A causal pathway of CYP2A6 activity and cigarette consumption involved in smoking-related lung cancer susceptibility

Mulong Du[1,2,*,†], Junyi Xin[3,†], Rui Zheng[4,5,†], Qianyu Yuan[1], Zhihui Wang[1], Hongliang Liu[6,7], Hanting Liu[4,5], Guoshuai Cai[8], Demetrius Albanes[9], Stephen Lam[10], Adonina Tardon[11], Chu Chen[12], Stig E. Bojesen[13,14,15], Maria Teresa Landi[16], Mattias Johansson[17], Angela Risch[18,19,20,21], Heike Bickeböller[22], H-Erich Wichmann[23,24,25], Gad Rennert[26], Susanne Arnold[27], Paul Brennan[28], John K. Field[29], Sanjay S. Shete[30], Loïc Le Marchand[31], Geoffrey Liu[32], Angeline S. Andrew[33], Lambertus A. Kiemeney[34], Shan Zienolddiny[35], Kjell Grankvist[36], Mikael Johansson[37], Neil E Caporaso[38], Angela Cox[39], Yun-Chul Hong[40], Jian-Min Yuan[41], Matthew B. Schabath[42], Melinda C. Aldrich[43], Meilin Wang[4,5], Hongbing Shen[44], Feng Chen[2], Zhengdong Zhang[4,5], Rayjean J. Hung[45], Christopher I. Amos[46], Qingyi Wei[6,7], Philip Lazarus[47], David C. Christiani[1,48,*]

1. Department of Environmental Health, Harvard T.H. Chan School of Public Health, 665 Huntington Avenue, Boston, MA, 02115, USA.
2. Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, 101 Longmian Avenue, Nanjing, 211166, Jiangsu, China.
3. Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, 101 Longmian Avenue, Nanjing, 211166, Jiangsu, China.
4. Department of Environmental Genomics, Jiangsu Key Laboratory of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, School of Public Health, Nanjing Medical University, 101 Longmian Avenue, Nanjing, 211166, Jiangsu, China.

5. Department of Genetic Toxicology, The Key Laboratory of Modern Toxicology of Ministry of Education, Center for Global Health, School of Public Health, Nanjing Medical University, 101 Longmian Avenue, Nanjing, 211166, Jiangsu, China.
6. Duke Cancer Institute, Duke University Medical Center, Durham, NC, 27710, USA.
7. Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, 27710, USA.
8. Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, Columbia, SC, 29208, USA.
9. Division of Cancer Epidemiology and Genetics, National Cancer Institute, US National Institutes of Health, Bethesda, Maryland, USA.
10. British Columbia Cancer Agency, Vancouver, British Columbia, Canada.
11. University of Oviedo, ISPA and CIBERESP, Faculty of Medicine, Oviedo, Spain.
12. Program in Epidemiology, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.
13. Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Copenhagen, Denmark.
14. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
15. Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen, Denmark.
16. Division of Cancer Epidemiology and Genetics, National Cancer Institute, US National Institutes of Health, Bethesda, Maryland, USA.
17. International Agency for Research on Cancer, World Health Organization, Lyon, France.
18. University of Salzburg and Cancer Cluster Salzburg, Salzburg, Austria.

19. Translational Lung Research Center Heidelberg (TLRC-H), Heidelberg, Germany.
20. German Center for Lung Research (DZL), Heidelberg, Germany.
21. German Cancer Research Center (DKFZ), Heidelberg, Germany.
22. Department of Genetic Epidemiology, University Medical Center, Georg August University Göttingen, Göttingen, Germany.
23. Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany.
24. Institute of Epidemiology II, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany.
25. Institute of Medical Statistics and Epidemiology, Technical University of Munich, Munich, Germany.
26. Clalit National Cancer Control Center at Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel.
27. Markey Cancer Center, University of Kentucky, Lexington, Kentucky, USA.
28. International Agency for Research on Cancer, World Health Organization, Lyon, France.
29. Department of Molecular and Clinical Cancer Medicine, Institute of Systems, Molecular & Integrative Biology, University of Liverpool, Liverpool, UK.
30. Department of Epidemiology, Division of Cancer Prevention and Population Science, The University of Texas, MD Anderson Cancer Center, Houston, Texas, USA.
31. Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, USA.
32. Princess Margaret Cancer Center, University of Toronto, Toronto, Ontario, Canada.
33. Norris Cotton Cancer Center, Geisel School of Medicine, Hanover, New Hampshire, USA.
34. Radboud University Medical Center, Nijmegen, the Netherlands.
35. National Institute of Occupational Health, Oslo, Norway.

36. Department of Medical Biosciences, Umeå University, Umeå, Sweden.

37. Department of Radiation Sciences, Umeå University, Umeå, Sweden.

38. Division of Cancer Epidemiology and Genetics, National Cancer Institute, US National Institutes of Health, Bethesda, Maryland, USA.

39. Department of Oncology, University of Sheffield, Sheffield, UK.

40. Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea.

41. UPMC Hillman Cancer Center and Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA.

42. Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA.

43. Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA.

44. Department of Epidemiology and Biostatistics, Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China.

45. Lunenfeld-Tanenbuaum Research Institute, Sinai Health System, University of Toronto, Toronto, Ontario, Canada.

46. Institute for Clinical and Translational Research, Baylor Medical College, Houston, Texas, USA.

47. Department of Pharmaceutical Sciences, College of Pharmacy and Pharmaceutical Sciences, Washington State University, Spokane, WA, 99210, USA.

48. Department of Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA, 02114, USA.

† These authors contributed equally to this work.

* Correspondence to:

David C. Christiani, dchris@hsph.harvard.edu, +1 617 726 9274, Department of Environmental Health, Harvard T.H. Chan School of Public Health, 665 Huntington Avenue, Boston, MA, 02115, USA.

Mulong Du, drdumulong@njmu.edu.cn, +86 25 86868423, Department of Biostatistics, Center for Global Health, School of Public Health, 101 Longmian Avenue, Nanjing, 211166, Jiangsu, China.

Running title: Causal pathway of *CYP2A6*, smoking, and lung tumorigenesis

Keywords: causal inference, *CYP2A6*, cigarette consumption, lung cancer

Additional information

Financial support: This work was partially supported by R01 ES025460-01 from National Institute of Environmental Health Sciences (PI: P. Lazarus), and U01 CA209414 from National Cancer Institute (PI: D. Christiani). CARET is funded by the National Cancer Institute, National Institutes of Health through grants U01-CA063673, UM1-CA167462, and U01-CA167462 (PI: C. Chen).

Conflict of Interest: The authors declare no potential conflicts of interest

Word count: 3454

Number of figures: 3

Number of tables: 1

Statement of Significance: We precisely quantify a causal pathway of *CYP2A6* genetic variant, activity, cigarette consumption, and lung cancer susceptibility in the smoking population. This study corroborates behavior modification intervention on disease in population-level prevention.

Abstract

Cigarette smoke, containing both nicotine and carcinogens, causes lung cancer. This study is aimed to delineate the mediation of metabolizing ability of tobacco carcinogens and smoking intensity in the causal pathway from genetic susceptibility to smoking-related lung tumorigenesis. We analyzed single-variant and gene-based associations of 43 tobacco carcinogen-metabolizing genes with lung cancer by using summary statistics and individual-level genetic data, followed by causal inference of Mendelian Randomization, Mediation analysis, and Structural Equation Modelling. Cigarette smoke-exposed cell models were used to detect gene expression patterns in relation to specific alleles. International Lung Cancer Consortium (29,266 cases and 56,450 controls) and UK Biobank (2,155 cases and 376,329 controls) supported that the genetic variant rs56113850 C>T located in intron 4 of *CYP2A6* was significantly associated with a decreased lung cancer risk among smokers [odds ratio (OR) = 0.88, 95% confidence interval = 0.85-0.91, $P = 2.18\times10^{-16}$], which might interact ($P_{interaction} = 0.028$) with and partially be mediated ($OR_{indirect} = 0.987$) by smoking status. Besides, smoking intensity accounted for 82.3% of the effect of CYP2A6 activity on lung cancer risk but entirely mediated the genetic effect of rs56113850. Mechanistically, rs56113850 T allele rescued the downregulation of *CYP2A6* caused by cigarette smoke exposure, potentially through preferential recruitment of transcription factor HLTF. Together, this study provides additional insights into the interplay between host susceptibility and carcinogen exposure involving smoking-related lung tumorigenesis.

Introduction

Lung cancer ranks second in cancer incidence but remains the leading cause of cancer-related death worldwide (1). In the United States, the annual decline in lung cancer incidence and mortality can be largely attributed to a significant decrease in smoking rates (2). However, it is important note that cigarette smoking remains the primary preventable cause of death, directly responsible for 82% of all deaths due to lung cancer (3,4).

Tobacco smoke comprises a toxic mixture of more than 7,000 chemicals, 70 of which are well known to cause cancers (5,6). Among these, polycyclic aromatic hydrocarbons (PAHs) and tobacco-specific nitrosamines (TSNAs) are recognized as significant contributors to lung carcinogenesis. They require bioactivation by key enzymes before binding to DNA and initiate genomic alterations (7,8). The interaction between host susceptibility and environmental exposure is widely acknowledged as a crucial factor in tumorigenesis, such as the interplay between fine particulate matter ($PM_{2.5}$) and genetic variants in colorectal cancer (9), as well as smoking and somatic mutations in lung cancer (10) reported in our previous studies. In addition, it has been suggested that genetic variation influenced the activity of cytochrome P450 family 2 subfamily A member 6 (CYP2A6), with high activity inducing more extensive and intense smoking, exposing the lungs to higher levels of carcinogens, and thus increasing lung cancer risk (11). Despite these findings, the causal relationship and underlying biological interpretation linking carcinogen exposure, toxic metabolism, and lung cancer remain unclear.

In this study, we postulate the existence of a causal cascade of tobacco carcinogen metabolism and dosage in smoking-related lung carcinogenesis. To investigate this hypothesis, we summarized 43 metabolizing enzymes involved in PAH and TSNA metabolism pathways and analyzed their genetic effects on lung cancer susceptibility, further performed

causal inference and function study to interpret the potential biological role in lung tumorigenesis.

Materials and Methods

Study subjects

Genome-wide association study (GWAS) summary statistics of lung cancer with 29,266 cases and 56,450 controls of European ancestry, as well as individual OncoArray genotyping data (imputed genotypes included) for 14,803 lung cancer cases and 12,262 controls for association analysis, of which outcome information for survival analysis was available for 6,129 cases, were obtained from the International Lung Cancer Consortium (ILCCO). Data for 378,484 available participants of European ancestry were obtained from the UK Biobank cohort, as conducted under Application #45611. This was a case-control study with a total of 2,155 incident and prevalent lung cancer cases and 376,329 controls. The details of both cohorts are described in **Supplementary Methods** and in previous studies (12,13). The study was conducted according to the principles of the Declaration of Helsinki. All research participants provided written informed consent, subject to oversight by the Institutional Review Board of all sites.

Gene and genetic variant selection

For genetic association analysis, 43 genes were carefully selected based on their known function in tobacco carcinogenesis metabolism pathways, including PAH and TSNA, as described in our previous studies (14,15). Selection of genes and single nucleotide polymorphisms (SNPs) and the corresponding quality control are described in **Supplementary Methods**.

Causal inference analytic framework

Causal mediation analysis via *med4way* command was implemented in STATA (16). Two-

sample Mendelian Randomization (MR) analysis was conducted using *TwoSampleMR* R package (17). CYP2A6 activity was assessed from the ratio of total trans-3'-hydroxycotinine (3HCOT) to cotinine (COT), as described in a previous study (11). Genetically instrumental variables (IVs) of each exposure [CYP2A6 activity and cigarettes per day (CigDay)] were obtained from previous large-scale GWAS summary statistics (11,18), and the corresponding F-statistic and statistical power of MR were calculated via mRnd online tool (19). The proportion of variance explained by IVs was calculated as in previous studies (20). The available lung cancer GWAS summary statistics as the outcome accompanied by the stratification by ever- and never-smoking status were obtained from ILCCO (12). Both MR Egger intercept test and MR Pleiotropy REsidual Sum and Outlier (MR-PRESSO) global test were used to detect horizontal pleiotropy (21). Reverse-direction MR analysis was also performed to assess potential reverse causal effects. Sobel test (22) was used to evaluate the mediation effect. We calculated the polygenic risk score (PRS) using IVs of CYP2A6 activity, which was used as a surrogate of CYP2A6 activity at the genetically predicted level in UK Biobank for ensuing analysis, as in our previous study (23). Structural equation modelling (SEM) by R package *lavaan* (24) and mediation analysis by R package *mediation* (25) were applied to predict the causal pathway. More details are described in **Supplementary Methods**.

CYP2A6 expression pattern in bulk tissues

CYP2A6 expression in tissues was analyzed in a manner similar to that in our previous studies (26,27) by using Genotype-Tissue Expression (GTEx) project, Human Protein Atlas (HPA), Functional Annotation of The Mammalian Genome (FANTOM5), and The Cancer Genome Atlas (TCGA) Pan-Cancer (PANCAN) cohort, as well as transcriptome in lung tissues between ever-smokers and non-smokers from Gene Expression Omnibus (GEO), including

GSE40419, GSE19667, GSE5058, GSE63127, and GSE7895. In addition, 31 tissues, including 18 lung cancer tissues and 13 matched adjacent tissues, were selected for RNA sequencing from Harvard Lung Cancer Biobank, which was a pilot study of lung cancer transcriptome analysis as a constituent of ILCCO. The Institutional Review Board of Massachusetts General Hospital and the Human Subjects Committee of the Harvard School of Public Health approved the study, and all participants signed consent forms. More details are described in **Supplementary Methods**.

Cigarette smoke exposure cell models

Cigarette smoke exposure cell models using human bronchial epithelial (HBE) cells subjected to 2% cigarette smoke extract (CSE) were carried out as in previous studies (28). The corresponding functional experiments are described in **Supplementary Methods**.

In silico analysis

Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) (29) and PhenomeXcan (30) provided UK Biobank-based resources to annotate the pleiotropy of both rs56113850 and assigned *CYP2A6* on multiple traits or diseases. HaploReg V4.1, FAVOR, and GeneCards were used for the functional annotation of candidate SNPs and genes. More details are described in **Supplementary Methods**.

Statistical analysis

Genetic association analyses were performed using logistic regression models with adjustments of the first three population structure principal components, as reported previously (12), and with adjustments of age, sex, and smoking status if appropriate. Multi-marker Analysis of GenoMic Annotation (MAGMA) was used to enrich the genetic effect of each SNP into a gene set for gene-level association with lung cancer risk based on summary SNP *P* values from a large-scale sample size (31). The *t* test and Wilcoxon rank sum test were

used for differential expression analysis as appropriate. Statistical analyses were performed with R (3.4.2), STATA (15), and PLINK (1.90). More details are described in **Supplementary Methods**.

Data availability

The data generated in this study are available upon request from the corresponding author.

Results

Evaluating genetic effects of tobacco carcinogen-metabolizing enzymes on lung cancer risk

The flowchart of this study is shown in **Supplementary Fig. 1A**. Among 43,483 SNPs located at 43 tobacco carcinogen metabolic genes, 5,423 SNPs passed the quality control; 4,140 of them were defined via the GWAS summary statistics from ILCCO. Notably, 44 SNPs reached statistical significance ($P < 0.05/4,140$; **Supplementary Table 1**), with eight having genome-wide significance ($P < 5\times10^{-8}$; **Supplementary Fig. 1B**) and located at *EPHX2* (rs11780471) and *CYP2A6* (seven SNPs distributed across three of seven linkage disequilibrium blocks; **Supplementary Fig. 1C**).

Deciphering genetic effects of rs56113850 on lung cancer risk by smoking status

Previous studies have indicated that rs56113850 C>T in *CYP2A6* and rs11780471 G>A in *EPHX2* are two well-defined SNPs associated with cigarette consumption and corresponding nicotine metabolism (18,32); therefore, we stratified genetic associations by smoking status. As shown in **Fig. 1A and Table 1**, the genome-wide significant association of rs56113850 with lung cancer risk remained in smokers [odds ratio (OR) = 0.88, 95% confidence interval (CI) = 0.85-0.91, $P = 4.35\times10^{-13}$] but entirely not in non-smokers ($P = 0.924$), with large heterogeneity ($P_{heterogeneity} = 0.002$; $I^2 = 79.8\%$). Subsequently, we carried out interaction analysis using individual-level genetic data for rs56113850 in 14,803 cases and 12,262

controls with smoking information. As expected, there was a significant interaction effect between rs56113850 and smoking status on lung cancer risk ($P_{interaction}$ = 0.028; **Fig. 1B and Supplementary Table 2**), and the protective effect of T allele was greater in smokers (ΔOR = -0.13) than in non-smokers (ΔOR = -0.01; **Fig. 1B**). In contrast, the genetic effect of rs11780471, as well as other SNPs, on lung cancer risk was diminished by stratification of smoking status, without heterogeneity ($P_{heterogeneity}$ = 0.348; I^2 = 0; **Fig. 1A and Table 1**). Moreover, we obtained a similar finding as in single-variant analysis that the genetic effect aggregated at *CYP2A6* gene was significantly associated with lung cancer risk (P < 0.05/43; **Supplementary Fig. 2A and Supplementary Table 3**), especially in smokers but not non-smokers. Independently, there was no association between rs56113850 and lung cancer survival [hazards ratio (HR) = 1.01, 95% CI = 0.97-1.06, P_{Cox} = 0.588, $P_{logrank}$ = 0.830; **Supplementary Fig. 2B**].

Furthermore, we performed four-way decomposition analysis to dissect the genetic effect of rs56113850 on lung cancer risk by smoking status (**Supplementary Table 4**). As illustrated in **Fig. 1C**, the total effect (TE) of rs56113850 was 0.876 (95% CI = 0.845-0.909), which could be divided into four parts: 1) the controlled direct effect (CDE; i.e., effect due to neither mediation nor interaction by fixing smoking status) was 0.857 (95% CI = 0.818-0.898) in smokers but not in never-smokers; 2) the reference interaction (INT_{ref}; i.e., additive interaction effect activated only if smoking status was present when in the presence of rs56113850, capturing interaction only) was 1.034 (95% CI = 1.022-1.046); 3) the mediated interaction (INT_{med}; i.e., additive interaction effect activated only if rs56113850 had an effect on smoking status, capturing both mediation and interaction) was 1.002 (95% CI = 1.001-1.003); and 4) the pure indirect effect (PIE; due to mediation only) via smoking status was 0.987 (95% CI = 0.980-0.994). These findings indicated the protective effect of the T allele of

rs56113850 against lung cancer development in smokers.

Estimating causal cascade of CYP2A6 activity and smoking intensity on lung tumorigenesis

Considering that CYP2A6 is a key enzyme metabolizing tobacco carcinogens (33) and that its activity is dramatically affected by rs56113850 (11) (**Supplementary Table 5**), we conducted causal inference to evaluate the causality of rs56113850, CYP2A6 activity, and smoking intensity on lung cancer risk, with particular emphasis on the smoking population (**Supplementary Fig. 1A**). In terms of summary statistics-based causal inference underlying MR (**Supplementary Fig. 3A**), we observed that high CYP2A6 activity was causally associated with increased smoking intensity (indicated CigDay; β_{IVW} = 0.267, SE = 0.094, $P_{IVW} = 4.61\times10^{-3}$, F-statistic = 275.35, power = 1.00; **Fig. 2A and Supplementary Table 6-7**), and that both high smoking intensity and elevated CYP2A6 activity dramatically increased the causal risk of lung cancer, especially in the smoking population (CYP2A6 activity: β_{IVW} = 0.333, SE = 0.048, $P_{IVW} = 5.67\times10^{-12}$, F-statistic = 331.04, power = 0.81; CigDay: β_{IVW} = 1.026, SE = 0.135, $P_{IVW} = 2.58\times10^{-14}$, F-statistic = 2,505.10, power = 1.00; **Fig. 2B-C and Supplementary Table 6-7**) but not in non-smokers (**Supplementary Fig. 3B-C and Supplementary Table 6-7**). There was no horizontal pleiotropy and no reverse causation among all MR analyses (**Supplementary Table 6**). Intriguingly, subsequent mediation analysis indicated that smoking intensity significantly mediated 82.3% of the causal effect of CYP2A6 activity on lung cancer risk in the smoking population (**Fig. 2D**).

Furthermore, we validated the above finding by using the individual-level genetic data from UK Biobank, including 2,155 cases and 376,329 controls (**Supplementary Table 8**), leveraging SEM and causal mediation. Similarly, rs56113850 remained the significant association with lung cancer risk only in the smoking population (OR = 0.89, 95% CI = 0.83-0.94, $P = 1.08\times10^{-4}$; OR_{meta} = 0.88, 95% CI_{meta} = 0.85-0.91, $P = 2.18\times10^{-16}$; **Table 1**). Moreover,

when including rs56113850 genotypes, CYP2A6 activity (surrogated by CYP2A6 PRS), and smoking intensity (indicated by pack-year of smoking) in SEM, we found that the effect of rs56113850 on lung cancer risk was totally amended through the pathway of CYP2A6 activity to smoking intensity (**Supplementary Fig. 3D**); and in the subsequent causal mediation, smoking intensity significantly mediated a 15.3% effect of genetically predicted CYP2A6 activity on the risk of lung cancer in the smoking population (**Supplementary Fig. 3E**).

Expression pattern of CYP2A6 and biological function of rs56113850 in lung tumorigenesis

Next, we detected the expression pattern of *CYP2A6* across human tissues and cells. *CYP2A6* was well expressed in the liver but relatively low in the lung derived from normal tissues of HPA, GTEx, FANTOM5, and TCGA (**Fig. 3A-B**) and was significantly decreased in tumors of liver and lung compared with the corresponding normal tissues from TCGA and Harvard Biobank datasets (**Fig. 3C**). At the tissue level, pulmonary *CYP2A6* expression was significantly downregulated in ever-smokers compared with non-smokers across each dataset (**Supplementary Fig. 4A**), and the following meta-analysis showed significantly and substantially decreased *CYP2A6* expression by 24% in ever-smokers from TCGA and GEO datasets (95% CI = 16-31%; $P = 2.71\times10^{-9}$; **Fig. 3D**). Similarly in 2% CSE-exposed HBE cell models, *CYP2A6* expression at both RNA and protein levels and its activity were downregulated compared with that observed in untreated cells (**Fig. 3E**); In pleiotropy analysis using the phenome-wide association study (PheWAS) strategy, *CYP2A6* expression in both liver and lung correlated significantly with more than 100 phenotypes, specifically those clustered into smoking status or lung-relevant traits in accordance with the above findings (**Supplementary Table 9**).

In terms of the genetic regulation, we observed high function scores of seven at-risk

SNPs in *CYP2A6* across multiple categories according to two functional annotation tools (**Supplementary Table 10**). Preferentially, we included rs56113850 for further function study not only for its top genetic association, but also it had high scores of protein function and local nucleotide diversity, along with five altered motifs. In addition, we found that the T allele of rs56113850 significantly decreased *CYP2A6* expression across tissues, especially in both lung and liver tissues (**Supplementary Fig. 4B**). Nevertheless, the T allele significantly rescued the downregulation of *CYP2A6* expression and activity caused by 2% CSE exposure when compared with the C allele (**Fig. 3F and Supplementary Fig. 4C**). Of note, helicase-like transcription factor (HLTF), one of five motifs assigned to TFs and involved in DNA damage/repair, preferentially bound to the rs56113850 T allele (**Supplementary Table 11**). Given this, we performed super-electrophoretic mobility shift assays with HLTF-containing nuclear extracts to independently verify the genetic regulation of rs56113850 on the TF binding. The supershift assays showed the preference of HLTF for the rs56113850 T allele probe (**Fig. 3G**). Taken together, this new evidence provided additional support for the regulatory function of rs56113850 and suggested the involvement of the transcription factor HLTF in mediating genetic effects on lung cancer susceptibility. Moreover, rs56113850 was dramatically associated with risks of cancer of the respiratory system and of the bronchus based on PheWAS scanning (**Supplementary Fig. 4D**).

Discussion

In this study, we found that rs56113850 played a critical role in affecting smoking-related lung cancer risk through both the cascade effect on CYP2A6 metabolic capacity to cigarette consumption and the genetic function on CYP2A6 activity against smoke exposure (**Fig. 3H**). These findings provide knowledge for cancer interventions based on susceptible populations.

Cigarette smoking is a heritable but modifiable social behavior related to various diseases, with 8% of SNP heritability for CigDay (18). Notably, CYP2A6 is a highly polymorphic and heritable biomarker, and its genetic variation dramatically modifies the genetic correlation between CigDay and lung cancer risk (33,34). This finding might be due to two metabolic pathways (i.e., metabolism and subsequent excretion of nicotine and simultaneous activation of TSNA) involved in the CYP2A6 enzyme.

Nicotine is the main psychoactive component in tobacco, producing temporary pleasurable effects in the brain (35). The nicotine metabolite ratio (NMR; ratio of 3HCOT/COT) is an established index of nicotine metabolic inactivated mainly by the CYP2A6 enzyme, which represents CYP2A6 activity in this study. Extensive research has emphasized that a higher NMR indicates higher CYP2A6 activity and faster nicotine inactivation (11,32), resulting in greater cigarette consumption and lower rates of smoking cessation (36,37). Furthermore, CYP2A6 activity is independently associated with increased lung cancer risk (38). In the causal inference analytic framework of this study, we advanced this observed association to a quantitative causal relationship between higher CYP2A6 activity and greater CigDay nicotine uptake, both causally and quantitatively increased the risk of lung cancer in smokers. This is likely due to the influence of genetics (on the nature side) on smoking behavior (on the nurture side), as NMR is dramatically heritable in nicotine metabolism with a heritability estimate of 81%; of note, rs56113850 in *CYP2A6* alone explains a considerable proportion (14-23%) of NMR variance (39). The data from this study suggest a causal inference that smoking status interacts and mediates the effect of rs56113850 on lung cancer risk as individuals carrying the rs56113850 T allele exhibit downregulated *CYP2A6* expression and activity, resulting in lower NMR, reduced smoking intensity, and lower exposure to tobacco carcinogens.

TSNAs are a class of procarcinogens known to be bioactivated by the CYP2A6 enzyme. N'-Nitrosonornicotine (NNN) and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) are the two most potent TSNAs present in unburned tobacco and tobacco smoke (6,40). They readily cause tumors in animal models and are classified by International Agency for Research on Cancer as "carcinogenic to humans". Population-based studies have revealed an associated cascade of CYP2A6 activity, TSNA bioactivation, and smoking-related lung cancer risk (41,42). The data from this study are consistent with the findings that smokers with lower CYP2A6 activity due to the presence of the rs56113850 T allele are exposed to less levels of carcinogens overall in tobacco smoke, including lower level of TSNA bioactivation, and hence have a decreased risk of lung cancer.

NNN and NNK both form DNA adducts, which are misrepaired or not repaired to constitute a necessary, though not sufficient, prerequisite for inducing cancer (43). It is worth noting that a balance between DNA adduct formation and removal exists because of the highly variable capacity of DNA adducts to induce DNA damage, including mutations and chromosomal aberrations (43). At both tissue and cell levels, we observed downregulated *CYP2A6* expression after smoke exposure, consistent with the findings of Gao *et al*. (44). These results suggest that DNA damage of *CYP2A6* occurs simultaneously in both the target organ (lung) and the metabolizing organ (liver) during carcinogenesis. In this study, we observed preferential binding capacity of the transcription factor HLTF at the rs56113850 T allele. HLTF plays a critical role in error-free post-replication repair of damaged DNA, maintaining genomic stability by acting as a ubiquitin ligase for 'Lys-63'-linked polyubiquitination of chromatin-bound proliferating cell nuclear antigen (45,46). Additionally, HLTF is inactivated in tumorigenesis due to promoter hypermethylation and truncated protein forms lacking functional domains, serving as a biomarker for lung cancer

prognosis (47,48). The data of the present study provide biological knowledge of the protective role of the rs56113850 T allele on smoking-related lung cancer risk by driving DNA repair of *CYP2A6* against smoke carcinogens via HLTF recruitment.

We acknowledge several limitations in this study. First, conclusive confirmation of all causal effects from MR may require a well-powered prospective cohort study or a well-designed randomized controlled trial of preventive interventions, especially that includes individual genetic data and CYP2A6 activity detection. Second, it remains to be determined whether many other genetic variants (such as indels) of tobacco carcinogen metabolic genes that were absent from the GWAS platform, far outside the -/+ 5 kb region, or less conserved based on association analysis also regulate relevant gene expression. Other driver genes, including but not limited to tumor-suppressor genes and transcription factors (e.g., *HER2*, *BRAF*, *PTEN*, *FGFR1*, *SOX2*), may also be causally related to smoking-related lung cancer. Thus, whole-genome and whole-exome sequencing based on next-generation sequencing technologies should be applied to identify novel driver genes and causal variants for lung cancer. Third, the sample size of the never-smokers is an order magnitude lower than for ever-smokers, which suggests that some of the effects observed may reflect differences in sample size, rather than true effects based on smoking status. Therefore, a large-scale population study focusing on never-smokers is essential to elucidate the genetic heterogeneity underlying lung cancer susceptibility. Forth, the direct or indirect biological mechanisms of reduced *CYP2A6* expression by cigarette smoke exposure and in tumors remain unclear. Accordingly, comprehensive biological evidence, potentially at multiple omics levels (e.g., abnormal DNA methylation as well as dysregulation of gene expression, protein expression, and protein activity by cigarette smoking) and in various models (e.g., multiple cell types or smoke mouse models), is essential to convincingly demonstrate such

an underlying mechanism.

In conclusion, rs56113850 and *CYP2A6* gene are causally associated with lung cancer risk depending on smoking status and intensity. These findings may bridge the gap between host susceptibility and individual behaviors for the biological interpretation of cancer prevention.

Acknowledgements

We thank the individuals who have contributed their samples and clinical data for this study and we also thank ILCCO members, who provided access to samples and clinical data.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A*, et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin **2021**
2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin **2023**;73:17-48
3. Islami F, Goding Sauer A, Miller KD, Siegel RL, Fedewa SA, Jacobs EJ*, et al.* Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. CA Cancer J Clin **2018**;68:31-54
4. Cornelius ME, Loretan CG, Jamal A, Davis Lynn BC, Mayer M, Alcantara IC*, et al.* Tobacco Product Use Among Adults - United States, 2021. MMWR Morb Mortal Wkly Rep **2023**;72:475-83
5. Le Foll B, Piper ME, Fowler CD, Tonstad S, Bierut L, Lu L*, et al.* Tobacco and nicotine use. Nat Rev Dis Primers **2022**;8:19
6. Soleimani F, Dobaradaran S, De-la-Torre GE, Schmidt TC, Saeedi R. Content of toxic components of cigarette, cigarette smoke vs cigarette butts: A comprehensive systematic review. Sci Total Environ **2022**;813:152667
7. Sarlak S, Lalou C, Amoedo ND, Rossignol R. Metabolic reprogramming by tobacco-specific nitrosamines (TSNAs) in cancer. Semin Cell Dev Biol **2020**;98:154-66
8. Moorthy B, Chu C, Carlin DJ. Polycyclic aromatic hydrocarbons: from metabolism to lung cancer. Toxicol Sci **2015**;145:5-15
9. Chu H, Xin J, Yuan Q, Wu Y, Du M, Zheng R*, et al.* A prospective study of the associations among fine particulate matter, genetic variants, and the risk of colorectal cancer. Environ Int **2021**;147:106309
10. Wang X, Ricciuti B, Nguyen T, Li X, Rabin MS, Awad MM*, et al.* Association between smoking history and tumor mutation burden in advanced non-small cell lung cancer. Cancer Res **2021**
11. Patel YM, Park SL, Han Y, Wilkens LR, Bickeboller H, Rosenberger A*, et al.* Novel Association of Genetic Markers Affecting CYP2A6 Activity and Lung Cancer Risk. Cancer Res **2016**;76:5768-76
12. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC*, et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet **2017**;49:1126-32
13. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J*, et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med **2015**;12:e1001779
14. Modesto JL, Hull A, Angstadt AY, Berg A, Gallagher CJ, Lazarus P*, et al.* NNK reduction pathway gene polymorphisms and risk of lung cancer. Mol Carcinog **2015**;54 Suppl 1:E94-E102
15. Liu H, Li G, Sturgis EM, Shete S, Dahlstrom KR, Du M*, et al.* Genetic variants in CYP2B6 and HSD17B12 associated with risk of squamous cell carcinoma of the head and neck. Int J Cancer **2022**;151:553-64
16. Discacciati A, Bellavia A, Lee JJ, Mazumdar M, Valeri L. Med4way: a Stata command to investigate mediating and interactive mechanisms using the four-way effect decomposition. Int J Epidemiol **2018**
17. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D*, et al.* The MR-Base platform supports systematic causal inference across the human phenome. Elife **2018**;7
18. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F*, et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet **2019**;51:237-44
19. Brion MJ, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian

randomization studies. Int J Epidemiol **2013**;42:1497-501
20. Yuan Z, Liu L, Guo P, Yan R, Xue F, Zhou X. Likelihood-based Mendelian randomization analysis with automated instrument selection and horizontal pleiotropic modeling. Sci Adv **2022**;8:eabl5744
21. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. Hum Mol Genet **2018**;27:R195-R208
22. Preacher KJ, Leonardelli GJ. Calculation for the Sobel test. Retrieved January **2001**;20:2009
23. Du M, Garcia JGN, Christie JD, Xin J, Cai G, Meyer NJ*, et al.* Integrative omics provide biological and clinical insights into acute respiratory distress syndrome. Intensive Care Med **2021**;47:761-71
24. Rosseel Y. lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software **2012**;48:1 - 36
25. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. Journal of Statistical Software **2014**;59:1 - 38
26. Du M, Cai G, Chen F, Christiani DC, Zhang Z, Wang M. Multiomics Evaluation of Gastrointestinal and Other Clinical Characteristics of COVID-19. Gastroenterology **2020**;158:2298-301 e7
27. Cai G, Bosse Y, Xiao F, Kheradmand F, Amos CI. Tobacco Smoking Increases the Lung Gene Expression of ACE2, the Receptor of SARS-CoV-2. Am J Respir Crit Care Med **2020**;201:1557-9
28. Ma H, Lu L, Xia H, Xiang Q, Sun J, Xue J*, et al.* Circ0061052 regulation of FoxC1/Snail pathway via miR-515-5p is involved in the epithelial-mesenchymal transition of epithelial cells during cigarette smoke-induced airway remodeling. Sci Total Environ **2020**;746:141181
29. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN*, et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet **2018**;50:1335-41
30. Pividori M, Rajagopal PS, Barbeira A, Liang Y, Melia O, Bastarache L*, et al.* PhenomeXcan: Mapping the genome to the phenome through the transcriptome. bioRxiv **2019**:833210
31. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol **2015**;11:e1004219
32. Buchwald J, Chenoweth MJ, Palviainen T, Zhu G, Benner C, Gordon S*, et al.* Genome-wide association meta-analysis of nicotine metabolism and cigarette consumption measures in smokers of European descent. Mol Psychiatry **2020**
33. Tanner JA, Tyndale RF. Variation in CYP2A6 Activity and Personalized Medicine. J Pers Med **2017**;7
34. Bray MJ, Chen LS, Fox L, Hancock DB, Culverhouse RC, Hartz SM*, et al.* Dissecting the genetic overlap of smoking behaviors, lung cancer, and chronic obstructive pulmonary disease: A focus on nicotinic receptors and nicotine metabolizing enzyme. Genet Epidemiol **2020**;44:748-58
35. Aubin HJ, Rollema H, Svensson TH, Winterer G. Smoking, quitting, and psychiatric disease: a review. Neurosci Biobehav Rev **2012**;36:271-84
36. Chenoweth MJ, Tyndale RF. Pharmacogenetic Optimization of Smoking Cessation Treatment. Trends Pharmacol Sci **2017**;38:55-66
37. Lerman C, Schnoll RA, Hawk LW, Jr., Cinciripini P, George TP, Wileyto EP*, et al.* Use of the nicotine metabolite ratio as a genetically informed biomarker of response to nicotine patch or varenicline for smoking cessation: a randomised, double-blind placebo-controlled trial. Lancet Respir Med **2015**;3:131-8
38. Park SL, Murphy SE, Wilkens LR, Stram DO, Hecht SS, Le Marchand L. Association of CYP2A6 activity with lung cancer incidence in smokers: The multiethnic cohort study. PLoS One **2017**;12:e0178435
39. Loukola A, Buchwald J, Gupta R, Palviainen T, Hallfors J, Tikkanen E*, et al.* A Genome-Wide Association Study of a Biomarker of Nicotine Metabolism. PLoS Genet **2015**;11:e1005498

40. Xia B, Blount BC, Guillot T, Brosius C, Li Y, Van Bemmel DM*, et al.* Tobacco-Specific Nitrosamines (NNAL, NNN, NAT, and NAB) Exposures in the US Population Assessment of Tobacco and Health (PATH) Study Wave 1 (2013-2014). Nicotine Tob Res **2021**;23:573-83
41. Zhu AZ, Binnington MJ, Renner CC, Lanier AP, Hatsukami DK, Stepanov I*, et al.* Alaska Native smokers and smokeless tobacco users with slower CYP2A6 activity have lower tobacco consumption, lower tobacco-specific nitrosamine exposure and lower tobacco-specific nitrosamine bioactivation. Carcinogenesis **2013**;34:93-101
42. Murphy SE, Park SL, Balbo S, Haiman CA, Hatsukami DK, Patel Y*, et al.* Tobacco biomarkers and genetic/epigenetic analysis to investigate ethnic/racial differences in lung cancer risk among smokers. NPJ Precis Oncol **2018**;2:17
43. Li Y, Hecht SS. Metabolism and DNA Adduct Formation of Tobacco-Specific N-Nitrosamines. Int J Mol Sci **2022**;23
44. Gao Y, Miksys S, Palmour RM, Tyndale RF. The Influence of Tobacco Smoke/Nicotine on CYP2A Expression in Human and African Green Monkey Lungs. Mol Pharmacol **2020**;98:658-68
45. Gallo D, Brown GW. Post-replication repair: Rad5/HLTF regulation, activity on undamaged templates, and relationship to cancer. Crit Rev Biochem Mol Biol **2019**;54:301-32
46. Elserafy M, Abugable AA, Atteya R, El-Khamisy SF. Rad5, HLTF, and SHPRH: A Fresh View of an Old Story. Trends Genet **2018**;34:574-7
47. Dhont L, Mascaux C, Belayew A. The helicase-like transcription factor (HLTF) in cancer: loss of function or oncomorphic conversion of a tumor suppressor? Cell Mol Life Sci **2016**;73:129-47
48. Dhont L, Pintilie M, Kaufman E, Navab R, Tam S, Burny A*, et al.* Helicase-like transcription factor expression is associated with a poor prognosis in Non-Small-Cell Lung Cancer (NSCLC). BMC Cancer **2018**;18:429

Table 1: Associations of two genetic variants in tobacco carcinogen metabolic genes achieving genome-wide significance with lung cancer risk.

SNP	CHR	Position	Reference/ Effect Allele	Gene	Study	Population	Cases	Controls	EAF	OR (95% CI)	P^{*}	$P_{het}{}^{**}$	I^{2**}	$P_{het}{}^{***}$	I^{2***}
rs11780471	8	27344719	G/A	3.6 kb 5' of *EPHX2*	ILCCO	Overall	29,266	56,450	0.060	0.87 (0.83-0.91)	1.69×10^{-8}	0.673	0.0		
						Ever-smoker	23,223	16,964	0.062	0.86 (0.81-0.92)	4.75×10^{-6}	0.251	22.4		
						Never-smoker	2,355	7,504	0.066	0.93 (0.81-1.07)	0.312	0.212	28.3	0.348	0.0
rs56113850	19	41353107	C/T	*CYP2A6* intronic	ILCCO	Overall	25,583	51,525	0.440	0.88 (0.86-0.91)	5.02×10^{-19}	0.511	0.0		
						Ever-smoker	19,706	13,322	0.435	0.88 (0.85-0.91)	4.35×10^{-13}	0.954	0.0		
						Never-smoker	2,196	6,251	0.437	1.00 (0.93-1.09)	0.924	0.963	0.0	0.002	79.8
					UK Biobank	Overall	2,155	376,329	0.393	0.89 (0.83-0.94)	1.08×10^{-4}				
						Ever-smoker	1,843	169,535	0.392	0.88 (0.82-0.94)	1.83×10^{-4}				
						Never-smoker	300	205,509	0.400	0.91 (0.78-1.08)	0.286			0.710	0.0
					ILCCO and UK Biobank	Overall				0.88 (0.86-0.90)	8.69×10^{-22}	0.746	0.0		
						Ever-smoker				0.88 (0.85-0.91)	2.18×10^{-16}	1.000	0.0		
						Never-smoker				0.98 (0.91-1.05)	0.612	0.307	4.1	0.006	86.7

SNP, single nucleotide polymorphism; CHR, chromosome; Position was mapped to GRCh37; EAF, effect allele frequency in all samples; OR, odds ratio; CI, confidence interval.

* *P* was obtained from lung cancer GWAS summary statistics of ILCCO and individual-level genetic data of UK Biobank, which was calculated using logistic regression model with adjustments of the first three population structure principal components, age, sex, and smoking status if appropriate.

** P_{het}, *P* for heterogeneity among the included sub-studies in ILCCO, as well as the meta-analysis of ILCCO and UK Biobank, along with I^2 (%).

*** P_{het}, *P* for heterogeneity between ever smoking and never smoking populations, along with I^2 (%).

Figure 1: Decomposition of the genetic effect of rs56113850 in *CYP2A6* on lung cancer risk by smoking status. (A) Manhattan plot for genetic effects of tobacco carcinogen metabolic genes on lung cancer risk stratified by smoking status. The X-axis represents each chromosome, with different colors assigned to each gene; the Y-axis represents association *P* values (-log10 transformed) with lung cancer risk, derived from lung cancer GWAS summary statistics in subgroups of smoking populations deposited in ILCCO; the red dashed horizontal line indicates a *P* value equal to GWAS significance at 5×10^{-8}. (B) Interaction effects between rs56113850 in *CYP2A6* and smoking status on lung cancer risk. Genotyping data of rs56113850 were acquired from ILCCO for 14,803 cases and 12,262 controls with individual smoking information. OR, odds ratio, calculated via logistic regression model underlying joint analysis approach. (C) Four-way decomposition analysis of the rs56113850 effect on lung cancer risk potentially mediated by smoking status. Y is the outcome: lung cancer; A is the exposure: rs56113850 genotypes obtained from ILCCO; M is the potential mediator: smoking status. The OR and corresponding 95% CI were calculated by mediation analysis with causal effects estimated for exposure and at the mean level of covariates. The controlled direct effect and the reference interaction were computed by fixing smoking status as never (M=0) or ever (M=1).

Figure 2: Causal inference for the causal pathway of CYP2A6 activity, cigarette consumption, and lung cancer risk. (A-C) Scatter plots for genetic associations across CYP2A6 activity, cigarettes per day, and lung cancer risk in the smoking population. The X-axis represents the per allele association of exposure-relevant SNPs and assigned outcomes, with the likelihood-based Mendelian Randomization estimate for genetic instrumental variables. (D) Directed acyclic graph for the causal mediation pathway of CYP2A6 activity, cigarettes per day, and lung cancer risk in the smoking population. E, Exposure; M, Mediator; O, Outcome.

IE, indirect effect. The Sobel test was used to evaluate the mediation effect upon the causal effect derived from the Mendelian Randomization estimate.

Figure 3: Expression pattern of *CYP2A6* across tissues, smoking status, and allele-specific manners. (A) *CYP2A6* expression pattern in the top 10 tissues using the consensus normalized expression value (NX) derived from HPA, GTEx, and FANTOM5. NX_{liver} = 199.5; NX_{lung} = 0.2; NX_{others} = 0. (B) *CYP2A6* expression pattern in normal tissues derived from TCGA PANCAN. The X-axis is assigned to tumor type, including BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; ESCA, esophageal carcinoma; HNSC, head and neck squamous cell carcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; NA, not available; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SARC, sarcoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma; THYM, thymoma; UCEC, uterine corpus endometrial carcinoma. The Y-axis represents *CYP2A6* normalized expression. (C) Differential expression analyses of *CYP2A6* between tumor and normal tissues derived from publicly available TCGA PANCAN (lung and liver cancers; left) and Harvard Biobank (lung cancer; right). An unpaired *t* test was applied for comparison of *CYP2A6* expression between tumor and normal tissues. (D) Forest plot for the effect of smoking status on *CYP2A6* pulmonary gene expression. The effect size of smoking status (ever-smoker vs. non-smoker) on *CYP2A6* expression was calculated via linear regression model, accompanied by the 95% CI. The size of the square is proportional to the weight, which is estimated by the standard

"inverse-variance" method for random-effects models in meta-analysis. (E) *CYP2A6* expression pattern at levels of RNA, protein, and activity after 2% CSE exposure in HBE cell line. Gene expression was normalized to that in cells treated with DMSO. An unpaired *t* test was applied for the group comparison. All experiments were performed in three biological replicates with three technical replicates each. (F) Allele-specific effect of rs56113850 on *CYP2A6* expression pattern at levels of RNA, protein, and activity after 2% CSE exposure in HBE cell line. Allele-specific constructs containing the putative activity region flanking rs56113850 were cloned into the pcDNA3.1-basic vector and transfected into HBE cells. Gene expression was normalized to that in cells treated with DMSO. All experiments were performed in three biological replicates with three technical replicates each. (G) Allele-specific effect of rs56113850 on TF HLTF binding affinity through super-electrophoretic mobility shift assays. (H) Graphical representation of the findings of this study. In smokers, a causal pathway model for relationships among *CYP2A6* variants (rs56113850 C>T included), CYP2A6 activity, smoking intensity, and lung cancer risk exists, which may be biologically interpreted by nicotine metabolism (indicated by NMR) and TSNA activation after CYP2A6 activity induced by cigarette smoke exposure in lung tumorigenesis.

Figure 1

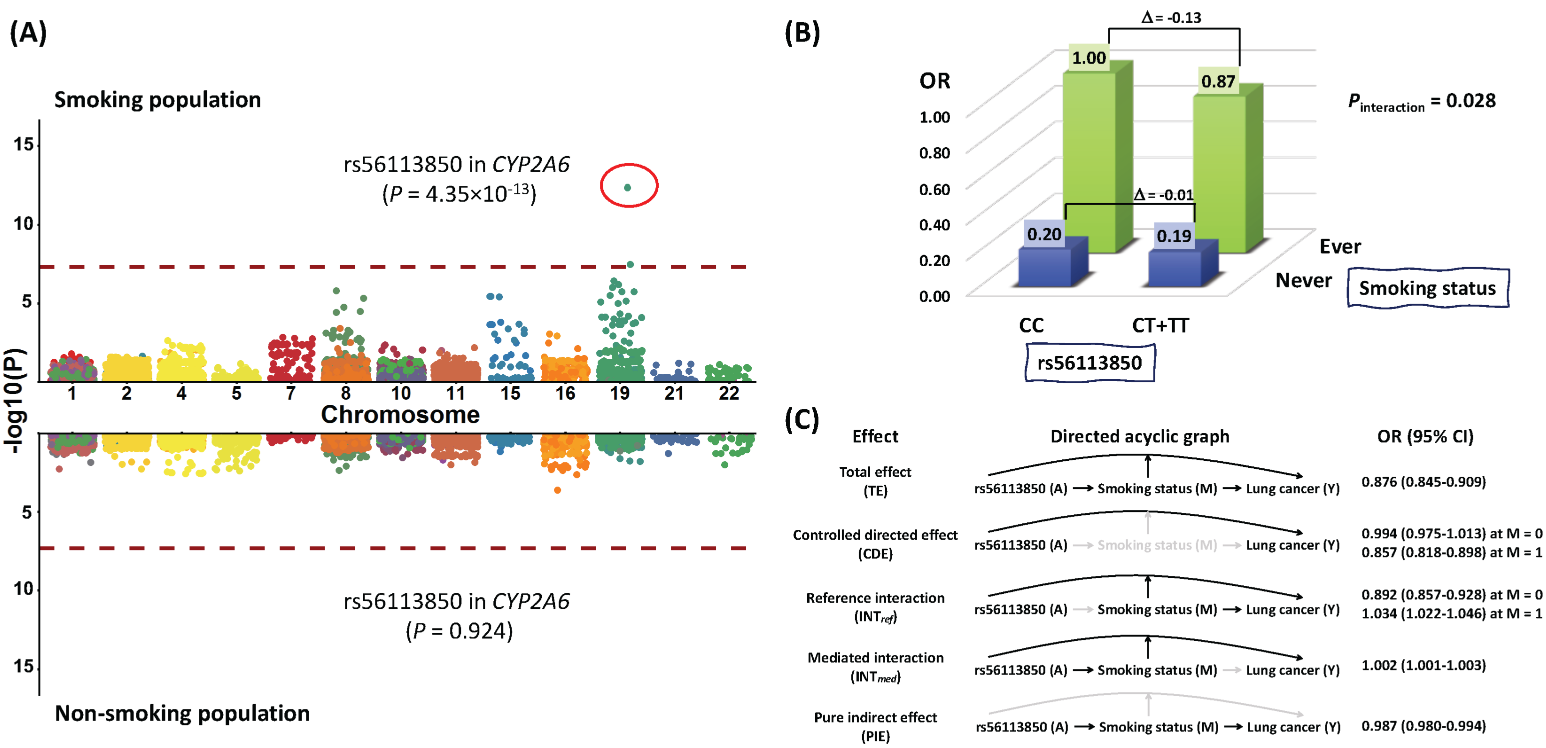

(A)
Smoking population
rs56113850 in CYP2A6
(P = 4.35×10-13)
15
10
5
-log10(P)
1
2
4
5
7
8
10
11
15
16
19
21
22
Chromosome
5
10
15
rs56113850 in CYP2A6
(P = 0.924)
Non-smoking population
(B)
Δ = -0.13
OR
1.00
0.87
Pinteraction = 0.028
1.00
0.80
0.60
0.40
0.20
0.00
Δ = -0.01
0.20
0.19
Ever
Never
Smoking status
CC
CT+TT
rs56113850
(C)
Effect
Directed acyclic graph
OR (95% CI)
Total effect (TE)
rs56113850 (A) → Smoking status (M) → Lung cancer (Y)
0.876 (0.845-0.909)
Controlled directed effect (CDE)
rs56113850 (A) → Smoking status (M) → Lung cancer (Y)
0.994 (0.975-1.013) at M = 0
0.857 (0.818-0.898) at M = 1
Reference interaction (INTref)
rs56113850 (A) → Smoking status (M) → Lung cancer (Y)
0.892 (0.857-0.928) at M = 0
1.034 (1.022-1.046) at M = 1
Mediated interaction (INTmed)
rs56113850 (A) → Smoking status (M) → Lung cancer (Y)
1.002 (1.001-1.003)
Pure indirect effect (PIE)
rs56113850 (A) → Smoking status (M) → Lung cancer (Y)
0.987 (0.980-0.994)

Figure 2

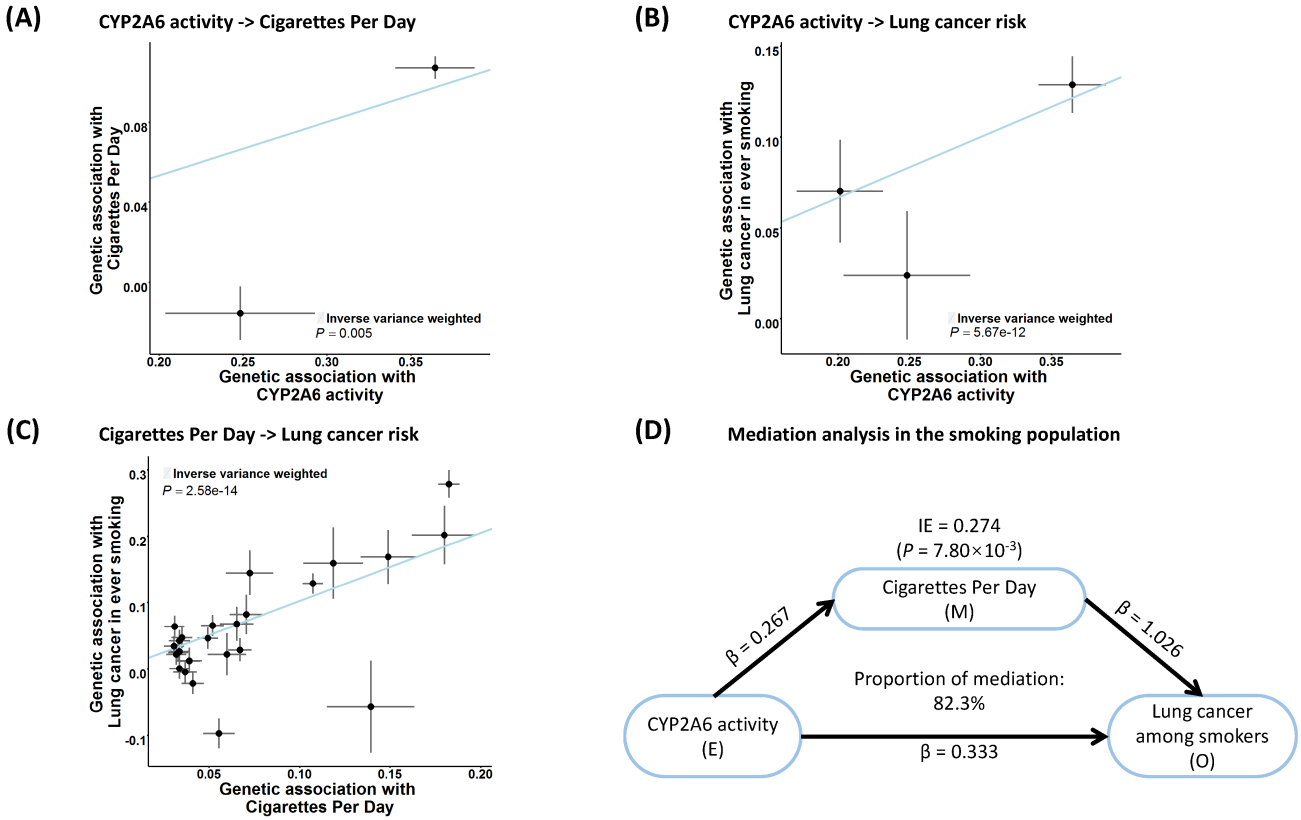
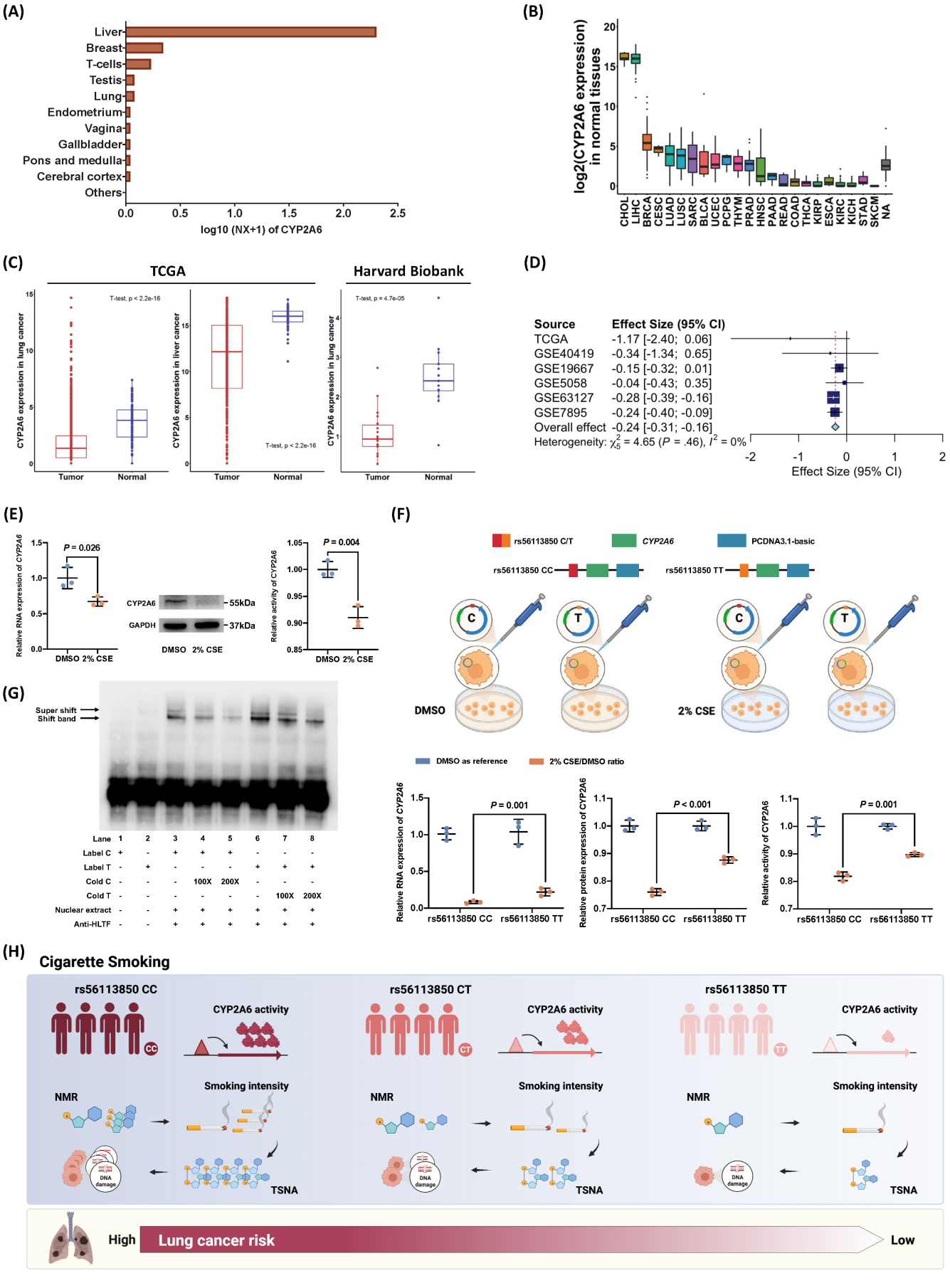

(A)
CYP2A6 activity -> Cigarettes Per Day
Genetic association with Cigarettes Per Day
Genetic association with CYP2A6 activity
Inverse variance weighted
P = 0.005
(B)
CYP2A6 activity -> Lung cancer risk
Genetic association with Lung cancer in ever smoking
Genetic association with CYP2A6 activity
Inverse variance weighted
P = 5.67e-12
(C)
Cigarettes Per Day -> Lung cancer risk
Genetic association with Lung cancer in ever smoking
Genetic association with Cigarettes Per Day
Inverse variance weighted
P = 2.58e-14
(D)
Mediation analysis in the smoking population
IE = 0.274
(P = 7.80×10-3)
Cigarettes Per Day
(M)
β = 0.267
β = 1.026
Proportion of mediation:
82.3%
CYP2A6 activity
(E)
β = 0.333
Lung cancer
among smokers
(O)

Figure 3


(A)
Liver
Breast
T-cells
Testis
Lung
Endometrium
Vagina
Gallbladder
Pons and medulla
Cerebral cortex
Others
0.0 0.5 1.0 1.5 2.0 2.5
log10 (NX+1) of CYP2A6
(B)
log2(CYP2A6 expression) in normal tissues
0 5 10 15 20
CHOL LIHC BRCA CESC LUAD LUSC SARC BLCA UCEC PCPG THYM PRAD HNSC PAAD READ COAD THCA KIRP ESCA KIRC KICH STAD SKCM NA
(C)
TCGA
Harvard Biobank
CYP2A6 expression in lung cancer
CYP2A6 expression in liver cancer
T-test, p < 2.2e-16
T-test, p = 4.7e-05
Tumor
Normal
(D)
Source Effect Size (95% CI)
TCGA -1.17 [-2.40; 0.06]
GSE40419 -0.34 [-1.34; 0.65]
GSE19667 -0.15 [-0.32; 0.01]
GSE5058 -0.04 [-0.43; 0.35]
GSE63127 -0.28 [-0.39; -0.16]
GSE7895 -0.24 [-0.40; -0.09]
Overall effect -0.24 [-0.31; -0.16]
Heterogeneity: χ^2_5 = 4.65 (P = .46), I^2 = 0%
-2 -1 0 1 2
Effect Size (95% CI)
(E)
Relative RNA expression of CYP2A6
P = 0.026
DMSO 2% CSE
CYP2A6 55kDa
GAPDH 37kDa
Relative activity of CYP2A6
P = 0.004
(F)
rs56113850 C/T
CYP2A6
PCDNA3.1-basic
rs56113850 CC
rs56113850 TT
DMSO
2% CSE
DMSO as reference
2% CSE/DMSO ratio
Relative RNA expression of CYP2A6
P = 0.001
Relative protein expression of CYP2A6
P < 0.001
Relative activity of CYP2A6
P = 0.001
rs56113850 CC
rs56113850 TT
(G)
Super shift
Shift band
Lane 1 2 3 4 5 6 7 8
Label C + - + + + - - -
Label T - + - - - + + +
Cold C - - - 100X 200X - - -
Cold T - - - - - - 100X 200X
Nuclear extract - - + + + + + +
Anti-HLTF - - + + + + + +
(H)
Cigarette Smoking
rs56113850 CC
rs56113850 CT
rs56113850 TT
CYP2A6 activity
NMR
Smoking intensity
DNA damage
TSNA
High
Lung cancer risk
Low