Lung cancer in ever- and never-smokers: findings from multi-population GWAS studies

Yafang Li1-3\*, Xiangjun Xiao1, Jianrong Li1, Younghun Han1-2, Chao Cheng1-3, Gail F Fernandes3-4, Shannon E Slewitzke1,3-4, Susan M Rosenberg3-4, Meng Zhu5, Jinyoung Byun1-2, Yohan Bossé6, James McKay7, Demetrios Albanes8, Stephen Lam9, Adonina Tardon10, Chu Chen11, Stig E Bojesen12-13, Maria T Landi8, Mattias Johansson7, Angela Risch14-16, Heike Bickeböller17, H-Erich Wichmann18, David C Christiani19, Gad Rennert20, Susanne Arnold21, Gary Goodman22, John K Field23, Michael PA Davies24, Sanjay S Shete24-25, Loic Le Marchand26, Geoffrey Liu27, Rayjean J Hung28-29, Angeline S Andrew30, Lambertus A Kiemeney31, Ryan Sun24, Shan Zienolddiny32, Kjell Grankvist33, Mikael Johansson34, Neil Caporaso7, Dawn M Teare35, Yun-Chul Hong36, Philip Lazarus37, Matthew B Schabath38, Melinda C Aldrich39, Ann G Schwartz40-41, Ivan Gorlov1-3, Kristen Purrington40-41, Ping Yang42, Yanhong Liu2-3, , Joan E Bailey-Wilson43, Susan M Pinney44, Diptasri Mandal45, James C Willey46, Colette Gaba47, Paul Brennan6, Jun Xia1,3-4, Hongbing Shen5, Christopher I Amos1-3\*. INTEGRAL-ILCCO lung cancer consortium.

\*Yafang Li and Christopher I Amos are the corresponding authors.

Yafang Li

Baylor College of Medicine

1 Baylor Plaza, Houston, TX 77030

Tel: 713-798-2804

Email: [yafang.li@bcm.edu](mailto:yafang.li@bcm.edu)

Christopher I. Amos

Baylor College of Medicine

1 Baylor Plaza, Houston, TX 77030

Tel: 713-798-2102

Email: chris.amos@bcm.edu

1. Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX.
2. Section of Epidemiology and Population Sciences, Department of Medicine, Baylor College of Medicine, Houston, TX.
3. Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX.
4. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.
5. 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, P.R. China.
6. Institut universitaire de cardiologie et de pneumologie de Québec, Department of Molecular Medicine, Laval University, Quebec City, Canada.
7. Section of Genetics, International Agency for Research on Cancer, World Health Organization, Lyon, France.
8. Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD.
9. Department of Integrative Oncology, University of British Columbia, Vancouver, BC, Canada.
10. Public Health Department, University of Oviedo, ISPA and CIBERESP, Asturias, Spain.
11. Program in Epidemiology, Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA.
12. Department of Clinical Biochemistry, Copenhagen University Hospital, Copenhagen, Denmark.
13. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
14. Thoraxklinik at University Hospital Heidelberg, Heidelberg, Germany.
15. Translational Lung Research Center Heidelberg (TLRC-H), Heidelberg, Germany.
16. University of Salzburg and Cancer Cluster Salzburg, Austria.
17. Department of Genetic Epidemiology, University Medical Center, Georg-August-University Göttingen, Germany.
18. Institute of Epidemiology, Helmholtz Munich, Germany.
19. Departments of Environmental Health and Epidemiology, Harvard TH Chan School of Public Health, Boston, MA.
20. Clalit National Cancer Control Center at Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel.
21. University of Kentucky, Markey Cancer Center, Lexington, Kentucky, USA.
22. Swedish Cancer Institute, Seattle, WA, USA.
23. Institute of Translational Medicine, University of Liverpool, Liverpool, United Kingdom.
24. Department of Biostatistics, The University of Texas, M.D. Anderson Cancer Center, Houston, TX.
25. Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX USA.
26. Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA.
27. University Health Network- The Princess Margaret Cancer Centre, Toronto, CA.
28. Luenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada.
29. Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Canada.
30. Departments of Epidemiology and Community and Family Medicine, Dartmouth College, Hanover, NH.
31. Radboud University Medical Center, Nijmegen, The Netherlands.
32. National Institute of Occupational Health, Oslo, Norway.
33. Department of Medical Biosciences, Umeå University, Umeå, Sweden.
34. Department of Radiation Sciences, Umeå University, Umeå, Sweden.
35. Newcastle University.
36. Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea.
37. Department of Pharmaceutical Sciences, College of Pharmacy and Pharmaceutical Sciences, Washington State University, Spokane, Washington, USA.
38. Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL.
39. Department of Thoracic Surgery, Division of Epidemiology, Vanderbilt University Medical Center.
40. Department of Oncology, Wayne State University School of Medicine, Detroit, MI.
41. Karmanos Cancer Institute, Detroit, MI.
42. Division of Epidemiology, Department of Health Sciences Research, Mayo Clinics.
43. National Human Genome Research Institute, NIH, Baltimore, MD.
44. University of Cincinnati College of Medicine, Cincinnati, OH.
45. Louisiana State University Health Sciences Center, New Orleans, LA.
46. College of Medicine and Life Sciences, University of Toledo, Toledo, OH.
47. The University of Toledo College of Medicine, Toledo, OH.

**Abstract**

Clinical, molecular, and genetic epidemiology studies displayed remarkable differences between ever- and never-smoking lung cancer. Association analysis stratified by smoking behavior has the potential to identify novel variants that confer risk in only ever- or never-smoking group and were missed by prior main-effect association studies. We conducted a stratified multi-population (European, East Asian and African-American) association study in 44,823 ever-smokers and 20,074 never-smokers. Five independent novel loci, including *GABRA4* from ever-smoking and *LRRC4C* and *LCNL1* from never-smoking lung cancer, were identified with association evidence from two or three populations (P < 5x10-8). In addition, we also validated the lung cancer risk effect for known variants at *VTI1A* and *ACVR1B* in never-smoking African American women. Further functional analysis provided multiple lines of evidence suggesting the variants affect lung cancer risk through excessive DNA damage (*GABRA4*) or cis-regulation of gene expression (*LCNL1*). For the variants with association effect in ever-smoking lung cancer (including the known variants), we examined their risk effect among never-, light- (packyear <= 20) and moderate-to-heavy- (MtoH) (packyear > 20) smokers with European ancestry. 9 out of 12 independent variants had increased lung cancer risk in MtoH-smokers compared with light-smokers. We also observed high correlation (79.19%, P=0.01) between ever- and never-smoking lung cancer suggesting the common variants shared by them. The high heritability (9.80%) of ever-smoking lung cancer suggests certain signaling pathways induced by smoking behavior in lung carcinoma; the low heritability (2.62%) of never-smoking lung cancer is aligned with the uncommon variants identified in this disease subtype.

**Introduction**

Genome-wide association studies (GWAS) have been fruitful in the past two decades and more than 50 susceptibility loci have been identified in lung cancer. However, they only account for a limited proportion of heritability implying the existence of additional susceptibility loci that are not yet revealed1. The missing variants may include low allele frequency variants (minor allele frequency < 0.01) and those that affect lung cancer risk through genetic/environmental interactions that cannot be disclosed by regular main-effect association studies2-3. Smoking is the number one environmental risk factor contributing to lung cancer and > 80% of lung cancer patients have a history of tobacco smoking4. Lung cancer in never-smokers, although much less common compared with lung cancer in ever-smokers, is still estimated to be the 7th leading cause of cancer-related deaths5. Remarkable differences have been identified in both clinical and molecular epidemiology studies between ever- and never-smoking lung cancer6. Quite a few genetic variants have been reported in ever-smoking lung cancer such as the well-known *CHRNA5/A3/B4* gene region, *TP63*, *TERT*, and *CYP2A6* gene, etc7-8. However, fewer studies have been focused on identifying genetic loci with smoking behavior subgroups.. Several susceptibility loci have also been identified in never-smoking lung cancer. For example, *VTI1A* and *ACVR1B* were found to be associated with lung cancer in Chinese and European never-smoking women9-10. Variants affecting expression of *hTERT* and *TP63* have also been associated with lung cancer in never-smokers11. These findings suggest the heterogeneity in genetic architecture between ever- and never-smoking lung cancer.

To date, the majority of GWAS studies have been conducted in European (EUR) and East Asian populations (EAS); African American (AA) populations have been under-represented. A multi-population GWAS including AA populations will help clarify the varying effects of smoking on risk for lung cancer among the major ancestral populations , identify novel variants with effect across multiple populations; and evluate the heterogeneity in lung cancer risk across ancestral groups.

One challenge in GWAS is to delineate the relationship between the genetic variants and the biological mechanisms underlying the statistical findings. Various functional annotation tools have been developed to infer the functional role of genetic findings such as CADD and RegulomeDB12-14. eQTL analysis has also been commonly used in GWAS to infer the cis-regulation of nearby gene expression for the variants15. Recently, DNA damage assays have also been applied in lung cancer GWAS to characterize candidate genes as lung cancer risk genes are enriched in the DNA damageome, proteins that can result in high DNA damage when overproduced16-17. For example, significantly increased DNA damage levels were observed in *CHEK2*, *ATM*, *POMC*, *MLNR*, and *MME*, and *PPIL6*, genes that were found to be associated with lung cancer, in DNA damage assay, suggesting that genetic variants may promote lung cancer through DNA damage regulation16-17. An integrative functional analysis has the potential to provide multi-layered evidence for a more comprehensive understanding of the GWAS findings.

In 2022, we performed a multi-population GWAS, including EUR, EAS and AA populations, and identified five novel susceptibility loci associated with lung cancer16. Leveraging this rich resource, we performed a comprehensive study of genetic variants associated with ever- and never-smoking lung cancer aiming to: 1, identify novel variants that confer risk in only ever- or never-smoking groups that were missed by prior regular GWAS studies; 2, perform heritability and correlation analysis between ever- and never-smoking lung cancer; 3, evaluate the association effect of known variants by smoking behavior strata in AA population; 4, provide evidence for the functional roles of the novel variants; 5, investigate the impact of tobacco smoking on risk effect of the genetic variants associated with ever-smoking lung cancer .

**Materials and Methods**

**Genotype data**

The imputed genotypes from the INTEGRAL (Integrative Analysis of Lung Cancer Etiology and Risk)-ILCCO (International Lung Cancer Consortium) lung cancer consortium were applied in this study. Detailed information about genotype imputation and data quality control can be found in our previous publication in 202216. About 9,000,000 high-quality imputed SNPs (information score >= 0.8) from a total of 64,897 individuals, including 44,823 ever-smokers and 20,074 never-smokers were analyzed in the study. The individuals came from 10 studies with diverse ancestry populations including EUR, EAS and AA (Table 1, Supplementary Table S1). 72.1% of the individuals are inferred with European ancestry (EUR, N=46,786), compared with 19.1% with Asian ancestry (EAS, N=12,423) and 8.8% with African-American ancestry (AA, N=5,688). About 35-40% of the ever-smoking lung cancer patients were diagnosed with lung adenocarcinoma (ADE) across the populations, and 25-34% of the patients were diagnosed with squamous carcinoma (SQC). ADE is the predominant subtype in never-smoking patients and accounts for > 57% of patients in all the populations (Supplementary Figure S2). Small-cell lung cancer (SCLC) is much less common compared with ADE and SQC in ever-smokers (9.79%) and very few in never-smokers (0.54%).

**Association analysis of lung cancer in ever- and never-smokers**

Smoking status was self-reported and was categorized into never-smokers (0) and ever-smokers (1, including both current smokers and former smokers). GWAS in the ever- and never-smoking groups were conducted in EUR, EAS and AA populations separately. Meta-analysis was performed to combine information from each population in the ever- and never-smoking strata separately. We adjusted for study sites in the analysis by including a categorical variable for each of the sites. Univariate chi-square tests showed the first three principal components were most significantly associated with disease status and were adjusted in the analysis as well. The significant SNPs were selected based on two criteria: 1, with same direction of risk effect and p value < 0.1 in two or three populations; 2, and with a joint p value < 5x10-8 in meta-analysis. For the significant variants with low allele frequency (MAF < 0.01), we further validated the signals with Firth logistic regression, a method designed for rare variants association test to reduce small-sample bias in regular logistic regression18.The variants that were not significant in Firth test were removed from the final report.The stratified GWAS analysis was conducted in overall lung cancer as well as ADE, SQC and SCLC subtypes. Genomic inflation factor, i.e., the lambda value was calculated to examine if there was inflated type I error rate in association analysis. The lambda value adjusted by sample size was also calculated using the formula: . PLINK 1.07 was used for GWAS and meta-analysis. R-4.0.2 and R package logistf 1.23 was applied for Firth logistic regression analysis.

To evaluate the smoking impact on the variants with association effect in ever-smoking lung cancer including the variants identified from prior reports and the novel ones identified in this study, we further examined their risk effect in never-smokers, light-smokers (pack year (packyr) <= 20) and moderate-to-heavy-smokers (MtoH) (pacyr > 20). We adjusted for the first three principal components and study sites in the analysis.

**Heritability and correlation analysis between ever- and never-smoking lung cancer**

The summary statistics from association analysis were provided as input and the heritability and correlation analysis were performed using LDSC program in EUR and EAS, respectively19-20. The European and East Asian LD scores were downloaded from 1000 Genomes. The analysis was conducted in overall lung cancer as well as lung adenocarcinoma and squamous lung cancer subtypes. The very small sample size in AA population (2420 in ever-smokers and 1407 in never-smokers strata) would not have enough power for the analysis and thus was removed from the analysis.

**Functional annotation analysis**

The web-based tool RegulomeDB was used to infer the regulatory potential of significant variants by integrating high-throughput, experimental data sets from ENCODE and other sources13. For each variant, it calculated a probability score indicating their likelihood of being a regulatory element or a sequence motif. Another webserver, RBPmap, was used to identify potential RNA binding protein (RBP) binding motifs in all transcripts overlapping with alternative and reference alleles14. A sequence of 61 bp, including 30 bp upstream/downstream of the candidate SNP were provided as the input for motif search. Transcription factor binding motifs or RBP binding motifs with p-value<0.05 for either the reference or the alternative allele were identified as putative binding sites.

**Colocalization analysis**

Genotype and gene expression rpkm (Reads Per Kilobase Million) data from 377 lung tissue samples with European ancestry were downloaded from GTEx (phs000424.GTEx.v7.p2). Average rpkm for the gene was used if there were duplicated samples. Individuals with rpkm < 0.25 were removed from the analysis. The SNPs from within +/- 250 kb of each candidate variant were retrieved from both GTEx and GWAS data. The z-score from association between genotype and gene expression data (GTEx) was plotted against those from the GWAS analysis for each retrieved SNP to examine the correlation between eQTL and GWAS studies. The eQTL analysis was conducted using program R-4.0.2.

**Human cell line, reagents, and DNA damage assays**

MRC5-SV40 human lung fibroblast cell line was maintained in DMEM, high glucose medium (Gibco, #11965118) with 10% FBS (Gibco, #10438034), 2mM L-glutamine, 100ul/ml streptomycin, and penicillin (Gibco, #10378016). The cell line had been authenticated via STR analysis (ATCC) and routinely checked for mycoplasma contamination (ABM, G238). Gating entry clones for each of the candidate genes, such as *GABRA4* (IOH27675) and *NF2F1* (IOH3781), were acquired from the Kenneth Scott cDNA library at Baylor College of Medicine. They were further subcloned into an N-terminal EmGFP tagged vector (pcDNA6.2/N-EmGFP-DEST, Invitrogen), using Gateway LR Clonase II Enzyme Mix (Invitrogen, #11791020). The previously cloned EmGFP-Tubulin was used as a control (PMID: 30633903).

Plasmid transfections were performed using GenJet In Vitro DNA Transfection Reagent Ver. II (SignaGen, #SL100489). To further characterize the candidate genes, flow-cytometric DNA damage assays were performed as previously described in MRC5-SV40 cell line with transient candidate gene overexpression21-22. Briefly, MRC5-SV40 human lung fibroblasts cells were fixed, permeabilized and, stained with γH2AX antibody (#05-636, Sigma), then samples were measured by a BD LSRFortessa flow cytometer and analyzed using the FlowJo software. For overproduction experiments, cells with mock transfection were used to set the threshold gating to determine the percentage of GFP− and γH2AX− cells, with 0.5% of control cells gated as the damage threshold as previously validated. The DNA-damage ratio caused by protein overproduction is defined by (Q2/Q3)/(Q1/Q4), where Q2 is the number of transfected damage-positive cells; Q3 is the number of transfected damage-negative cells; Q1 is the number of untransfected damage positive cells, and Q4 is the number of untransfected damage-negative cells.

DNA damage assays with benzo[a]pyrene (Bap; #48564, Sigma) were carried out under similar conditions that do not involve in exogenous agent exposure. Briefly, BaP (8uM) were added when cells were transfected with plasmids, and incubated for 72 hours, followed by flow-cytometric DNA damage assays as described above.

**Results**

**Variants with association effect in ever- or never-smoking lung cancer**

Genome-wide association analyses were conducted in ever- and never-smokers in overall lung cancer as well as other lung cancer subtypes. Figure 1A displayed the Manhattan plots of the signals from the stratified analysis. QQ-plots of the p values from the association analysis and adjusted genomic inflation values (lambda values) by sample size displayed no inflated type I error rate in the analysis (Figure 1A 1-4 right). We identified quite a few significant variants in ever- and never-smoking lung cancer. The significant variants from known genes, such as *AK5*, *TP63*, and *TERT, etc.*, were summarized in Supplementary Table S3 (Labeled in black in Figure 1A). Table 2 listed the variants conferring risk in only ever- or never-smoking individuals including the well-known 15q25.1 region. Six novel variants were identified in the study, but one of the variants, with association effect in the EUR and AA population and with MAF < 0.01, did not reach genome-wide significance in the Firth test check and was removed from the final report (P\_firth=4.00x10-7, Supplementary Table S4). The other five novel variants, rs62303696 from *GABRA4*, rs58778970 from intergenic region, rs4756620 from *LRRC4C*, rs1383429 from *LINC01088* and rs968516 from *LCNL1*, had genome-wide significant findings (labeled in red at Figure 1A).

We further examined the regional association surrounding the five variants (Figure 1B 1-5). Multiple supporting variants in strong LD (r2>= 0.8) with the target SNPs were identified indicating the reliability of the signals except for SNP rs4756620, for which only one supporting variant with r2 of 0.6 was detected in the region (Figure 1B 3). To check the authenticity of the signal at rs4756620, we further checked the imputation quality of this SNP and found that this SNP was genotyped in four of the 10 studies (Supplementary Table S5). We examined the association using only genotyped data from these four studies and rs4756620 had p values of 9.79x10-7 (OR=0.61, N=7132) EAS and 6.49x10-2 (OR=0.70, N=1387) in AA population. We believe the association at rs4756620 was reliable and we reported it as a novel susceptibility locus associated with never-smoking lung adenocarcinoma.

Table 2 displays the detailed information for the variants associated with ever- or never-smoking lung cancer. rs62303696, located at 3’ UTR (untranslated region) of *GABRA4*, was identified in ever-smoking overall lung cancer with a joint p value of 1.22x10-9 and OR (Odds Ratio) of 1.18. The evidence of association was detected in all three continental populations with P values of 2.71x10-7, 4.81x10-3 and 6.08x10-2 from EUR, EAS and AA populations, respectively. rs58778970 was identified in ever-smoking small cell lung cancer (P=1.58x10-8, OR=1.34). The association evidence came from both European (P=1.50x10-7, OR=1.33) and AA populations (P=2.40x10-2, OR=1.53). Three SNPs, rs4756620 (P=6.51x10-10, OR=0.59), rs1383429 (P=6.44x10-9, OR=0.67) and rs968516 (P=8.19x10-10, OR=0.34) were identified in never-smoking lung cancer. It was noted that all these three variants achieved genome-wide significance in the EAS population (P < 5x10-8) and replicated in either EUR or AA population. We compared the risk effect between ever- and never-smoking groups for the newly identified variants. It was observed that all five of these novel variants were significant in either ever- or never-smoking group and not significant in non-stratified analysis which explains why these variants were not discovered in prior GWAS studies (Figure 2A).

**Genetic correlation analysis in lung cancer between ever- and never-smokers**

The summary statistics from ~ 750K SNPs were applied for heritability and correlation analysis between ever- and never-smoking overall lung cancer as well as other subtypes in CEU population. The heritability is the highest in ever-smoking lung adenocarcinoma (11.63%), then followed by overall lung cancer (9.80%) and squamous carcinoma (8.74%) (Table 3). The heritability in never-smoking lung cancer is 8.57% in lung squamous carcinoma followed by 5.34% in lung adenocarcinoma and 2.62% in overall lung cancer. The genetic correlations between ever- and never-smoking lung cancer is the highest in lung adenocarcinoma (79.19%, P=0.01) and followed by overall lung cancer (67.22%, P=0.10) and then squamous subtype (25.60%, P=0.22).

We also performed the analysis in EAS population in ever- and never-smoking population. The estimated heritability in never-smoking overall lung cancer is 16.15% and 8.68% in ever-smoking lung adenocarcinoma, higher than 8.57% and 5.34% from corresponding CEU subgroups. We were not able to get reliable estimate from ever-smoking lung cancer due to the small sample size in ever-smokers (N=2,527, Supplementary Table S8) thus no valid estimate for correlation between ever- and never-smoking lung cancer in Asian population. The sample size in YRI population did not have power for this analysis.

**Some of the known variants confer lung cancer risk in only ever- or never-smoking population**

Aside from the novel findings, the stratified analysis also found that some of the previously identified susceptibility loci conferred lung cancer risk in only ever- or never-smoking group. Our previous study found evidence for association between rs5767055 at *IKZF2* and squamous lung cancer in East Asian population (OR=0.23, P=8.39x10-11, Figure 2A)16. Further stratified analysis displayed this variant was more significant in never-smoking squamous lung cancer in EAS population (OR=0.19, P=1.51x10-11) and not significant in ever-smoking group (OR=1.05, P=0.37). It is a very rare variant with minor allele frequency (MAF) of 0.091 in our EAS population. Our collaborator at Nanjing, China further validated this signal using data from six independent study sites in China, including a total of 8,407 never-smokers, and the final joint analysis showed an OR of 0.56 and p value of 7.77x10-12 in never-smoking squamous lung cancer (Figure 2B, supplementary Table S6)23. Five of the study sites has MAF varying from 0.003 to 0.006 and one study site with MAF of 0.012.

rs17879961, a rare variant located in exon of *CHEK2* gene, has been reported to be associated with squamous lung cancer but no further study was conducted regarding the smoking status of the patients16,24. The results from our study showed that it was more significant in ever-smoking lung squamous group (OR=0.25, p=2.93x10-11, Figure 2A) than in non-stratified analysis (OR=0.26, P=5.86x10-11), and there was no suggestive evidence from non-smoking group (OR=0.59, P=0.56). These findings implied rs17879961 might confer lung cancer risk in only ever-smoking individuals.

**Validation of lung cancer susceptibility loci in never-smoking women using data from African Americans**

*VTI1A* and *ACVR1B* were reported to be associated with never-smoking lung cancer in both Asian and European women about a decade ago10-11. However, there is no report about association from AA due to the under-represented AA participants in previous lung cancer GWAS studies. In our analysis, rs12265047, from *VTI1A*, had an OR of 0.63 (P=4.64x10-5), 0.77 (P=4.53x10-13) and 0.63 (P=3.29x10-3) in never-smoking women from the EUR, EAS and AA population, respectively (Table 2). rs7962469, located in *ACVR1B*, had an increased risk effect for lung adenocarcinoma in never-smoking women and the OR is much higher in AA population (OR=1.74, P= 3.14x10-3) compared with EUR (OR=1.12, P=5.61x10-2) and EAS (OR=1.18, P=1.63x10-6) populations implying the existence of heterogeneity of risk effect in lung cancer across continental populations.

**Evaluation of impact of smoking on lung cancer** **risk**

For the variants with association effect in ever-smoking lung cancer, including the known variants, we compared their lung cancer risk in never-, light- (packyr <= 20), and moderate-to-heavy-smokers (MtoH, packyr >20) in EUR, EAS and AA population, respectively. Due to the smaller sample size in EAS and AA population, there was limited power for most of the variants for these two populations, so we focused on the results from EUR population (Supplementary material II). The bar chart in Figure 2C displayed the ORs in different smoking groups for variants from 12 independent regions. The variants can be divided into four groups: 1, variants associated with lung cancer in only ever-smokers, including the rare variants in *BRCA2*, *CHEK2* and the well-known *CHRNA5*. rs17879961 at *CHEK2* had OR of 0.10 and P value of 5.18x10-3 in light-smokers vs. OR of 0.27 and P value of 5.68x10-9 in MtoH-smokers (Figure 2C). For rs55781567 in *CHRNA5*, we found it had slightly higher lung cancer risk in MtoH-smokers (OR=1.30, P=6.17x10-39) than light-smokers (OR=1.25, P=3.19x10-14. And we observed the similar pattern in AA population, OR=1.29 and P=9.68x10-4 in light-smokers vs. OR=1.33 and P=1.28x10-4 in MtoH-smokers (Figure 2D left). Group 2, variants associated with lung cancer in both never- and ever-smokers including the variants in *TP63*, *TERT* and *CLPTM1L*, shared a pattern of increasing risk effect for lung cancer in never-, light- and MtoH-smokers. For example, rs2853677 in *TERT* had OR of 0.81 OR and P value of 4.61x10-9 in never-smokers, OR of 0.84 and P value of 3.15x10-9 in light-smokers and OR of 0.89 and P value of 3.71x10-9 in MtoH-smokers. We observed a similar risk pattern in EAS population, OR of 0.73 OR and P value of 9.59x10-21 vs. of 0.80 and P value of 3.24x10-2 vs. OR of 0.84 and P value of 3.71x10-2 (Figure 2D middle). Group 3 included the variants rs9489152 at *ROS1* and rs62303696 at *GABRA4*, which had consistent risk effect across the different smoking groups. And we observed the similar risk pattern in both EAS and AA population for rs9489152 although the estimations were not significant (P>0.05, Figure 2D right). rs12337510 at *MTAP* is an exceptional example and the OR decreases from 1.37 (P=6.28x10-7) in never-smokers to 1.14 (P=1.28x10-3) in MtoH-smokers. However, we did not see similar pattern in either EAS or AA population although it was significant in the other two populations (Supplementary Figure S7).

**Functional analysis of identified novel variants**

We first conducted functional annotation analysis using RegulomeDB to evaluate how these identified variants affect lung cancer risk. All the five new variants are located in non-coding regions such as 3’ or 5’ UTR, intronic and inter-genetic regions. The query from regulomeDB database showed that all five variants were located within peaks from more than one CHIP-seq, DNase-seq or FAIRE-seq experiment suggesting that they were located within regulatory DNA regions (Supplementary Table S8). Two SNPs, rs62303696 located at the 3’ UTR in GABRA4 gene and rs1383429 located in the intronic region in LINC01088, are predicted to be regulatory variants with probability > 0.6. CHIP-seq peaks are also detected at both of these two SNPs suggesting they were located in binding sites for regulatory proteins such as transcription factor, histone modifications, etc (Figure 3A). Position weight matrix (PWM) analysis predicted that rs1383429 was a highly conserved SNP in sequence motifs (Figure 3B).

We also evaluated and compared the RNA binding proteins (RBPs) with significant sequence motifs between reference and alternative alleles. Figure 3C displays the RBPs with significant motifs (P < 0.05) for novel variants located within coding genes. rs58778970 was located in an intergenic region and thus was removed from the analysis. We noticed different RBPs with significant motifs between reference and alternative alleles for the variants. For example, there were 13 RBPs for the reference allele of rs1383429 while only two for the alternative allele. rs4756620 had 2 RBPs for the alternative allele but three additional RBPs for the reference allele. These findings, combined with the results from RegulomeDB, suggest that the two variants might regulate lung cancer risk by interacting with different regulatory proteins such as transcription factors and RBPs.

eQTL analysis was conducted to evaluate the association between lung cancer risk and nearby gene expression for each of the five novel variants. The z-score from association between genotype and nearby gene expression data (GTEx) was plotted against the z-score from GWAS analysis. A strong association was identified between lung cancer risk and *LCNL1* gene expression for rs968516 and ~ 2,200 surrounding SNPs that were in strong LD with it (r2> 0.8). These results suggested rs968516 could affect lung cancer risk in never-smokers through regulation of *LCNL1* gene expression (Figure 3D).

We performed DNA damage assays on each candidate genes following the procedures as displayed in Figure 4A1. We found that overproduced EmGFP fusions of *GABRA4* and *NR2F1* promoted DNA damage, measured by sensitive flow cytometric assays (Figure 4A 2-6). BaP is one of the cigarette smoke carcinogens involved in lung tumorigenesis. Because *GABRA4* was nominated from the lung cancer smoking analysis, we hypothesized that BaP exposure might enhance *GABRA4*-induced DNA damage. BaP exposure for 72 hours significantly increased *GABRA4*-induced DNA double strand breaks, but not in tubulin overproducing cells (Figure 4B 1-2). This observation supports the hypothesis that low-dose environmental mutagens can further titrate out DNA repair and cause amplified DNA damage in cells that have elevated endogenous DNA damage (Figure 4B 3).

**Discussion**

Differences in genomic features have been identified in lung cancer between ever- and never-smokers such as genetic variants, gene mutation, gene expression and DNA methylation profiles, etc.6 For example, the well-known *CHRNA5/A3/B4* gene region was associated with nicotine dependence and lung cancer in ever-smokers7,15,24-25. Variants in *VTI1A* and *ACVR1B* were found significantly associated with lung cancer in never-smoking women9-10. Leveraging the genotype from three continental populations, we identified five novel susceptibility loci in ever- or never- smoking lung cancer. All five variants have significant association effects in one smoking group and no effect in the other. These findings display heterogeneity in genetic predisposition to lung cancer between different smoking groups and highlight the complicated genetic architecture in this deadly disease. We also examined their interaction effect with smoking status in lung cancer rick using genotype data from CEU in Oncoarray study, the study with the largest sample size of European individuals (N=29,905), and none of them was significant (P< 0.05) (Supplementary Table S9). The re-evaluation of known variants in *IKZF2* and *CHEK2* showed the variants conferred risk in only ever- or never-smoking lung cancer rather than non-stratified lung cancer. These findings illustrated that stratified GWAS was imperative for identification of novel variants with effect only in subgroups that cannot be revealed by regular GWAS or genome-wide interaction studies and for prioritizing likely causal mechanisms as well.

One challenge in a GWAS studies is that the variants identified in one population have failed to be replicated in other populations. In our multi-population study, the identified novel variants had association evidence from at least two ancestry groups, and three of the variants (rs62303696, rs58778970, and rs4756620) had association effects in African-descent populations. *VTI1A* was first discovered associated with lung cancer in Asian never-smoking women and then validated with nominal significance in European never-smoking women; *ACVR1B* was first reported in lung adenocarcinoma in European never-smokers and then reported in Asian women never-smokers9-11,26-27. Little is known about their association with lung cancer in AA population. We successfully validated their association effect in people with AA ancestry for the first time as far as we known. These two variants, together with the novel variant rs62303696 at *GABRA4*, are the only three variants associated with ever- or never-smoking lung cancer in all the three continental populations (Table 2). These findings demonstrate that inclusion of AAs in the multi-population GWAS is crucial for a better understanding of genomic and environmental variations underpinning lung cancer. However, the AA sample size is still not large (N=5,688) which limits our ability to identify novel variants in this population.

For the variants with association effect in ever-smokers, including the variants identified from previous studies, we also performed a comprehensive comparison or their risk effect among never-, light- and MtoH-smokers with European ancestry. Among the 12 tested variants selected from independent associated regions, 9 of them had increased lung cancer risk in MtoH-smokers compared with light-smokers, 2 of them with consistent effect across the smoking groups (Figure 2C). These observations provide evidence from genetic association analysis that both tobacco smoking and genetic factors contribute to lung cancer risk and smoking behavior can modulate lung cancer risk for genetic variants, i.e., there are genetic interactions between smoking and genetic susceptibility loci.

Tobacco smoking is the number one risk factor associated with lung cancer. The high heritability of ever-smoking lung cancer compared with never-smoking lung cancer (11.63% vs 5.34% in lung adenocarcinoma, 9.80% vs. 2.62% in overall lung cancer, Figure 2D), suggests certain signaling pathways induced by smoking behavior in lung carcinoma. The high correlations between ever- and never-smoking overall lung cancer (67.22, P=0.10) and lung adenocarcinoma (79.19%, P=0.01) suggest ever- and never-smoking lung cancer share some common susceptibility variants in tumorigenesis such as the variants in *TERT*, *TP63* and *CLPTM1L*, etc, that are shown to have association effect in both ever- and never-smoking lung cancer. On the other hand, the relatively low heritability of never-smoking lung cancer implies varied molecular mechanisms implicated in never-smoking lung cancer which is aligned with the observation that most of the identified susceptibility loci associated with never-smoking lung cancer are uncommon variants with minor allele frequency < 0.1. The heritability is about 8% in both ever- and never-smoking squamous carcinoma. However, the correlation is very low (25.6%, P=0.22) suggesting very distinct molecular mechanisms between ever- and never-smoking squamous lung cancer.

As we step into the post-GWAS era, the ultimate goal is to understand the biological consequences of the statistical associations. We adopted multiple approaches for functional inference and obtained multi-layers of evidence supporting the regulatory role of the identified novel variants in ever- and never-smoking lung cancer. For example, rs968516, identified in never-smoking squamous lung cancer, was shown to affect lung cancer risk through regulation of nearby *LCNL1* gene expression. It is also an eQTL in multiple tissues including lung (Supplementary Figure S10). rs62303696, identified in ever-smoking lung cancer, is located in 3’UTR region of *GABRA4*, a gene that has been reported to be related with alcohol use disorder in European population28. A systematic study showed that ~ 3% of GWAS hits were located within 3’ UTR region29. Genetic variations in 3’ UTR may change the binding sites for RBPs and miRNAs and lead to differential gene expression. DNase-seq and CHIP-seq experiments showed that rs62303696 was located within regions sensitive to cleavage by DNase I and DNA binding sites for transcription factors *NR2F1* and *JUNB* (Figure 3A). Further RBP analysis showed that the reference allele of rs62303696 enabled a binding motif for RBM6 while the alternative allele didn’t (Figure 3C). Aside from being reported as an alternative splicing factor and a putative tumor suppressor gene, RBM6 has been identified as a regulator involved in repair of DNA double-strand breaks in a recent study30-33. We further discovered *GABRA4* induced DNA damage in lung fibroblast cell line which offered one mechanistic explanation for lung cancer: increased DNA damage and mutagenesis caused by upregulation of *GABRA4* may underlie tumorigenesis and poor clinical prognosis. These integrated results suggest rs62303696 could affect lung cancer risk in smokers through increased DNA damage and genome instability (Figure 4).

In summary, we performed a multi-population GWAS stratified by smoking status in lung cancer, and we identified five novel variants associated with ever- or never-smoking lung cancer. The extensive functional analysis provided evidence for the functional roles of the identified variants and provided insights about the molecular mechanism underlying lung carcinogenesis. We further evaluated the impact of smoking quantity on lung cancer risk for the variants associated with ever-smoking lung cancer and the results provided evidence from genetic association analysis that both smoking behavior and genetic variants contribute to lung cancer and smoking behavior modulates the risk effect of the genetic variants. The results from heritability and correlation analysis between ever- and never-smoking lung cancer are well aligned with findings from GWAS studies. Our study highlighted the genetic heterogeneity between ever- and never-smoking lung cancer and provided helpful etiological insights for the complicated genetic architecture in this deadly disease.

**Competing interests**

The authors declare no competing interests.

**Acknowledgement**

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: dbGaP accession number phs000424.v7.p2.

**Funding**

This work was supported by U19-CA203654 (NIH), RR170048 (Cancer Prevention Research Institute of Texas (CPRIT)), RR160097T (CPRIT), RR180061 (CPRIT), U01 CA243483-02 (Genetic Epidemiology of Lung Cancer Consortium (GELCC)), X01HG007491, Department of Health and Human Services contracts HHSN26820100007C, HHSN268201700012C, 75N92020C00001, and R21CA235464 (NIH) for support. CARET is funded by the National Cancer Institute, National Institutes of Health through grants U01-CA063673, UM1-CA167462, and U01-CA167462. The Harvard Lung Cancer Study is funded by NIH (NCI) grant U01CA209414 (Boston Lung Cancer Survival Cohort). The Asian validation study was funded by National Natural Science Foundation of China (81820108028). The Liverpool Lung Project is supported by MPAD. Roy Castle Lung Foundation (UK). The EAGLE study was supported by the intramural program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH. Joan E. The ReSoLuCENT study is funded by the Weston Park Hospital Cancer Charity. Acknowledgement to the George Isaac Family Fund for Cancer Research for grant support. Bailey-Wilson was supported by the Intramural Program of the National Human Genome Research Institute, NIH. Jun Xia was supported by the National Institute Of Environmental Health Sciences (K99ES033259). Susan Rosenberg was supported by NIH grants (R01CA250905 and DP1-AG072751). This project was also supported by the Cytometry and Cell Sorting Core at Baylor College of Medicine with funding from the CPRIT Core Facility Support Award (CPRIT-RP180672), the NIH (CA125123 and RR024574) and the assistance of Joel M Sederstrom. Chris Amos and Chao Cheng are Research Scholars of the Cancer Prevention Institute of Texas.

**Author Contributions**

Yafang Li designed and planned the study, presented the results and wrote the manuscript. Xiangjun Xiao assisted the data preparation and genome-wide gene-sex interaction analysis; Jianrong Li and Chao Cheng assisted the motif analysis in functional annotation; Jun Xia, Gail F Fernandes, Shannon E Slewitzke conducted the DNA damage assay; Meng Zhu performed the validation analysis of the variants using independent samples from Asian population; Jun Xia and Chao Cheng contributed to the writing of the original manuscript; Younghun Han assisted the figure preparation. Dipstasri Mandal, Gail F Fernandes, Ann G Schwartz, Meng Zhu, Ping Yang, Chu Chen, Joan E Bailey-Wilson, Philip Lazarus, Yohan Bossé and Heike Bickeböller contributed to reviewing and editing of the manuscript. Christopher I. Amos conceived and supervised the study. The other authors contributed to data collection. All authors discussed the results and commented on the manuscript.

**Materials & Correspondence**

Correspondence should be addressed to Yafang Li or Christopher I. Amos. And material requests should be addressed to Christopher I. Amos.

**Reference**

1. Bosse Y, Amos C. A decade of GWAS results in lung cancer. Cancer. Epidemiol. Biomarkers. Prev. 2018; 27(4):363-379.
2. Momozawa Y and Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. J Hum Genetic 2021; 66:11-23.
3. Zhang Y, Shen S, Wei Y, Zhu Y, Li Y, Et al. A Large-Scale Genome-Wide Gene-Gene Interaction Study of Lung Cancer Susceptibility in Europeans With a Trans-Ethnic Validation in Asians. J Thorac Oncol 2022; in press. <https://doi.org/10.1016/j.jtho.2022.04.011>
4. Walser T, Cui X, Yanagawa J, Lee JM, Heinrich E, Lee G, Sharma S, and Dubinett SM. Smoking and Lung Cancer--The Role of Inflammation. Proc. Am. Thorac. Soc. 2008; 5(8): 811–815.
5. Rivera GA, Wakelee H. Lung Cancer in Never Smokers. Adv Ep Med Biol 2016; 893:43-57. doi: 10.1007/978-3-319-24223-1\_3.
6. Sun S, Schiller JH, Gazdar AF. Lung cancer in never-smokers- a different disease. Nat Rev Can 2007; 7: 778-790.
7. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 2008; 40(5): 616-622.
8. Patel YM, Park SL, Han Y, Wilkens LR, Bickeböller H, et al. Novel association of genetic markers affecting CYP2A6 activity and lung cancer risk. Cancer Res 2016; 76(19): 5768-5776. doi: 10.1158/0008-5472.CAN-16-0446.
9. Lan Q, Hsiung CA, Keitaro Matsuo K, Hong YC, Seow A, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. Nat Genet 2012; 44(12): 1330-1335.
10. Spitz MR,Gorlov IP, Amos CI, Dong Q, Chen W, et al. Variants in inflammation genes are implicated in risk of lung cancer in never smokers exposed to second-hand smoke. Cancer Discov 2011; 1(5): 420-429.
11. Hung RJ, Spitz MR, Houlston RS, Schwartz AG, Field JK, et al. Lung cancer risk in never-smokers of European descent is associated with genetic variation in the 5p15.33 TERT-CLPTM1Ll region. J Thorac Oncol 2019; 14(8):1360-1369. doi: 10.1016/j.jtho.2019.04.008.
12. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 47(D1), D886-D894. doi: 10.1093/nar/gky1016 (2019).
13. Boyle, A.P., et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research 22(9), 1790-1797. PMID: 22955989 (2012).
14. Paz I, Kosti I, Ares M Jr, Cline M, Mandel-Gutfreund Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res., 2014; 42(Web Server issue): W361–W367. doi: 10.1093/nar/gku406.
15. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet 2017; 49(7):1126-1132. doi: 10.1038/ng.3892.
16. Byun J, Han Y, Li Y, Xia J, Xiao X, et al. Trans-ethnic genome-wide meta-analysis of 35,732 cases and 34,424 controls identifies novel genomic cross-ancestry loci contributing to lung cancer susceptibility. Nature Genetics in press.
17. Liu Y, Xia J, McKay J, Tsavachidis S, Xiao X, et al. Rare deleterious germline variants and risk of lung cancer. NPJ Precis Oncol 2021; 5(12).
18. Wang X. Firth logistic regression for rare variant association tests. *Front Genet* 2014; **5**,187.
19. Bulik-Sullivan, B., et al. An atlas of genetic correlations across Human diseases and traits. Nat Genets 2015; 47(11):1236-41.
20. Finucane, HK, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet, 2015; 47(11):1228-35.
21. Xia J, Chiu L-Y, Nehring RB, Núñez MAB, Mei A, et al. Bacteria-to-Human protein networks reveal origins of endogenous DNA damage. Cell 2019; 176(1-2):127-143.e24. doi: 10.1016/j.cell.2018.12.008
22. Bossé Y, Li Z, Xia J, Manem V, Carreras-Torres R, et al. Transcriptome-wide association study reveals candidate causal genes for lung cancer. Int J Cancer 2020; 146(7):1862-1878. doi: 10.1002/ijc.32771.
23. Dai J, Lv J, Zhu M, Wang Y, Qin N, et al. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. Lancet Respir Med 2019;7(10):881-891.
24. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet 2014; 46(7):736-41.
25. Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF et al. Novel genes identified in a high-density genome wide association study for nicotine dependence. Hum Mol Genet 2007; 16: 24–35.
26. Chen LS, Saccone NL, Culverhouse RC, Bracci PM, Chen CH, Dueker N et al. Smoking and genetic risk variation across populations of European, Asian, and African American ancestry—a meta-analysis of chromosome 15q25. Genet Epidemiol 2012; 36: 340–351.
27. Wang Z, Seow W, Shiraishi K, Hsiung CA, Matsuo K, et al. Meta-analysis of genome-wide association studies identifies multiple lung cancer susceptibility loci in never-smoking Asian women. Hum Mol Genet 2016; 25(3): 620-629.
28. Zhou H, Sealock JM, Sanchez-Roige S, Clarke TK, Levey DF, et al. Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. Nat Neurosci 2020; 23:809–818
29. Steri M, Idda ML, Whalen MB, Orrù V. Genetic Variants in mRNA Untranslated Regions. Wiley Interdiscip Rev RNA 2018; 9(4): e1474.
30. Heath E, Sablitzky F, Morgan GT. Subnuclear targeting of the RNA-binding motif protein RBM6 to splicing speckles and nascent transcripts. Chromosome Res. 2010; 18:851–872
31. Bechara EG, Sebestyen E, Bernardis I, Eyras E, Valcarcel J. RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. Mol Cell 2013; 52:720–733.
32. Wang Q, Wang F, Zhong W, Ling H, Wang J, Cui J, Xie T, Wen S, Chen J. RNA-binding protein RBM6 as a tumor suppressor gene represses the growth and progression in laryngocarcinoma. Gene. 2019; 697:26–34
33. Wistuba II, Behrens C, Virmani AK, Mele G, Milchgrub S, et al. High resolution chromosome 3p allelotyping of human lung cancer and preneoplastic/preinvasive bronchial epithelium reveals multiple, discontinuous sites of 3p allele loss and three regions of frequent breakpoints. Cancer Res. 2000; 60:1949–1960.