Evidence for Large-Scale Gene-by-Smoking Interaction Effects on Pulmonary Function

Hugues Aschard[*,1,2], Martin D. Tobin[3,4], Dana B. Hancock[5], David Skurnik[6], Akshay Sood[7], Alan James[8,9], Albert Vernon Smith[10,11], Ani W. Manichaikul[12,13], Archie Campbell[14,15], Bram P. Prins[16], Caroline Hayward[17], Daan W. Loth[18], David J. Porteous[14,15], David P. Strachan[19], Eleftheria Zeggini[16], George O'Connor[20,21], Guy G. Brusselle[18,22,23], H. Marike Boezen[24,25], Holger Schulz[26,27], Ian J. Deary[28,29], Ian P. Hall[30], Igor Rudan[31], Jaakko Kaprio[32,33,34], James F Wilson[35,36], Jemma Wilk[37], Jennifer E. Huffman[17], Jing Hua Zhao[38,39], Kim de Jong[24,25], Leo-Pekka Lyytikäinen[40,41], Louise V. Wain[3,4], Marjo-Riitta Jarvelin[42,43,44,45], Mika Kähönen[46], Myriam Fornage[47], Ozren Polasek[31,48], Patricia A. Cassano[49,50], R. Graham Barr[51], Rajesh Rawal[52,53,54], Sarah E. Harris[17,28], Sina A. Gharib[55], Stefan Enroth[56], Susan R. Heckbert[57], Terho Lehtimäki[40,41], Ulf Gyllensten[56], Understanding Society Scientific Group, Victoria E. Jackson[3], Vilmundur Gudnason[10,11], Wenbo Tang[49,58], Josée Dupuis[20,59], María Soler Artigas[3], Amit D. Joshi[1,2,60], Stephanie J. London[ǂ,5,61], Peter Kraft[ǂ,1,2]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[2]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[3]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester, United Kingdom.

[4]National Institute for Health Research, Leicester Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester, UK

[5]Behavioral and Urban Health Program, Behavioral Health and Criminal Justice Research Division, Research Triangle Institute (RTI) International, Research Triangle Park, NC, USA.

[6]Division of Infectious Diseases, Brigham and Women Hospital, Harvard Medical School, Boston, MA, USA.

[7]Division of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM, USA.

[8]Department of Pulmonary Physiology and Sleep Medicine, Sir Charles Gairdner Hospital, Nedlands, Australia.

[9]School of Medicine and Pharmacology, University of Western Australia, Crawley, Australia.

[10]Icelandic Heart Association, Kopavogur, Iceland.

[11]Faculty of Medicine, University of Iceland, Reykjavik, Iceland.

[12]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA.

[13]Department of Public Health Sciences, Division of Biostatistics and Epidemiology, University of Virginia, Charlottesville, VA, USA.

[14]Centre for Genomic & Experimental Medicine, Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom.

[15]Generation Scotland, Centre for Genomic and Experimental Medicine, University of Edinburgh, United Kingdom.

[16]Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom.

[17]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom.

[18]Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands.

[19]Population Health Research Institute, St George's University of London, London, United Kingdom.

[20]The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, USA.

[21]The Pulmonary Center, Department of Medicine, Boston University School of Medicine, Boston, MA, USA.

[22]Department of Respiratory Medicine, Ghent University Hospital, Ghent, Belgium.

[23]Department of Respiratory Medicine, Erasmus Medical Center, Rotterdam, the Netherlands.

[24]University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, the Netherlands.

[25]University of Groningen, University Medical Center Groningen, Groningen Research Institute for Asthma and COPD (GRIAC), Groningen, the Netherlands.

[26]Institute of Epidemiology I, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany.

[27]Comprehensive Pneumology Center Munich (CPC-M), Member of the German Center for Lung Research, Munich, Germany.

[28]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, United Kingdom.

[29]Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom.

[30]Division of Respiratory Medicine, University of Nottingham, Queen's Medical Centre, Nottingham, United Kingdom.

[31]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, United Kingdom.

[32]Department of Public Health, University of Helsinki, Helsinki, Finland.

[33]Institute for Molecular Medicine, University of Helsinki, Helsinki, Finland.

[34]National Institute for Health and Welfare, Department of Health, Helsinki, Finland.

[35]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, United Kingdom.

[36]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom.

[37]Human Genetics & Computational Biomedicine, Pfizer Worldwide Research and Development, Cambridge, MA, USA.

[38]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge United Kingdom.

[39]Institute of Metabolic Science, Biomedical Campus, Cambridge, United Kingdom.

[40]Department of Clinical Chemistry, Fimlab Laboratories, Tampere, Finland.

[41]Department of Clinical Chemistry, University of Tampere School of Medicine, Tampere, Finland.

[42]Department of Epidemiology and Biostatistics, MRC–PHE Centre for Environment & Health, School of Public Health, Imperial College London, United Kingdom.

[43] Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland.

[44]Biocenter Oulu, University of Oulu, Oulu, Finland.

[45]Unit of Primary Care, Oulu University Hospital, Oulu, Finland.

[46]Department of Clinical Physiology, University of Tampere and Tampere University Hospital, Tampere, Finland.

[47]Brown Foundation Institute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, TX, USA.

[48]Faculty of Medicine, University of Split, Split, Croatia.

[49]Division of Nutritional Sciences, Cornell University, Ithaca, NY, USA.

[50]Department of Healthcare Policy and Research, Weill Cornell Medical College, NY, NY, USA.

[51]Departments of Medicine and Epidemiology, Columbia University Medical Center.

[52]Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany.

[53]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany.

[54]Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany.

[55]Computational Medicine Core at Center for Lung Biology, Division of Pulmonary & Critical Care Medicine, University of Washington, Seattle, WA.

[56]Department of Immunology, Genetics and Pathology, Uppsala Universitet, Science for Life Laboratory, Uppsala, Sweden.

[57]Cardiovascular Health Research Unit and Department of Epidemiology, University of Washington, Seattle, WA, USA.

[58]Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, USA.

[59]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.

[60]Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, USA. Human Services, Research Triangle Park, NC, USA.

[61]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, U.S. Department of Health and Human Services, Research Triangle Park, NC, USA.

[ǂ] These authors contributed equally

[*]**Corresponding author**:
Hugues Aschard
Harvard School of Public Health
Department of Epidemiology
Building 2, Room 205
665 Huntington Avenue
Boston, MA 02115
haschard@hsph.harvard.edu
Phone : 1-617-820-0648
Fax : 1-617-432-1722

Word count : 3,737

Abstract

Background: Smoking is the strongest environmental risk factor for reduced pulmonary function. The genetic component of various pulmonary traits has also been demonstrated, and at least 26 loci have been reproducibly associated with either FEV_1 (forced expiratory volume in 1 second) or FEV_1/FVC (FEV_1/forced vital capacity). Although the main effects of smoking and genetic loci are well established, the question of potential gene-by-smoking interaction effect remains unanswered. The aim of the present study was to assess, using a genetic risk score approach, whether the effect of these 26 loci on pulmonary function is influenced by smoking.

Methods: We evaluated the interaction between smoking exposure, considered as either ever vs. never or pack-years, and a 26 SNPs genetic risk score in relation to FEV_1 or FEV_1/FVC in 50 047 participants of European ancestry from the CHARGE and SpiroMeta consortia.

Results: We identified an interaction ($\beta_{int} = -0.036$, 95% confidence interval, -0.040 – -0.032, *P*=0.00057) between an unweighted 26 SNPs genetic risk score and smoking status (ever/never) on the FEV_1/FVC ratio. In interpreting this interaction, we showed that the genetic risk of falling below the FEV_1/FVC threshold used to diagnose chronic obstructive pulmonary disease is higher among ever smokers than among never smokers.

Conclusions: This study highlights the benefit of using genetic risk scores for identifying interactions missed when studying individual SNPs, and shows for the first time that persons with the highest genetic risk for low FEV_1/FVC may be more susceptible to the deleterious effects of smoking.

Key words

FEV_1/FVC, smoking, gene-environment interaction, genetic risk score

Key messages

- Spirometric measures of pulmonary function are influenced by both smoking and genetics. This paper identified a genetic risk score-by-ever smoking interaction on FEV_1/FVC (forced expiratory volume in 1 second / forced vital capacity).
- In individuals of European ancestry, the reduction in FEV_1/FVC due to smoking was greater among individuals who are genetically predisposed to lower FEV_1/FVC ratio.
- Genetic risk score-by-ever smoking interaction can allow the identification of subgroups in the population whose genetic background makes them more susceptible to the deleterious effects of smoking.

Introduction

Spirometric measures of pulmonary function, such as the forced expiratory volume in one second (FEV_1) or its ratio with the forced vital capacity (FEV_1/FVC), form the basis of the diagnosis of chronic obstructive pulmonary disease (COPD).[1-3] Pulmonary function measures are also used clinically to monitor severity and control of asthma and other respiratory diseases, and are independent risk factors for mortality.[1-3] Pulmonary function is strongly influenced by cigarette smoking and by multiple low-penetrance genetic variants. Indeed, genome-wide association studies (GWAS) of marginal genetic effects (i.e. not including interaction effects between genetic variants and smoking) have identified at least 26 loci associated with FEV_1 or FEV_1/FVC in the general population.[4] However, the interplay between genetic factors and environmental exposures has not been well established for pulmonary function or its associated traits. More broadly, while considerable efforts have been made to identify interaction effects between genetic variants and environmental exposures across the wide range of human traits and diseases,[5, 6] such investigations have been mostly unsuccessful in detecting robust gene-environment interactions.[5, 7] The well-established effect of cigarette smoking on numerous human health outcomes[8] makes it a serious candidate for identification of novel gene-environment interactions, especially for pulmonary traits.

Hypothesizing the presence of single nucleotide polymorphism (SNP)-by-smoking interaction, Hancock et al.[9] performed a genome-wide interaction study of pulmonary function, modeling single SNP main effects and their interactions with smoking in 50 047 participants of European ancestry across 19 studies within the Cohorts for Heart and Aging Research in

Genomic Epidemiology (CHARGE)[10] and SpiroMeta consortia[11] –the largest genome-wide interaction study of pulmonary function as modified by smoking to date. However, rather than focusing on the interaction effects *per se*, they performed a meta-analysis of the joint test of SNP main effects and SNP-by-smoking interaction effects, in order to improve power for identifying genetic variants associated with pulmonary function.[12, 13] While they reported new candidate variants based on this joint test, the study did not identify any SNPs with genome-wide significant interaction with smoking.

Here, we explored gene-by-smoking interaction effects limited to genetic variants previously found to be associated with pulmonary function in standard marginal effects GWAS,[4] therefore not including the new variants reported by Hancock et al.[9] based on the joint test of main effects plus interaction. Specifically, we aimed to determine whether smoking modifies the effect of established genetic variants, when considered singly or combined using a genetic risk score summarizing the genetic predisposition to abnormal pulmonary function. The primary motivation for using genetic risk score is statistical power[14, 15]. Indeed, several genetic risk score-by-exposure interactions have already been identified in cases where single SNPs did not show evidence for statistically significant interactions[16-21]. Genetic risk score-by-exposure interaction testing expands on the principle of omnibus test while leveraging the assumption that, for a given choice of coded alleles, most interaction effects will have the same direction. This is similar to burden tests that have been widely used for rare variant analysis[22] where a single parameter can accumulate evidence for association without increasing the number of degrees of freedom. When interaction effects are null on average (i.e. if interaction effects are both negative and positive so that the sum of interaction coefficients tend to zero), the single

SNP approach will generally outperform the risk score-based approach. Conversely, if interaction effects tend to be in the same direction, the risk score-based approach can have dramatically higher power.[14]

Methods

Single SNP-by-smoking interaction

The present analysis relies on the Hancock et al.[9] genome-wide screening for main genetic effects plus interaction effects with smoking in relation to pulmonary function among 50 047 participants (56% women) of European ancestry from 19 studies. The mean age was 53 years at the time of pulmonary function testing. Approximately 15% were current-smokers and 56% were ever smokers. Among ever smokers, the average pack-years of smoking was 21. **Supplementary Table 3** (available as Supplementary data at IJE online) provides the main characteristics of the studies included, while complete details of study-specific pulmonary function testing protocols can be found in previous work[4]. For studies with spirometry at a single visit, we analyzed FEV_1/FVC and FEV_1 measured at that visit. For studies with spirometry at more than one visit, we analyzed measurements from the baseline visit or the most recent examination with spirometry data. Smoking history (current, former, and never smoking) was ascertained by questionnaire at the time of pulmonary function testing. Pack-years of smoking was calculated for current and past smokers by multiplying smoking amount (packs/day) and duration (years smoked). Approximately 2.5 million autosomal SNPs were tested for interaction with smoking status (ever smoking vs never smoking) and pack-years, for two outcomes, $\mathbf{FEV_1}$

and FEV_1/FVC, using the following model (see **Supplementary Note**, available as Supplementary data at IJE online):

$$Y \sim \beta_0 + \beta_G G + \beta_{GE_k} G E_k + \sum_{l=1\ldots3} \beta_{E_l} E_l \qquad \text{(Equation 1)}$$

where β_G and β_{E_l} are the main effect of the SNP G and exposure E_l, β_{GE_k} is the interaction effect between G and exposure E_k, and β_0 the intercept.

Multivariate interaction analysis overview

First, we considered an unweighted genetic risk score-by-smoking interaction where the risk score simply sums the number of risk alleles (i.e. alleles associated with a lower pulmonary function). This unweighted genetic risk score is most powerful when the interaction effects have the same direction as marginal SNP effects, i.e. the harmful effects of smoking are magnified in individuals with a genetic predisposition to reduced pulmonary function. Second, we used a weighted genetic risk score where SNPs were weighted by the absolute value of their marginal effect estimates obtained from stage 1 screening of FEV_1 and FEV_1/FVC from Soler Artigas et al.[4] (**Supplementary Table 1** available as Supplementary data at IJE online). This weighting scheme is most powerful when the magnitude of interaction effects is proportional to the SNP marginal effects. Finally, for our third multivariate analysis, we derived a standard omnibus test of all interaction effects. This test will retain power in the presence of effects in both directions or of different magnitudes. Although there is strong correlation among the 12 tests performed (these three models, considering interaction with two different smoking

metrics, ever/never smoking or pack-years, for the two pulmonary function metrics FEV1 and FEV1/FVC), we used a stringent Bonferroni *p*-value correction threshold of 4×10^{-3} to account for multiple testing.

When raw data are available, the weighted genetic risk score (GRS) is usually expressed as $GRS = \sum_m [w_i \times G_i]$ where m is the number of SNPs included in the genetic risk score and $w = (w_1, .. w_m)$ are the weights attributed to each single SNP. Following previous notation, the test of interaction between the genetic risk score and the exposure E_k can be applied using the following model:

$$Y \sim \gamma_0 + \gamma_{GRS} \times GRS + \gamma_{INT} \times GRS \times E_k + \sum_{l=1...3} \gamma_{E_l} \times E_l \qquad \text{(Equation 2)}$$

where γ_0, γ_{GRS}, γ_{E_l} and γ_{INT} are the intercept, the main effect of the genetic risk score, the main effect of the exposure E_l and the interaction effect between E_k and the genetic risk score, respectively. However, as raw data were not directly available, we performed the test of γ_{INT} from summary statistics of interaction effects using an inverse-variance weighted sum as proposed by Aschard.[14] The chi-square for the interaction term γ_{INT} was derived as follows:

$$\chi^2_{int} = \frac{\left(\sum_{i=1...m} \frac{w_i \times \hat{\beta}_{G_i E_k}}{\hat{\sigma}^2_{\beta_{G_i E_k}}} \right)^2}{\sum_{i=1...m} \frac{w_i^2}{\hat{\sigma}^2_{\beta_{G_i E_k}}}}$$

where $\hat{\beta}_{G_i \times E_k}$ and $\hat{\sigma}^2_{\beta_{G_i \times E_k}}$ are the estimated effects and variance of the interaction between the exposure E_k and the SNP G_i obtained from *Equation 1*; and w_i is the weight applied to SNP

G_i . Under the null hypothesis of no interaction effect, χ^2_{int} follows a chi-squared distribution with one degree of freedom.

The standard omnibus test of all interaction effects consisted in evaluating jointly $\boldsymbol{\alpha_{G \times E_k}} = (\alpha_{G_1 \times E_k}, \ldots, \alpha_{G_m \times E_k})$ from the model:

$$Y \sim \alpha_0 + \sum_{i=1\ldots m} \left[\alpha_{G_i} \times G_i\right] + \sum_{i=1\ldots m} \left[\alpha_{G_i \times E_k} \times G_i \times E_k\right] + \sum_{l=1\ldots 3} \alpha_{E_l} \times E_l$$

where α_0, α_{G_i}, α_{E_l} and $\alpha_{G_i \times E_k}$ are the intercept, the main effects of SNP G_i and the exposure E_l, and the interaction effect between G_i and E_k. Leveraging the independence between the SNPs considered (a single SNP was selected for each independent locus), we also derived the omnibus test using summary statistics. Under this independence assumption, the $G_i \times E_k$ interaction terms would also be independents,[14] so that it can be performed by summing the chi-square from each univariate interaction test to form a chi-square with m degree of freedom as follows:

$$\chi^2_{joint} = \sum_{i=1\ldots m} \frac{\hat{\beta}^2_{G_i \times E_k}}{\hat{\sigma}^2_{\beta_{G_i \times E_k}}}$$

where $\hat{\beta}_{G_i \times E_k}$ and $\hat{\sigma}^2_{\beta_{G_i \times E_k}}$ are the estimated effects and variance of the interaction between the exposure E_k and the SNP G_i obtained from *Equation 1*.

Relative risk in ever smokers versus never smokers

Genetic risk score interaction effects can further be translated in terms of risk prediction. For pulmonary function, low FEV_1 or FEV_1/FVC increases the risk of death[23] and together form the basis for the diagnosis of COPD.[1-3] COPD stage 2 or higher are defined by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) as having FEV_1/FVC < 0.70 and FEV_1 < 80% of the predicted value. According to recent studies,[2, 24] between 5% and 20% of European ancestry adults are expected to have FEV_1/FVC <0.70, depending on smoking characteristics and age distribution. Moreover, several studies argue for a more stringent threshold to define COPD[24, 25] based on lower limit of normal predicted value, rather than a fixed absolute value, to prevent disease misclassification.

To explore the impact of interaction effect on the risk of disease, we derived the relative risk (RR) of having FEV_1/FVC below a given threshold (1%, 5% and 20%) in ever smokers versus never smokers conditional on the unweighted genetic risk score. This quantity is defined as the joint probability of having both FEV_1/FVC in the interval $[-\infty, \mathrm{FEV}_1/\mathrm{FVC}_{up}]$ and the genetic risk score (GRS) in the interval $\left[GRS_{low}, GRS_{up}\right]$. This can be expressed as the following integral:

$$\int_{-\infty}^{\mathrm{FEV}_1/\mathrm{FVC}_{up}} \int_{GRS_{low}}^{GRS_{up}} f_1(y|g,e) \times f_2(g|e)\, dy\, dg$$

where y, e and g are FEV_1/FVC, smoking status and the genetic risk score, respectively, and f_1 and f_2 are the probability density function of y and g. The detailed derivation of the above integral is available as Supplementary data at IJE online.

Results

We selected 26 loci previously found to be associated with FEV_1 and/or FEV_1/FVC at genome-wide significance ($P < 5\times10^{-8}$) in marginal association tests[4, 11, 26] (i.e. not including interaction effects with smoking exposures), and replicated in the GWAS by Soler Artigas et al.,[4] the largest meta-analysis of marginal genetic effect conducted for these two traits in the general population. Additional loci for these two phenotypes have been identified in two recent studies.[27, 28] However, these new loci were not included in our analysis because both these studies used a large cohort ascertained through smoking status. For each of the 26 selected loci, we choose the SNP with the strongest evidence for association (i.e. smallest *p*-value) with each of these phenotypes. The final list included 26 SNPs per phenotype, with only two SNPs being different between FEV_1 and FEV_1/FVC as previously reported[4] (**Supplementary Table 1** available as Supplementary data at IJE online). Estimated interaction effects of these SNPs were extracted from the meta-analysis summary statistics for the four tests performed in the Hancock et al.[9] analysis: SNP-by-smoking status (ever smoking vs never smoking) interaction effect on FEV_1 and FEV_1/FVC; and SNP-by-smoking pack-years interaction effect on FEV_1/FVC and FEV_1. As showed in **Supplementary Table 2** (available as Supplementary data at IJE online), nine SNPs showed nominal significance ($P < 0.05$) out of the 104 tests performed ; however, none remained significant after accounting for multiple testing (Bonferroni corrected p-value threshold of 5×10^{-4}). The minimum *p*-value was observed for the interaction between rs993925, near the TGFβ2 gene, and smoking status on FEV_1 ($\beta_{int} = -0.036$, 95% confidence interval (CI), -0.009 – -0.032, P=0.007).

Next, using these data, we conducted three multivariate (as opposed to single SNP) interaction analyses, testing jointly for the interaction effects between those SNPs and either smoking status or pack-years on the two phenotypes (FEV_1 and FEV_1/FVC) for a total of 12 tests. As shown in **Table 1**, none of the multivariate interaction tests with pack-years were significant. However, four of the six multivariate interaction tests with smoking status (ever versus never) showed nominal significance, and two tests for FEV_1/FVC had a *p*-value below the Bonferroni significance level (12 tests, $P<4\times10^{-3}$). . The strongest signal was observed for the unweighted genetic risk score by smoking status interaction effect on FEV_1/FVC ($\beta_{int} = -0.036$, 95% CI, -0.040 – -0.032, *P*=0.00057). The Cochran's Q test for heterogeneity of the interaction effect across studies was not significant (*P*=0.97), and the forest plot of study-specific results did not display any obvious outlier (**Supplementary Figure 1,** available as Supplementary data at IJE online).

The contrast between this significant risk score interaction and the absence of strong single SNP interaction effects can be explained by looking at the distribution of the single SNP interaction effect estimates. **Figure 1** shows this distribution for the alleles associated with decreased FEV_1/FVC. It highlights that while the 95% CI of most single SNP interaction effects encompass the null (and therefore the absence of significant single SNP interaction effect), there is an enrichment for negative interaction effects. Indeed, even a binomial test can be used to confirm the unbalanced direction of interaction effects (18 out of 26 interactions are negative leading to a *p*-value of 0.014 for a binomial test with an expected equiprobable distribution of 0.5). The genetic risk score-based interaction test exploits such enrichment by testing for the average interaction effect across all SNPs.[14] As with any multivariate approach

based on a composite null hypothesis, this result indicates that at least a subset of these 26 SNPs interact with smoking status, but does not allow us to determine which SNP(s) or how many SNPs are driving the genetic risk score-by smoking interaction. The three other sets of single-SNP interaction tests showed a similar (but not significant after correction for multiple testing) trend with enrichment for negative interactions (**Supplementary Figs. 2-4** available as Supplementary data at IJE online). We summarized the contribution of the unweighted genetic risk score-by-smoking interaction on FEV_1/FVC in **Table 2** and **Figure 2A**. This indicates that the deleterious effect of smoking is enhanced among carriers of the risk alleles or equivalently that the deleterious effect of smoking is reduced among subjects carrying the protective alleles.

We used two datasets of 8859 unrelated individuals and 9457 related individuals, respectively to test for independent replication of our results (**Supplementary Note** available as Supplementary data at IJE online). Both replication samples showed consistent negative GRS-by-ever smoking interaction effect on FEV1/FVC ($\hat{\beta}_{int} = -0.0025$, 95% CI -0.0165, 0.0115, *P*=0.72 and $\hat{\beta}_{int} = -0.0030$, 95% CI -0.0214 0.0154, *P*=0.74, and overall interaction effect in the combined replication datasets $\hat{\beta}_{int} = -0.0027$, 95% CI -0.0136, 0.0082 *P*=0.63), and a Cochran's Q test for heterogeneity showed no significant difference in the three effect estimates (*P*=0.51). However, despite having total N=18,316 individuals, our combined replication sample remains underpowered (<50% power at nominal significance of 5%) and analyses including more participants would be necessary to strength the replication evidence for interaction effects.

To quantify the impact of this result from a public health perspective, we estimated the impact of the genetic risk score-by-smoking interaction on having FEV_1/FVC below 1%, 5% and 20% in the lower tails of the distribution in the population. Specifically, we derived the relative risk (RR) of having FEV_1/FVC below these cutoff points (1%, 5% and 20%) in ever smokers compared to never smokers. **Figure 2B** quantifies the excess RR (i.e. the RR minus one) of individuals across five genetic risk score quintiles. It highlights the higher risk associated with smoking among individuals carrying risk alleles (i.e. alleles associated with poorer pulmonary function) as compared to individuals carrying protective alleles (i.e. alleles associated with better pulmonary function). For example, among individuals with a genetic risk score above the 80^{th} percentile, smokers have on average a 26% excess RR of having FEV_1/FVC in the lowest 1% of the population distribution, whereas ever smokers with a genetic risk score below the 20^{th} percentile have on average a 18% excess RR of falling in that same FEV_1/FVC category compared to never smokers. Applying the same approach for FEV_1, we observed a similar pattern (**Supplementary Figs. 5** available as Supplementary data at IJE online). However, as expected, the lower magnitude of the genetic risk score-by-ever smoking interaction on FEV_1 implied a lower difference in RR between ever smokers and never smokers.

Discussion

Using the largest dataset to date of European ancestry participants from the general population with pulmonary function (FEV_1/FVC and FEV_1), smoking, and genetic data, we identified a gene-by-smoking interaction effect on FEV_1/FVC by using a genetic risk score

composed of 26 SNPs identified and replicated in a prior GWAS meta-analysis of marginal genetic effects. Replication study showed interaction effect estimates in the same direction as the discovery study, though analyses including more participants would be necessary to strength the replication evidence for interaction effects. To our knowledge, our study is the first to report a synergistic action of genes and smoking on pulmonary function, i.e. the reduction in FEV_1/FVC due to smoking is greater among individuals who are genetically predisposed to lower FEV_1/FVC ratio. Our study also highlights the importance of developing and applying alternative strategies to evaluate interaction effects for lung phenotypes along with other complex traits and diseases. The genetic risk score-based approach enabled us to identify an interaction when the standard univariate test (i.e. evaluating each single genetic variant for interaction independently) failed to identify any interactions.

Genetic risk score-by-exposure interaction can have higher clinical value than the identification of single SNP-by-exposure interaction as it can capture a wealth of information in a single measure to identify subgroups in the population whose genetic background makes them more susceptible to the deleterious effects of smoking.[19, 29, 30] Indeed, if single SNP-by-smoking interactions are distributed unconditionally on the marginal genetic effect (i.e. interaction effects have equal chances to be positive or negative given the coded alleles are the risk alleles), the genetic effect will be similar between ever and never smokers on average. The enrichment for negative interactions we identified through our genetic risk score approach reveals a stronger genetic component among the ever smoker subgroup in the population, and can allow the implementation of more efficient implementation of prevention strategies. For example, in the public health setting, programs targeting smoking cessation campaigns to

individuals who are genetically predisposed to low pulmonary function may have a stronger impact in preventing COPD.

Our results may also elucidate biological mechanisms underlying the interplay between genes and smoking in pulmonary function. In particular, the higher statistical power for the genetic risk score-based interaction test points towards the potential presence of an unmeasured intermediate biomarker mediating the effect of the 26 loci on FEV_1/FVC. As shown in **Figure 3**, the most parsimonious model (i.e. the less complex following Occam's razor) that would explain multiple interactions going in the same direction (**Figure 1**) implies the genetic variants together influence an intermediate biomarker which itself interacts with smoking. Future studies with extended genomic data, including transcriptomic, proteomic, and/or metabolomic data, might be able to further assess such an hypothesis by evaluating i) the effect of the genetic risk score on those biomarkers, and ii) testing for interactions between smoking and the candidate biomarkers identified at step i).

This study has some limitations. The 26 selected variants together explain a relatively small proportion of the additive genetic variance in FEV_1/FVC and in FEV_1.[4] However, GWAS with increasing sample sizes will likely continue to provide additional associated genetic variants to further assess the role of SNP-by-smoking interaction effects on pulmonary phenotypes, and may increase the gap between smokers and never smokers to allow for a significant impact in the clinic or at the population level. Moreover, we focused on genetic variants previously found to be associated at genome-wide significance level, but future studies might consider less stringent criteria to select genetic variants, including those with only

suggestive evidence, or alternatively candidate variants with functional annotation relevant to the outcomes and exposures in question. Obviously, the signal to noise ratio might decrease when relaxing the constraint on the SNP selection. However, as we recently showed, additional gain in statistical power might be achieved even if a substantial proportion of the variants do not interact with the exposure.[14] Finally, investigation of interaction effects with other environmental exposures such as secondhand smoke, air pollution, asbestos, or occupational risks may lead to a more comprehensive understanding of the biological and epidemiological significance of these variants.

In summary, the identification of interaction effects between genetic variants and environmental exposures in human traits is recognized as extremely challenging, and this quest has been mostly unsuccessful so far. In this study, we discovered novel gene-by-smoking interactions using risk scores, that were not observed at the level of individual genetic variants. This risk score analysis suggests that persons with a greater genetic predisposition to low pulmonary function are more susceptible to the deleterious effects of smoking. By extension, the use of a genetic risk score may help predict which smokers will fall below thresholds that establish the diagnosis of COPD.

Author Contributions

H.A., P.K., S.J.L., M.T., D.B.H. and A.J. were involved in designing the study. M.D.T., D.B.H., A.S., A.J., A.V.S., A.W.M., D.W.L., D.P.S., G.O.C., R.G.B., G.G.B., I.P.H., J.K.P., J.F.W., J.W., J.H.Z., K.d.J., L.V.W., M.S.A., H.M.B., M-R.J., M.F., P.A.C., S.A.G., S.R.H., V.G., W.T. S.J.L., I.R., O.P., J.E.H.,, C.H., A.C., D.J.P., S.E.H., I.J.D., S.E., U.G., LP.L., T.L., E.Z., B.P.P., and V.E.J. were involved in participant recruitment, sample collection or genotyping. H.A. performed analyses from the discovery study. H.A., V.E.J. and M.S.A performed the replication analysis. H.A. drafted the paper, with substantial editorial input from P.K., S.J.L., J.D., D.S., M.T., D.B.H. and A.J. All authors have reviewed and approved the final draft. This material has not been published previously in a substantively similar form.

Conflict of interest

Jaakko Kaprio consulted for Pfizer on nicotine dependence. Other authors declare no competing financial interest.

Acknowledgments

We thank the many colleagues who contributed to collection and phenotypic characterization of the clinical sampling and genotyping of the data. We especially thank those who kindly agreed to participate in the studies. The research undertaken by M.D.T., L.V.W. was partly funded by

the National Institute for Health Research (NIHR). The views expressed are those of theauthor(s) and not necessarily those of the NHS, the NIHR or the Department of Health. A full list of studies and authors source of funding source and acknowledgments is provided in the Supplementary Data.

References

1. Rabe KF, Hurd S, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American journal of respiratory and critical care medicine* 2007 Sep 15;**176**(6): 532-55.
2. Lange P, Celli B, Agusti A, et al. Lung-Function Trajectories Leading to Chronic Obstructive Pulmonary Disease. *The New England journal of medicine* 2015 Jul 9;**373**(2): 111-22.
3. Vestbo J, Hurd SS, Agusti AG, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American journal of respiratory and critical care medicine* 2013 Feb 15;**187**(4): 347-65.
4. Soler Artigas M, Loth DW, Wain LV, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nature genetics* 2011 Nov;**43**(11): 1082-90.
5. Aschard H, Lutz S, Maus B, et al. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. *Human genetics* 2012 Oct;**131**(10): 1591-613.
6. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. *American journal of epidemiology* 2009 Jan 15;**169**(2): 227-30; discussion 34-5.
7. Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM, Tank NCIG-ET. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. *Genetic epidemiology* 2013 Nov;**37**(7): 643-57.
8. *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General*. Atlanta (GA); 2010.
9. Hancock DB, Artigas MS, Gharib SA, et al. Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLoS genetics* 2012;**8**(12): e1003098.
10. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circulation Cardiovascular genetics* 2009 Feb;**2**(1): 73-80.
11. Repapi E, Sayers I, Wain LV, et al. Genome-wide association study identifies five loci associated with lung function. *Nature genetics* 2010 Jan;**42**(1): 36-44.
12. Aschard H, Hancock DB, London SJ, Kraft P. Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. *Hum Hered* 2011;**70**(4): 292-300.
13. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered* 2007;**63**(2): 111-9.
14. Aschard H. A perspective on interaction effects in genetic association studies. *Genetic epidemiology* in press.

15. Marigorta UM, Gibson G. A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. *Frontiers in Genetics* 2014;**5**.
16. Pollin TI, Isakova T, Jablonski KA, et al. Genetic modulation of lipid profiles following lifestyle modification or metformin treatment: the Diabetes Prevention Program. *PLoS genetics* 2012;**8**(8): e1002895.
17. Qi L, Cornelis MC, Zhang C, van Dam RM, Hu FB. Genetic predisposition, Western dietary pattern, and the risk of type 2 diabetes in men. *The American journal of clinical nutrition* 2009 May;**89**(5): 1453-8.
18. Ahmad S, Rukh G, Varga TV, et al. Gene x physical activity interactions in obesity: combined analysis of 111,421 individuals of European ancestry. *PLoS genetics* 2013;**9**(7): e1003607.
19. Qi Q, Chu AY, Kang JH, et al. Sugar-sweetened beverages and genetic risk of obesity. *The New England journal of medicine* 2012 Oct 11;**367**(15): 1387-96.
20. Fu Z, Shrubsole MJ, Li G, et al. Interaction of cigarette smoking and carcinogen-metabolizing polymorphisms in the risk of colorectal polyps. *Carcinogenesis* 2013 Apr;**34**(4): 779-86.
21. Langenberg C, Sharp SJ, Franks PW, et al. Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study. *PLoS medicine* 2014 May;**11**(5): e1001647.
22. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *American journal of human genetics* 2014 Jul 3;**95**(1): 5-23.
23. Viegi G, Pistelli F, Sherrill DL, Maio S, Baldacci S, Carrozzi L. Definition, epidemiology and natural history of COPD. *The European respiratory journal* 2007 Nov;**30**(5): 993-1013.
24. Roche N, Dalmay F, Perez T, et al. FEV1/FVC and FEV1 for the assessment of chronic airflow obstruction in prevalence studies: do prediction equations need revision? *Respiratory medicine* 2008 Nov;**102**(11): 1568-74.
25. Swanney MP, Ruppel G, Enright PL, et al. Using the lower limit of normal for the FEV1/FVC ratio reduces the misclassification of airway obstruction. *Thorax* 2008 Dec;**63**(12): 1046-51.
26. Hancock DB, Eijgelsheim M, Wilk JB, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nature genetics* 2010 Jan;**42**(1): 45-52.
27. Artigas MS, Wain LV, Miller S, et al. Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nature communications* 2015;**6**: 8658.
28. 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 Oct;**3**(10): 769-81.
29. Aschard H, Zaitlen N, Lindstrom S, Kraft P. Variation in predictive ability of common genetic variants by established strata: the example of breast cancer and age. *Epidemiology* 2015 Jan;**26**(1): 51-8.
30. Aschard H, Chen J, Cornelis MC, Chibnik LB, Karlson EW, Kraft P. Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *American journal of human genetics* 2012 Jun 8;**90**(6): 962-72.

Figure Legends

Figure 1. Distribution of interaction effects on FEV_1/FVC.

Single SNP risk allele-by-smoking status (ever/never) interaction effect estimates (β_{int}) and 95% confidence intervals are plotted by increasing values. Negative and positive interactions are in dark blue and light blue, respectively. The unweighted genetic risk score-by-smoking status interaction is plotted in purple.

Figure 2. Overview of the unweighted genetic risk score-by-smoking interaction effect on FEV_1/FVC.

Upper panel (A) presents the distribution of the unweighted genetic risk score (GRS, grey density plot) and the relationship between the unweighted GRS and standardized FEV_1/FVC in ever smokers (red line) and never smokers (black line). Lower panel (B) shows the excess relative risk (RR) of having FEV_1/FVC in the lowest 1%, 5% and 20% of the population for ever smokers compared to never smokers, as stratified by GRS quintiles.

Figure 3. Underlying causal model.

Potential causal diagrams underlying the gene and smoking interaction effects on FEV_1/FVC. Panel (A) presents a scenario where each genetic variant influences the outcome through a SNP-specific pathway, and interactions with the environmental exposure take place along these

pathways. Panel B) presents an alternative (and simpler) model where multiple genetic variants influence an unmeasured intermediate biomarker, which effect on FEV_1/FVC depends on smoking. In scenario (A), the single SNP-by-smoking interaction test is the optimal approach, while in scenario (B), the single SNP-by-smoking interaction test can become inefficient and interaction would be easier to detect using a genetic risk score-by-smoking interaction test, as it summarizes all interaction effects in a single test.

Table 1. Multivariate interaction tests of the 26 loci associated with pulmonary function.

Outcome	Exposure	Test	$\widehat{\beta}_{int}$ *(CI)*	p-value
FEV_1	Smoking status*	*uGRS*	-0.0055 *(-0.011, 2.7x10^{-5})*	0.051
		wGRS	-0.21 *(-0.40,-0.033)*	**0.020**
		CHISQ	- -	0.49
FEV_1	Pack-years	*uGRS*	-1.6x10^{-5} *(-4.6x10^{-5}, 1.4x10^{-5})*	0.30
		wGRS	-6.5x10^{-4} *(-1.6x10^{-3},3.3x10^{-4})*	0.19
		CHISQ	- -	0.46
FEV_1/FVC	Smoking status	*uGRS*	-0.0099 *(-0.016, -0.0043)*	**0.00057**†
		wGRS	-0.21 *(-0.33,-0.073)*	**0.0022**†
		CHISQ	- -	**0.026**
FEV_1/FVC	Pack-years	*uGRS*	-4.4e-06 *(-3.6x10^{-5}, 2.7x10^{-5})*	0.78
		wGRS	-6.5x10^{-5} *(-8.0x10^{-4},6.6x10^{-4})*	0.85
		CHISQ	- -	0.53

uGRS is the genetic risk score using equal weights to all SNPs; wGRS is the genetic risk score weighted by effect estimates from the marginal screening; CHISQ is the omnibus test of all interaction effects; $\hat{\beta}_{int}$ *is the estimated interaction effect between the GRS and the outcome; and CI is the confidence interval of that estimate.*

Nominally significant tests are indicated in bold.

*Smoking status is defined as never smokers versus ever smokers.

†*Significant* p-*value after Bonferroni correction.*

Table 2. Summary of effect estimates for genetic risk score-by-smoking status interaction on FEV_1/FVC.

Predictors	beta	SD	*p*-value
From the marginal exposure model			
Pack-years	-0.0030	0.00017	1.2×10^{-71}
Current smoking	-0.040	0.0047	7.7×10^{-18}
Smoking status*	-0.0023	0.0046	0.61
From the interaction model			
GRS	-0.0363	0.0021	3.9×10^{-64}
GRS x Smoking status*	-0.0099	0.0029	5.7×10^{-4}

GRS is the unweighted genetic risk score ; beta is the effect estimates of each predictor; and SD the standard deviation of the each beta.

**Smoking status was defined as never smokers versus ever smokers*