nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-13855-2 OPEN

# Immune-mediated genetic pathways resulting in pulmonary function impairment increase lung cancer susceptibility

Linda Kachuri et al.[#]

Impaired lung function is often caused by cigarette smoking, making it challenging to disentangle its role in lung cancer susceptibility. Investigation of the shared genetic basis of these phenotypes in the UK Biobank and International Lung Cancer Consortium (29,266 cases, 56,450 controls) shows that lung cancer is genetically correlated with reduced forced expiratory volume in one second ($FEV_1$: $r_g = 0.098$, $p = 2.3 \times 10^{-8}$) and the ratio of $FEV_1$ to forced vital capacity ($FEV_1$/FVC: $r_g = 0.137$, $p = 2.0 \times 10^{-12}$). Mendelian randomization analyses demonstrate that reduced $FEV_1$ increases squamous cell carcinoma risk (odds ratio (OR) = 1.51, 95% confidence intervals: 1.21–1.88), while reduced $FEV_1$/FVC increases the risk of adenocarcinoma (OR = 1.17, 1.01–1.35) and lung cancer in never smokers (OR = 1.56, 1.05–2.30). These findings support a causal role of pulmonary impairment in lung cancer etiology. Integrative analyses reveal that pulmonary function instruments, including 73 novel variants, influence lung tissue gene expression and implicate immune-related pathways in mediating the observed effects on lung carcinogenesis.

[#] A full list of authors and their affiliations appears at the end of the paper.

Lung cancer is the most commonly diagnosed cancer worldwide and the leading cause of cancer mortality[1]. Although tobacco smoking remains the predominant risk factor for lung cancer, clinical observations and epidemiological studies have consistently shown that individuals with airflow limitation, particularly those with chronic obstructive pulmonary disease (COPD), have a significantly higher risk of developing lung cancer[2–7]. Several lines of evidence suggest that biological processes resulting in pulmonary impairment warrant consideration as independent lung cancer risk factors, including observations that previous lung diseases influence lung cancer risk independently of tobacco use[6,8–10], and overlap in genetic susceptibility loci for lung cancer and chronic obstructive pulmonary disease (COPD) on 4q24 (*FAM13A*), 4q31 (*HHIP*), 5q.32 (*HTR4*), the 6p21 region, and 15q25 (*CHRNA3*/*CHRNA5*)[11–14]. Inflammation and oxidative stress have been proposed as key mechanisms promoting lung carcinogenesis in individuals affected by COPD or other non-neoplastic lung pathologies[9,11,15].

Despite an accumulation of observational findings, previous epidemiological studies have been unable to conclusively establish a causal link between indicators of impaired pulmonary function and lung cancer risk due to the interrelated nature of these conditions[7]. Lung cancer and obstructive pulmonary disease share multiple etiological factors, such as cigarette smoking, occupational inhalation hazards, and air pollution, and 50–70% of lung cancer patients present with co-existing COPD or airflow obstruction[6]. Furthermore, reverse causality remains a concern since pulmonary symptoms may be early manifestations of lung cancer or acquired lung diseases in patients whose immune system has already been compromised by undiagnosed cancer.

Disentangling the role of pulmonary impairment in lung cancer development is important from an etiological perspective, for refining disease susceptibility mechanisms, and for informing precision prevention and risk stratification strategies. In this study we comprehensively assess the shared genetic basis of impaired lung function and lung cancer risk by conducting genome-wide association analyses in the UK Biobank cohort to identify genetic determinants of three pulmonary phenotypes, forced expiratory volume in 1s ($FEV_1$), forced vital capacity (FVC), and $FEV_1$/FVC. We examine the genetic correlation between pulmonary function phenotypes and lung cancer, followed by Mendelian randomization (MR) using novel genetic instruments to formally test the causal relevance of impaired pulmonary function, using the largest available dataset of 29,266 lung cancer cases and 56,450 controls from the OncoArray lung cancer collaboration[16].

## Results

**Heritability and genetic correlation**. Array-based, or narrow-sense, heritability ($h_g$) estimates for all lung phenotypes were obtained using LD score regression[17] based on summary statistics from our GWAS of the UKB cohort ($n = 372{,}750$ for $FEV_1$, $n = 370{,}638$ for FVC, $n = 368{,}817$ for $FEV_1$/FVC; Supplementary Fig. 1) are presented in Table 1. Heritability estimates based on UKB-specific LD scores ($n = 7{,}567{,}036$ variants) were consistently lower but more precise than those based on the 1000 Genomes (1000G) Phase 3 reference population ($n = 1{,}095{,}408$ variants). For $FEV_1$, $h_g = 0.163$ (SE = 0.006) and $h_g = 0.201$ (SE = 0.008), based on UKB and 1000 G LD scores, respectively. Estimates for FVC were $h_g = 0.175$ (SE = 0.007) and $h_g = 0.214$ (SE = 0.010). Heritability was lower for $FEV_1$/FVC: $h_g = 0.128$ (SE = 0.006) and 0.157 (SE = 0.010), based on internal and 1000 G reference panels, respectively. For all phenotypes, $h_g$ did not differ by smoking status and estimates were not affected by excluding the major histocompatibility complex (MHC) region.

Partitioning heritability by functional annotation identified large and statistically significant ($p < 8.5 \times 10^{-4}$) enrichments for multiple categories (Fig. 1; Supplementary Tables 1–3). A total of 35 categories, corresponding to 22 distinct annotations, were significantly enriched for all three pulmonary phenotypes, including annotations that were not previously reported[18]. Large enrichment, defined as the proportion of heritability accounted for by a specific category relative to the proportion of SNPs in that category, was observed for elements conserved in primates[19,20] (17.6% of SNPs, 54.7–58.5% of $h_g$), McVicker background selection statistic[21,22] (17.8% of SNPs, 22.6–25.1% of $h_g$), flanking bivalent transcription starting sites (TSS)/enhancers from Roadmap[20,23] (1.4% of SNPs, 11.1–13.2% of $h_g$), and super enhancers (16.7% of SNPs, 33.9–38.6% of $h_g$). We also replicated previously reported significant enrichments for histone methylation and acetylation marks H3K4me1, H3K9Ac, and H3K27Ac[18,24].

Substantial genetic correlation was observed for pulmonary phenotypes with body composition and smoking traits, mirroring phenotypic correlations in epidemiologic studies (Fig. 2). Large positive correlations with height were observed for $FEV_1$ ($r_g = 0.568$, $p = 2.5 \times 10^{-567}$) and FVC ($r_g = 0.652$, $p = 1.8 \times 10^{-864}$). Higher adiposity was negatively correlated with $FEV_1$ (BMI: $r_g = -0.216$, $p = 4.2 \times 10^{-74}$; percent body fat: $r_g = -0.221$, $p = 1.7 \times 10^{-66}$), FVC (BMI: $r_g = -0.262$, $p = 1.6 \times 10^{-114}$; percent body fat: $r_g = -0.254$, $p = 1.2 \times 10^{-88}$). Smoking status (ever vs. never) was significantly correlated with all lung function phenotypes ($FEV_1$ $r_g = -0.221$, $p = 8.1 \times 10^{-78}$; FVC $r_g = -0.091$, $1.0 \times 10^{-16}$; $FEV_1$/FVC $r_g = -0.360$, $p = 7.5 \times 10^{-130}$). Cigarette pack-years and impaired lung function in smokers were also significantly genetically correlated with $FEV_1$ ($r_g = -0.287$ $p = 1.1 \times 10^{-35}$), FVC ($r_g = -0.253$, $p = 1.9 \times 10^{-30}$), and $FEV_1$/FVC ($r_g = -0.108$, $p = 3.0 \times 10^{-4}$). As a positive control, we verified that $FEV_1$ and FVC were genetically correlated with each other ($r_g = 0.922$) and with $FEV_1$/FVC ($FEV_1$: $r_g = 0.232$, $p = 4.1 \times 10^{-32}$; FVC: $r_g = -0.167$, $p = 1.0 \times 10^{-19}$).

Genetic correlations between lung function phenotypes and lung cancer are presented in Fig. 3. For simplicity of interpretation coefficients were rescaled to represent genetic correlation with impaired (decreasing) lung function. Impaired $FEV_1$ was positively correlated with lung cancer overall ($r_g = 0.098$, $p = 2.3 \times 10^{-8}$), squamous cell carcinoma ($r_g = 0.137$, $p = 7.6 \times 10^{-9}$), and lung cancer in smokers ($r_g = 0.140$, $p = 1.2 \times 10^{-7}$). Genetic correlations were attenuated for adenocarcinoma histology ($r_g = 0.041$, $p = 0.044$) and null for never smokers ($r_g = -0.002$, $p = 0.96$). A similar pattern of associations was observed for FVC. Reduced $FEV_1$/FVC was positively correlated with all lung cancer subgroups (overall: $r_g = 0.137$, $p = 2.0 \times 10^{-12}$; squamous carcinoma: $r_g = 0.137$, $p = 4.3 \times 10^{-8}$; adenocarcinoma: $r_g = 0.125$, $p = 7.2 \times 10^{-9}$; smokers: $r_g = 0.185$, $p = 1.4 \times 10^{-10}$), except for never smokers ($r_g = 0.031$, $p = 0.51$).

Exploring the functional underpinnings of these genetic correlations revealed three functional categories that were significantly enriched for lung cancer (Supplementary Table 4), and have not been previously reported[25]. All these categories were also significantly enriched for pulmonary traits. CpG dinucleotide content[22] included only 1% of SNPs, but had a strong enrichment signal for lung cancer ($p = 2.1 \times 10^{-7}$), $FEV_1$ ($p = 7.7 \times 10^{-24}$), FVC ($p = 2.3 \times 10^{-23}$, and $FEV_1$/FVC ($p = 3.8 \times 10^{-17}$). Other shared features included background selection (lung cancer: $p = 1.0 \times 10^{-6}$, $FEV_1$: $p = 1.9 \times 10^{-20}$, FVC: $p = 6.9 \times 10^{-23}$, $FEV_1$/FVC: $p = 1.5 \times 10^{-15}$) and super enhancers (lung cancer: $p = 4.4 \times 10^{-6}$, $FEV_1$: $p = 3.4 \times 10^{-24}$, FVC: $p = 5.1 \times 10^{-20}$, $FEV_1$/FVC: $p = 9.6 \times 10^{-22}$).

**Table 1 Array-based heritability for $FEV_1$, FVC, and $FEV_1$/FVC.**

| | $FEV_1$ | | FVC | | $FEV_1$/FVC | |
|---|---|---|---|---|---|---|
| UKB LD scores | $h_g$ | (SE) | $h_g$ | (SE) | $h_g$ | (SE) |
| Overall | 0.163 | (0.006) | 0.175 | (0.007) | 0.128 | (0.006) |
| Never smokers | 0.163 | (0.007) | 0.169 | (0.007) | 0.126 | (0.008) |
| Smokers | 0.159 | (0.007) | 0.172 | (0.009) | 0.129 | (0.008) |
| Overall no MHC | 0.162 | (0.006) | 0.175 | (0.007) | 0.125 | (0.006) |
| 1000G LD scores | | | | | | |
| Overall | 0.201 | (0.008) | 0.214 | (0.010) | 0.157 | (0.010) |
| Never smokers | 0.209 | (0.010) | 0.215 | (0.011) | 0.159 | (0.011) |
| Smokers | 0.208 | (0.010) | 0.221 | (0.011) | 0.166 | (0.010) |

Estimates were obtained using LD score regression applied to genome-wide summary statistics from the UK Biobank (UKB). Two types of LD scores were used: LD scores estimated using UK Biobank (internal reference population) and pre-computed LD scores based on the 1000 Genomes Phase 3 reference population

**Genome-wide association analysis for instrument development**. Based on the results of our GWAS in the UK Biobank, we identified 207 independent instruments for $FEV_1$ ($P < 5 \times 10^{-8}$, replication $P < 0.05$; LD $r^2 < 0.05$ within 10,000 kb), 162 for FVC, and 297 for $FEV_1$/FVC. We confirmed that our findings were not affected by spirometry performance quality, with a nearly perfect correlation between effect sizes ($R^2 = 0.995$, $p = 2.5 \times 10^{-196}$) in the main discovery analysis and after excluding individuals with potential blow acceptability issues (Field 3061 ≠ 0; $n = 60{,}299$). After applying these variants to the lung cancer OncoArray dataset and selecting LD proxies ($r^2 > 0.90$) for unavailable variants, the final set of instruments consisted of 193 variants for $FEV_1$, 144 for FVC, and 264 SNPs for $FEV_1$/FVC (Supplementary Data 1–3), for a total of 601 instruments. The proportion of trait variation accounted for by each set of instruments was estimated in the UKB replication sample consisting of over 110,00 individuals (Supplementary Fig. 1), and corresponded to 3.13% for $FEV_1$, 2.27% for FVC, and 5.83% for $FEV_1$/FVC. We also developed instruments specifically for never smokers based on a separate GWAS of this population, which yielded 76 instruments for $FEV_1$, 112 for $FEV_1$/FVC, and 57 for FVC, accounting for 2.06%, 4.21%, and 1.36% of phenotype variation, respectively (Supplementary Data 4–6).

After removing overlapping instruments between pulmonary phenotypes and LD-filtering ($r^2 < 0.05$) across the three traits, 447 of the 601 variants were associated with at least one of $FEV_1$, FVC, or $FEV_1$/FVC ($P < 5 \times 10^{-8}$, replication $P < 0.05$). We compared these 447 independent variants to the 279 lung function variants recently reported by Shrine et al.[18] based on an analysis of the UK Biobank and SpiroMeta consortium, by performing clumping with respect to these index variants (LD $r^2 < 0.05$ within 10,000 kb). Our set of instruments included an additional 73 independent variants, 69 outside the MHC region (Supplementary Table 5), that achieved replication at the Bonferroni-corrected threshold for each trait (maximum replication $P = 2.0 \times 10^{-4}$).

Our instruments included additional independent signals in known lung function loci and variants in genes newly linked to lung function, such as *HORMAD2* at 22q12.1 (rs6006399: $P_{FEV1} = 1.9 \times 10^{-18}$), which is involved in synapsis surveillance in meiotic prophase, and *RIPOR1* at 16q22.1 (rs7196853: $P_{FEV1/FVC} = 1.3 \times 10^{-16}$), which plays a role in cell polarity and directional migration. Several new variants further support the importance of the tumor growth factor beta (TGF-β) signaling pathway, including *CRIM1* (rs1179500: $P_{FEV1/FVC} = 3.6 \times 10^{-17}$) and *FGF18* (rs11745375: $P_{FEV1/FVC} = 1.6 \times 10^{-11}$). Another novel gene, *PIEZO1* (rs750739: $P_{FEV1} = 1.8 \times 10^{-10}$), encodes a mechano-sensory ion channel, supports adaptation to stretch of the lung epithelium and endothelium, and promotes repair after alveolar injury[26,27]. In never smokers a signal was identified at 6q15 in *BACH2* (rs58453446: $P_{FEV1/FVC\text{-}nvsmk} = 8.9 \times 10^{-10}$), a gene required for pulmonary surfactant homeostasis. Last, two lung function variants mapped to genes somatically mutated in lung cancer: *EML4* (rs12466981: $P_{FEV1/FVC} = 2.7 \times 10^{-14}$) and *BRAF* (rs13227429: $P_{FVC} = 5.6 \times 10^{-9}$).

**Mendelian randomization**. The causal relevance of impaired pulmonary function was investigated by applying genetic instruments developed in the UK Biobank to the OncoArray lung cancer dataset, comprised of 29,266 lung cancer cases and 56,450 controls (Supplementary Table 6). Primary analyses were based on the maximum likelihood (ML) and inverse variance weighted (IVW) multiplicative random-effects estimators[28,29]. Sensitivity analyses were conducted using the weighted median (WM) and robust adjusted profile score (RAPS) estimators[30,31]. A genetically predicted decrease in $FEV_1$ was significantly associated with increased risk of lung cancer overall ($OR_{ML} = 1.28$, 95% CI: 1.12–1.47, $p = 3.4 \times 10^{-4}$) and squamous carcinoma ($OR_{ML} = 2.04$, 1.64–2.54, $p = 1.2 \times 10^{-10}$), but not adenocarcinoma ($OR_{ML} = 0.99$, 0.83–1.19, $p = 0.96$) (Fig. 4; Supplementary Table 7). The association with lung cancer was not significant across all estimators ($OR_{WM} = 1.06$, $p = 0.57$; $OR_{RAPS} = 1.13$, $p = 0.26$). There was no evidence of directional pleiotropy based on the MR Egger intercept test ($\beta_{0\ Egger} \neq 0$, $p < 0.05$), but significant heterogeneity among SNP-specific causal effect estimates was observed, which may be indicative of balanced horizontal pleiotropy (lung cancer: $P_Q = 2.1 \times 10^{-41}$; adenocarcinoma: $P_Q = 3.4 \times 10^{-9}$; squamous carcinoma: $P_Q = 1.1 \times 10^{-30}$). After excluding outlier variants contributing to this heterogeneity, 36 for lung cancer and 34 for squamous carcinoma, the association with $FEV_1$ diminished for both phenotypes (lung cancer: $OR_{ML} = OR_{IVW} = 1.12$, $p = 0.13$), but remained statistically significant for squamous carcinoma ($OR_{IVW} = 1.51$, 1.21–1.88, $p = 2.2 \times 10^{-4}$), with comparable effects observed using other estimators ($OR_{ML} = 1.50$, $p = 6.7 \times 10^{-4}$; ORRAPS = 1.48, $p = 1.7 \times 10^{-3}$; $OR_{WM} = 1.44$, $p = 0.040$).

Genetic predisposition to reduced FVC was inconsistently associated with squamous carcinoma risk ($OR_{ML} = 1.68$, $p = 1.8 \times 10^{-4}$; $OR_{WM} = 1.19$, $p = 0.38$). Effects became attenuated and more similar after removing outliers ($OR_{ML} = 1.27$, $p = 0.10$; $OR_{RAPS} = 1.25$, $p = 0.14$) (Fig. 4; Supplementary Table 8). A genetically predicted 10% decrease in $FEV_1$/FVC was associated with an elevated risk of lung cancer in some models ($OR_{ML} = 1.18$, 1.07–1.31, $p = 1.6 \times 10^{-3}$), but not others ($OR_{WM} = 1.10$, $p = 0.30$; $OR_{RAPS} = 1.11$, $p = 0.14$) (Fig. 4; Supplementary Table 9). The association with squamous carcinoma was also inconsistent across estimators. After removing outliers contributing to significant effect heterogeneity (lung cancer: $P_Q = 1.2 \times 10^{-28}$; adenocarcinoma: $P_Q = 3.4 \times 10^{-9}$; squamous carcinoma: $P_Q = 5.3 \times 10^{-15}$), the association

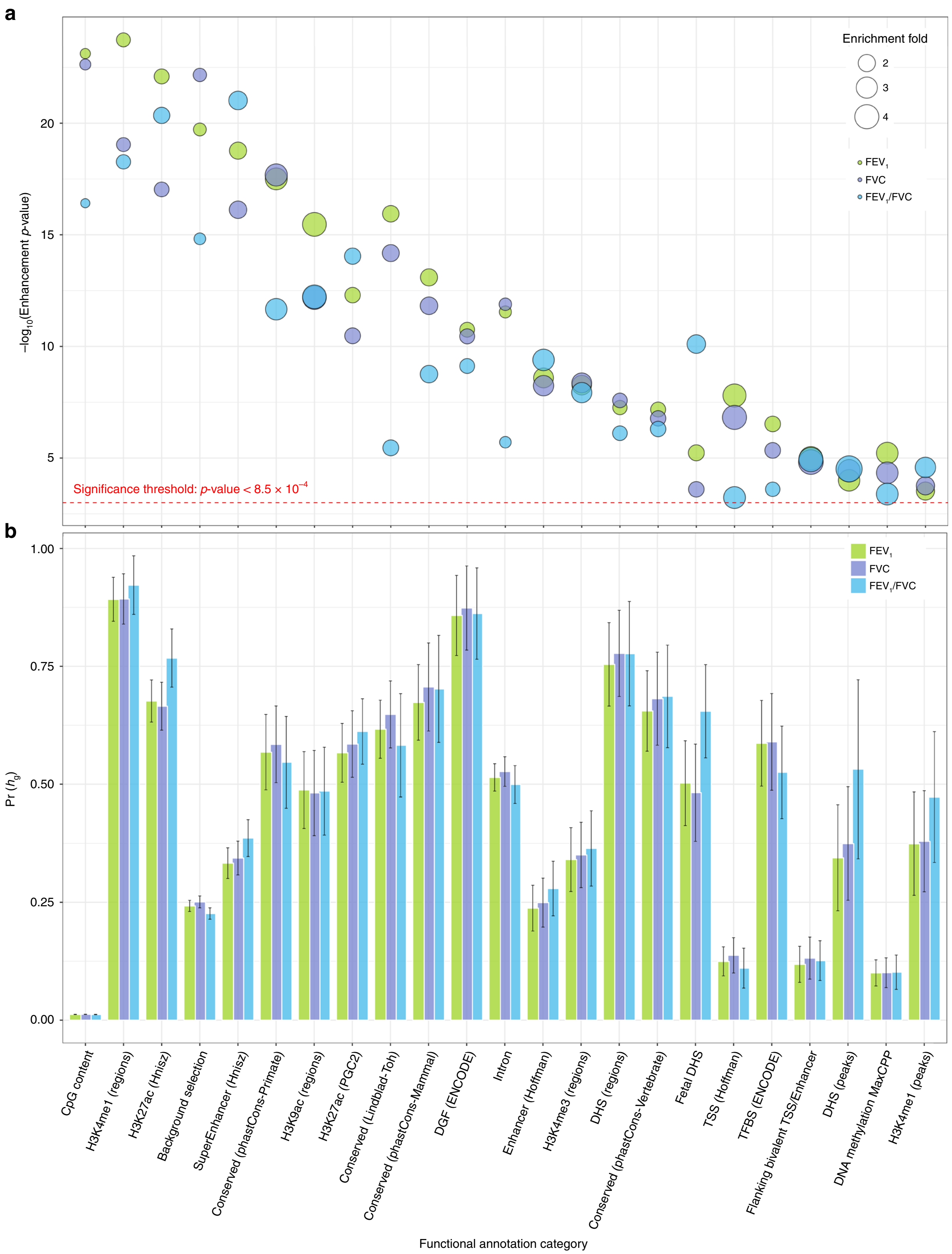
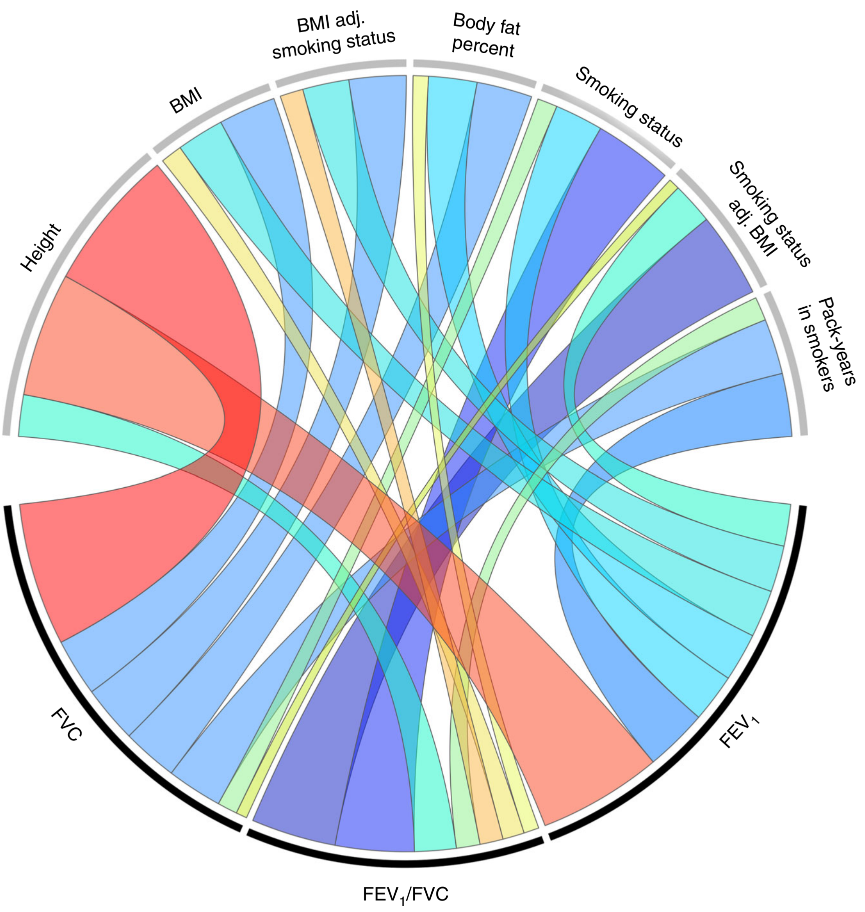


**Fig. 1 Functional partitioning of array-based heritability for each pulmonary function phenotype. a** The magnitude of category-specific enrichment and corresponding −log$_{10}$(*p*-value) for 22 distinct functional annotations that were significantly enriched for all three phenotypes (FEV$_1$, FVC, FEV$_1$/FVC). **b** The proportion of heritability, Pr($h_g$), accounted for by each functional annotation with corresponding standard errors. Functional annotation categories are not mutually exclusive.



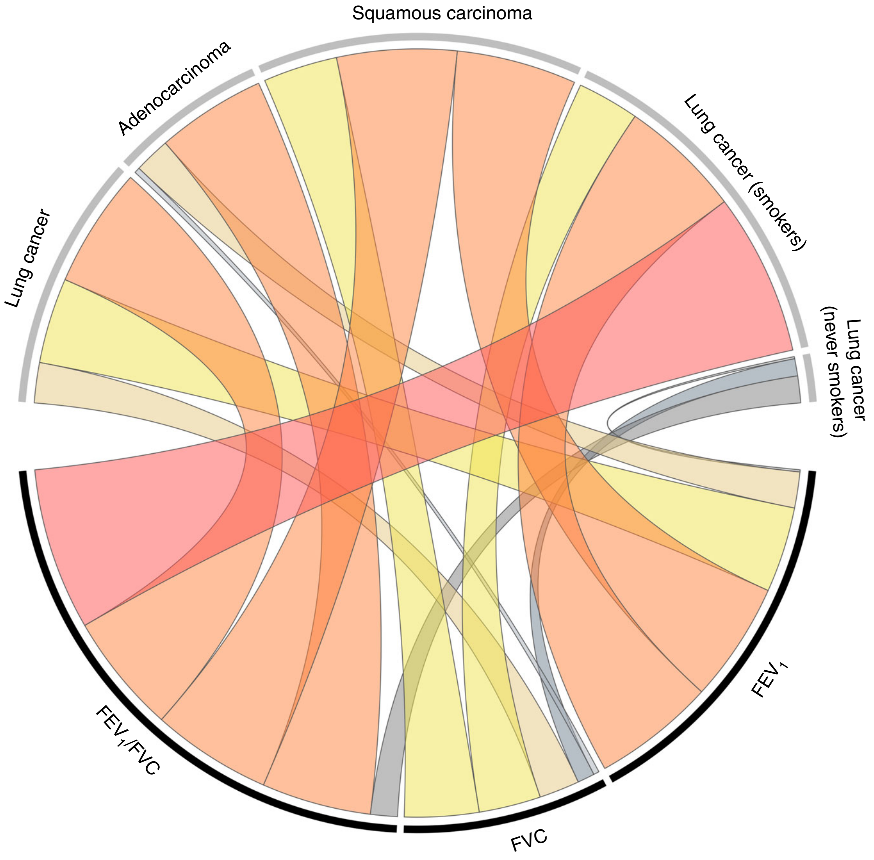
| | FVC | | $FEV_1$/FVC | | $FEV_1$ | |
|---|---|---|---|---|---|---|
| | $r_g$ | p-value | $r_g$ | p-value | $r_g$ | p-value |
| Height | 0.652 | $1.8 \times 10^{-864}$ | −0.190 | $1.9 \times 10^{-34}$ | 0.568 | $2.5 \times 10^{-567}$ |
| BMI | −0.262 | $1.6 \times 10^{-114}$ | 0.099 | $4.2 \times 10^{-14}$ | −0.216 | $4.2 \times 10^{-74}$ |
| BMI adjusted for smoking status | −0.261 | $3.3 \times 10^{-112}$ | 0.106 | $9.1 \times 10^{-16}$ | −0.212 | $1.7 \times 10^{-70}$ |
| Body fat percent | −0.254 | $1.2 \times 10^{-88}$ | 0.070 | $4.4 \times 10^{-8}$ | −0.221 | $1.7 \times 10^{-66}$ |
| Smoking status | −0.091 | $1.0 \times 10^{-16}$ | −0.360 | $7.5 \times 10^{-130}$ | −0.221 | $8.1 \times 10^{-78}$ |
| Smoking status adjsted for BMI | −0.051 | $1.7 \times 10^{-6}$ | −0.388 | $5.8 \times 10^{-145}$ | −0.191 | $7.0 \times 10^{-60}$ |
| Cigarette pack-years in smokers | −0.253 | $1.9 \times 10^{-30}$ | −0.108 | $3.0 \times 10^{-4}$ | −0.287 | $1.1 \times 10^{-35}$ |

**Fig. 2 Genetic correlation ($r_g$) between pulmonary function phenotypes and anthropometric and smoking phenotypes.** Estimates for $r_g$ are based on UK Biobank-specific LD scores. The colors in the Circos plot correspond to the direction of genetic correlation, with warm shades denoting positive relationships and cool tones depicting negative correlations. The width of each band in the Circos plot is proportional to the magnitude of the absolute value of the $r_g$ estimate.

with adenocarcinoma strengthened ($OR_{ML} = 1.17$, 1.01–1.35; $OR_{RAPS} = 1.18$, 1.02–1.38), while associations for lung cancer and squamous carcinoma became attenuated.

We examined the cancer risk in never smokers, by applying genetic instruments developed specifically in this population, to 2355 cases and 7504 controls (Fig. 5; Supplementary Table 10). A genetically predicted 1-SD decrease in $FEV_1$ and FVC was not associated with lung cancer risk in never smokers. However, a 10% reduction in $FEV_1$/FVC was associated with a 61% increased risk ($OR_{ML} = 1.61$, 1.10–2.35, $p = 0.014$; $OR_{IVW} = 1.60$, $p = 0.030$). Outlier filtering did not have an appreciable impact on the results ($OR_{ML} = 1.56$, 1.05–2.30, $p = 0.027$; $OR_{IVW} = 1.55$, 1.05–2.28, $p = 0.028$). A sensitivity analysis applied to 264 $FEV_1$/FVC instruments not specific to never smokers yielded an attenuated estimate ($OR_{IVW} = 1.35$, 1.03–1.75, $p = 0.027$), but confirmed the impact of $FEV_1$/FVC reduction on lung cancer risk.

For completeness, we also present MR estimates for the effect of impaired pulmonary function on lung cancer risk in smokers (Supplementary Table 11). Despite the larger sample size (23,223 cases and 16,964 controls) compared to never smokers, a



| | FEV$_1$/FVC | | FVC | | FEV$_1$ | |
|---|---|---|---|---|---|---|
| | $r_g$ | *p*-value | $r_g$ | *p*-value | $r_g$ | *p*-value |
| Lung cancer | 0.137 | $2.0 \times 10^{-12}$ | 0.046 | $5.3 \times 10^{-3}$ | 0.098 | $2.3 \times 10^{-8}$ |
| Adenocarcinoma | 0.125 | $7.2 \times 10^{-9}$ | −0.006 | 0.75 | 0.041 | 0.044 |
| Squamous cell carcinoma | 0.137 | $4.3 \times 10^{-8}$ | 0.085 | $7.9 \times 10^{-5}$ | 0.137 | $7.6 \times 10^{-9}$ |
| Lung cancer (smokers) | 0.185 | $1.4 \times 10^{-10}$ | 0.071 | $2.6 \times 10^{-3}$ | 0.140 | $1.2 \times 10^{-7}$ |
| Lung cancer (never smokers) | 0.031 | 0.51 | −0.020 | 0.65 | 0.002 | 0.96 |

**Fig. 3 Genetic correlation ($r_g$) between pulmonary function phenotypes and lung cancer subtypes.** Estimates of $r_g$ are based on genome-wide summary statistics from the UK Biobank cohort for pulmonary traits, and the International Lung Cancer Consortium OncoArray study for lung cancer. Genetic correlations have been rescaled to depict associations between impaired (reduced) pulmonary function and lung cancer risk. The colors in the Circos plot correspond to the direction of genetic correlation, with warm shades depicting positive correlations between impaired pulmonary function and lung cancer risk, and gray tones corresponding to inverse and null correlations. The width of each band in the Circos plot is proportional to the magnitude of the absolute value of the $r_g$ estimate.

genetically predicted 10% reduction in FEV$_1$/FVC was weakly and inconsistently associated with lung cancer risk ($OR_{IVW} = 1.15$, $p = 0.038$; $OR_{RAPS} = 1.08$, $p = 0.488$). Genetic predisposition to FEV$_1$ and FVC impairment did not appear to confer an increased risk among smokers.

Extensive MR diagnostics are summarized in Supplementary Table 12. All analyses used strong instruments (F-statistic > 40) and did not appear to be weakened by violations of the no measurement error (NOME) assumption ($I^2_{GX}$ statistic > 0.97). MR Steiger test[32] was used to orient the causal effects and confirmed that instruments for pulmonary function were affecting lung cancer susceptibility, not the reverse, and this direction of effect was highly robust. No instruments were removed based on Steiger filtering. We also confirmed that none of the genetic instruments were associated with nicotine dependence phenotypes ($P < 1 \times 10^{-5}$), such as time to first cigarette, difficulty in quitting smoking, and number of quit attempts, which were available for a subset of individuals in the UKB. All MR analyses were adequately powered, with >80% power to detect a minimum OR of 1.25 for FEV$_1$ and FEV$_1$/FVC (Supplementary Fig. 2). For never smokers, we had 80% power to detect a minimum OR of 1.40 for FEV$_1$/FVC and 1.60 for FEV$_1$.

Given the genetic correlation observed for pulmonary phenotypes cigarette smoking and adiposity, we conducted several sensitivity analyses to further address any potential confounding by these phenotypes. The finding for squamous

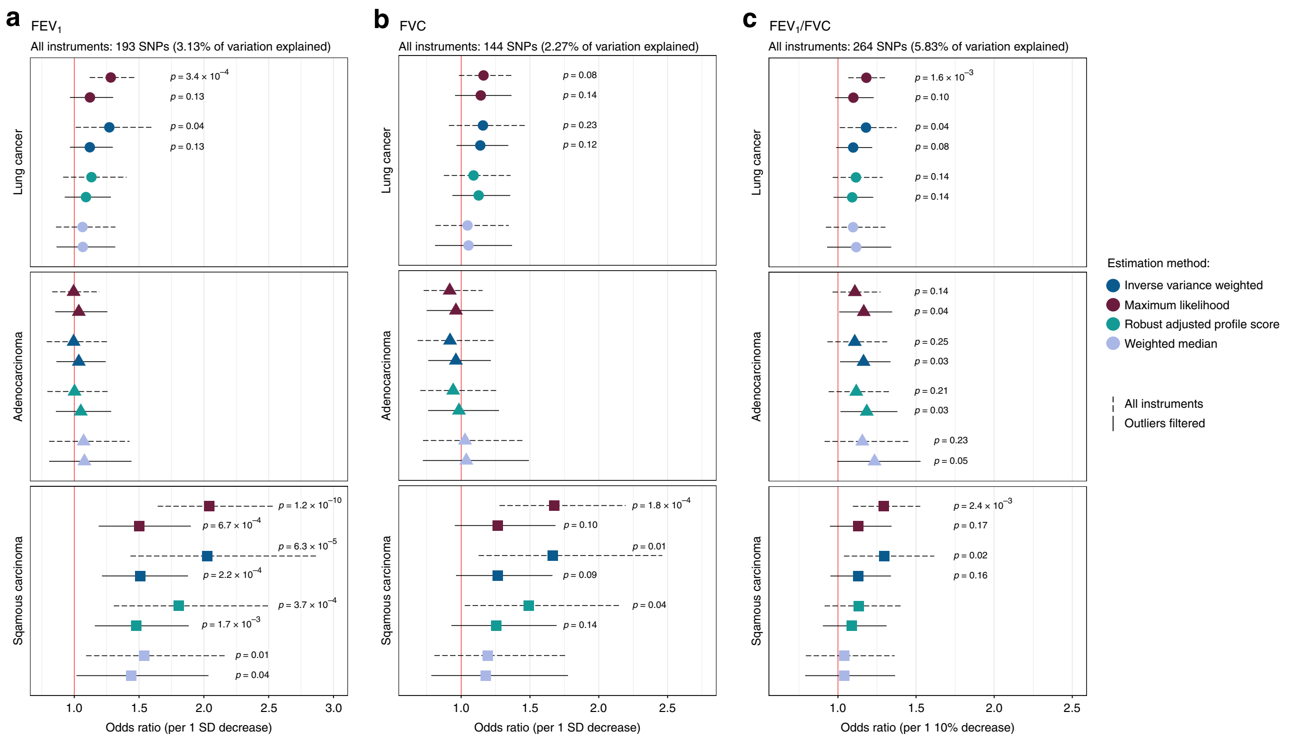
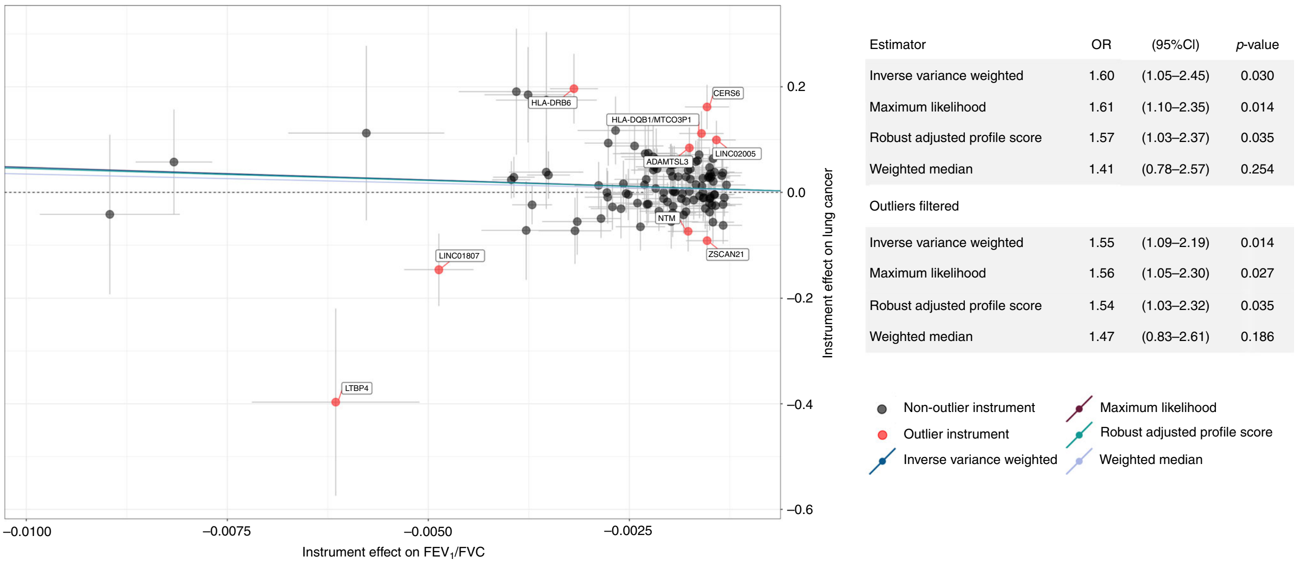


**Fig. 4 Odds ratios (OR) and 95% confidence intervals for the effect of impaired pulmonary function on lung cancer risk, estimated using Mendelian randomization (MR).** Multiple MR estimation methods were applied to the International Lung Cancer Consortium OncoArray dataset, comprised of 29,266 lung cancer cases (11,273 adenocarcinoma, 7426 squamous cell carcinoma) and 56,450 controls to assess the causal relevance of impaired FEV$_1$ (**a**), FVC (**b**), and FEV$_1$/FVC (**c**). MR estimates based on the full set of genetic instruments are compared to estimates after excluding outliers suspected of violating MR assumptions. Only associations with *p*-values < 0.25 are labeled. Proportion of variation explained by the genetic instruments was estimated in a separate replication sample of over 110,000 individuals from the UK Biobank.



| Estimator | OR | (95%CI) | *p*-value |
|---|---|---|---|
| Inverse variance weighted | 1.60 | (1.05–2.45) | 0.030 |
| Maximum likelihood | 1.61 | (1.10–2.35) | 0.014 |
| Robust adjusted profile score | 1.57 | (1.03–2.37) | 0.035 |
| Weighted median | 1.41 | (0.78–2.57) | 0.254 |
| Outliers filtered | | | |
| Inverse variance weighted | 1.55 | (1.09–2.19) | 0.014 |
| Maximum likelihood | 1.56 | (1.05–2.30) | 0.027 |
| Robust adjusted profile score | 1.54 | (1.03–2.32) | 0.035 |
| Weighted median | 1.47 | (0.83–2.61) | 0.186 |

**Fig. 5 Scatterplot depicting the Mendelian randomization (MR) results for FEV$_1$/FVC in never smokers and lung cancer in never smokers (2355 cases, 7504 controls).** The scatterplot illustrates the effects of individual instruments on FEV$_1$/FVC and lung cancer risk, highlighting potentially invalid outlier instruments that were filtered. Individual instrument effects in the scatterplot correspond to a 1-unit decrease in FEV$_1$/FVC, but the summary odds ratios (ORs) for lung cancer have been rescaled to correspond to a 10% decrease in FEV$_1$/FVC. Summary log(ORs) based on different MR estimators correspond to the slope of the lines in scatterplot.

carcinoma and $FEV_1$ was further interrogated using multivariable MR (MVMR) by incorporating genetic instruments for BMI[33] and smoking behavior[34] to estimate the direct effect of $FEV_1$ on squamous carcinoma risk. MVMR using all instruments yielded an OR of 1.95 (95% CI: 1.36–2.80, $p = 2.8 \times 10^{-4}$) per 1-SD decrease in $FEV_1$ and an OR of 1.63 (95% CI: 1.20–2.23, $p = 1.8 \times 10^{-3}$) after filtering outlier instruments.

We confirmed that none of the genetic instruments were associated with smoking status (ever/never), cigarette pack-years (continuous), or adiposity (body fat percentage) at the $P < 5 \times 10^{-8}$ level. However, several variants were associated based on a $P < 1 \times 10^{-5}$ threshold (25 for $FEV_1$ and 18 for $FEV_1$/FVC). We repeated MR analyses after removing these variants (Supplementary Table 13) and confirmed that our results remained robust for $FEV_1$ and squamous cell carcinoma ($OR_{IVW} = 2.02$, 1.40–2.92, $p = 1.9 \times 10^{-4}$) and $FEV_1$/FVC and adenocarcinoma ($OR_{IVW} = 1.19$, 1.01–1.40, $p = 0.04$). However, there was still significant heterogeneity among the causal effect estimates. After filtering the remaining outliers, the effect of a 10% decrease in $FEV_1$/FVC on adenocarcinoma strengthened ($OR_{IVW} = 1.24$, 1.08–1.43, $p = 2.4 \times 10^{-3}$), while estimates attenuated slightly for FEV1 and squamous carcinoma ($OR_{IVW} = 1.46$, 1.14–1.87, $p = 2.7 \times 10^{-3}$).

We also considered the possibility of residual confounding in our GWAS due to insufficient adjustment for smoking-related factors. We thus re-estimated SNP effects on $FEV_1$, FVC, and $FEV_1$/FVC with adjustment for continuous cigarette pack-years and years since quitting. The distribution of effect sizes did not differ between the two analyses ($p > 0.05$), and the correlation with our original instrument weights was strong for all phenotypes (Pearson's $r \geq 0.87$, $p < 1 \times 10^{-40}$) (Supplementary Fig. 3).

Last, we examined the association between $FEV_1$ and $FEV_1$/FVC genetic instruments and COPD, defined as $FEV_1/FVC < 0.70$. Among $FEV_1$ instruments, 64% (123 variants) were associated with COPD at $p < 0.05$ and 16% (31 variants) at $p < 5 \times 10^{-8}$ (Supplementary Fig. 4). All instruments for $FEV_1$/FVC were associated with COPD at the nominal level, and 40% (105 variants) reached genome-wide significance. Using weights from estimated associations between the 105 instruments and COPD log(OR), we observed a modestly increased risk of lung adenocarcinoma ($OR_{IVW} = 1.08$, 1.01–1.15, $p = 0.015$), which parallels our findings based on instruments developed for the continuous $FEV_1$/FVC phenotype.

**Functional characterization of lung function instruments**. To gain insight into biological mechanisms mediating the observed effects of impaired pulmonary function on lung cancer risk, we conducted in silico analyses of functional features associated with the genetic instruments for each lung phenotype.

We identified 185 statistically significant (Bonferroni $p < 0.05$) *cis*-eQTLs for 101 genes among the genetic instruments for $FEV_1$ and $FEV_1$/FVC based on lung tissue gene expression data from the Laval biobank[35] (Supplementary Data 7). Predicted expression of seven genes was significantly ($p < 5.0 \times 10^{-4}$) associated with lung cancer risk: *SECISBP2L*, *HLA-L*, *DISP2*, *MAPT*, *KANSL1-AS1*, *LRRC37A4P*, and *PLEKHM1* (Supplementary Fig. 5). Of these, *SECISBP2L* (OR = 0.80, $p = 5.2 \times 10^{-8}$), *HLA-L* (OR = 0.84, $p = 1.6 \times 10^{-6}$), and *DISP2* (OR = 1.25, $p = 1.6 \times 10^{-4}$) displayed consistent directions of effect for pulmonary function and lung cancer risk, whereby alleles associated with increased expression were associated with impaired $FEV_1$ or $FEV_1$/FVC and increased cancer risk (or conversely, positively associated with pulmonary function and inversely associated with cancer risk). Gene expression associations with inconsistent effects are more likely to indicate pleiotropic pathways not operating primarily through pulmonary impairment. Differences by histology were observed for *SECISBP2L*, which was associated with adenocarcinoma (OR = 0.54, $p = 3.1 \times 10^{-14}$), but not squamous cell carcinoma (OR = 1.05, $p = 0.44$). Effects observed for *DISP2* (OR = 1.21, $p = 0.021$) and *HLA-L* (OR = 0.90, $p = 0.034$) were attenuated for adenocarcinoma, but not for squamous carcinoma (*DISP2*: OR = 1.30, $p = 6.2 \times 10^{-3}$; *HLA-L*: OR = 0.75, $p = 1.6 \times 10^{-6}$).

A total of 70 lung function instruments were mapped to genome-wide significant ($p < 5.0 \times 10^{-8}$) protein quantitative trait loci (pQTL) affecting the plasma levels of 64 different proteins (Supplementary Data 8), based on data from the Human Plasma Proteome Atlas[36]. Many of these pQTL targets are involved in regulation of immune and inflammatory responses, such as interleukins (IL21, IL1R1, IL17RD, IL18R1), MHC class I polypeptide-related sequences, transmembrane glycoproteins expressed by natural killer cells, and members of the tumor necrosis receptor superfamily (TNFSF12, TNFRSF6B, TR19L). Other notable associations include NAD(P)H dehydrogenase [quinone] 1 (NQO1) a detoxification enzyme involved in protecting lung tissues in response to reactive oxidative stress (ROS) and promoting p53 stability[37]. NQO1 is a target of the NFE2-related factor 2 (NRF2), a master regulator of cellular antioxidant response that has generated considerable interest as a chemoprevention target[38,39].

Next, we analyzed genes where the lung function instruments were localized using curated pathways from the Reactome database. Significant enrichment (FDR $q < 0.05$) was observed only for $FEV_1$/FVC instruments in never smokers, with an over-representation of pathways involved in adaptive immunity and cytokine signaling (Supplementary Fig. 6). Top-ranking pathways with $q = 2.2 \times 10^{-6}$ included translocation of ZAP-70 to immunological synapse, phosphorylation of CD3 and TCR zeta chains, and PD-1 signaling. These findings are in line with the predominance of immune-related pQTL associations. Examining all instruments for $FEV_1$ and $FEV_1$/FVC identified significant over-representation (FDR $q < 0.05$) of six immunologic signatures from the ImmuneSigDB collection[40], including pathways implicated in host response to infection and immunization (Supplementary Fig. 7).

## Discussion

Despite a substantial body of observational literature demonstrating an increased risk of lung cancer in individuals with pulmonary dysfunction[2–7,41], confounding by shared environmental risk factors and high co-occurrence of lung cancer and airflow obstruction created uncertainty regarding the causal nature of this relationship. We comprehensively investigated this by characterizing shared genetic profiles between lung cancer and lung function, and interrogated causal hypotheses using Mendelian randomization, which overcomes many limitations of observational studies. We also provide insight into biological pathways underlying the observed associations by incorporating functional annotations into heritability analyses, assessing eQTL and pQTL effects of lung function instruments, and conducting pathway enrichment analyses.

The large sample size of the UK Biobank allowed us to successfully create instruments for three pulmonary function phenotypes, $FEV_1$, $FEV_1$/FVC, and FVC. Although these phenotypes are closely related, they capture different aspects of pulmonary impairment, with $FEV_1$ and $FEV_1$/FVC used for diagnostic purposes in clinical setting. Our genetic instruments captured known and novel mechanisms involved in pulmonary function. Of the 73 novel variants identified here, many were in loci implicated in immune-related functions and pathologies. Examples include

*HORMAD2*, which has been previously linked to inflammatory bowel disease[42,43] and tonsillitis[44], and *RIPOR1* (also known as *FAM65A*), which is part of a gene expression signature for atopy[45]. *PIEZO1* is primarily involved in mechano-transduction and tissue differentiation during embryonic development[46–48], however, recent evidence has emerged delineating its role in optimal T-cell receptor activation and immune regulation[49]. *BACH2*, the new signal for $FEV_1$/FVC in never smokers, is involved in alveolar macrophage function[50], as well as selection-mediated *TP53* regulation and checkpoint control[51]. The lead variant identified here is independent ($r^2 < 0.05$) of *BACH2* loci nominally associated with lung function decline in a candidate gene study of COPD patients[52], suggesting there may be differences in the genetic architecture of pulmonary traits in never smokers.

Our genetic correlation analyses indicate shared genetic determinants between pulmonary function with anthropometric traits and cigarette smoking. Our results are in contrast with the recent findings of Wyss et al.[24], who did not observe statistically significant genetic correlations for any pulmonary function phenotypes with height and smoking, as well FVC and $FEV_1$/FVC, using publicly available summary statistics from the UKB and other studies of European ancestry individuals. In this respect, assessing genetic correlation within a single well-characterized population provides improved power while minimizing potential for bias and heterogeneity when combining data from multiple sources.

We observed statistically significant genetic correlations between pulmonary function impairment and lung cancer susceptibility for all lung cancer subtypes, except for never smokers. Reduced $FEV_1$/FVC was significantly correlated with increased risk of lung cancer overall, squamous cell carcinoma, and adenocarcinoma. Significant genetic correlations with $FEV_1$ and FVC were observed for lung cancer overall, in smokers, and for tumors with squamous cell histology, but not adenocarcinoma. Jiang et al.[25] reported a similar magnitude of genetic correlation with $FEV_1$/FVC, but did not observe an association with FVC, and did not assess $FEV_1$. Differences in our results may be attributable to their use of GWAS summary statistics for pulmonary phenotypes from the interim UK Biobank release. Our findings demonstrate substantial overlap in the genetic architecture of obstructive and neoplastic lung disease, particularly for highly conserved variants that are likely to be subject to natural selection, and super enhancers. However, genetic correlations do not support a causal interpretation, especially considering the shared heritability with potentially confounding traits, such as smoking and obesity.

On the other hand, Mendelian randomization analyses revealed histology-specific effects of reduced $FEV_1$ and $FEV_1$/FVC on lung cancer susceptibility, suggesting that these indicators of impaired pulmonary function may be causal risk factors. Genetic predisposition to $FEV_1$ impairment conferred an increased risk of lung cancer overall, particularly for squamous carcinoma. This relationship persisted after filtering potentially pleiotropic instruments and performing other sensitivity analyses, including multivariable Mendelian randomization and manual filtering of variants associated with smoking or adiposity. $FEV_1$/FVC reduction appeared to increase the risk of lung adenocarcinoma, as well as lung cancer among never smokers. The latter finding is particularly compelling since it precludes confounding by smoking-related factors and demonstrates an association with the most clinically relevant pulmonary phenotype. The increased lung cancer risk in never smokers was also observed using genetic instruments developed specifically in never smokers and in sensitivity analyses using instruments from the population that also includes smokers. We hypothesize that the effects of pulmonary obstruction are mediated by chronic inflammation and immune response, which is supported by the over-representation of adaptive immunity and cytokine signaling pathways and pQTL effects among $FEV_1$ and $FEV_1$/FVC instruments.

Examining lung eQTL effects of our genetic instruments identified additional relevant mechanisms, including gene expression of *SECISBP2L* and *DISP2*. *SECISBP2L* at 15q21 is essential for ciliary function[53] and has an inhibitory effect on lung tumor growth by suppressing cell proliferation and inactivation of Aurora kinase A[54]. This gene was among several susceptibility regions identified in the most recent lung cancer GWAS[16], and now we more conclusively establish impaired pulmonary function as the mechanism mediating *SECISBP2L* effects on risk of lung cancer overall, particularly adenocarcinoma. Less is known about *DISP2*, although it has been implicated in the conserved Hedgehog signaling pathway essential for embryonic development and cell differentiation[55].

One of the main challenges and outstanding questions in previous epidemiologic studies has been clarifying how smoking fits into the causal pathway between impaired pulmonary function and lung cancer risk. Are indicators of airway obstruction simply proxies for smoking-induced carcinogenesis? The association between reduced $FEV_1$/FVC and risk of adenocarcinoma and lung cancer in never smokers observed in our Mendelian randomization analysis and in previous studies[8,9], argues against this simplistic explanation and points to alternative pathways. Chronic airway inflammation fosters a lung microenvironment with altered signaling pathways, aberrant expression of cytokines, chemokines, growth factors, and DNA damage-promoting agents, all of which promote cancer initaiton[15]. This mechanism may be particularly relevant for adenocarcinoma, which is the most common lung cancer histology in never smokers, arising from the peripheral alveolar epithelium that has less direct contact with inhaled carcinogens.

Dysregulated immune function is a hallmark of lung cancer and COPD, with both diseases sharing similar inflammatory cell profiles characterized by macrophages, neutrophils, and CD4+ and CD8+ lymphocytes. Immune cells in COPD and emphysema exhibit T helper 1 (Th1)/Th17 polarization, decreased programmed death ligand-1 (PD-L1) expression in alveolar macrophages, and increased production of interferon (IFN)-γ by CD8+ T cells[56], a phenotype believed to prevail at tumor initiation, whereas established tumors are dominated by Th2/M2-like macrophages[57]. These putative mechanisms were highlighted in our pathway analysis, with an enrichment of genes involved in INF-γ, PD-1 and IL-1 signaling among $FEV_1$/FVC genetic instruments, and over-representation of pQTL targets in these pathways. Furthermore, a study of *trans*-thoracically implanted tumors in an emphysema mouse model demonstrates how this pulmonary phenotype results in impaired antitumor T-cell responses at a critical point when nascent cancer cells evade detection and elimination by the immune system resulting in enhanced tumor growth[58].

Other relevant pathways implicating pulmonary dysfunction in lung cancer development include lung tissue destruction via matrix degrading enzymes and increased genotoxic and apoptotic stress resulting from cigarette smoke in conjunction with macrophage- and neutrophil-derived ROS[15,59]. This may explain our findings for $FEV_1$ and squamous carcinoma, for which cigarette smoking is a particularly dominant risk factor. Genetic predisposition to impaired $FEV_1$ may create a milieu that promotes malignant transformation and susceptibility to external carcinogens and tissue damage, rather than increasing the likelihood of cigarette smoking. In our analysis we attempted to isolate the former pathway from the latter by carefully instrumenting pulmonary phenotypes and confirming that they are not associated with behavioral aspects of nicotine dependence. However, residual confounding by smoking cannot be entirely precluded, given its high genetic and phenotypic correlation with $FEV_1$.

The causal interpretation of our results critically depends on the validity of fundamental Mendelian randomization assumptions. We employed a range of estimation techniques with different underlying assumptions, as well as diagnostic tests, to interrogate the robustness of our results with respect to confounding, horizontal pleiotropy, and weak instrument bias. However, despite these efforts, residual confounding by related phenotypes, such as smoking, or subtle effects of population structure cannot be ruled out. In evaluating the contribution of our findings, several limitations should be acknowledged. Our approach to outlier removal based on Cochran's Q-statistic with modified second order weights may have been overly stringent; however, manually pruning based on such a large set of genetic instruments may not be feasible and may introduce additional bias, thus we feel this systematic conservative approach is justified. Furthermore, outlier removal did not have an adverse impact on instrument strength and precision of the MR analysis.

In addition to pleiotropy, selection bias may also undermine the validity of a Mendelian Randomization study, particularly in the form of collider bias, if selection is a function of the exposure or outcome. In the context of the UKB, low participation (5.5%) may have resulted in an unrepresentative study population[60,61]. Although enrollment in the cohort was not explicitly contingent on cancer status or pulmonary function, it is likely that individuals who did not complete a spirometry assessment were more likely to be smokers and have poor lung function. Simulations by Gkatzionis and Burgess[61] demonstrate that when the effect of a risk factor on selection is mild to moderate (odds of selection: 0.82–0.61), the type I error rate remains reasonable at 5.0–6.6%. The direction of the resulting bias depends on the direction and strength of the exposure (lung function)–confounder (smoking) relationship. In the context of our study, the causal effect may be underestimated since the confounder and exposure are both likely to increase non-participation or result in missing spirometry data.

Another limitation is that we did not assess the relationship between the velocity of lung function decline and lung cancer risk, which may also prove to be a risk factor and capture a different dimension of pulmonary dysfunction. Furthermore, since our study includes the largest GWAS of lung cancer cases in never smokers, this precludes a well-powered replication study in an independent European ancestry population. In addition, dichotomous stratification by smoking status does not permit an evaluation of the relationship between pulmonary impairment and lung cancer risk across more granular levels of smoking. Last, in our efforts to present the most comprehensive assessment of pulmonary function impairment and lung cancer risk, a number of analyses were conducted, and it may be possible that some inconsistently observed associations were due to chance.

Despite these limitations, important strengths of this work include the large sample size for instrument development and causal hypothesis testing. Our Mendelian randomization approach leveraged a large number of genetic instruments, including variants specifically associated with lung function in never smokers, while balancing the concerns related to genetic confounding and pleiotropy. By triangulating evidence from gene expression and plasma protein levels, we also provide a more enriched interpretation of the genetic effects of pulmonary function loci on lung cancer risk, which implicate immune-mediated pathways. Despite the small individual SNP effect sizes, combining multiple instruments revealed meaningful increases in lung cancer risk. A genetically predicted 10% reduction in $FEV_1$/FVC confers an ~55% increased risk of lung cancer in never smokers, and a similar magnitude of effect was observed for $FEV_1$ and squamous carcinoma. However, effects of $FEV_1$/FVC on adenocarcinoma were more modest (16–23% increase). Taken together, these findings provide more robust etiological insight than previous studies that relied on using observed lung function phenotypes directly as putatively casual factors.

As our understanding of the shared genetic and molecular pathways between lung cancer and pulmonary disease continues to evolve, identification of new susceptibility loci for pulmonary function and lung cancer risk may have important implications for future precision prevention and screening endeavors. Multiple genetic determinants of lung function are in pathways that contain druggable targets, based on our pQTL findings and previous reports[18], which may open new avenues for chemoprevention or targeted therapies for lung cancers with an obstructive pulmonary etiology. In addition, with accumulating evidence supporting the effectiveness of low-dose computed tomography for lung cancer[62,63], impairment in $FEV_1$ and $FEV_1$/FVC and their genetic determinants may provide additional information for refining risk stratification and screening eligibility criteria.

## Methods

**Study populations**. The UK Biobank (UKB) is a population-based prospective cohort of over 500,000 individuals aged 40–69 years at enrollment in 2006–2010 who completed extensive questionnaires on health-related factors, physical assessments, and provided blood samples[64]. Participants were genotyped on the UK Biobank Affymetrix Axiom array (89%) or the UK BiLEVE array (11%)[64]. Genotype imputation was performed using the Haplotype Reference Consortium data as the main reference panel as well as using the merged UK10K and 1000 Genomes (1000G) phase 3 reference panels[64]. Our analyses were restricted to individuals of predominantly European ancestry based on self-report and after excluding samples with either of the first two genetic ancestry principal components (PCs) outside of 5 standard deviations (SD) of the population mean. Samples with discordant self-reported and genetic sex were removed. Using a subset of genotyped autosomal variants with minor allele frequency (MAF) ≥0.01 and call rate ≥97%, we filtered samples with call rates <97% or heterozygosity >5 standard deviations (SD) from the mean. First-degree relatives were identified using KING[65] and one sample from each pair was excluded, leaving at total of 413,810 individuals available for analysis.

We further excluded 36,461 individuals without spirometry data, 207 individuals who only completed one blow ($n = 207$), for whom reproducibility could not be assessed (Supplementary Fig. 1). For the remaining subjects, we examined the difference between the maximum value per individual (referred to as the best measure) and all other blows. Values differing by more than 0.15 L were considered non-reproducible, based on standard spirometry guidelines[66], and were excluded. Our analyses thus included 372,750 and 370,638 individuals for of $FEV_1$ and FVC, respectively. The best per individual measure among the reproducible blows was used to derive $FEV_1$/FVC, resulting in 368,817 individuals. $FEV_1$ and FVC values were then converted to standardized Z-scores with a mean of 0 and standard deviation (SD) of 1.

The OncoArray Lung Cancer study has been previously described[16]. Briefly, this dataset consists of genome-wide summary statistics based on 29,266 lung cancer cases (11,273 adenocarcinoma, 7426 squamous carcinoma) and 56,450 controls of predominantly European ancestry (≥80%) assembled from studies part of the International Lung Cancer Consortium. Summary statistics from the lung cancer GWAS were adjusted for appropriate covariates, including genetic ancestry PCs, and showed no signs of genomic inflation for lung cancer overall ($\lambda_{GC} = 1.0035$) or for any subtypes, including adenocarcinoma ($\lambda_{GC} = 1.0050$), squamous carcinoma ($\lambda_{GC} = 1.0051$), and lung cancer in never smokers ($\lambda_{GC} = 1.0060$).

Informed consent was obtained from study participants in the UK Biobank and studies contributing data to the OncoArray Lung Cancer collaboration. UK Biobank received ethics approval from the Research Ethics Committee (REC reference: 11/NW/0382). Approval for OncoArray studies was obtained from each of the participating institutional research ethics review boards.

**Genome-wide association analysis**. Genome-wide association analyses of pulmonary function phenotypes in the UK Biobank cohort were conducted using PLINK 2.0 (October 2017 version). We excluded variants out of with Hardy–Weinberg equilibrium at $p < 1 \times 10^{-5}$ in cancer-free individuals, call rate <95% (alternate allele dosage required to be within 0.1 of the nearest hard call to be non-missing), imputation quality INFO < 0.30, and MAF < 0.005. To minimize potential for reverse causation, prevalent lung cancer cases, defined as diagnoses occurring up to 5 years before cohort entry and incident cases occurring within 2 years of enrollment, were excluded ($n = 738$). Linear regression models for pulmonary function phenotypes (standardized Z-scores for $FEV_1$ and FVC; untransformed $FEV_1$/FVC ratio bounded by 0 and 1) were adjusted for age, age2, sex, genotyping array and 15 PCs to permit an assessment of heritability ($h_g$) and genetic correlation ($r_g$) with height, smoking (status and pack-years), and anthropometric traits.

**Heritability and genetic correlation**. LD Score regression[17] was used to estimate $h_g$ for each lung phenotype and $r_g$ with lung cancer and other traits. To better capture LD patterns present in the UKB data, we generated LD scores for all variants that passed QC with MAF > 0.0001 using a random sample of 10,000 UKB participants. UKB LD scores were used to estimate $h_g$ for each lung phenotype and $r_g$ with other non-cancer traits. Genetic correlation with lung cancer was estimated using publicly available LD scores based on the 1000G phase 3 reference population ($n = 1,095,408$ variants).

To assess the importance of specific functional annotations in SNP-heritability, we partitioned trait-specific heritability using stratified-LDSC[67]. The analysis was performed using 86 annotations (baseline-LD model v2.1), which incorporated MAF-adjustment and other LD-related annotations, such as predicted allele age and recombination rate[20,22]. The MHC region was excluded from partitioned heritability analyses. Enrichment was considered statistically significant if $p < 8.5 \times 10^{-4}$, which reflects Bonferroni correction for 59 annotations (functional categories with and without a 500 bp window around it were considered as the same annotation).

**Development of genetic instruments for pulmonary function**. For the purpose of instrument development, a two-stage genome-wide analysis was employed, with a randomly sampled 70% of the cohort used for discovery and the remaining 30% reserved for replication. In addition to age, age2, sex, genotyping array and 15 PC's, models were adjusted for covariates that explain a substantial proportion of variation in pulmonary phenotypes, such as smoking and height, in order to decrease the residual variance and help isolate the relevant genetic signals. Specifically, we adjusted for height, height2, and cigarette pack-year categories (0, corresponding to never smokers, >0–10, >10–20, >20–30, >30–40, and >40). Other covariates, such as UKB assessment center (Field 54), use of an inhaler prior to spirometry (Field 3090), and blow acceptability (Field 3061) were considered. However, these covariates did not explain a substantial proportion of phenotype variation and had low variable importance metrics (lmg < 0.01), and thus were not included in our final models. Instruments were selected from independent associated variants (LD $r^2 < 0.05$ in a clumping window of 10,000 kb) with $P < 5 \times 10^{-8}$ in the discovery stage and $P < 0.05$ and consistent direction of effect in the replication stage. Since the primary goal of our GWAS was to develop a comprehensive set of genetic instruments we applied a less stringent replication threshold in anticipation of subsequent filtering based on potential violation of Mendelian randomization assumptions.

**Mendelian randomization**. Mendelian randomization (MR) analyses were carried out to investigate the potential causal relationship between impaired pulmonary function and lung cancer risk. Genetic instruments excluded multi-allelic and non-inferable palindromic variants with intermediate allele frequencies (MAF > 0.42). Odds ratios (OR) and corresponding 95% confidence intervals were obtained using the maximum likelihood and inverse variance weighted multiplicative random-effects (IVW-RE) estimators[28,29]. Effects for $FEV_1$ and FVC were estimated for a genetically predicted 1-SD decrease in the standardized Z-score. For $FEV_1$/FVC, we modeled cancer risk corresponding to a 10% decrease in the ratio. Sensitivity analyses included the weighted median (WM) estimator[30], which provides unbiased estimates when up to 50% of the weights are from invalid instruments, and MR RAPS (Robust Adjusted Profile Score), which incorporates random effect and robust loss functions to limit the influence of potentially pleiotropic instruments. MR RAPS assumes balanced (mean 0) horizontal pleiotropy. In contrast to IVW-RE, MR RAPS models idiosyncratic and systematic pleiotropy effects as additive, rather than multiplicative[31]. Using MR estimation techniques with different underlying statistical models allows for a more comprehensive assessment of the robustness of our results with respect to violations of MR assumptions. We also applied the following diagnostic tests: (i) significant ($p < 0.05$) deviation of the MR Egger intercept ($\beta_{0\ \mathrm{Egger}}$) from 0, as a test for directional pleiotropy[68]; (ii) $I^2_{GX}$ statistic < 0.90 indicative of regression dilution bias and inflation in the MR Egger pleiotropy test due to violation of the no measurement error (NOME) assumption[68]; (iii) Cochran's Q-statistic with modified second order weights to asses heterogeneity ($p$-value < 0.05) indicative of (balanced) horizontal pleiotropy[69].

All statistical analyses were conducted using R (version 3.6.1). Mendelian randomization analyses were conducted using the TwoSampleMR R package (version 0.4.23).

**Functional characterization of lung function instruments**. In order to characterize functional pathways that are represented by the genetic instruments for $FEV_1$ and $FEV_1$/FVC, we examined effects on gene expression in lung tissues from 409 subjects from the Laval eQTL study[35]. Lung function instruments with significant (Bonferroni $p$-value < 0.05) eQTL effects were used as instruments to estimate the effect of the gene expression on lung cancer risk. For genes with multiple eQTLs, independent variants (LD $r^2 < 0.05$) were used to obtain IVW estimates of the predicted effects of increased gene expression on lung cancer risk. For genes with a single eQTL, OR estimates were obtained using the Wald method. Next, we examined data from the genetic atlas of the human plasma proteome[36], queried using PhenoScanner[70], to assess whether any of the genetic instruments for $FEV_1$ and $FEV_1$/FVC had significant ($p < 5 \times 10^{-8}$) effects on intracellular protein levels. Last, we summarized the pathways represented by the genes where the lung function instruments were localized using pathway enrichment analysis via the Reactome database and ImmuneSigDB (collection C7 from MSigDB).

**URLs**. PLINK 2.0: https://www.cog-genomics.org/plink/2.0/

LDSC (version 1.0.0) from: https://github.com/bulik/ldsc/

LDSC functional annotations available from: https://data.broadinstitute.org/alkesgroup/LDSCORE/1000G_Phase3_EUR_baselineLD_v2.1_ldscores.tgz

R package for Circos plots (version 0.4.7): https://github.com/jokergoo/circlize

R package for Mendelian Randomization (version 0.4.23): https://github.com/MRCIEU/TwoSampleMR

R package for PhenoScanner (version 1.0): https://github.com/phenoscanner/phenoscanner

R packages for pathway analysis: https://bioconductor.org/packages/release/bioc/html/ReactomePA.html and https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html

ImmuneSigDB (C7): http://software.broadinstitute.org/gsea/msigdb/collections.jsp.

**Reporting summary**. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The datasets generated during and/or analyzed during the current study are available from the authors on request. Genotype data for the Oncoarray Consortium Lung Cancer studies have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession: phs001273.v2.p2. Readers interested in obtaining a copy of the lung cancer GWAS summary statistics can do so by completing the proposal request form at http://oncoarray.dartmouth.edu/. The UK Biobank in an open access resource, available at https://www.ukbiobank.ac.uk/researchers/. This research was conducted with approved access to UK Biobank data under applications number 14105 and 23261. All data supporting the findings of this study are available within the article and its supplementary information files, and from the corresponding authors upon reasonable request. A reporting summary for this article is available as a Supplementary file.

Received: 19 May 2019; Accepted: 29 November 2019;
Published online: 07 January 2020

## References

1. Ferlay, J. et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J. Cancer* **144**, 1941–1953 (2019).
2. Wasswa-Kintu, S., Gan, W. Q., Man, S. F., Pare, P. D. & Sin, D. D. Relationship between reduced forced expiratory volume in one second and the risk of lung cancer: a systematic review and meta-analysis. *Thorax* **60**, 570–575 (2005).
3. Calabro, E. et al. Lung function predicts lung cancer risk in smokers: a tool for targeting screening programmes. *Eur. Respir. J.* **35**, 146–151 (2010).
4. Fry, J. S., Hamling, J. S. & Lee, P. N. Systematic review with meta-analysis of the epidemiological evidence relating FEV1 decline to lung cancer risk. *BMC Cancer* **12**, 498 (2012).
5. Mannino, D. M., Aguayo, S. M., Petty, T. L. & Redd, S. C. Low lung function and incident lung cancer in the United States: data From the First National Health and Nutrition Examination Survey follow-up. *Arch. Intern. Med.* **163**, 1475–1480 (2003).
6. Young, R. P. et al. COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. *Eur. Respir. J.* **34**, 380–386 (2009).
7. Zhai, R., Yu, X., Wei, Y., Su, L. & Christiani, D. C. Smoking and smoking cessation in relation to the development of co-existing non-small cell lung cancer with chronic obstructive pulmonary disease. *Int J. Cancer* **134**, 961–970 (2014).
8. Brenner, D. R., McLaughlin, J. R. & Hung, R. J. Previous lung diseases and lung cancer risk: a systematic review and meta-analysis. *PLoS One* **6**, e17479 (2011).
9. Brenner, D. R. et al. Previous lung diseases and lung cancer risk: a pooled analysis from the International Lung Cancer Consortium. *Am. J. Epidemiol.* **176**, 573–585 (2012).
10. Denholm, R. et al. Is previous respiratory disease a risk factor for lung cancer? *Am. J. Respir. Crit. Care Med.* **190**, 549–559 (2014).
11. Durham, A. L. & Adcock, I. M. The relationship between COPD and lung cancer. *Lung Cancer* **90**, 121–127 (2015).

12. Yang, I. A., Holloway, J. W. & Fong, K. M. Genetic susceptibility to lung cancer and co-morbidities. *J. Thorac. Dis.* **5**, S454–S462 (2013). **Suppl 5**.
13. Young, R. P. et al. Chromosome 4q31 locus in COPD is also associated with lung cancer. *Eur. Respir. J.* **36**, 1375–1382 (2010).
14. Hancock, D. B. et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.* **42**, 45–52 (2010).
15. Houghton, A. M. Mechanistic links between COPD and lung cancer. *Nat. Rev. Cancer* **13**, 233–245 (2013).
16. McKay, J. D. et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**, 1126–1132 (2017).
17. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
18. Shrine, N. et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**, 481–493 (2019).
19. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
20. Gazal, S. et al. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
21. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
22. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
23. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
24. Wyss, A. B. et al. Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. *Nat. Commun.* **9**, 2976 (2018).
25. Jiang, X. et al. Shared heritability and functional enrichment across six solid cancers. *Nat. Commun.* **10**, 431 (2019).
26. Gudipaty, S. A. et al. Mechanical stretch triggers rapid epithelial cell division through Piezo1. *Nature* **543**, 118–121 (2017).
27. Zhong, M., Komarova, Y., Rehman, J. & Malik, A. B. Mechanosensing Piezo channels in tissue homeostasis including their role in lungs. *Pulm. Circ.* **8**, 2045894018767393 (2018).
28. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol.* **37**, 658–665 (2013).
29. Bowden, J. et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).
30. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol.* **40**, 304–314 (2016).
31. Zhao, Q., Wang, J., Hemani, G., Bowden, J. & Small, D. S. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. Preprint at https://arxiv.org/pdf/1801.09652v09653.pdf (2019).
32. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).
33. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
34. Tobacco, GeneticsC. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).
35. Hao, K. et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8**, e1003029 (2012).
36. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
37. Asher, G., Lotem, J., Cohen, B., Sachs, L. & Shaul, Y. Regulation of p53 stability and p53-dependent apoptosis by NADH quinone oxidoreductase 1. *Proc. Natl Acad. Sci. USA* **98**, 1188–1193 (2001).
38. Sporn, M. B. & Liby, K. T. NRF2 and cancer: the good, the bad and the importance of context. *Nat. Rev. Cancer* **12**, 564–571 (2012).
39. Rojo de la Vega, M., Chapman, E. & Zhang, D. D. NRF2 and the Hallmarks of Cancer. *Cancer Cell* **34**, 21–43 (2018).
40. Godec, J. et al. Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. *Immunity* **44**, 194–206 (2016).
41. Wilson, D. O. et al. Association of radiographic emphysema and airflow obstruction with lung cancer. *Am. J. Respir. Crit. Care Med.* **178**, 738–744 (2008).
42. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
43. Imielinski, M. et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat. Genet.* **41**, 1335–1340 (2009).
44. Feenstra, B. et al. Genome-wide association study identifies variants in HORMAD2 associated with tonsillectomy. *J. Med. Genet.* **54**, 358–364 (2017).
45. Howrylak, J. A. et al. Gene expression profiling of asthma phenotypes demonstrates molecular signatures of atopy and asthma control. *J. Allergy Clin. Immunol.* **137**, 1390–1397 e1396 (2016).
46. Li, J. et al. Piezo1 integration of vascular architecture with physiological force. *Nature* **515**, 279–282 (2014).
47. Lewis, A. H., Cui, A. F., McDonald, M. F. & Grandl, J. Transduction of repetitive mechanical stimuli by Piezo1 and Piezo2 ion channels. *Cell Rep.* **19**, 2572–2585 (2017).
48. Andolfo, I. et al. Multiple clinical forms of dehydrated hereditary stomatocytosis arise from mutations in PIEZO1. *Blood* **121**, 3925–3935 (2013). S3921-3912.
49. Liu, C. S. C. et al. Cutting edge: Piezo1 mechanosensors optimize human T cell activation. *J. Immunol.* **200**, 1255–1260 (2018).
50. Nakamura, A. et al. Transcription repressor Bach2 is required for pulmonary surfactant homeostasis and alveolar macrophage function. *J. Exp. Med.* **210**, 2191–2204 (2013).
51. Swaminathan, S. et al. BACH2 mediates negative selection and p53-dependent tumor suppression at the pre-B cell receptor checkpoint. *Nat. Med.* **19**, 1014–1022 (2013).
52. Sandford, A. J. et al. NFE2L2 pathway polymorphisms and lung function decline in chronic obstructive pulmonary disease. *Physiol. Genomics* **44**, 754–763 (2012).
53. Boldt, K. et al. An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nat. Commun.* **7**, 11491 (2016).
54. Yu, C. T. et al. The novel protein suppressed in lung cancer down-regulated in lung cancer tissues retards cell proliferation and inhibits the oncokinase Aurora-A. *J. Thorac. Oncol.* **6**, 988–997 (2011).
55. Katoh, Y. & Katoh, M. Hedgehog signaling pathway and gastric cancer. *Cancer Biol. Ther.* **4**, 1050–1054 (2005).
56. Grumelli, S. et al. An immune basis for lung parenchymal destruction in chronic obstructive pulmonary disease and emphysema. *PLoS Med.* **1**, e8 (2004).
57. Conway, E. M. et al. Macrophages, Inflammation, and Lung Cancer. *Am. J. Respir. Crit. Care Med.* **193**, 116–130 (2016).
58. Kerdidani, D. et al. Cigarette smoke-induced emphysema exhausts early cytotoxic CD8(+) T cell responses against nascent lung cancer cells. *J. Immunol.* **201**, 1558–1569 (2018).
59. Haqqani, A. S., Sandhu, J. K. & Birnboim, H. C. Expression of interleukin-8 promotes neutrophil infiltration and genetic instability in mutatect tumors. *Neoplasia* **2**, 561–568 (2000).
60. Manolio, T. A. et al. New models for large prospective studies: is there a better way? *Am. J. Epidemiol.* **175**, 859–866 (2012).
61. Gkatzionis, A. & Burgess, S. Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *Int. J. Epidemiol.*. **48**, 691–701 (2018).
62. National Lung Screening Trial Research T. et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395–409 (2011).
63. De Koning, H., Van Der Aalst, C., Ten Haaf, K. & Oudkerk, M. PL02.05 effects of volume CT lung cancer screening: mortality results of the NELSON randomised-controlled population based trial. *J. Thorac. Oncol.* **13**, S185 (2018).
64. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
65. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
66. Miller, M. R. et al. Standardisation of spirometry. *Eur. Respir. J.* **26**, 319–338 (2005).
67. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
68. Bowden, J. et al. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J. Epidemiol.* **45**, 1961–1974 (2016).
69. Bowden, J. et al. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int. J. Epidemiol.* **48**, 728–742 (2018).
70. Kamat, M. A. et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**,4851–4853 (2019).

## Acknowledgements

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization. This research was supported by funding from the National Institutes of Health (US NCI R25T CA112355 and R01 CA201358; PI: Witte) and the Canadian Institute for Health Research (Foundation grant FDN 167273, PI: Hung; Canada Research Chair, PI: Hung). The OncoArray project was supported by NIH U19 CA203654 (MPI: Hung, Amos, Brennan, Lin). The Boston Lung Cancer Study was funded by NIH (NCI) U01CA209414 (PI: Christiani). The lung eQTL study at Laval University was supported by the Fondation de l'Institut universitaire de cardiologie et de pneumologie de Québec and the Canadian Institutes of Health Research (MOP − 123369). Y.B. holds a Canada Research Chair in Genomics of Heart and Lung Diseases. The EAGLE study was supported by the Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS. The Multiethnic Cohort Study is supported by National Institutes of Health (CA164973). The CARET study was supported by the National Institutes of Health/National Cancer Institute: UM1 CA167462 (PI: Goodman), U01 CA6367307 (PIs: Omen, Goodman); R01 CA111703 (PI: Chen), 5R01 CA151989-01A1 (PI: Doherty) and U01 CA167462 (PI: Chen). M.B.S. was supported in part by a Cancer Center Support Grant (P30 CA076292) and by NIH P50 CA119997. R.M.M. is supported by a CRUK programme grant, the Integrative Cancer Epidemiology Programme (C18281/A19169), and by the National Institute for Health Research (NIHR) Bristol Biomedical Research Centre based at University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. Dr Haycock is supported by CRUK Population Research Postdoctoral Fellowship C52724/A20138. M.P.D. is supported by the Roy Castle Lung Cancer Foundation UK.

## Author contributions

Study conception: L.K., R.J.H., M.J., and P.B.; Statistical analysis: L.K.; L.K. drafted the paper with input from R.J.H. and J.S.W.; Project coordination: R.J.H., P.B., M.J., and J.S.W.; UK Biobank genotype and sample quality control: S.R.R. and R.E.G.; Lung tissue eQTL analysis: Y. Bossé and V.M.; Development of analytic strategy: L.K., S.R.R., R.E.G., R.J.H., J.S.W., M.J., R.M.M., C.R., G.D.S., and P.C.H.; Data acquisition and development of lung cancer epidemiological studies: R.J.H., C.I.A., P.B., M.J., J.S.W., N.E.C., M.T.L., D.C.C., P.V., G.L., G.S., D.Z., S.S.S., D.A., M.C.A., A.T., G.R., C.C., G.E.G, J.A.D., H.B., J.K.F, M.P.D., M.D.T., L.A.K., S.E.B., A.H., S.Z., S.L., L.L.M., I.C., M.B.S., E.J.D., A.S.A., J.M., P.L., S.A., J.D.M., N.C.E., M.T.W., Y.B., M.O., and Y. Bossé. All authors contributed to the interpretation of the results and provided critical feedback on the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41467-019-13855-2.

**Correspondence** and requests for materials should be addressed to J.S.W. or R.J.H.

**Peer review information** *Nature Communications* thanks Bjorn Olav Asvold and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Linda Kachuri[1,2], Mattias Johansson[3], Sara R. Rashkin[1], Rebecca E. Graff[1], Yohan Bossé[4], Venkata Manem[4], Neil E. Caporaso[5], Maria Teresa Landi[5], David C. Christiani[6], Paolo Vineis[7], Geoffrey Liu[8], Ghislaine Scelo[3], David Zaridze[9], Sanjay S. Shete[10], Demetrius Albanes[5], Melinda C. Aldrich[11], Adonina Tardón[12], Gad Rennert[13], Chu Chen[14], Gary E. Goodman[14], Jennifer A. Doherty[15], Heike Bickeböller[16], John K. Field[17], Michael P. Davies[17], M. Dawn Teare[18], Lambertus A. Kiemeney[19], Stig E. Bojesen[20], Aage Haugen[21], Shanbeh Zienolddiny[21], Stephen Lam[22], Loïc Le Marchand[23], Iona Cheng[1], Matthew B. Schabath[24], Eric J. Duell[25], Angeline S. Andrew[26], Jonas Manjer[27], Philip Lazarus[28], Susanne Arnold[29], James D. McKay[3], Nima C. Emami[1], Matthew T. Warkentin[2,30], Yonathan Brhane[2], Ma'en Obeidat[31], Richard M. Martin[32,33,34], Caroline Relton[32,33], George Davey Smith[32,33], Philip C. Haycock[32,33], Christopher I. Amos[35], Paul Brennan[3], John S. Witte[1,36]⋆ & Rayjean J. Hung[2,30,36]⋆

[1]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA, USA. [2]Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada. [3]International Agency for Research on Cancer, Lyon, France. [4]Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval, Quebec City, Canada. [5]Division of Cancer Epidemiology & Genetics, US NCI, Bethesda, MD, USA. [6]Departments of Environmental Health and Epidemiology, Harvard TH Chan School of Public Health, Boston, MA, USA. [7]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [8]Princess Margaret Cancer Center, University Health Network, Toronto, ON, Canada. [9]Russian N.N. Blokhin Cancer Research Centre, Moscow, Russian Federation. [10]Department of Biostatistics, Division of Basic Sciences, MD Anderson Cancer Center, Houston, TX, USA. [11]Department of Thoracic Surgery and Division of Epidemiology, Vanderbilt University Medical Center, Nashville, TN, USA. [12]Faculty of Medicine, University of Oviedo and ISPA and CIBERESP, Campus del Cristo, Oviedo, Spain. [13]Clalit National Cancer Control Center, Technion Faculty of Medicine, Haifa, Israel. [14]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [15]Department of Population Health Sciences, Huntsman Cancer Institute, Salt Lake City, UT, USA. [16]Department of Genetic Epidemiology, University Medical Center, Georg-August-Universität Göttingen, Göttingen, Germany. [17]Roy Castle Lung Cancer Research Programme, Department of Molecular and Clinical Cancer Medicine, The University of Liverpool, London, UK.

[18]Biostatistics Research Group, Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK. [19]Radboud Institute for Health Sciences, Radboud University Medical Centre, Nijmegen, The Netherlands. [20]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [21]The National Institute of Occupational Health, Oslo, Norway. [22]BC Cancer Agency, Vancouver, BC, Canada. [23]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA. [24]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL, USA. [25]Unit of Biomarkers and Susceptibility, Oncology Data Analytics Program, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain. [26]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA. [27]Skåne University Hospital, Lund University, Lund, Sweden. [28]Department of Pharmaceutical Sciences, College of Pharmacy and Pharmaceutical Sciences, Washington State University, Spokane, WA, USA. [29]Markey Cancer Center, University of Kentucky, Lexington, KY, USA. [30]Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada. [31]University of British Columbia, Centre for Heart Lung Innovation, Vancouver, BC, Canada. [32]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [33]Bristol Medical School, Population Health Sciences, University of Bristol, Bristol, UK. [34]National Institute for Health Research (NIHR) Bristol Biomedical Research Centre, University Hospitals Bristol NHS Foundation Trust and the University of Bristol, Bristol, UK. [35]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA. [36]These authors jointly supervised this work: John S. Witte, Rayjean J. Hung. *email: jwitte@ucsf.edu; rayjean.hung@lunenfeld.ca