# A large-scale genome-wide gene-gene interaction study of lung cancer susceptibility in Europeans with a trans-ethnic validation in Asians

Sipeng Shen1,2,3,4#, Yongyue Wei1,2,3#, Ruyang Zhang1,2,3#\*, Ying Zhu1, Yi Li5, Jiajin Chen1, Jinxing Guan1, Yuzhuo Wang6,7, Meng Zhu6,7, Junxing Xie6, Xiangjun Xiao8, Dakai Zhu8, Yafang Li8, Demetrios Albanes9, Maria Teresa Landi9, Neil Caporaso9, Stephen Lam10, Adonina Tardon11, Chu Chen12, Stig E. Bojesen13, Mattias Johansson14, Angela Risch15, Heike Bickeböller16, H-Erich Wichmann17, Gadi Rennert18, Susanne Arnold19, Paul Brennan14, James McKay14, John K. Field20, Sanjay S. Shete21, Loic Le Marchand22, Olle Melander23, Hans Brunnström23, Geoffrey Liu24, Angeline Andrew25, Lambertus A. Kiemeney26, Shan Zienolddiny-Narui27, Kjell Grankvist28, Mikael Johansson28, Angie Cox29, Philip Lazarus30, Matthew B. Schabath31, Melinda C. Aldrich32, Yang Zhao1, Zhibin Hu3,6,7\*, Feng Chen1,3,7\*, Hongbing Shen3,6,7§, Rayjean Hung33§, Christopher I. Amos8§, David C. Christiani2,34\*§

1Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China

2Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA

3China International Cooperation Center (CICC) for Environment and Human Health, Nanjing Medical University, Nanjing, Jiangsu, China

4State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing, Jiangsu, China

5Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

6Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China

7Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Cancer Center, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, Jiangsu, China

8The Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA

9National Cancer Institute, NIH, Bethesda, Maryland, USA

10British Columbia Cancer Agency, Vancouver, Canada

11University of Oviedo and CIBERESP, Faculty of Medicine, Oviedo, Spain.

12Division Program in Epidemiology, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

13Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Denmark

14International Agency for Research on Cancer, World Health Organization, Lyon, France

15University of Salzburg and Cancer Cluster Salzburg, Salzburg, Austria

16Department of Genetic Epidemiology, University Medical Center, Georg August University Göttingen, Göttingen, Germany

17Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany

18Carmel Medical Center,Israel Instiute of Technology, Faculty of Medicine Carmel Haifa, USA

19Markey Cancer Center, University of Kentucky, Lexington, Kentucky, USA

20Department of Molecular and Clinical Cancer Medicine, Institute of Translational Medicine, Liverpool University, UK

21The University of Texas MD Anderson Cancer Center. Houston, Texas, USA

22Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA

23Department of Clinical Sciences Malmö, Lund University, Sweden

24Princess Margaret Cancer Centre, Toronto, Canada

25Department of Epidemiology, Geisel School of Medicine, Hanover, New Hampshire, USA

26Radboud university medical center, Nijmegen, Netherlands

27National Institute of Occupational Health, Oslo, Norway

28Department of Medical Biosciences, Umeå University, Sweden

29Department of Oncology, University of Sheffield, Sheffield, UK

30Washington State University, College of Pharmacy, USA

31Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, USA

32Vanderbilt University Medical Center, Department of Thoracic Surgery and Division of Epidemiology, USA

33Lunenfeld-Tanenbaum Research Institute, Sinai Health, and University of Toronto, Toronto, Ontario, Canada

34Pulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

#These authors contributed equally to this work.

§Joint last and senior authors who supervised the work.

**\*Send correspondence to:**

Ruyang Zhang, SPH Building Room 406, 101 Longmian Avenue, Nanjing, Jiangsu, China, 211166; email: zhangruyang@njmu.edu.cn

Feng Chen, SPH Building Room 412, 101 Longmian Avenue, Nanjing, Jiangsu, China, 211166; email: fengchen@njmu.edu.cn

Zhibin Hu, SPH Building Room 426, 101 Longmian Avenue, Nanjing, Jiangsu, China, 211166; email: zhibin\_hu@njmu.edu.cn

David C. Christiani, Building I Room 1401, 665 Huntington Avenue, Boston, MA, USA, 02115; email: dchris@hsph.harvard.edu

# Abstract

Although genome-wide association studies have been conducted broadly to investigate genetic variation of lung tumorigenesis, little remains known of gene-gene (G×G) interactions that may influence the risk of non-small cell lung cancer (NSCLC). Leveraging a total of 445,221 European-descent participants from the International Lung Cancer Consortium OncoArray project, Transdisciplinary Research in Cancer of the Lung and UK Biobank, we performed a large-scale genome-wide G×G interaction study on European NSCLC risk via a series of analyses. First, we used BiForce to evaluate and rank over 58 billion G×G interactions from 340,958 SNPs. Then, the top interactions were further tested by demographically adjusted logistic regression models. Finally, we used the selected interactions to build lung cancer screening models of NSCLC, separately, for never and ever smokers. With the Bonferroni correction, we identified eight statistically significant pairs of SNPs, which predominantly appeared in the 6p21.32 and 5p15.33 regions (e.g., rs521828*C6orf10* and rs204999*PRRT1*, *OR*interaction = 1.17, *P* = 6.57×10-13; rs3135369*BTNL2* and rs2858859*HLA-DQA1*, *OR*interaction = 1.17, *P* = 2.43×10-13; rs2858859*HLA-DQA1* and rs9275572*HLA-DQA2*, *OR*interaction = 1.15, *P* = 2.84×10-13; rs2853668*TERT* and rs62329694*CLPTM1L*, *OR*interaction = 0.73, *P* = 2.70×10-13). Notably, even with much genetic heterogeneity across ethnicities, three pairs of SNPs in the 6p21.32 region identified from the European-ancestry population remained significant among an Asian population from the Nanjing Medical University Global Screening Array project (rs521828*C6orf10* and rs204999*PRRT1*, *OR*interaction = 1.13, *P* = 0.008; rs3135369*BTNL2* and rs2858859*HLA-DQA1*, *OR*interaction = 1.11, *P* = 5.23×10-4; rs3135369*BTNL2* and rs9271300*HLA-DQA1*, *OR*interaction = 0.89, *P* = 0.006). The interaction empowered polygenetic risk score (iPRS) that integrated classical polygenetic risk score and G×G information score was remarkable in lung cancer risk stratification. In conclusion, significant G×G interactions were identified and enriched in the 5p15.33 and 6p21.32 regions, which may enhance lung cancer screening models.

**Keywords**: gene-gene interaction, GWAS, lung cancer, single nucleotide polymorphism, cancer risk, genetic screening model, genetic epidemiology

# Introduction

Lung cancer, as the leading cause of deaths worldwide 1, has become a global epidemic. Non-small cell lung cancer (NSCLC) accounts for 85% of all lung cancer cases 2. It is well known that, in addition to environmental exposures (e.g., tobacco smoking) 3, genetic variants also contribute to NSCLC susceptibility 4,5. Although many susceptible single nucleotide polymorphisms (SNPs) have been identified in genome-wide association studies (GWAS) during the last decade 6, together they explain only a small proportion of variation in the risk of NSCLC 7. Hence, recent research efforts have expanded to studies of rare variants 8, copy number variants 9, gene-environmental (G×E) interactions 10 and gene-gene (G×G) interactions 11.

G×E interaction studies of NSCLC have provided additional genetic evidence of pathogenesis, including gene-smoking interactions 10, gene-asbestos interactions 12, and gene-occupation interactions 13. G×G interactions or epistasis may also explain the missing heritability of NSCLC 14,15. However, due to computationally intensive G×G interaction analyses on a genome-wide scale, only a handful of G×G interaction studies have been conducted for prostate cancer 16-18, colorectal cancer 19, breast carcinoma 20 and nasopharyngeal cancer 21. For NSCLC, to our knowledge, we were among the first few who ever performed genome-wide G×G interaction analyses for lung cancer susceptibility among a Han Chinese population, and identified a significant interaction between two SNPs in the 2p32.2 region 6,11. For Europeans, by focusing on significant index SNPs within the 15q25.1 region, we scanned the entire genome to identify SNPs that interacted with those 15q25.1 index SNPs, and detected evidence for G×G interactions involved in lung cancer etiology 22. However, there is still a paucity of genome-wide G×G interaction studies among European population, and the genetic architecture of lung cancer risk under a genome-wide G×G interaction framework remains largely unclear.

Leveraging a total of 445,221 European participants from several international consortia, this study registers the first attempt to conduct a genome-wide G×G interaction study of lung cancer risk. Specifically, the study population includes 28,353 participants from the International Lung Cancer Consortium OncoArray project (ILCCO-OncoArray) 23, and 7,253 participants from the Transdisciplinary Research in Cancer of the Lung (TRICL) 24, and 409,615 participants from the UK Biobank 25. We conducted a series of analyses (a two-phase study, meta-analysis and stratified analysis) to identify significant G×G interactions, followed by trans-ethnic validation of significant G×G interactions using 19,546 Asian participants from the Nanjing Medical University (NJMU) Global Screening Array (GSA) project (NJMU-GSA) 26. We further developed lung cancer screening models utilizing both classic polygenetic risk score (PRS) and the detected G×G interactions for screening high-risk subpopulations.

# Results

## Two significant G×G interactions identified by a two-phase study among Europeans

Table 1 presents the characteristics of NSCLC cases and controls in ILCCO-OncoArray (15,157 cases and 13,196 controls), TRICL (3,288 cases and 3,965 controls) and UK Biobank (3,017 cases and 406,598 controls). We adopted a two-phase (discovery and validation) study design to identify important G×G interactions, while controlling the number of false positives (Figure 1).

The discovery phase involves a screening step followed by a testing step in order to efficiently extract candidate G×G interactions. With the participants from ILCCO-OncoArray, the screening step utilizes BiForce 27, an entropy-based method, to screen over 58 billion interactions and select the top epistatic pairs with a software-default threshold of *P* value, 1.00×10-6. In the testing step, we first estimated the effects of these top epistatic pairs using logistic regression model adjusted for age, gender, smoking status and the top three principal components (PCs) derived from GWAS data for population structure correction, among the participants from ILCCO-OncoArray and TRICL, separately, and then performed a meta-analysis to combine the results and obtain more robust and efficient estimates of these G×G interactions. We observed that two pairs of SNPs (rs521828, intronic of *C6orf10* at 6p21.32 and rs204999, 6.2 kb 3' of *PRRT1* at 6p21.32, *OR*interaction = 1.20, 95% CI: 1.14-1.26, *P* = 6.10×10-13; rs2853668, 4.8 kb 5' of *TERT* at 5p15.33 and rs62329694, intronic of *CLPTM1L* at 5p15.33, *OR*interaction = 0.69, 95% CI: 0.63-0.77, *P* = 6.08×10-13) reached the Bonferroni-corrected significance level (*P* < 8.60×10-13).

In the validation phase, we validated these two G×G signals by using independent participants from UK Biobank and confirmed their significance (rs521828 and rs204999: *OR*interaction = 1.09, 95% CI: 1.00-1.18, *P* = 0.044; rs2853668 and rs62329694: *OR*interaction = 0.83, 95% CI: 0.69-0.98, *P* = 0.034).

To understand better the interaction between rs521828 and rs204999, we also evaluated the association of rs521828 with NSCLC risk stratified by rs204999 using all three cohorts combined. The A allele of rs521828 was significantly associated with a lower risk among subjects carrying the wild genotype (AA) of rs204999 (*OR* = 0.86, 95% CI: 0.80-0.92, *P* = 1.64×10-5), the effect was reversed among those carrying the heterozygous AG genotype of rs204999 (*OR* =1.09, 95% CI: 1.01-1.17, *P* = 2.09×10-2), and the effect became more detrimental among those with the homozygous GG genotype of rs204999 (*OR* = 1.23, 95% CI: 1.06-1.43, *P* = 5.24×10-3). Thus, the effect of rs521828 on NSCLC was modified by rs204999, clearly indicating the existence of their interaction. The pattern was further investigated by a series of stratified analyses (Figure 2A). Similar patterns were observed between rs2853668 and rs62329694. The G allele of rs2853668 was associated with a higher NSCLC risk (*OR* = 1.30, 95% CI: 1.14-1.49, *P* = 1.10×10-4) among subjects carrying GG genotype of rs62329694. But the effect was reversed among subjects carrying the GA (*OR* = 0.87, 95% CI: 0.78-0.99, *P* = 2.74×10-2) and AA (*OR* = 0.75, 95% CI: 0.59-0.99, *P* = 2.64×10-2) genotype of rs2853668, respectively. The pattern was confirmed by sensitivity analyses (Figure 2B).

In addition, we evaluated the interaction pattern (synergetic *vs* antagonistic) for these two pairs of SNPs coded in genetic dominant model (Supplementary Table 1-2). For rs521828 and rs204999, subjects carrying the two wild genotypes (GG genotype of rs521828 and AA genotype of rs204999) were set to be the reference group. The main effects of GA/AA genotype of rs521828 and AG/GG genotype of rs204999 were protective, with an OR of 0.90 and 0.89, respectively. However, their joint effect was harmful, with an OR of 1.04, greater than the product of two main effects (0.90×0.89 = 0.80), indicating an antagonistic effect between rs521828 and rs204999 (*OR*intreaction = 1.28, 95% CI: 1.18-1.39, *P*interaction = 1.23×10-9). Similarly, for rs2853668 and rs62329694, their joint effect conferred an OR of 1.03, which was significantly less than the product of their main effects (1.30×1.16 = 1.51), also indicating an antagonistic effect between them (*OR*intreaction = 0.68, 95% CI: 0.61-0.76, *P*interaction = 1.18×10-10).

All SNPs within the ~500 KB flanking regions of the significant epistatic pairs were further tested by logistic regression models, which detected a cluster of G×G signals enriched in close proximity to the identified pairs (Figure 2C-D).

## Six more significant G×G interactions identified by meta-analysis among Europeans

To identify interactions with weak to moderate effects, we used fixed effect meta-analysis to synthesize results from ILCCO-OncoArray, TRICL and UK Biobank. We also conducted the same analysis among NSCLC subgroups defined by lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), male, female, never smoker, ever smoker (current and former smoker), as well as all histological types of lung cancers, respectively. G×G signals with *P*interaction<5×10-8 were summarized in Supplementary Table S3-S10. A total of 8 pairs of SNPs reached the Bonferroni-corrected threshold (*P* < 8.60×10−13) in various subpopulations (Table 2). Among them, 2 pairs of SNPs were the same as those identified by the two-phase study. Also, among the 6 newly detected G×G interactions, 4 pairs appeared in the 6p21.32 region, including rs3135369 and rs9271300, rs3135369 and rs2858859, rs2858859 and rs9275572, rs2858859 and rs9275596 (Supplementary Figure 1). With a moderate level of linkage disequilibrium (LD) between rs9271300 and rs2858859 (*r*2 = 0.66, *D*’ = 0.996), as well as rs9275596 and rs9275572 (*r*2 = 0.72, *D*’ = 0.998), these 4 G×G signals were likely to be as the result of three SNPs: rs3135369, rs2858859 and rs9275572. All other SNPs in the 6p21.32 region were relatively independent of each other, regardless of LD-*r*2 or *D*’ statistics (Supplementary Table 11). The other 2 pairs of SNPs resided in different regions, including rs28591443 in 8p23.3 and rs9265981 in 6p25.2, and rs589027 in 1q32.2 and rs713395 in 2p24.2.

Although each of these 6 SNP pairs was identified from a specific subpopulation, all of them exhibited nominal significance across all of the subpopulations considered but never smokers with limited sample size (Supplementary Figure 2-4), except for one pair (rs589027 and rs713395) which seemed to be significant only for female (Supplementary Figure 4B).

## Successful trans-ethnic validation of significant G×G interactions in Asians and Europeans

First, we evaluated the 8 G×G interactions identified from the European population by using an external Asian population (10,248 cases and 9,298 controls) from NJMU-GSA (Supplementary Table 12). We were able to validate 3 pairs of SNPs in the 6p21.32 region in several of its subpopulations. They included rs521828 and rs204999 among NSCLC (*OR*interaction = 1.13, 95% CI: 1.03-1.24, *P* = 0.008, *q-FDR* = 0.022), rs3135369 and rs9271300 among NSCLC (*OR*interaction = 0.89, 95% CI: 0.83-0.96, *P* = 0.006, *q-FDR* = 0.022) and Smoker (*OR*interaction = 0.82, 95% CI: 0.72-0.92, *P* = 0.001, *q-FDR* = 0.005), rs3135369 and rs2858859 in NSCLC (*OR*interaction = 1.11, 95% CI: 1.04-1.17, *P* = 5.23×10-4, *q-FDR* = 0.005) (Supplementary Table 13). No significant results were available for the other pairs, possibly due to differences in effect allele frequency (EAF) for SNPs between these two ethnic populations (Figure 3).

Conversely, we validated the only pair of SNPs at 2p32.2 (rs16832404 and rs2562796) that reached genome-wide significance among the Asian population 19, using the European population. This pair indeed exhibited a significant G×G interaction effect on NSCLC risk among the European population (*OR*interaction = 1.11, 95% CI: 1.01-1.22, *P* = 0.028) (Supplementary Table 14).

## SNPs with G×G interactions potentially involved in biological regulatory function

We predicted the functional relevance of 8 pairs of SNPs by SNPinfo, RegulomeDB, and HaploReg v4.1. In the RegulomeDB database, the abundant biological regulatory function was observed for 10 out of 13 SNPs, including expression quantitative trait loci (eQTL), transcription factor binding site or DNase peak. Numerous enhancer histone marks and motifs changed were observed for the SNPs (Supplementary Table 15). In the eQTL analysis using the GTEx database of lung tissues, abundant regulatory relations in the human leukocyte antigen (HLA) region were identified for all 8 SNPs in 6p21.32 and 6p25.2 (Supplementary Table 16), while no significant eQTLs were found for the others. Further, we performed differential expression analysis with the Gene Expression Omnibus (GEO) repository, namely, GSE43458 and GSE12428. For the three genes (*C6orf10*, *CLPTM1L* and *TERT*) identified in the two-phase study, their expression levels were significantly up-regulated in tumor tissues (Supplementary Figure 5). In addition, *BTNL2,* which was additionally identified by the meta-analysis, was significantly differentially expressed between lung cancer tumor and normal tissues (Supplementary Figure 6).

## G×G interaction score effectively distinguish population at high risk in UK Biobank

We developed lung cancer screening models among never smokers and ever smokers, respectively, due to their substantially different genetic backgrounds. For each subgroup, considered for risk screening were: (i) the PRS comprising 128 SNPs with significant marginal effects identified by GWAS in Europeans so far (Supplementary Table 17), (ii) G×G interactions identified in this study and reached nominal significant (*P*<0.05) in never or ever smoking subgroup by meta-analysis of ILCCO-OncoArray, TRICL and UK Biobank, and (iii) G×G interactions selected by group least absolute shrinkage and selection operator (groupLASSO) using ILCCO-OncoArray (training set) with the largest sample size of cases in our study (Supplementary Table 18-19). The interaction empowered polygenetic risk score (iPRS) which incorporating both PRS and G×G interaction score (Supplementary Table 20), outperformed PRS while we categorized subjects into 10 groups by the deciles of the score in ILCCO-OncoArray (Supplementary Figure 7).

iPRS was externally validated among 162,316 never smokers and 245,998 ever smokers with available follow-up time of lung cancer incidence from UK Biobank. The follow-up time was defined as the time lag between entry to cohort and lung cancer diagnosis date or the last follow-up date, whichever came first. Its median was 9.45 years and the inter-quantile range was from 8.48 to 10.52 years. Each subject was assigned an iPRS score (see Methods), and all subjects were categorized into 10 groups by the deciles of the score. Subjects at the high risk group (top 10%) had a significantly higher risk of lung cancer than those at the low risk group (bottom 10%), with a hazard ratio (HR) = 5.31 (95% CI: 3.11-9.07, *P* = 8.60×10-10, Figure 4A) for never smokers, and HR = 6.21 (95% CI: 5.01-7.70, *P* < 2.2×10-16, Figure 4B) for ever smokers. Compared to PRS, iPRS showed a better discrimination power. Meanwhile, we validated the lung cancer screening model composed of demographic variables (age, gender, and packyears) and iPRS, of which weights of SNPs were retained from the training set. The cumulative lung cancer risk curves distinguished obviously from each other across the 5 groups categorized by the quintiles of the overall risk scores of (*P* < 0.001), indicating that the iPRS enhanced model served a good risk classifier (Figure 4C-D).

Age and smoking pack years were two well-recognized factors used to define the high risk population for low-dose computed tomography (LDCT) screening of lung cancer 28. Therefore, we illustrated the absolute incidence of lung cancer in various subpopulations classified by iPRS, age and packyears (Figure 4E). Clinically, the iPRS enhanced model may change the practice of lung cancer screening. For example, subjects aged < 55 years or smoked < 30 pack years (including never smokers), but with a high iPRS, should be re-defined as the high risk population for lung cancer screening; for those with a high iPRS (top 20%) and smoked > 60 pack years, lung cancer screening should start as early as 50 years old; and for those with a low iPRS, screening can be postponed (Figure 4E).

## Genetic variants significantly enriched in biological pathways

To biologically understand the genes mapped to epistatic SNPs in the screening models, we performed gene enrichment pathway analyses with the KEGG database for ever and never smokers separately. A total of 16 pathways were significant among never smokers, such as Cell adhesion molecules and Allograft rejection (Supplementary Figure 8). For ever smokers, 22 pathways were identified, including the well-known pathways such as Th1 and Th2 cell differentiation, Notch signaling pathway, Leishmaniasis, indicating more biological pathways were involved in smoking behaviors leading to tumorigenesis (Supplementary Figure 9).

# Discussion

To our knowledge, this is the largest and the most comprehensive G×G interaction study of NSCLC risk on the genome-wide scale. We identified a total of 8 pairs of SNPs that predominantly appeared in the 6p21.32 and 5p15.33 regions. Even with ethnic differences between the European and Asian populations, our trans-ethnic validation found that 3 out of 5 pairs of SNPs in the 6p21.32 region remained significant in both populations. Furthermore, we developed an iPRS enhanced lung cancer screening model by incorporating G×G signals, which outperformed the classic model with PRS only, and can facilitate screening high risk subpopulations.

Due to computational constraints, very few genome-wide G×G interaction studies are available for a limited number of diseases (Supplementary Table 21), including prostate cancer 16-18, colorectal cancer 19, nasopharyngeal cancer 21, breast cancer 20 and our previous Asian lung cancer study 11. To address the computing challenge, we adopted a ***Screening before Testing*** strategy 16,18,29 to efficiently extract G×G signals from over 58 billion of SNP pairs, while maintaining the type I error and increasing the statistical power. Though focused on lung cancer, we envision broad applications of this strategy to the other diseases.

A multi-phase study design is another commonly used strategy to increase the reproducibility of association results. We used a two-phase study, which involves discovery and validation phases, to identify two pairs of SNPs. Moreover, to detect G×G interactions with weak to moderate effect sizes, we resorted to meta-analysis by pooling subpopulations of interest from all three cohorts to boost substantially statistical power 30. As a result, we identified six more pairs of SNPs, of which 5 pairs exhibited acceptable significance across all of the subpopulations considered, except for one female-specific epistasis, the mechanism of which warrants further research.

PRS is a popular approach for identifying individual-level genetic risks of lung cancer 26,31. However, with weak marginal effects of individual SNPs, the stratification performance of PRS based models is generally unsatisfactory 32, resulting in a severe missing heritability issue. By incorporating two-way interactions into the screening model, the discrimination ability has improved much, as confirmed in all three independent cohorts. Therefore, as pointed out in our previous prognostic prediction of lung cancer 33, complex association patterns (e.g., G×G interactions) among multiple factors should be factored in for studies of complex diseases (e.g., lung cancer).

Six out of eight significant pairs of SNPs were found to be located in the 6p21.32 and 5p15.33 regions. For example, one pair mapped to *TERT* and *CLPTM1L* are in 5p15.33, a well-known region reported by GWAS of lung cancer risk in Asians 34,35, African-Americans 36, European 37, as well as for lung cancer prognosis 38. A GWAS by McKay et al suggested two genes, telomerase reverse transcriptase (*TERT*) and cleft lip and palate transmembrane 1 like (*CLPTM1L*) play a role in the development of lung cancer 39. This current study reported their interaction effect for the first time. Interestingly, the two genes are all involved in tumor anti-apoptosis 40,41. *TERT* plays a role in cellular senescence as it is normally repressed in postnatal somatic cells, resulting in shortening of telomeres, and, therefore, aging and anti-apoptosis. Deregulation of telomerase expression in somatic cells may be involved in oncogenesis 42. *CLPTM1L* is a commonly overexpressed anti-apoptotic factor in lung tumors, and is associated with DNA damage measured by bulky aromatic/hydrophobic DNA adducts 43. Knockdown of *CLPTM1L* transcript in NSCLC cells results in increased sensitivity to genotoxic stress-mediated apoptotic killing, and diminishes the expression of Bcl-xL in a manner depending on *CLPTM1L* expression 44.

Another five pairs of SNPs reside in 6p21.32, where *HLA-DQA1, HLA-DQA2,* and *BTNL2* are located. This region was reported to be associated with lung cancer risk among Asians 45,46, and we now report the G×G signals in this region for Europeans. Proline rich transmembrane protein 1 (*PRRT1*) and *C6orf10* [also known as testis expressed basic protein 1 (*TSBP1*)] are located on the major histocompatibility complex (MHC) region, widely recognized as an important regulatory region for multiple diseases 47. The two SNPs (rs3135369 and rs2858859) also showed abundant eQTL relationship with the genes in HLA, known as a critical mediator in disease defense through presenting intra- or extra-cellular peptides on the cell surface in a form, which can be recognized by the T cell receptors (TCR) and activate a specific T cell response 48. Thus, identifying polymorphism signals controlling the expression of specific HLA molecules and affecting the peptide binding groove or the contact surface with the TCR may help disentangle lung cancer MHC associations, shedding new light on cancer risk and possible immunotherapy targets 49.

The last two pairs of SNPs (rs9265981 and rs28591443; rs589027 and rs713395) were mapped to *HLA-B*, *CSMD1*, *HHAT* and *AC008069.1* in four different regions. *HLA-B* belongs to the HLA class I heavy chain paralogues; deregulation of *CSMD1* is associated with cancer progression and poor survival through the NF-κB pathway in gastric cancer 50; *HHAT* regulates the proliferation of estrogen receptor cells in breast cancer and the *HHAT* inhibitor plays a critical role for therapeutic benefits 51. However, biological functions of lncRNA *AC008069.1* remain unknown.

Our study has several strengths. First, this is perhaps the largest G×G interaction study of lung cancer risk by utilizing consortium resources, and also the first G×G interaction study among the European population, providing evidence beyond an Asian population 11. Second, to address the issue of analyzing an extremely large number of G×G interactions, we performed data mining by integrating various statistical and machine learning tools; and to investigate the robustness of the results, we conducted a series of stratified analyses. While we used the conservative Bonferroni method to control the false positives and to ensure the reproducibility of the results, our stringent procedure detected eight significant pairs of SNPs. Third, even with ethnic differences between the European and Asian populations, we performed trans-ethnic validation of significant G×G signals identified in this study that focuses on Europeans and a previous Asian study, and found that four pairs of SNPs maintained statistical significance in both populations. Finally, we developed an iPRS enhanced lung cancer screening model with independent validation in UK Biobank among never and ever smokers, respectively. The model may lay a theoretical groundwork for precision prevention of lung cancer among Europeans.

There are some limitations with this study. First, we only focused on two-way interactions in the study, as the computation burden of high-order interactions is prohibitive (e.g., there are 6,606 trillion three-way interactions from the SNPs considered in this work) and the interpretation of high-order interactions is more complex. Second, we did not verify the biological mechanisms of the SNPs involved in the identified G×G interactions, which may warrant further functional studies. Third, since this study was primarily designed for a European population with the majority of participants being Europeans in ILCCO-OncoArray, TRICL and UK Biobank, future G×G interaction studies on subjects with African-American ancestry are needed. Fourth, these G×G interactions included in the screening models were selected by groupLASSO in ILCCO-OncoArray, without being further validated UK Biobank.

# Methods

## Study population in global consortiums

**ILCCO-OncoArray**: OncoArray Consortium is a network created to increase understanding of the genetic architecture of common cancers. The OncoArray GWAS was originally designed to profiled genotype information of 57,775 participants, obtained from 29 studies across North America, Europe, and Asia 23. All participants signed the informed consent, and the studies were approved by the local internal review boards or ethics committees and administered by trained personnel.

**TRICL**: TRICL Research Team is part of the Genetic Associations and MEchanisms in ONcology (GAME-ON) Consortium 24. Tumors from patients were classified as adenocarcinomas, squamous carcinomas, large-cell carcinomas, mixed adenosquamous carcinomas and other NSCLC histologies following either the International Classification of Diseases for Oncology (ICD-O) or World Health Organisation coding. The TRICL GWAS was originally designed to profiled genotype information of 12,651 participants. All participants provided informed written consent. All studies were reviewed and approved by institutional ethics review committees.

All the duplicated samples between ILCCO-OncoArray and TRICL have been removed from ILCCO-OncoArray dataset.

**UK Biobank**: The UK Biobank is a large prospective study of individuals aged 40-70 years at assessment 25, who attended assessment centers between 2006 and 2010, and contributed blood samples for genotyping and blood analysis and answered questionnaires about medical history and environmental exposures. In the years since assessment, health outcome data for these individuals (e.g., diagnoses of cancer) have been accruing via UK national registries and hospital records. Lung cancer cases were collected based on the ICD10 code of cancer diagnosis (Filed ID: 40006, 41202) or self-reported lung cancer histology (Filed ID: 20001).

## Genotyping and quality control of GWAS data

Genotyping of 533,631 SNPs in ILCCO-OncoArray was completed at the Center for Inherited Disease Research, the Beijing Genome Institute, the Helmholtz Zentrum München, Copenhagen University Hospital and the University of Cambridge in Illumina Infinium OncoArray platform. Details of quality control (QC) procedures were described in a previous study 52. Briefly, before standard QC, we removed the intentionally duplicated samples and samples from unrelated OncoArray studies and HapMap control individuals of European, African, Chinese and Japanese origins. Further excluded were those who lacked disease status, were second-degree relatives or closer having identity by descent (IBD) > 0.2 or had low-quality DNA (call rate < 95%), or sex inconsistency, or were non-European. SNPs were removed if meeting any of the following criteria: (1) sex chromosome, (2) minor allele frequency (MAF) < 0.05, (3) call rate < 95%, and (4) Hardy-Weinberg equilibrium (HWE) test *P* < 1.00×10−7 in controls or *P* < 1.00×10−12 in cases. Finally, a total of 28,353 participants (15,157 cases and 13,196 controls) with 340,958 qualified SNPs remained in the subsequent association analysis. To explore the potential functional variants, we extracted the genotyped data in the flanking regions with the imputed data 52.

The genotype data of TRICL were generated from the Affymetrix Axiom Array, which contained 414,504 markers. To estimate missing genotype information, we phased haplotypes with Eagle v2.3 using 1000 Genomes Project data (phase 3) as a reference panel 53 and then performed imputations using the Minimac (version 3) software. SNPs with an imputation quality score *R2* < 0.4, MAF < 0.01, or *P* < 1×10-6 for the HWE test were excluded from analyses.

We analyzed the imputed genetic data from the full UK Biobank cohort, consisting of 488,377 individuals genotyped on the Affymetrix UK BiLEVE and UK Biobank Axiom arrays, and applied the same quality control procedures. We included 409,615 European participants (3,017 cases and 406,598 controls) as the independent validation set.

## A two-phase study of G×G interaction in Europeans

We adopt a two-phase design to identify G×G interactions. Significant G×G signals identified in the discovery phase using ILCCO-OncoArray and TRICL, were further confirmed in the validation phase using UK Biobank (Figure 1). In view of over 58 billion possible G×G interactions considered in our study, we employed ***Screening before Testing*** for dimensional reduction in the discovery phase.

***Screening step:*** BiForce, a Java program that integrates bitwise computing with multithreaded parallelization and allows rapid full pairwise genome scans, was applied to scan billions of GxG interactions exhaustively in ILCCO-OncoArray 27, and select potential G×G interactions by using the criterion of the log likelihood difference between two log-linear models with and without the interaction term, defined as, where *n* is the sample size,  is the observed frequency of subjects with SNP1 coded *i* (0,1, and 2), SNP2 coded *j* (0,1, and 2) and disease status coded *k* (0 and 1). And, was the expected frequency under null hypothesis. BiForce utilized Kirkwood superposition approximation (KSA) instead of likelihood estimation to calculate. KSA, without an iterative process, enables BiForce to quickly scan all pairs of SNP combinations, while capturing positive signals to the extent possible.

***Testing step:*** Because of computational constraints, it was unrealistic to use logistic models directly to exhaustively test all 58 billion G×G interactions.Instead, weused the top SNP pairs selected by BiForce 27, with the default setting to filter noises (*P*BiForce ≤ 1.00×10-6). The top pairs were retested through LRM adjusted for covariates.



Where, *β*1, *β*2 and *β*3 were the main effects of SNP1 and SNP2, and their interaction effect, respectively, and *Covi* were age, gender, smoking status and top three PCs derived from GWAS data for population structure correction.

The interaction effect was estimated in ILCCO-OncoArray and TRICL, respectively. Meta-analysis pooled the estimates from ILCCO-OncoArray and TRICL for a more robust and efficient estimate. Normally, the genome-wide significance level using the Bonferroni correction method was defined as 8.60×10-13 = 0.05/*C*(340958, 2), where *C(n,r)* is the combination formula and 340,958 was the number of qualified SNPs that passed QC. All the significant G×G interactions in the discovery phase will be independently validated in the validation phase. SNP pairs with a *P* value ≤ 0.05 and a consistent direction in the validation phase were defined as overall significant G×G signals.

*Meta-analysis of G×G interaction in* *Europeans*

Meta-analysis of ILCCO-OncoArray, TRICL and UK Biobank using fixed effect model was applied to detect weak to moderate effects of G×G interactions in serval NSCLC subgroups, including LUAD, LUSC, male, female, never smoker, ever smoker and all histological types of lung cancers.

## Trans-ethnic validation of significant G×G signals in Asians and Europeans

We extracted all SNPs having significant G×G interaction in a Han Chinese population from NJMU-GSA, including 19,546 participants (10,248 cases and 9,298 controls) 26. QC procedures for genotypes were similar to those in OncoArray, except for the HWE test with *P* < 1.00×10-5 in all participants. G×G interactions were analyzed through logistic regression models adjusted for the same covariates aforementioned. FDR correction using the Benjamini & Hochberg method 54, was applied to adjust *P* values for multiple comparisons.

## In-silico functional validation of the SNPs with G×G interaction

We used an *in-silico* approach through SNPinfo 55, RegulomeDB 56, and HaploReg v4.1 57, to predict potential functions of the identified SNPs. eQTL was analyzed using the 578 lung tissue in GTEx project 58. To concordantly analyze expression profiles from tumor and health lung tissues, we collected data from GEO repository, including GSE43458 (80 cases and 30 controls) and GSE12428 (34 cases and 28 controls). We statistically normalized the data before analysis and used Student’s *t*-test to compare the differences between tumor and normal tissues.

## Development of an iPRS enhanced lung cancer screening model in ILCCO-OncoArray population

The iPRS enhanced lung cancer screening model was first developed in never and ever smokers, respectively, using ILCCO-OncoArray as training set. Besides age, gender and pack-years of smoking, three components were utilized to screen lung cancer risk. (i) PRS, constructed as the sum of the number of minor alleles of SNPs, weighted by their effect size as ln(odds ratio). These SNPs are collected from the known susceptibility loci of lung cancer and conditions related to lung cancer (such as lung function impairment) previously identified through literature curation and NHGRI-EBI GWAS Catalog, and additional loci that passed the suggestive significance-level in GWAS studies. When correlation exists, variants representing independent loci with the strongest statistical significance were retained. Finally, 128 SNPs were used to generate PRS 59. (ii) G×G interactions identified in this study and reached nominal significant (*P*<0.05) in never or ever smoking subgroups by fixed effect meta-analysis of ILCCO-OncoArray, TRILC and UK Biobank. (iii) G×G interactions additionally selected among SNP pairs with *P*interaction ≤ 5×10-8 in the meta-analysis by groupLASSO with the tuning parameter lambda determined via 5-fold cross validation 60, and the coefficients were estimated by adjusted logistic regression models. G×G interaction score were defined by following scoring process:



The iPRS composed of classic PRS and G×G interaction score were utilized to generate a lung cancer screening model together with age, gender and pack-years of smoking.

## Gene enrichment pathway analysis

We collected the pathway information with gene sets from the KEGG database, containing a total of 186 pathways up to July 2021. All enrichment analyses were performed using the R package *clusterProfiler* 61.

## Validation of the iPRS enhanced screening model in UK Biobank population

With follow-up of lung cancer incidence available in the UK Biobank cohort, we used Cox proportional hazards models to test the association between genetic risk score and lung cancer incidence, after adjusting for age, sex, source of region, and smoking status. We classified participants into ten groups by the deciles of genetic risk score, and measured by hazard ratio the difference in lung cancer risk among these groups.

All statistical analyses were performed using R version 3.6.3 (The R Foundation). *P* values were two-sided, and *P* < 0.05 was considered statistically significant.

# Data Availability

ILCCO-Oncoarray data is available from: <https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001273.v3.p2>.

TRICL data is available from: <https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001681.v1.p1>.

UK Biobank data is available from <https://www.ukbiobank.ac.uk/>.

KEGG database is available from <http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=CP:KEGG>.

# Code Availability

The R software codes that support our findings are available from the corresponding author by a reasonable request.

## Acknowledgments

This study was funded by National Key Research and Development Program of China (2016YFE0204900 to F.C.), National Natural Science Foundation of China (82103946 to S.S. and 81973142 to Y.W.), Natural Science Foundation of Jiangsu Province (BK20191354 to R.Z.), China Postdoctoral Science Foundation (2020M681671 to S.S.), Jiangsu Planned Projects for Postdoctoral Research Funds (2020Z019 to S.S.), the US National Institutes of Health (CA209414, CA092824 and ES000002 to D.C.C., AG056764 to Y.L.), Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). R.Z. was partially supported by the Qing Lan Project of the Higher Education Institutions of Jiangsu Province and the Outstanding Young Teachers Training Program of Nanjing Medical University.

## Author contributions

R.Z., Y. Wei, S.S., F.C., Z.H., H.S, R.H., C.I.A. and D.C.C. designed the study. R.Z., Y. Wei, Y. Zhu, S.S., Y. Zhao, Y. Wang, M.Z., X.X., D.Z. and Yafang L. contributed to data management and statistical analysis. R.Z., S.S., Y. Zhu, Y. Wei and Yi L. wrote the manuscript. D.A., M.T.L, N.C., S.L., A.T., C.C., S.E.B., M.J., A.R., H.B., H.E.W., G.R., S.A., P.B., J.M., J.K.F., S.S.S., L.L.M., O.M., H.B., G.L., A.A., L.A.K., S.Z.N., K.G., M.J., A.C., P.L., M.B.S., M.C.A., R.H., C.I.A. and D.C.C. contributed to Sample collection and development of the epidemiological studies. All authors conducted data preparation, discussed the results and revised the manuscript.

## Conflict of interest statement

The authors declare no potential conflicts of interest.

# Reference

1 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA: a cancer journal for clinicians* **69**, 7-34, doi:10.3322/caac.21551 (2019).

2 Shen, S. *et al.* A multi-omics study links TNS3 and SEPT7 to long-term former smoking NSCLC survival. *NPJ Precis Oncol* **5**, 39, doi:10.1038/s41698-021-00182-3 (2021).

3 Mucci, L. A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *Jama* **315**, 68-76, doi:10.1001/jama.2015.17703 (2016).

4 Timofeeva, M. N. *et al.* Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet* **21**, 4980-4995, doi:10.1093/hmg/dds334 (2012).

5 Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet* **46**, 736-741, doi:10.1038/ng.3002 (2014).

6 Bosse, Y. & Amos, C. I. A Decade of GWAS Results in Lung Cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **27**, 363-379, doi:10.1158/1055-9965.Epi-16-0794 (2018).

7 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 (2009).

8 Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40**, 695-701, doi:10.1038/ng.f.136 (2008).

9 Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253-1260, doi:10.1038/ng.237 (2008).

10 Zhang, R. *et al.* A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. *Carcinogenesis* **35**, 1528-1535, doi:10.1093/carcin/bgu076 (2014).

11 Chu, M. *et al.* A genome-wide gene-gene interaction analysis identifies an epistatic gene pair for lung cancer susceptibility in Han Chinese. *Carcinogenesis* **35**, 572-577, doi:10.1093/carcin/bgt400 (2014).

12 Liu, C. Y. *et al.* Genome-wide Gene-Asbestos Exposure Interaction Association Study Identifies a Common Susceptibility Variant on 22q13.31 Associated with Lung Cancer Risk. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **24**, 1564-1573, doi:10.1158/1055-9965.Epi-15-0021 (2015).

13 Malhotra, J. *et al.* Effect of occupational exposures on lung cancer susceptibility: a study of gene-environment interaction analysis. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **24**, 570-579, doi:10.1158/1055-9965.Epi-14-1143-t (2015).

14 Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 1193-1198, doi:10.1073/pnas.1119675109 (2012).

15 Ashworth, A., Lord, C. J. & Reis-Filho, J. S. Genetic interactions in cancer progression and treatment. *Cell* **145**, 30-38, doi:10.1016/j.cell.2011.03.020 (2011).

16 Tao, S. *et al.* Genome-wide two-locus epistasis scans in prostate cancer using two European populations. *Human genetics* **131**, 1225-1234, doi:10.1007/s00439-012-1148-4 (2012).

17 Ciampa, J. *et al.* Large-scale exploration of gene-gene interactions in prostate cancer using a multistage genome-wide association study. *Cancer Res* **71**, 3287-3295, doi:10.1158/0008-5472.Can-10-2646 (2011).

18 Shen, J., Li, Z., Song, Z., Chen, J. & Shi, Y. Genome-wide two-locus interaction analysis identifies multiple epistatic SNP pairs that confer risk of prostate cancer: A cross-population study. *Int J Cancer* **140**, 2075-2084, doi:10.1002/ijc.30622 (2017).

19 Jiao, S. *et al.* Genome-wide search for gene-gene interactions in colorectal cancer. *PloS one* **7**, e52535, doi:10.1371/journal.pone.0052535 (2012).

20 Milne, R. L. *et al.* A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46,450 cases and 42,461 controls from the breast cancer association consortium. *Hum Mol Genet* **23**, 1934-1946, doi:10.1093/hmg/ddt581 (2014).

21 Su, W. H. *et al.* How genome-wide SNP-SNP interactions relate to nasopharyngeal carcinoma susceptibility. *PloS one* **8**, e83034, doi:10.1371/journal.pone.0083034 (2013).

22 Ji, X. *et al.* Identification of susceptibility pathways for the role of chromosome 15q25.1 in modifying lung cancer risk. *Nat Commun* **9**, 3221, doi:10.1038/s41467-018-05074-y (2018).

23 McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* **49**, 1126-1132, doi:10.1038/ng.3892 (2017).

24 Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet* **46**, 736-741, doi:10.1038/ng.3002 (2014).

25 Sudlow, C. *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* **12**, e1001779, doi:10.1371/journal.pmed.1001779 (2015).

26 Dai, J. *et al.* Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. *The Lancet. Respiratory medicine* **7**, 881-891, doi:10.1016/s2213-2600(19)30144-4 (2019).

27 Gyenesei, A. *et al.* BiForce Toolbox: powerful high-throughput computational analysis of gene-gene interactions in genome-wide association studies. *Nucleic Acids Res* **40**, W628-632, doi:10.1093/nar/gks550 (2012).

28 Mazzone, P. J. *et al.* Screening for Lung Cancer: CHEST Guideline and Expert Panel Report. *Chest*, doi:10.1016/j.chest.2021.06.063 (2021).

29 Milne, R. L. *et al.* A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46,450 cases and 42,461 controls from the breast cancer association consortium. *Hum Mol Genet* **23**, 1934-1946, doi:10.1093/hmg/ddt581 (2014).

30 McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* **49**, 1126-1132, doi:10.1038/ng.3892 (2017).

31 Lambert, S. A. *et al.* The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet* **53**, 420-425, doi:10.1038/s41588-021-00783-5 (2021).

32 Choi, S. W., Mak, T. S. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15**, 2759-2772, doi:10.1038/s41596-020-0353-1 (2020).

33 Zhang, R. *et al.* Independent Validation of Early-Stage Non-Small Cell Lung Cancer Prognostic Scores Incorporating Epigenetic and Transcriptional Biomarkers With Gene-Gene Interactions and Main Effects. *Chest* **158**, 808-819, doi:10.1016/j.chest.2020.01.048 (2020).

34 Dong, J. *et al.* Fine mapping of chromosome 5p15.33 identifies novel lung cancer susceptibility loci in Han Chinese. *Int J Cancer* **141**, 447-456, doi:10.1002/ijc.30702 (2017).

35 Hu, Z. *et al.* A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat Genet* **43**, 792-796, doi:10.1038/ng.875 (2011).

36 Zanetti, K. A. *et al.* Genome-wide association study confirms lung cancer susceptibility loci on chromosomes 5p15 and 15q25 in an African-American population. *Lung Cancer* **98**, 33-42, doi:10.1016/j.lungcan.2016.05.008 (2016).

37 Pande, M. *et al.* Novel genetic variants in the chromosome 5p15.33 region associate with lung cancer risk. *Carcinogenesis* **32**, 1493-1499, doi:10.1093/carcin/bgr136 (2011).

38 Chen, Z. *et al.* The associations of TERT-CLPTM1L variants and TERT mRNA expression with the prognosis of early stage non-small cell lung cancer. *Cancer Gene Ther* **24**, 20-27, doi:10.1038/cgt.2016.74 (2017).

39 McKay, J. D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**, 1404-1406, doi:10.1038/ng.254 (2008).

40 Liu, R. *et al.* Therapeutic targeting of FOS in mutant TERT cancers through removing TERT suppression of apoptosis via regulating survivin and TRAIL-R2. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2022779118 (2021).

41 James, M. A., Vikis, H. G., Tate, E., Rymaszewski, A. L. & You, M. CRR9/CLPTM1L regulates cell survival signaling and is required for Ras transformation and lung tumorigenesis. *Cancer research* **74**, 1116-1127, doi:10.1158/0008-5472.CAN-13-1617 (2014).

42 Calado, R. T. & Chen, J. Telomerase: not just for the elongation of telomeres. *Bioessays* **28**, 109-112, doi:10.1002/bies.20365 (2006).

43 Zienolddiny, S. *et al.* The TERT-CLPTM1L lung cancer susceptibility variant associates with higher DNA adduct formation in the lung. *Carcinogenesis* **30**, 1368-1371, doi:10.1093/carcin/bgp131 (2009).

44 Gealy, R., Zhang, L., Siegfried, J. M., Luketich, J. D. & Keohavong, P. Comparison of mutations in the p53 and K-ras genes in lung carcinomas from smoking and nonsmoking women. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **8**, 297-302 (1999).

45 Lan, Q. *et al.* Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet* **44**, 1330-1335, doi:10.1038/ng.2456 (2012).

46 Wang, G. *et al.* Daily cooking duration and its joint effects with genetic polymorphisms on lung cancer incidence: Results from a Chinese prospective cohort study. *Environ Res* **179**, 108747, doi:10.1016/j.envres.2019.108747 (2019).

47 Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nature reviews. Immunology* **18**, 325-339, doi:10.1038/nri.2017.143 (2018).

48 Ferreiro-Iglesias, A. *et al.* Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. *Nat Commun* **9**, 3927, doi:10.1038/s41467-018-05890-2 (2018).

49 Beatty, G. L. & Gladney, W. L. Immune escape mechanisms as a guide for cancer immunotherapy. *Clinical cancer research : an official journal of the American Association for Cancer Research* **21**, 687-692, doi:10.1158/1078-0432.CCR-14-1860 (2015).

50 Chen, X. L. *et al.* Deregulation of CSMD1 targeted by microRNA-10b drives gastric cancer progression through the NF-kappaB pathway. *Int J Biol Sci* **15**, 2075-2086, doi:10.7150/ijbs.23802 (2019).

51 Matevossian, A. & Resh, M. D. Hedgehog Acyltransferase as a target in estrogen receptor positive, HER2 amplified, and tamoxifen resistant breast cancer cells. *Mol Cancer* **14**, 72, doi:10.1186/s12943-015-0345-x (2015).

52 Amos, C. I. *et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **26**, 126-135, doi:10.1158/1055-9965.Epi-16-0106 (2017).

53 Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics* **48**, 1443 (2016).

54 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B: Methodological* **57**, 289-300 (1995).

55 Xu, Z. & Taylor, J. A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic acids research* **37**, W600-W605 (2009).

56 Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome research* **22**, 1790-1797 (2012).

57 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research* **40**, D930-D934 (2011).

58 Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).

59 Hung, R. J. *et al.* Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model. *Cancer research* **81**, 1607-1615, doi:10.1158/0008-5472.CAN-20-1237 (2021).

60 Breheny, P. & Huang, J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing* **25**, 173-187, doi:10.1007/s11222-013-9424-2 (2015).

61 Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* **16**, 284-287 (2012).

# Tables

## **Table 1.** Demographic and clinical descriptions of non-small cell lung cancer cases and controls in ILCCO-OncoArray, TRICL and UK Biobank.

|  |  |  |  |
| --- | --- | --- | --- |
| Characteristics | ILCCO-OncoArray | TRICL | UK Biobank |
| Case (*N* =15,157) | Control (*N* =13,196) | Case (*N* =3,288) | Control (*N* =3,965) | Case (*N* =3,017) | Control (*N* =406,598) |
| Age | 63.66 ± 10.75 | 61.72 ± 11.38 | 61.76 ± 10.56 | 58.7 ± 9.53 | 61.89 ± 5.88 | 56.88 ± 8.00 |
| Gender (%) |  |  |  |  |  |  |
| Male | 9,778 (64.5) | 7,967 (60.4) | 1,643 (50.0) | 2,028 (51.1) | 1,398 (46.3) | 219,979 (54.1) |
| Female | 5,376 (35.5) | 5,228 (39.6) | 1,641 (50.0) | 1,937 (48.9) | 1,619 (53.7) | 186,619 (45.9) |
| Smoking status (%) |  |  |  |  |  |  |
| Never  | 1,403 (9.3) | 3,981 (30.2) | 264 (8.0) | 1,023 (25.8) | 357 (11.8) | 161,959 (39.8) |
| Ever | 13,461 (88.8) | 8,908 (67.5) | 3,024 (92.0) | 2,942 (74.2) | 2,642 (87.6) | 243,356 (59.9) |
| Smoking pack years (mean ± SD) | 35.84 ± 34.77 | 16.42 ± 27.41 | 40.96 ± 30.9 | 26.26 ± 26.96 | 39.76 ± 24.93 | 23.18 ± 18.55 |
| Histology (%) |  |  |  |  |  |  |
| NSCLC | 10,997 (72.6) | - | 1,952 (59.3) | - | 1,731 (57.3) | - |
| LUAD | 6,158 (40.6) | - | 1,296 (39.4) | - | 944 (31.3) | - |
| LUSC | 3,886 (25.6) | - | 513 (15.6) | - | 569 (18.9) | - |
| LSCC | 1,564 (10.3) | - | 310 (9.4) | - | 252 (8.4) | - |

NSCLC: non-small cell lung cancer; LUAD: lung adenocarcinoma; LUSC: lung squamous carcinoma; LSCC: lung small cell carcinoma.

Ever smoker was composed of former and current smokers.

## Table 2. The 8 pairs of SNPs that reached the Bonferroni-corrected significance threshold in the meta-analysis of ILCCO-OncoArray, TRICL and UK Biobank.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SNP 1 | SNP 2 | Population | OR | 95% CI | *P* |
| Region | SNP | Nearest Gene | EAF | Region | SNP | Nearest Gene | EAF |
| 5p15.33 | rs2853668 | *TERT* | 0.276 | 5p15.33 | rs62329694 | *CLPTM1L* | 0.051 | NSCLC# | 0.73 | 0.67-0.79 | 2.70×10-13 |
| All† | 0.74 | 0.68-0.80 | 5.39×10-13 |
| 6p21.32 | rs521828 | *C6orf10* | 0.290 | 6p21.32 | rs204999 | *PRRT1* | 0.257 | NSCLC# | 1.17 | 1.12-1.22 | 6.57×10-13 |
| 6p21.32 | rs3135369 | *BTNL2* | 0.266 | 6p21.32 | rs9271300 | *HLA-DQA1* | 0.447 | NSCLC | 0.86 | 0.82-0.89 | 1.93×10-14 |
| All† | 0.86 | 0.83-0.90 | 1.59×10-13 |
| Smoker | 0.84 | 0.80-0.88 | 6.12×10-15 |
| 6p21.32 | rs3135369 | *BTNL2* | 0.266 | 6p21.32 | rs2858859 | *HLA-DQA1* | 0.452 | NSCLC | 1.17 | 1.12-1.21 | 2.43×10-13 |
| All† | 1.16 | 1.11-1.20 | 8.51×10-13 |
| 6p21.32 | rs2858859 | *HLA-DQA1* | 0.452 | 6p21.32 | rs9275572 | *HLA-DQA2* | 0.394 | Smoker | 1.15 | 1.11-1.20 | 2.84×10-13 |
| 6p21.32 | rs2858859 | *HLA-DQA1* | 0.452 | 6p21.32 | rs9275596 | *HLA-DQA2* | 0.318 | Smoker | 1.16 | 1.11-1.21 | 4.41×10-13 |
| 8p23.3 | rs28591443 | *CSMD1* | 0.066 | 6p25.2 | rs9265981 | *HLA-B* | 0.275 | LUAD | 1.50 | 1.35-1.68 | 6.11×10-13 |
| 1q32.2 | rs589027 | *HHAT* | 0.328 | 2p24.2 | rs713395 | *AC008069.1* | 0.251 | Female | 0.78 | 0.73-0.83 | 6.85×10-13 |

#These G×G signals reached the Bonferroni-corrected significance threshold (*P* < 8.60×10−13) in the discovery phase by meta-analysis of ILCCO-OncoArray and TRICL, and remained significant (*P* < 0.05) in the validation phase using UK Biobank.

EAF: effect allele frequency; OR: odds ratio; 95% CI: 95% confidence interval; LUAD: lung adenocarcinoma.

†All includes lung cancer cases with all histological types.

# Figure Legends

**Figure 1**. The work flow diagram of this study. We adopt a two-phase design in genome-wide G×G interaction study. In discovery phase, a two-step strategy, ***Screening before Testing***, was used for high-dimensionality reduction using European participants from ILCCO-OncoArray and TRICL. In validation phase, Bonferroni-corrected significant G×G interactions were further confirmed in UK Biobank. Meanwhile, meta-analysis of ILCCO-OncoArray, TRICL and UK Biobank, as well as stratified analysis, was additionally performed to identify weak effect G×G signals. Trans-ethnic validation of G×G interactions was conducted using Asian participants from NJMU-GSA. An improved lung cancer screening model incorporating polygenetic risk score and G×G interaction score was also developed.

**Figure 2**. Forest plot of G×G interactions for (**A**) rs204999×rs521828 and (**B**) rs2853668×rs62329694 using European participants from ILCCO-OncoArray, TRICL and UK Biobank. The 3D G×G interaction signal map for association results of all epistatic pairs upstream and downstream of the identified G×G interaction using imputed data in (**C**) 6p21.32 and (**D**) 5p15.33 regions. The *P* values were derived from the LRM adjusted for covariates and pooled by meta-analysis of ILCCO-OncoArray, TRICL and UK Biobank. *P* values were plotted on a negative log10-transformed scale.

**Figure 3**. The comparison of G×G interaction association results and effect allele frequency between Europeans and Asians. Star symbol (\*) indicates that G×G interaction is significant in both Europeans and Asians.

**Figure 4**. Participants in UK Biobank were divided into ten equal groups according to the RPS and iPRS, respectively. HR and 95% CI of each group were derived from Cox proportional hazards model adjusted for covariates by setting the lowest group as reference for never (**A**) and ever smokers (**B**). Cumulative lung cancer incidence curves were illustrated for subjects at different G×G groups calculated from demographic variables (age, gender, and packyears) and iPRS for never (**C**) and ever smokers (**D**). HR and 95% CI were derived from proportional hazards model adjusted for covariates by setting the lowest group as reference. The absolute lung cancer incidence rates were presented for subjects at different iPRS, pack years and age groups (**E**).