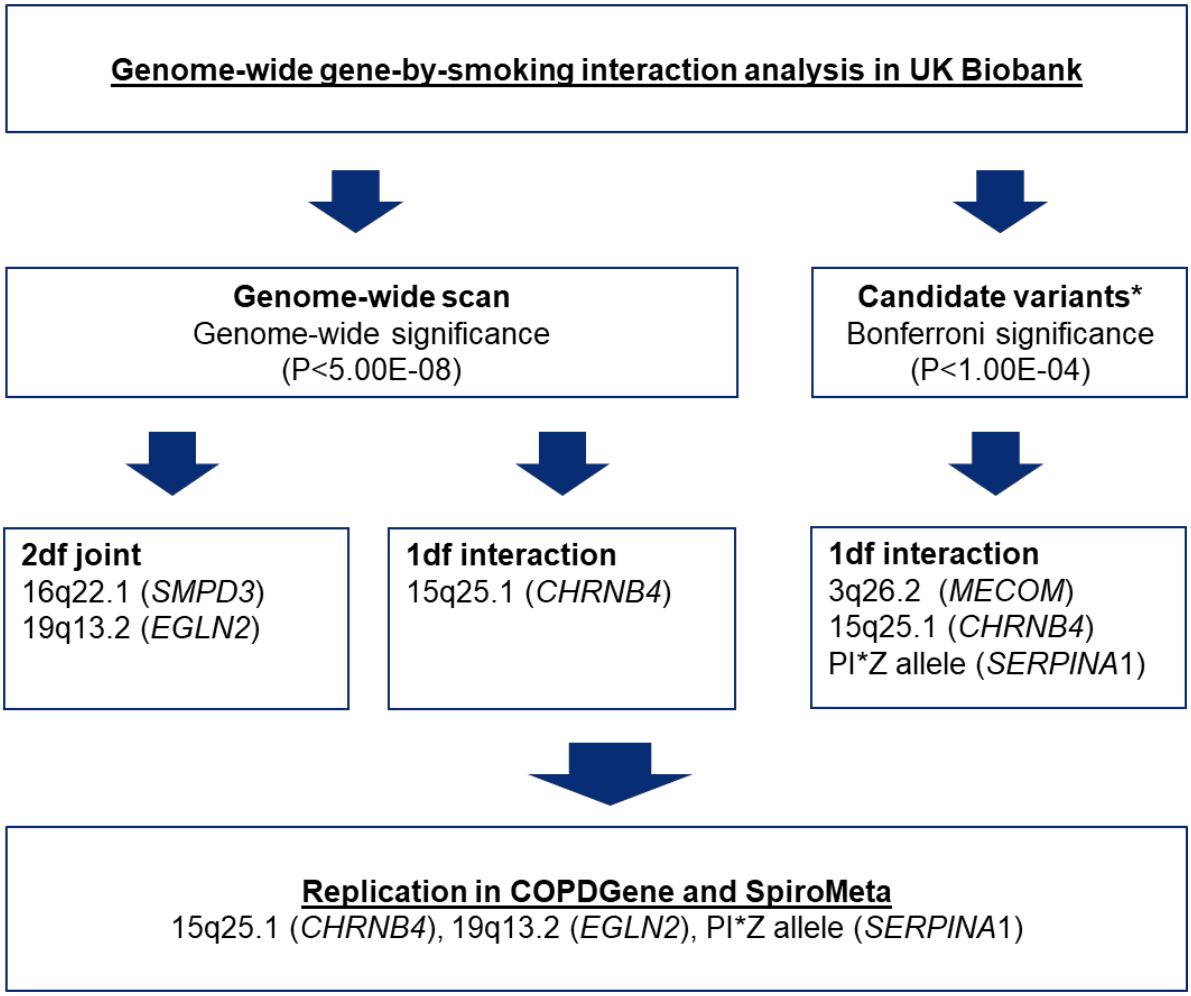
Table 3. Replications of selected SNPs in COPDGene and SpiroMeta

						COPDGene NHW		SpiroMeta (FEV1)					
						Marginal Association		Never Smoker		Ever Smoker		Interaction	2DF Joint
rsID	**Chr:Position**	**Nearest Gene**	**Effect/Ref. Allele**	**EAF**	**Smoking Exposure**	**OR (95% CI)**	**P**	**OR (95% CI)**	**P**	**OR (95% CI)**	**P**	**P**	**P**
Main effects													
rs141322661	16:68398875	*SMPD3*	G/A	0.01	Ever Smoking	0.99 (0.69-1.41)	9.52E-01	0.01 (-0.05, 0.08)	6.45E-01	0.06 (0, 0.13)	6.03E-02	1.07E-01	1.54E-01
rs2604894	19:41292404	*EGLN2*	A/G	0.45	Current Smoking	0.9 (0.82-0.98)	1.62E-02	0.001 (-0.02, 0.02)	9.12E-01	0.003 (-0.01, 0.02)	6.91E-01	7.66E-01	9.18E-01
Evidence with interaction effects													
rs7642001	3:168746145	*MECOM*	A/G	0.37	Current Smoking	1.1 (1.01-1.2)	3.63E-02	-0.04 (-0.05, -0.02)	3.31E-05	-0.01 (-0.03, 0.01)	1.93E-01	1.07E-03	7.79E-05
rs28929474	14:94844947	*SERPINA1*	T/C	0.02	Ever Smoking	1.34 (1-1.81)	5.08E-02	0.04 (-0.02, 0.09)	2.04E-01	0.02 (-0.04, 0.08)	4.65E-01	5.53E-01	3.41E-01
rs12440014	15:78926726	*CHRNB4*	G/C	0.24	Ever Smoking	0.76 (0.69-0.85)	3.41E-07	0.01 (-0.01, 0.03)	4.95E-01	0.03 (0.01, 0.05)	1.46E-03	7.33E-03	5.00E-03

EAF: Effect Allele Frequency; Interaction: Interaction test with 1 degree of freedom; 2DF Joint: Joint test with 2 degrees of freedom of genetic main and interaction effects; Marginal association between each selected SNP and COPD was tested in COPDGene; Lookup of selected SNPs

in GWAS of spirometric measures of lung function stratified by never- and ever-smoker groups in SpiroMeta; EAFs from COPDGene and SpiroMeta were similar.

Figure 1. Analysis workflow



*Candidate variants selected from reported interactions and GWAS of COPD and lung function

Figure 2. Regional plots of 16q22.1 and 19q13.2 regions based on 2df joint test

Figure 2a. 16q22.1 region

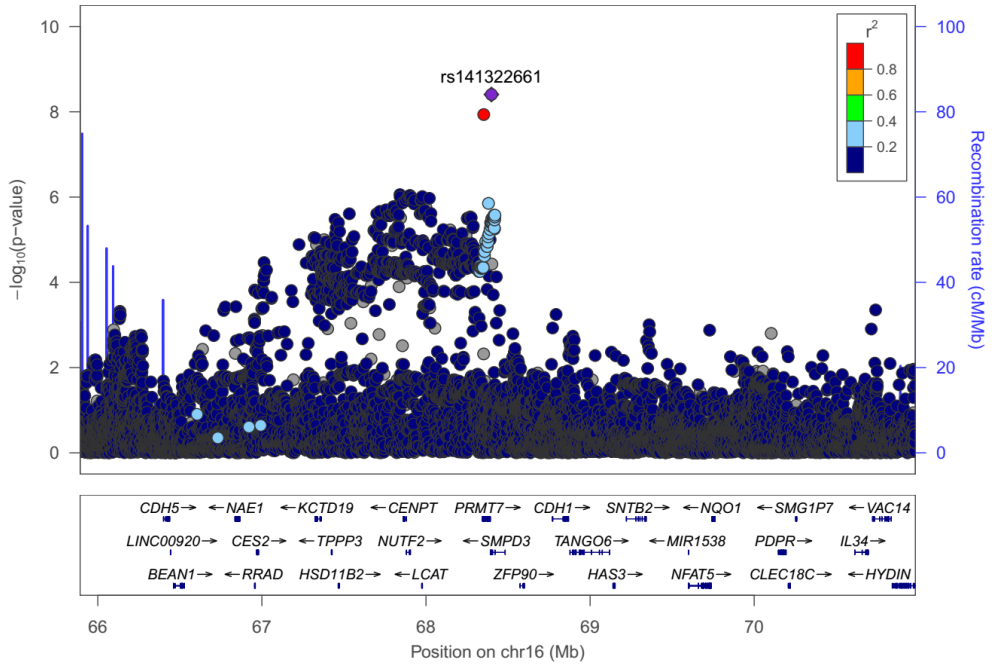


Figure 2b. 19q13.2 region

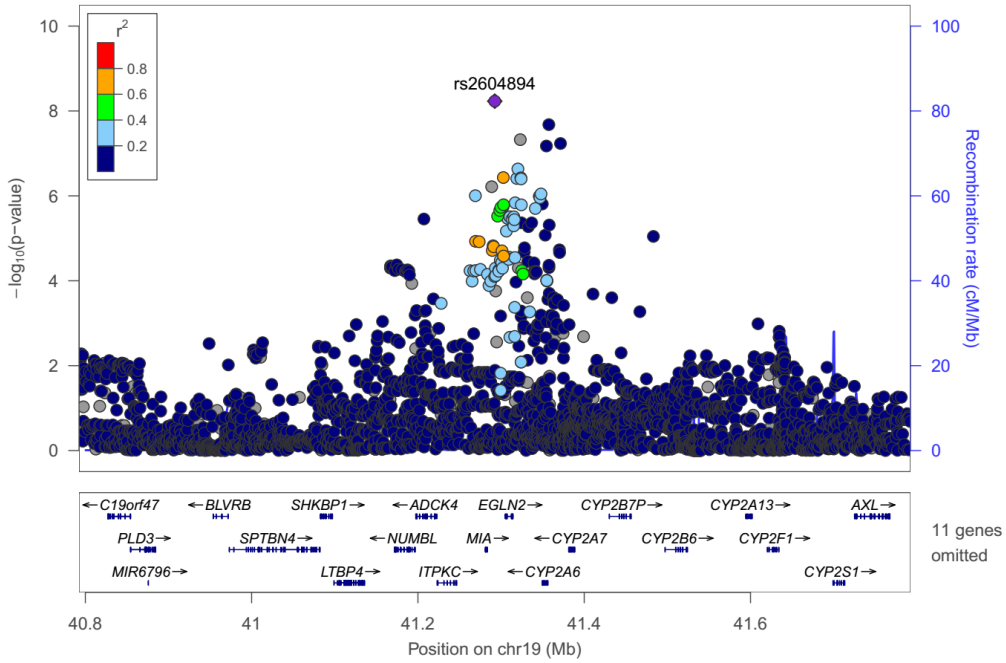


Figure 3. Manhattan plot and regional plot of 15q25 region based on 1df interaction test

Figure 3a. Manhattan plot of 15q25 region

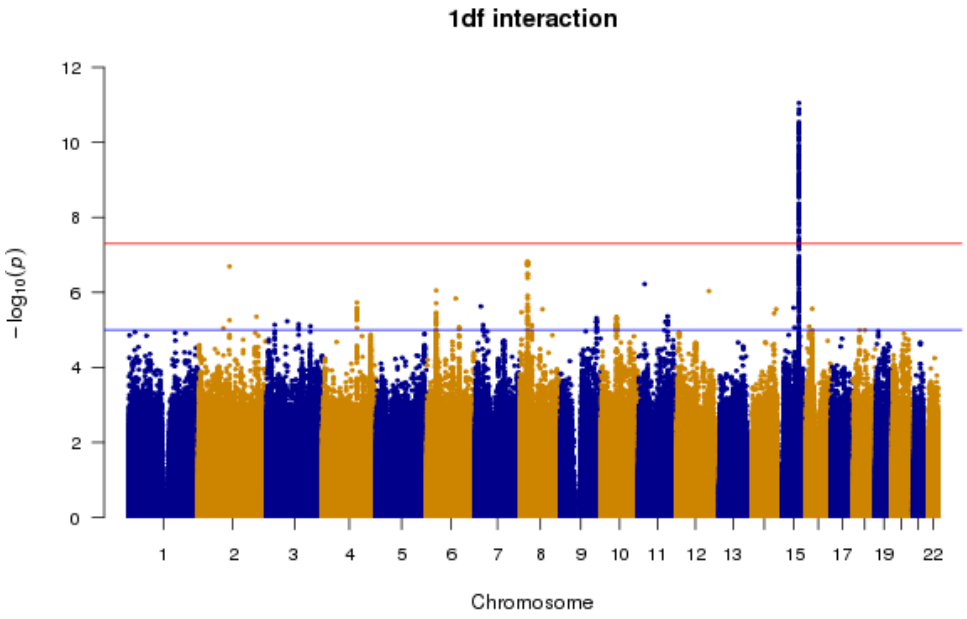


Figure 3b. Regional plot of 15q25 region

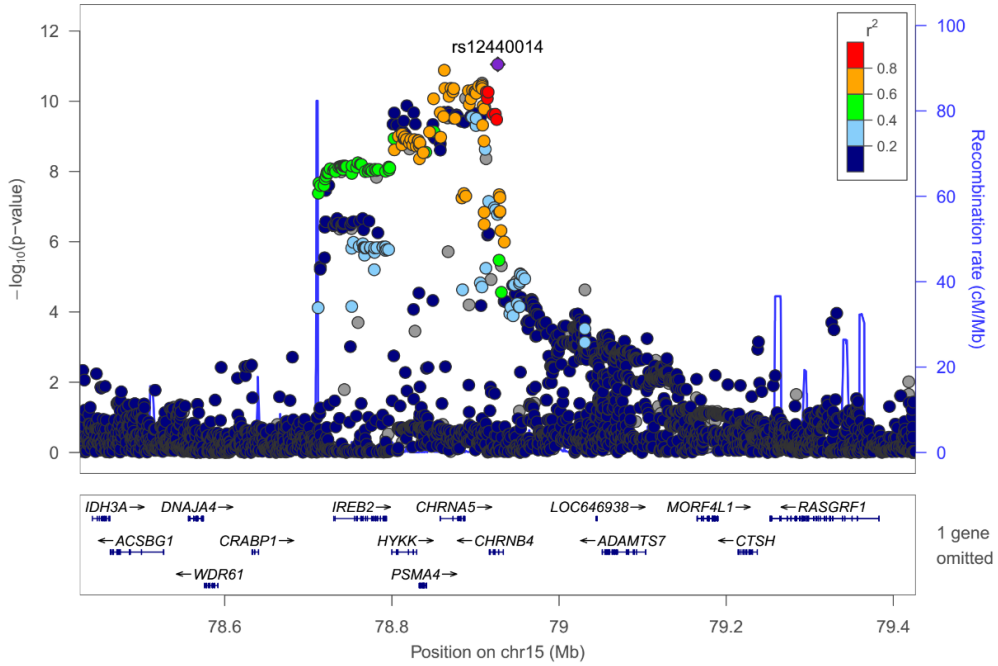


Figure 4. Statistical significances of selected SNPs