ORIGINAL ARTICLE

A Large-Scale Genome-Wide Gene-Gene Interaction Study of Lung Cancer Susceptibility in Europeans With a Trans-Ethnic Validation in Asians

Ruyang Zhang, PhD,[a,b,c] Sipeng Shen, PhD,[a,b,c,d] Yongyue Wei, PhD,[a,b,c]
Ying Zhu, PhD,[a] Yi Li, PhD,[e] Jiajin Chen, BSc,[a] Jinxing Guan, BSc,[a]
Zoucheng Pan, BSc,[a] Yuzhuo Wang, PhD,[f,g] Meng Zhu, PhD,[f,g] Junxing Xie, MMed,[f]
Xiangjun Xiao, MSc,[h] Dakai Zhu, MSc, MBA,[h] Yafang Li, PhD,[h]
Demetrios Albanes, MD,[i] Maria Teresa Landi, MD, PhD,[i] Neil E. Caporaso, MD,[i]
Stephen Lam, MD,[j] Adonina Tardon, PhD,[k] Chu Chen, PhD,[l] Stig E. Bojesen, MD,[m]
Mattias Johansson, PhD,[n] Angela Risch, PhD,[o] Heike Bickeböller, PhD,[p]
H-Erich Wichmann, MD, PhD,[q] Gadi Rennert, MD, PhD,[r] Susanne Arnold, MD,[s]
Paul Brennan, PhD,[n] James D. McKay, PhD,[n] John K. Field, PhD,[t]
Sanjay S. Shete, PhD,[u] Loic Le Marchand, MD, PhD,[v] Geoffrey Liu, MD, MSc,[w]
Angeline S. Andrew, PhD,[x] Lambertus A. Kiemeney, PhD,[y]
Shan Zienolddiny-Narui, PhD,[z] Annelie Behndig, MD, PhD,[aa]
Mikael Johansson, MD, PhD,[bb] Angela Cox, PhD,[cc] Philip Lazarus, PhD,[dd]
Matthew B. Schabath, PhD,[ee] Melinda C. Aldrich, PhD,[ff] Juncheng Dai, PhD,[f,g]
Hongxia Ma, PhD,[f,g] Yang Zhao, PhD,[a] Zhibin Hu, PhD,[c,f,g] Rayjean J. Hung, PhD,[gg]
Christopher I. Amos, PhD,[h] Hongbing Shen, PhD,[c,f,g] Feng Chen, PhD,[a,c,g,*]
David C. Christiani, MD, MPH[b,hh]

[a]*Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, People's Republic of China*
[b]*Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, Massachusetts*
[c]*China International Cooperation Center (CICC) for Environment and Human Health, Nanjing Medical University, Nanjing, People's Republic of China*
[d]*State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing, People's Republic of China*
[e]*Department of Biostatistics, University of Michigan, Ann Arbor, Michigan*
[f]*Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing, People's Republic of China*
[g]*Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Cancer Center, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, People's Republic of China*
[h]*The Institute for Clinical and Translational Research, Department of Medicine, Baylor College of Medicine, Houston, Texas*
[i]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland*
[j]*Department of Medicine, British Columbia Cancer Agency, University of British Columbia, Vancouver, Canada*
[k]*Faculty of Medicine, University of Oviedo and CIBERESP, Oviedo, Spain*
[l]*Department of Epidemiology, University of Washington School of Public Health, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington*
[m]*Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Copenhagen, Denmark*
[n]*Section of Genetics, International Agency for Research on Cancer, World Health Organization, Lyon, France*
[o]*Department of Biosciences and Cancer Cluster Salzburg, University of Salzburg, Salzburg, Austria*
[p]*Department of Genetic Epidemiology, University Medical Center, Georg August University Göttingen, Göttingen, Germany*

*Corresponding author.

.Drs. R. Zhang, S. Shen, and Y. Wei contributed equally to this work.

Disclosure: The authors declare no conflict of interest.

Address for correspondence: Feng Chen, PhD, Department of Biostatistics, School of Public Health, Nanjing Medical University, 101 Longmian Avenue, Nanjing 211166, People's Republic of China. E-mail: fengchen@njmu.edu.cn or hbshen@njmu.edu.cn or dchris@hsph.harvard.edu.

ISSN: 1556-0864

https://doi.org/10.1016/j.jtho.2022.04.011

[q]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany
[r]Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Carmel, Haifa, Israel
[s]Markey Cancer Center, University of Kentucky, Lexington, Kentucky
[t]Department of Molecular and Clinical Cancer Medicine, Institute of Translational Medicine, University of Liverpool, Liverpool, United Kingdom
[u]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas
[v]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii
[w]Princess Margaret Cancer Centre, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[x]Department of Epidemiology, Department of Community and Family Medicine, Dartmouth Geisel School of Medicine, Hanover, New Hampshire
[y]Department for Health Evidence, Department of Urology, Radboud University Medical Center, Nijmegen, The Netherlands
[z]National Institute of Occupational Health, Oslo, Norway
[aa]Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden
[bb]Department of Radiation Sciences, Umeå University, Umeå, Sweden
[cc]Department of Oncology and Metabolism, The Medical School, University of Sheffield, Sheffield, United Kingdom
[dd]Department of Pharmaceutical Sciences, College of Pharmacy, Washington State University, Spokane, Washington
[ee]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida
[ff]Department of Thoracic Surgery and Division of Epidemiology, Vanderbilt University Medical Center, Nashville, Tennessee
[gg]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
[hh]Pulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts

Received 16 October 2021; revised 13 April 2022; accepted 20 April 2022
Available online - XXX

ABSTRACT

Introduction: Although genome-wide association studies have been conducted to investigate genetic variation of lung tumorigenesis, little is known about gene-gene (G × G) interactions that may influence the risk of non-small cell lung cancer (NSCLC).

Methods: 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 by a series of analyses. First, we used BiForce to evaluate and rank more than 58 billion G × G interactions from 340,958 single-nucleotide polymorphisms (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.

Results: 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., $\text{rs521828}_{C6orf10}$ and rs204999_{PRRT1}, $OR_{\text{interaction}} = 1.17$, $p = 6.57 \times 10^{-13}$; rs3135369_{BTNL2} and $\text{rs2858859}_{HLA\text{-}DQA1}$, $OR_{\text{interaction}} = 1.17$, $p = 2.43 \times 10^{-13}$; $\text{rs2858859}_{HLA\text{-}DQA1}$ and $\text{rs9275572}_{HLA\text{-}DQA2}$, $OR_{\text{interaction}} = 1.15$, $p = 2.84 \times 10^{-13}$; rs2853668_{TERT} and $\text{rs62329694}_{CLPTM1L}$, $OR_{\text{interaction}} = 0.73$, $p = 2.70 \times 10^{-13}$). 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 ($\text{rs521828}_{C6orf10}$ and rs204999_{PRRT1}, $OR_{\text{interaction}} = 1.13$, $p = 0.008$; rs3135369_{BTNL2} and $\text{rs2858859}_{HLA\text{-}DQA1}$, $OR_{\text{interaction}} = 1.11$, $p = 5.23 \times 10^{-4}$; rs3135369_{BTNL2} and $\text{rs9271300}_{HLA\text{-}DQA1}$, $OR_{\text{interaction}} = 0.89$, $p = 0.006$). The interaction-empowered polygenetic risk score that integrated classical polygenetic risk score and G × G information score was remarkable in lung cancer risk stratification.

Conclusions: Important 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

Introduction

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

The statistical interaction effect between two factors on NSCLC risk can be defined as the marginal effect of one factor on NSCLC risk varying across strata of another factor.[11] G × E interaction studies of NSCLC have provided additional genetic evidence of pathogenesis, including gene-smoking interactions,[9,12] gene-asbestos interactions,[13] and gene-occupation interactions.[14] G × G interactions or epistasis may also explain the missing heritability of NSCLC.[15,16] Nevertheless, owing to computationally intensive G × G interaction analyses on a genome-wide scale, only a small number of G × G interaction studies have been conducted for prostate cancer,[17–19] colorectal cancer,[20] breast carcinoma,[21] and nasopharyngeal cancer.[22] 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.[5,10] For population of European ancestry, 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 susceptibility.[23] Nevertheless, there is still a paucity of genome-wide G × G interaction studies among European-descent 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-descent 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),[24] 7253 participants from the Transdisciplinary Research in Cancer of the Lung (TRICL),[4] 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 using both classic polygenetic risk score (PRS) and the detected G × G interactions for screening high-risk subpopulations.

Materials and 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.[24] 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.[4] Tumors from patients were classified as adenocarcinomas, squamous carcinomas, large-cell carcinomas, mixed adenosquamous carcinomas, and other NSCLC histological types following either the International Classification of Diseases for Oncology (ICD-O) or WHO 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 data set.

UK Biobank. The UK Biobank is a large prospective study of individuals aged 40 to 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 through UK national registries and hospital records. Lung cancer cases were collected on the basis of the International Classification of Diseases, Tenth Revision, code of cancer diagnosis (filed ID: 40006, 41202) or self-reported lung cancer histological type (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.[27] 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 more than 0.2 or had low-quality DNA (call rate $< 95\%$), or sex inconsistency, or were non-European-ancestry. SNPs were removed if meeting any of the following criteria: (1) sex chromosome, (2) minor allele frequency less than 0.05, (3) call rate less than 95%, and (4) Hardy-Weinberg equilibrium (HWE) test p less than 1.00×10^{-7} in controls or p less than 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.[27]

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 version 2.3 using 1000 Genomes Project data (phase 3) as a reference panel[28] and then performed imputations using the Minimac (version 3) software. SNPs with an imputation quality score R^2 less than 0.4, minor allele frequency less than 0.01, or p less than 1×10^{-6} for the HWE test were excluded from the 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 (3017 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 (discovery and validation) study design to identify G × G interactions, while controlling the number of false positives (Fig. 1). Significant G × G signals identified in the discovery phase using ILCCO-OncoArray and TRICL were further confirmed in the validation phase using UK Biobank. In view of more than 58 billion possible G × G interactions considered in our study, we used *Screening before Testing* for dimensional reduction in the discovery phase.

Screening Step. BiForce is an entropy-based method, and it is implemented in a Java program that integrates bitwise computing with multithreaded parallelization and allows rapid full pairwise genome scans.[29] BiForce was applied to scan billions of G × G interactions exhaustively in ILCCO-OncoArray 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 $n\sum_{i,j,k}\widehat{\pi}_{ijk}\log(\widehat{\pi}_{ijk}/\widehat{p}_{ijk})$, where n is the sample size, $\widehat{\pi}_{ijk}$ is the observed frequency of subjects with SNP_1 coded i (0, 1, and 2), SNP_2 coded j (0, 1, and 2), and disease status coded k (0 and 1). In addition, $\widehat{p}_{ijk}$ was the expected frequency under null hypothesis. BiForce used Kirkwood superposition approximation (KSA) instead of likelihood estimation to calculate $\widehat{p}_{ijk}$. 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 regression models directly to exhaustively test all 58 billion G × G interactions. Instead, we used the top SNP pairs selected by BiForce,[29] with the default setting to filter noises ($P_{\mathrm{BiForce}} \leq 1.00 \times 10^{-6}$). The top pairs were retested through logistic regression model adjusted for covariates.

$$\begin{aligned}\mathrm{logit}(\pi) = {} & \beta_0 + \beta_1 \times SNP_1 + \beta_2 \times SNP_2 \\ & + \beta_3 \times SNP_1 \times SNP_2 + \sum \alpha_i \times Cov_i\end{aligned}$$

Where, β_1, β_2, and β_3 were the main effects of SNP_1 and SNP_2 and their interaction effect, respectively. Specifically, following the common practice as in the genomic studies of lung cancer risk,[24,25] we included age, sex, smoking status, and the top three principal components (derived from GWAS data for population structure) in the logistic regression model. In another sense, the interaction is the deviation between the joint effect of two SNPs and sum of their main effects, indicating a synergistic or antagonistic effect. Through a logistic regression model, we can have the estimate of interaction effect (β_3).

$$\begin{aligned}\beta_3 &= \ln\left(\frac{OR_{\text{joint effect}}}{OR_{\text{main effect, 1}} \times OR_{\text{main effect, 2}}}\right) \\ &= \beta_{\text{joint effect}} - (\beta_1 + \beta_2)\end{aligned}$$

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 $8.60 \times 10^{-13} = 0.05 \div C(340{,}958, 2)$, 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 less than or equal to 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

Owing to the population heterogeneity caused by different demographic and clinical characteristics (e.g.,

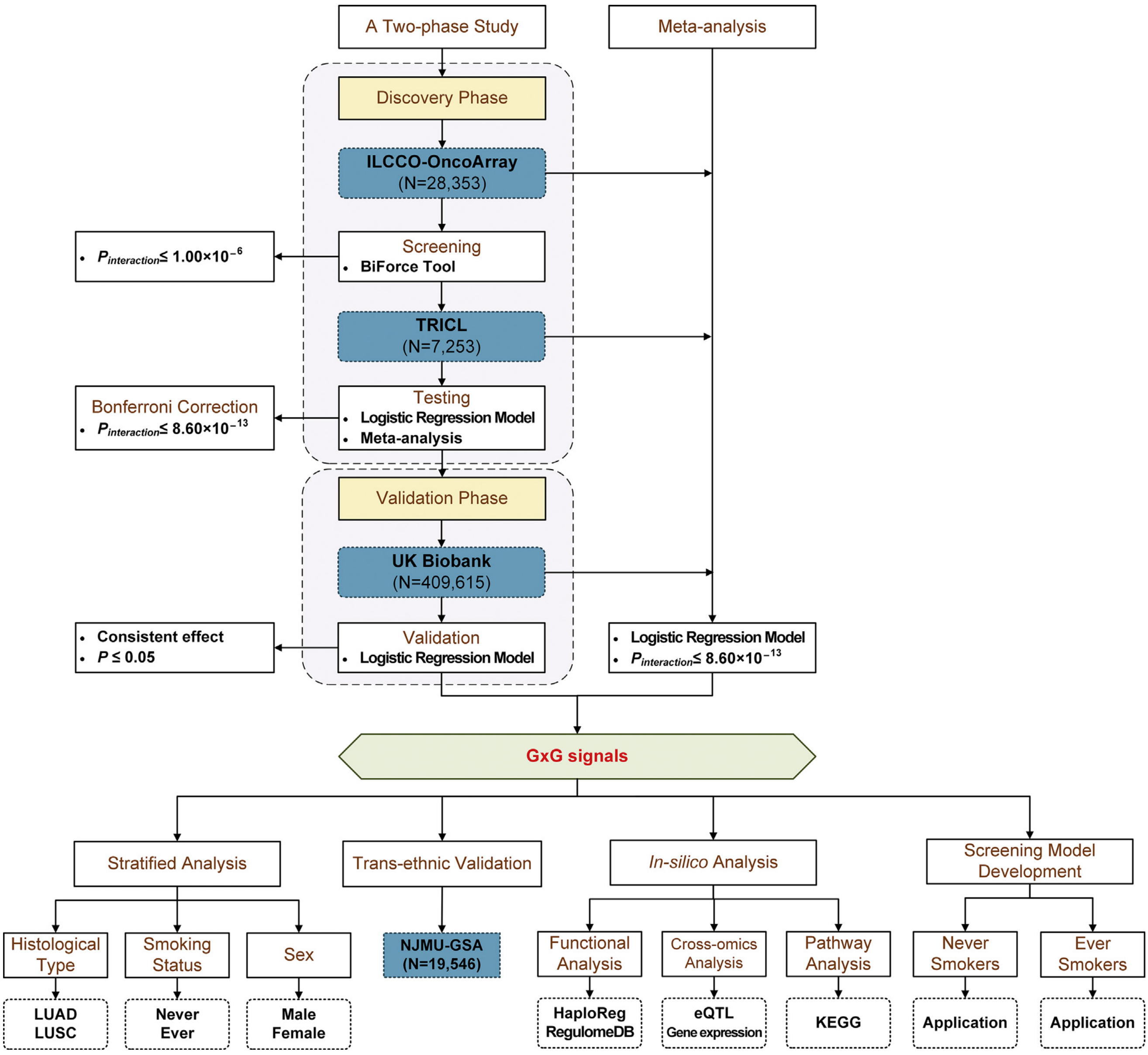

Figure 1. The workflow diagram of this study. We adopt a two-phase design in genome-wide G × G interaction study. In the discovery phase, a two-step strategy, *Screening before Testing*, was used for high-dimensionality reduction using European-ancestry participants from ILCCO-OncoArray and TRICL. In the validation phase, Bonferroni-corrected significant G × G interactions were further confirmed in the UK Biobank. Meanwhile, meta-analysis of ILCCO-OncoArray, TRICL, and UK Biobank and stratified analysis were 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. eQTL, expression quantitative trait loci; ILCCO-OncoArray, International Lung Cancer Consortium OncoArray project; KEGG, Kyoto Encyclopedia of Genes and Genomes; LUAD, lung adenocarcinoma; LUSC, lung squamous carcinoma; NJMU-GSA, Nanjing Medical University-Global Screening Array; TRICL, Transdisciplinary Research in Cancer of the Lung.

smoking versus nonsmoking), there might exist different types of lung cancer etiology.[24,30] In addition to detection of G × G signals for NSCLC, we aim to explore subpopulation-specific signals. Thus, to detect G × G interactions having weak-to-moderate effect sizes among these subpopulations with limited sample sizes, we meta-analyzed the ILCCO-OncoArray, TRICL, and UK Biobank cohorts by using fixed effect model in several NSCLC subgroups, including lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), male, female, never smoker, ever (current and former) 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 9298 controls).[26] QC procedures for genotypes were similar to those in OncoArray, except for the HWE test with p less than 1.00×10^{-5} in all participants. G × G interactions were analyzed through logistic regression models adjusted for the same covariates aforementioned. False discovery rate (FDR) correction using the Benjamini and Hochberg method[31] 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,[32] RegulomeDB,[33] and HaploReg version 4.1,[34] to predict potential functions of the identified SNPs. Expression quantitative trait loci (eQTL) was analyzed using the 578 lung tissues in the GTEx project.[35] To concordantly analyze expression profiles from tumor and healthy lung tissues, we collected data from Gene Expression Omnibus (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

PRS is constructed as the sum of the number of minor alleles one carries, weighted by effect coefficient as the per allele logarithmic odds ratio (OR).[36] In this study, we developed an enhanced lung cancer screening model incorporating demographic factors and interaction-empowered polygenetic risk score (iPRS) in never and ever smokers, respectively. iPRS was a linear combination of the following three components: (1) PRS constructed by 128 SNPs.[36] 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 GWASs. When correlation exists, variants representing independent loci with the strongest statistical significance were retained. (2) Score of significant G × G interactions identified in the two-phase study and meta-analysis of ILCCO-OncoArray, TRILC, and UK Biobank, meanwhile, reached nominal significance level ($p < 0.05$) in never or ever smoking subgroups by fixed effect meta-analysis. (3) Score of G × G interactions also selected among SNP pairs with $p_{\text{interaction}}$ less than or equal to 5×10^{-8} in the meta-analysis by group least absolute shrinkage and selection operator (groupLASSO) with the tuning parameter lambda determined by five-fold cross-validation,[37] and the coefficients were estimated by demographically adjusted logistic regression models. Score of G × G interaction in (2) and (3) was defined by the following scoring process, where β_{1i} and β_{2i} denoted the main effects of the two G × G SNPs and β_{3i} denoted their interaction effect:

$$Score_{G\times G} = \sum_{i=1}^{K}(\beta_{1i} \times SNP_{1i} + \beta_{2i} \times SNP_{2i} + \beta_{3i} \times SNP_{1i} \times SNP_{2i} big)$$

The iPRS, composed of classic PRS and G × G interaction score, was used to generate an enhanced lung cancer screening model together with age, sex, and pack-years of smoking.

First, with the ILCCO-OncoArray population, we categorized the continuous iPRS score to a 10-level categorical variable by its decile values and included the discretized iPRS in the demographically adjusted logistic regression model and computed the ORs and 95% confidence intervals (CIs) for each level with the lowest group set as the reference. We then repeated the same analysis for PRS scores and compared the performance of iPRS and PRS by using their ORs across 10 groups.

Validation of the iPRS Enhanced Lung Cancer Screening Model in UK Biobank Population

We adopted the same weights as used in ILCCO-OncoArray for SNPs to generate iPRS and PRS with the UK Biobank population and compared their performances similarly by using a Cox proportional hazards regression model:

$$h(t, G, C_i) = h_0(t)\exp\left(\beta_G \times G + \sum \beta_i \times C_i\right)$$

where G indicated the genetic risk score (iPRS or PRS) and C_i indicated the covariates, including age, sex, source of region, and smoking. For the UK Biobank survival analysis, time zero for each patient was defined to be the date of baseline attendance, and the follow-up time was defined from the date of baseline attendance to the date of diagnosis of an invasive primary lung cancer or censoring date (September 30, 2021), whichever occurred first.

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*.[38]

All statistical analyses were performed using R version 3.6.3 (The R Foundation for Statistical Computing, Vienna, Austria). p values were two-sided, and p less than 0.05 was considered statistically significant, unless otherwise specified.

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 (3288 cases and 3965 controls), and UK Biobank (3017 cases and 406,598 controls). In the discovery phase, 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, $OR_{\text{interaction}} = 1.20$, 95% CI: 1.14–1.26, $p = 6.10 \times 10^{-13}$; rs2853668, 4.8 kb 5′ of *TERT* at 5p15.33 and rs62329694, intronic of *CLPTM1L* at 5p15.33, $OR_{\text{interaction}} = 0.69$, 95% CI: 0.63–0.77, $p = 6.08 \times 10^{-13}$) reached the Bonferroni-corrected significance level ($p < 8.60 \times 10^{-13}$) using subjects from ILCCO-OncoArray and TRICL. In the validation phase, we confirmed the significance of these two G × G signals by using independent participants from UK Biobank (rs521828 and rs204999: $OR_{\text{interaction}} = 1.09$, 95% CI: 1.00–1.18, $p = 0.044$; rs2853668 and rs62329694: $OR_{\text{interaction}} = 0.83$, 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 odds among subjects carrying the wild genotype (AA) of rs204999 ($OR = 0.86$, 95% CI: 0.80–0.92, $p = 1.64 \times 10^{-5}$); the effect was reversed among those carrying the heterozygous AG genotype of rs204999 ($OR = 1.09$, 95% CI: 1.01–1.17, $p = 2.09 \times 10^{-2}$), and the effect became more detrimental among those with the homozygous GG genotype of rs204999 ($OR = 1.23$, 95% CI: 1.06–1.43, $p = 5.24 \times 10^{-3}$). 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 (Fig. 2*A*). Similar patterns were observed between rs2853668 and rs62329694. The G allele of rs2853668 was associated with a higher NSCLC odds ($OR = 1.30$, 95% CI: 1.14–1.49, $p = 1.10 \times 10^{-4}$) 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, $p = 2.74 \times 10^{-2}$) and AA ($OR = 0.75$, 95% CI: 0.59–0.99, $p = 2.64 \times 10^{-2}$) genotypes of rs2853668, respectively. The pattern was confirmed by sensitivity analyses (Fig. 2*B*).

In addition, we evaluated the interaction pattern (synergetic versus antagonistic) for these two pairs of SNPs coded in the genetic dominant model (Supplementary Tables 1 and 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 or AA genotype of rs521828, and AG or GG genotype of rs204999 were protective, with an OR of 0.90 and 0.89, respectively. Nevertheless, their

Table 1. Demographic and Clinical Descriptions of NSCLC Cases and Controls in ILCCO-OncoArray, TRICL, and UK Biobank

	ILCCO-OncoArray		TRICL		UK Biobank	
Characteristics	Case (n = 15,157)	Control (n = 13,196)	Case (n = 3288)	Control (n = 3965)	Case (n = 3017)	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
Sex (%)						
Male	9778 (64.5)	7967 (60.4)	1643 (50.0)	2028 (51.1)	1398 (46.3)	219,979 (54.1)
Female	5376 (35.5)	5228 (39.6)	1641 (50.0)	1937 (48.9)	1619 (53.7)	186,619 (45.9)
Smoking status (%)						
Never	1403 (9.4)	3981 (30.9)	264 (8.0)	1023 (25.8)	357 (11.9)	161,959 (40.0)
Ever	13,461 (90.6)	8908 (69.1)	3024 (92.0)	2942 (74.2)	2642 (88.1)	243,356 (60.0)
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
Histological type (%)						
NSCLC	10,997 (87.5)	-	1952 (86.3)	-	1731 (87.3)	-
LUAD	6158 (49.0)	-	1296 (57.3)	-	944 (47.6)	-
LUSC	3886 (30.9)	-	513 (22.7)	-	569 (28.7)	-
Others	953 (7.6)	-	143 (6.3)	-	218 (11.0)	
LSCC	1564 (12.5)	-	310 (13.7)	-	252 (12.7)	-

Ever smoker was composed of former and current smokers.

ILCCO-OncoArray, International Lung Cancer Consortium OncoArray project; LSCC, lung small cell carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous carcinoma; TRICL, Transdisciplinary Research in Cancer of the Lung.

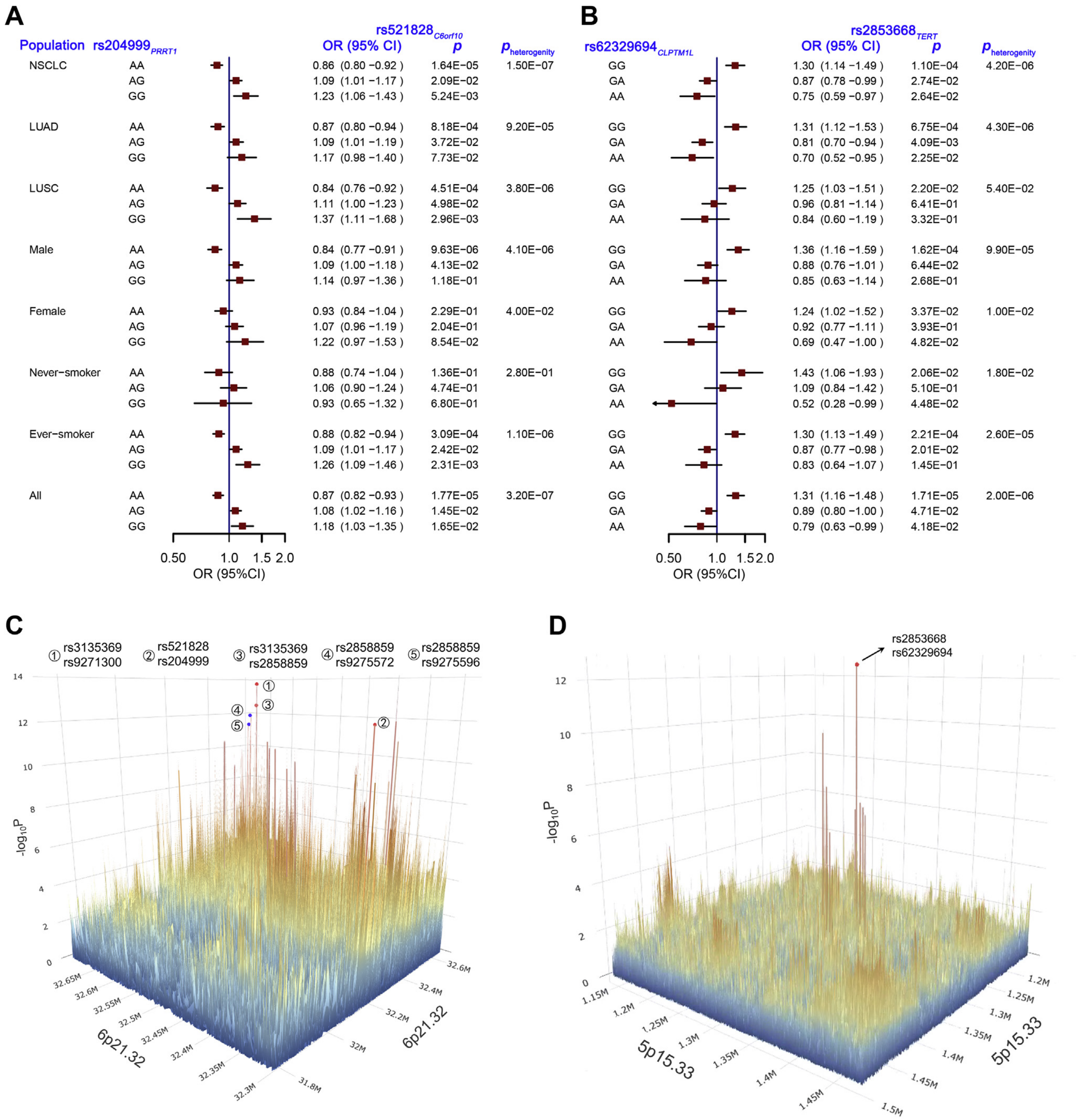

Figure 2. Forest plot of G × G interactions for (*A*) rs204999 × rs521828 and (*B*) rs2853668 × rs62329694 using the European-ancestry participants from ILCCO-OncoArray, TRICL, and UK Biobank. The three-dimensional G × G interaction signal map for association results of all epistatic pairs upstream and 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 logistic regresssion model adjusted for covariates and pooled by meta-analysis of ILCCO-OncoArray, TRICL, and UK Biobank. *p* values were plotted on a negative $\log_{10}$-transformed scale. CI, confidence interval; ILCCO-OncoArray, International Lung Cancer Consortium OncoArray project; TRICL, Transdisciplinary Research in Cancer of the Lung.

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 ($OR_{\text{interaction}} = 1.28$, 95% CI: 1.18–1.39, $p_{\text{interaction}} = 1.23 \times 10^{-9}$). 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 ($OR_{\text{interaction}} = 0.68$, 95% CI: 0.61–0.76, $p_{\text{interaction}} = 1.18 \times 10^{-10}$).

All SNPs within the approximately 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 (Fig. 2*C* and *D*).

Six More Significant G × G Interactions Identified by Meta-Analysis Among Europeans

G × G signals with $p_{\text{interaction}}$ less than 5×10^{-8} derived from meta-analysis from different subpopulations were summarized in Supplementary Tables 3 to 10. A total of eight pairs of SNPs reached the Bonferroni-corrected threshold ($p < 8.60 \times 10^{-13}$) in various subpopulations (Table 2). Among them, two pairs of SNPs were the same as those identified by the two-phase study. Furthermore, among the six newly detected G × G interactions, four pairs appeared in the 6p21.32 region, including rs3135369 and rs9271300, rs3135369 and rs2858859, rs2858859 and rs9275572, rs2858859 and rs9275596 (Supplementary Fig. 1). With a moderate level of linkage disequilibrium (LD) between rs9271300 and rs2858859 ($r^2 = 0.66$, $D' = 0.996$) and rs9275596 and rs9275572 ($r^2 = 0.72$, $D' = 0.998$), these four G × G signals were likely to be the result of the following 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 two pairs of SNPs resided in different regions, including rs28591443 in 8p23.3 and rs9265981 in 6p25.2, rs589027 in 1q32.2, and rs713395 in 2p24.2.

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

Sensitivity Analyses

We further performed sensitivity analyses to evaluate these eight G × G interactions. (1) We evaluated the unadjusted effects of the eight G × G signals by not including any other covariates in the logistic regression model and found that all G × G interactions still reached a significance level with p less than 5×10^{-7} in various subpopulations (Supplementary Table 12). (2) To account for type I error inflation caused by imbalance of cases and controls in the UK Biobank population, we applied SAIGE (version 0.44.6.5) in the validation phase to reconfirm these eight signals. SAIGE uses saddlepoint approximation to account for case-control imbalance, which can efficiently analyze large sample data, controlling for case-control imbalance and sample relatedness.[39,40] All G × G interactions remained nominally significant (Supplementary Table 13), except one pair (rs521828 and rs204999) that was marginally significant ($p = 0.056$). These results by sensitivity analyses indicated a satisfactory robustness of the eight G × G interactions.

Successful Trans-Ethnic Validation of Significant G × G Interactions in Asians and Europeans

First, we evaluated the eight G × G interactions identified from the European-ancestry population by using an

Table 2. The Eight 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							
Region	SNP	Nearest Gene	EAF	Region	SNP	Nearest Gene	EAF	Population	OR	95% CI	*p*
5p15.33	rs2853668	*TERT*	0.276	5p15.33	rs62329694	*CLPTM1L*	0.051	NSCLC[a]	0.73	0.67-0.79	2.70×10^{-13}
								All[b]	0.74	0.68-0.80	5.39×10^{-13}
6p21.32	rs521828	*C6orf10*	0.290	6p21.32	rs204999	*PRRT1*	0.257	NSCLC[a]	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[b]	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[b]	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}

[a] These G × G signals reached the Bonferroni-corrected significance threshold ($p < 8.60 \times 10^{-13}$) 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.

[b] All includes lung cancer cases with all histological types.

CI, confidence interval; EAF, effect allele frequency; ILCCO-OncoArray, International Lung Cancer Consortium OncoArray project; LUAD, lung adenocarcinoma; SNP, single-nucleotide polymorphism; TRICL, Transdisciplinary Research in Cancer of the Lung.

external Asian population from NJMU-GSA (Supplementary Table 14). We were able to validate three pairs of SNPs in the 6p21.32 region in several of its subpopulations. They included rs521828 and rs204999 among NSCLC ($OR_{\text{interaction}} = 1.13$, 95% CI: 1.03–1.24, $p = 0.008$, $q\text{-}FDR = 0.022$), rs3135369 and rs9271300 among NSCLC ($OR_{\text{interaction}} = 0.89$, 95% CI: 0.83–0.96, $p = 0.006$, $q\text{-}FDR = 0.022$) and smoker ($OR_{\text{interaction}} = 0.82$, 95% CI: 0.72–0.92, $p = 0.001$, $q\text{-}FDR = 0.005$), and rs3135369 and rs2858859 in NSCLC ($OR_{\text{interaction}} = 1.11$, 95% CI: 1.04–1.17, $p = 5.23 \times 10^{-4}$, $q\text{-}FDR = 0.005$) (Supplementary Table 15). No significant results were available for the other pairs, possibly owing to differences in effect allele frequency for SNPs between these two ethnic populations (Fig. 3).

Conversely, we validated the only pair of SNPs at 2p32.2 (rs16832404 and rs2562796) that reached genome-wide significance among the Asian population,[20] using the European-ancestry population. This pair indeed exhibited a significant G × G interaction effect on NSCLC odds among the European-ancestry population ($OR_{\text{interaction}} = 1.11$, 95% CI: 1.01–1.22, $p = 0.028$) (Supplementary Table 16).

SNPs With G × G Interactions Potentially Involved in Biological Regulatory Function

In the RegulomeDB database, the abundant biological regulatory function was observed for 10 of 13 SNPs, including eQTL, transcription factor binding site, or DNase peak. Numerous enhancer histone marks and motifs changed were observed for the SNPs (Supplementary Table 17). 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 eight SNPs in 6p21.32 and 6p25.2 (Supplementary Table 18), whereas no significant eQTLs were found for the others. Furthermore, we performed differential expression analysis with the GEO repository. For the three genes (*C6orf10*, *CLPTM1L*, and *TERT*) identified in the two-phase study, their expression levels were significantly up-regulated in the tumor tissues (Supplementary Fig. 5). In addition, *BTNL2*, which was also identified by the meta-analysis, was significantly differentially expressed between lung cancer tumor and normal tissues (Supplementary Fig. 6). By tumor mutational burden analysis of these 10 genes in tumor tissues with the LUAD- and LUSC-TCGA databases, on the basis of somatic mutations by next-generation sequencing, we have found that three genes, specifically, *TERT*, *CLPTM1L*, and *CSMD1*, presented high proportions of somatic mutations in the tumor cells (Supplementary Fig. 7). The findings may inspire novel targeted therapies of lung cancer.

G × G Interaction Score Effectively Distinguishes Population at High Risk in UK Biobank

We developed lung cancer screening models among never smokers and ever smokers because of their substantially different genetic backgrounds. For each subgroup, considered for risk screening were the following: (1) the PRS comprising 128 SNPs with significant marginal effects identified by GWAS in Europeans so far (Supplementary Table 19); (2) score of G × G interactions (Table 2), of which p is less than 0.05 in never or ever smoking subgroups by meta-analysis of three data sets; and (3) score of G × G interactions selected by groupLASSO using ILCCO-OncoArray (training set) with the largest sample size of cases in our study (Supplementary Tables 20 and 21). The iPRS (Supplementary Table 22) has remarkable stratification performance while we categorized subjects into 10 groups by the deciles of the score in ILCCO-OncoArray (Supplementary Fig. 8).

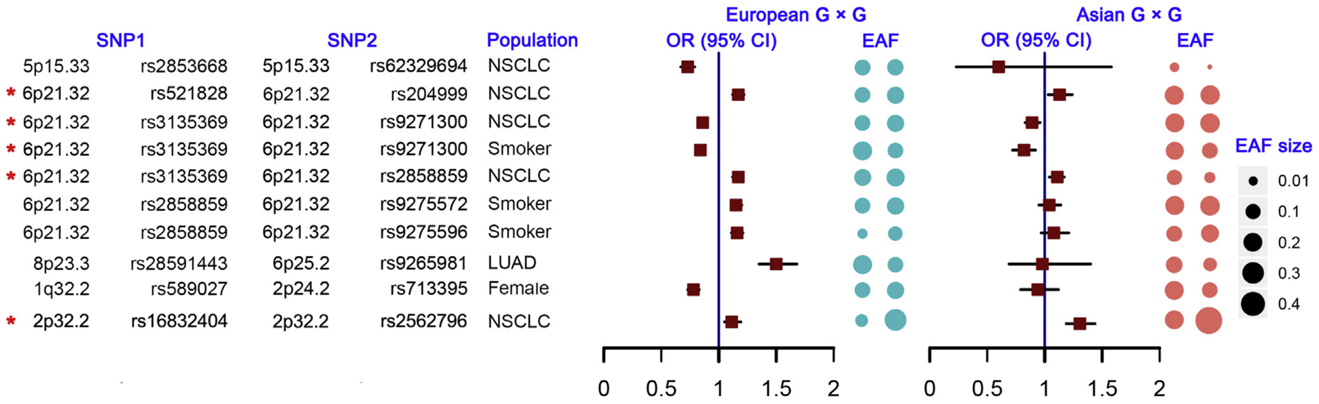

Figure 3. The comparison of G × G interaction association results and effect of allele frequency between Europeans and Asians. Star symbol (*) indicates that G × G interaction is significant in both Europeans and Asians. CI, confidence interval; EAF, effect allele frequency; LUAD, lung adenocarcinoma.

iPRS was externally validated among 162,316 never smokers and 245,998 ever smokers with available follow-up time of lung cancer from UK Biobank. The median of follow-up time was 9.45 years, and its interquantile range was from 8.48 to 10.52 years. Each subject was assigned an iPRS score, and all subjects

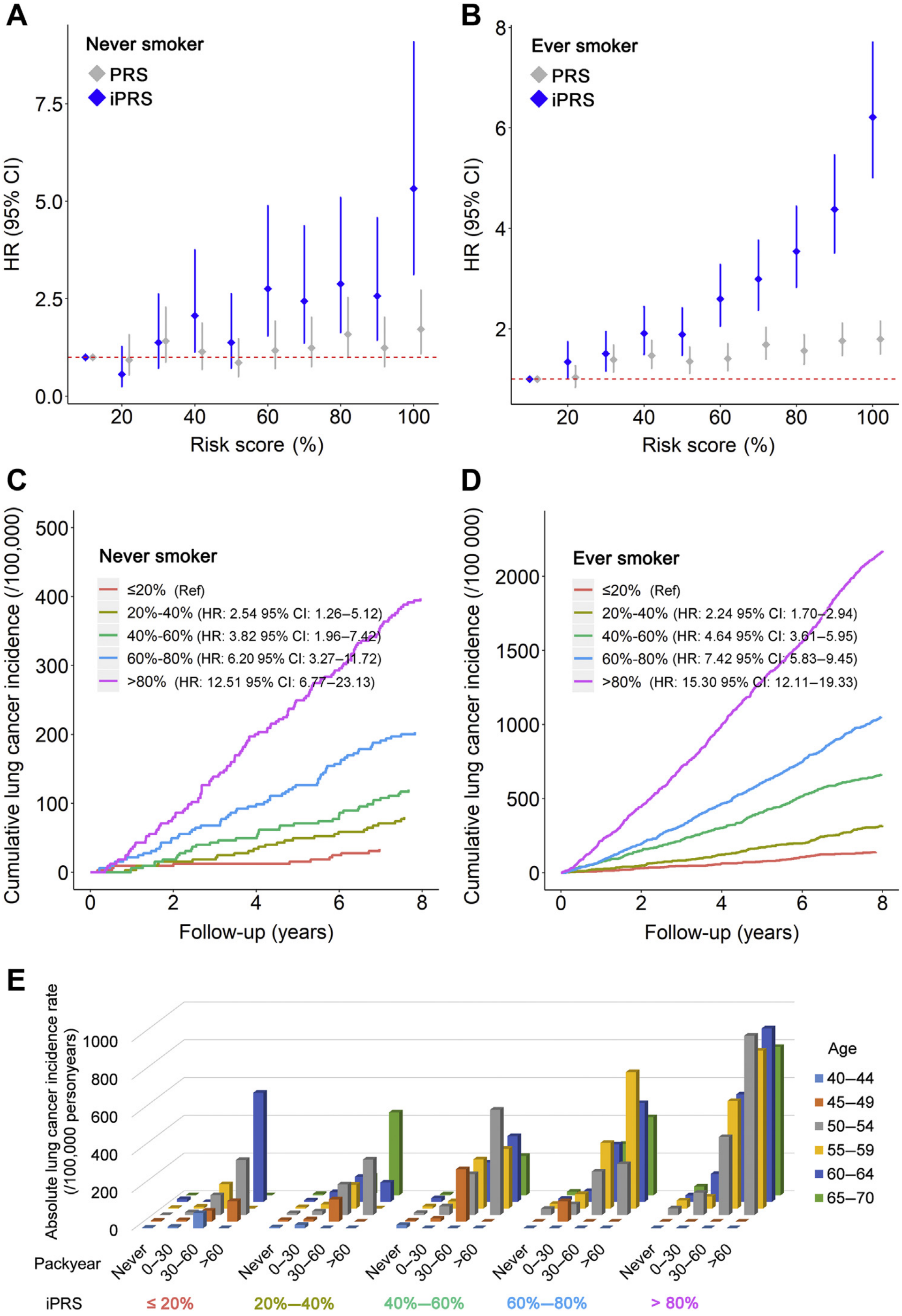

Figure 4. Participants in the UK Biobank were divided into 10 equal groups according to the PRS and iPRS, respectively. HR and 95% CI of each group were derived from Cox proportional hazards model adjusted for covariates by setting the lowest group as reference for never (*A*) and ever smokers (*B*). Cumulative lung cancer incidence curves were illustrated for subjects at different overall risk score groups calculated from demographic variables (age, sex, and pack-years) 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*). CI, confidence interval; HR, hazard ratio; iPRS, interaction-empowered polygenetic risk score; PRS, polygenetic risk score; Ref, reference group.

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) equals to 5.31 (95% CI: 3.11–9.07, $p = 8.60 \times 10^{-10}$, Fig. 4*A*) for never smokers and *HR* equals to 6.21 (95% CI: 5.01–7.70, $p < 2.2 \times 10^{-16}$, Fig. 4*B*) for ever smokers. Compared with PRS [top 10% versus bottom 10%: $HR = 1.72$ (95% CI: 1.09–2.72) for never smokers; $HR = 1.80$ (95% CI: 1.49–2.15) for ever smokers],[26] iPRS was found to have a better discrimination power. Meanwhile, we validated the lung cancer screening model composed of demographic variables (age, sex, and pack-years of smoking) 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 five groups categorized by the quintiles of the overall risk scores of ($p < 0.001$), indicating the iPRS enhanced model served as a good risk classifier (Fig. 4*C* and *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.[41] Therefore, we illustrated the absolute incidence of lung cancer in various subpopulations classified by iPRS, age, and pack-years of smoking (Fig. 4*E*). Clinically, the iPRS enhanced model may change the practice of lung cancer screening. For example, subjects aged less than 55 years or smoked less than 30 pack-years (including never smokers), but with a high iPRS, may be suggested as the high-risk population for lung cancer screening; for those with a high iPRS (top 20%) and smoked more than 60 pack-years, lung cancer screening may be better to start as early as 50 years old; and for those with a low iPRS, screening can be postponed (Fig. 4*E*).

The Liverpool Lung Project lung cancer risk model version 3 (LLPv3), using several demographic and clinical factors (age, sex, smoking duration, history of respiratory disease, previous malignancy, family history of lung cancer, and exposure to asbestos), is a well-known and validated model for lung cancer risk stratification.[42] In addition, the iPRS could further stratify subjects into different risk groups even at the same subgroup stratified by LLPv3 (Supplementary Fig. 9), indicating iPRS could enhance the screening ability of classical lung cancer risk model.

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 Fig. 10). For ever smokers, 22 pathways were identified, including the well-known pathways such as TH1 and TH2 cell differentiation, Notch signaling pathway, and leishmaniasis, indicating more biological pathways were involved in smoking behaviors leading to tumorigenesis (Supplementary Fig. 11).

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 eight 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 three of five 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.

Strategy of Data Analysis and Controlling of False Positives

Owing to computational constraints, very few genome-wide G × G interaction studies are available for a limited number of diseases (Supplementary Table 23), including prostate cancer,[17–19] colorectal cancer,[20] nasopharyngeal cancer,[22] breast cancer,[21] and our previous Asian lung cancer study.[10] To address the computing challenge, we adopted a *Screening before Testing* strategy[17,19,21] to efficiently extract G × G signals from more than 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 multiphase study design is another most often used strategy to increase the reproducibility of association results. Thus, we used a two-phase study, which involves discovery and validation phases, to identify two pairs of SNPs associated with NSCLC risk. To detect subpopulation-specific 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.[24,43] As a result, we identified six more pairs of SNPs, of which five pairs exhibited acceptable significance across all the subpopulations considered, except for one female-specific epistasis, the mechanism of which warrants further research.

To address the multiple comparison issue, first, we detected that the number of “independent” (LD-$r^2 < 0.1$) SNPs is 108,951 by using the PLINK prune function.[44,45] Then, considering eight possible subsets (e.g., LUAD, LUSC) in our G × G interaction study, the theoretical genome-wide significance level should be $1.05 \times 10^{-12} = 0.05 \div C(108{,}951, 2) \div 8$, which was greater than 8.60×10^{-13}. Thus, the overall false positive rate should be well controlled in the entire study if we stick to the level of 8.60×10^{-13}.

Independent Contribution by G × G Interactions to Lung Cancer Screening

Because most patients in the study were smokers, either former or current, that we did not observe significant G × G interactions in never smokers might be due to a small sample size of lifelong nonsmokers, which did not equip the subgroup analysis with enough power. Nevertheless, when fitting the regression models to assess the significance of the G × G interactions on the basis of the entire population, we did adjust for the smoking status. We further performed the chi-square test to assess the associations between the SNPs involved in these G × G interactions and the smoking status in the ILCCO-OncoArray, TRICL, and UK Biobank populations, respectively. For these 13 SNPs reported in Table 2, 11 were independent with the smoking status (Supplementary Table 24).

Because smoking is a well-established risk factor for lung cancer risks and mortality, smoking cessation is never too late for smokers and the sooner the better; our results did not contradict this. Nevertheless, because lung cancer is such a complex disease, it is driven by multiple environmental, clinical, and genetic factors.[46] Hence, smoking cessation alone may not be sufficient for lung cancer prevention. For example, previous respiratory diseases can also increase lung cancer risks, including emphysema and chronic obstructive pulmonary disease.[47,48] Our study identified 212 SNPs that were associated with respiratory diseases (e.g., chronic obstructive pulmonary disease and asthma) using the GWAS Catalog database (Supplementary Table 25). These identified biomarkers can share the same role in the common pathogenetic pathways both for respiratory diseases and lung cancer,[49] contributing to lung cancer risks in the presence or the absence of smoking.

PRS is a popular approach for identifying individual-level genetic risks of lung cancer.[26,50] Nevertheless, with weak marginal effects of individual SNPs, the stratification performance of PRS based models is generally unsatisfactory,[51] resulting in a severe missing heritability issue. By incorporating two-way G × G 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,[52] complex association patterns (e.g., G × G interactions) among multiple factors should be factored in for studies of complex diseases (e.g., lung cancer).

Benefit and Prospect of Applying iPRS in Lung Cancer Screening

Confirmed by a recent randomized trial,[53] LDCT screening of asymptomatic subjects with high lung cancer risk is a well-recognized way to reduce cancer morbidity and mortality by detecting very early stage cases or those predisposed to lung cancer and then leading to early treatment and intervention strategies. Nevertheless, identifying suitable subpopulations for LDCT screening is quite essential to maximize the cost-effectiveness of the screening project.[54,55] The Centers for Medicare and Medicaid Services and the U.S. Preventive Services Task Force guideline recommends to target subjects merely on the basis of their age and smoking history,[56] which is convenient and effective in real-word practice but still lacks precision. Therefore, more than 20 lung cancer risk prediction models have been created in the last couple of decades,[57] by incorporating more clinical factors to distinguish high-risk populations (e.g., LLPv3[42]). In recent years, emerging evidence has revealed that PRS unitizing genetic factors improved the ability of targeting subjects at high risk of lung cancer.[26] Because the proposed iPRS outperformed PRS (Supplementary Fig. 8 and Fig. 4*A* and *B*), iPRS possessed the additional capability to substantially enhance the guideline- and model-based lung cancer screening strategies (Fig. 4*E* and Supplementary Fig. 9). Though genome-wide SNP genotyping is not widely applied in real-word clinical practice currently, we envision that simple and fast biotechnology for the detection of target genes with SNPs is opening up genetic research and diagnostics beyond laboratory settings.[58] Therefore, as time and cost of SNP genotyping dramatically reduced in the future, we suspect that a custom-designed chip may make iPRS readily accessible and eventually maximize cost-effectiveness of lung cancer screening.

Potential Biological Functions of Genes to Which These Identified SNPs Were Mapped

Of eight significant pairs of SNPs, six were found to be located in the 6p21.32 and 5p15.33 regions. For example, one pair mapped to *TERT* and *CLPTM1L* is in 5p15.33, a well-known region reported by GWAS of lung cancer risk in Asians, African Americans,[59] European,[60] and for lung cancer prognosis.[61] A GWAS by McKay et al.[62] suggested two genes, *TERT* and *CLPTM1L*, that play a role in the development of lung cancer. This current study reported their interaction effect for the first time. Interestingly, the two genes are all involved in tumor antiapoptosis.[63,64] *TERT* plays a role in cellular senescence because it is normally repressed in postnatal somatic cells, resulting in shortening of telomeres, and, therefore, aging and antiapoptosis. Deregulation of telomerase expression in somatic cells may be involved in oncogenesis.[65] *CLPTM1L* is a most often overexpressed antiapoptotic factor in lung tumors and is associated with DNA damage measured by bulky aromatic and hydrophobic DNA adducts.[66] 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.[67]

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,[68,69] and we now report the G × G signals in this region for Europeans. *PRRT1* and *C6orf10* (also known as *TSBP1*) are located on the major histocompatibility complex (MHC) region, widely recognized as an important regulatory region for multiple diseases.[70] The two SNPs (rs3135369 and rs2858859) also have 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 (TCRs) and activate a specific T-cell response.[71] 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.[72]

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[73]; *HHAT* regulates the proliferation of estrogen receptor cells in breast cancer and the *HHAT* inhibitor plays a critical role for therapeutic benefits.[74] Nevertheless, biological functions of lncRNA *AC008069.1* remain unknown.

Strengths and Limitations

Our study has several strengths. First, this is perhaps the largest G × G interaction study of lung cancer risk by using consortium resources and the first G × G interaction study among the European-ancestry population, providing evidence beyond an Asian population.[10] 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. Although 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 the UK Biobank among never and ever smokers. 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 6606 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, because this study was primarily designed for a European-ancestry population with most participants being Europeans in ILCCO-OncoArray, TRICL, and UK Biobank, future G × G interaction studies on subjects with African American ancestry are needed. Fourth, some of these G × G interactions included in the screening models were selected by groupLASSO in ILCCO-OncoArray, without being further validated in the UK Biobank.

In summary, we have identified several novel G × G interactions, which were internally and externally validated by multiethnic populations. The developed iPRS may enrich the screening tool box for physicians.

CRediT Authorship Contribution Statement

Ruyang Zhang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Sipeng Shen, Yongyue Wei: Methodology, Software, Validation, Formal analysis, Data Curation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition.

Ying Zhu: Software, Formal analysis, Writing - original draft.

Yi Li: Methodology, Writing - original draft, Writing - review & editing, Funding acquisition.

Jiajin Chen, Jinxing Guan, Zoucheng Pan, Yuzhuo Wang, Meng Zhu, Junxing Xie, Xiangjun Xiao, Dakai Zhu, Yafang Li, Juncheng Dai, Hongxia Ma, Yang Zhao: Software, Validation, Data Curation, Visualization.

Demetrios Albanes, Maria Teresa Landi, Neil E. Caporaso, Stephen Lam, Adonina Tardon, Chu Chen, Stig E. Bojesen, Mattias Johansson, Angela Risch, Heike Bickeböller, H-Erich Wichmann, Gadi Rennert, Susanne Arnold, Paul Brennan, James D. McKay, John K. Field, Sanjay S. Shete, Loic Le Marchand, Geoffrey Liu, Angeline S. Andrew, Lambertus A. Kiemeney, Shan Zienolddiny-Narui, Annelie Behndig, Mikael Johansson, Angela Cox, Philip Lazarus, Matthew B. Schabath, Melinda C. Aldrich: Validation, Resources, Data curation, Writing - review & editing, Supervision, Project administration.

Feng Chen, Hongbing Shen, David C. Christiani, Christopher I. Amos, Rayjean J. Hung, Zhibin Hu: Conceptualization, Methodology, Validation, Resources, Data curation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Data Availability

ILCCO-Oncoarray data are available from: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001273.v3.p2. TRICL data are available from: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001681.v1.p1. UK Biobank data are available from https://www.ukbiobank.ac.uk/. KEGG database are 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 on reasonable request.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (81820108028 to Dr. H. Shen, 81973142 to Dr. Y. Wei, 82103946 to Dr. S. Shen, 81922061 and 81973123 to Dr. H. Ma), National Key Research and Development Program of China (2016YFE0204900 to Dr. F. Chen), Natural Science Foundation of Jiangsu Province (BK20191354 to Dr. R. Zhang), China Postdoctoral Science Foundation (2020M681671 to Dr. S. Shen), Jiangsu Planned Projects for Postdoctoral Research Funds (2020Z019 to Dr. S. Shen), the U.S. National Institutes of Health (CA209414, CA092824, and ES000002 to Dr. D.C. Christiani, AG056764 to Dr. Y. Li), and Priority Academic Program Development of Jiangsu Higher Education Institutions. Dr. R. Zhang was partially supported by the Qing Lan Project of the Higher Education Institutions of Jiangsu Province and the Outstanding Young Level Academic Leadership Training Program of Nanjing Medical University. Where authors are identified as personnel of the International Agency for Research on Cancer, WHO, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer, WHO.

Supplementary Data

Note: To access the supplementary material accompanying this article, visit the online version of the *Journal of Thoracic Oncology* at www.jto.org and at https://doi.org/10.1016/j.jtho.2022.04.011.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin*. 2021;71:7-33.
2. Mucci LA, Hjelmborg JB, Harris JR, et al. Familial risk and heritability of cancer among twins in Nordic countries. *JAMA*. 2016;315:68-76.
3. Timofeeva MN, Hung RJ, Rafnar T, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet*. 2012;21:4980-4995.
4. Wang Y, McKay JD, Rafnar T, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet*. 2014;46:736-741.
5. Bossé Y, Amos CI. A decade of GWAS results in lung cancer. *Cancer Epidemiol Biomarkers Prev*. 2018;27:363-379.
6. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747-753.
7. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*. 2008;40:695-701.
8. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*. 2008;40:1253-1260.
9. Zhang R, Chu M, Zhao Y, et al. A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. *Carcinogenesis*. 2014;35:1528-1535.
10. Chu M, Zhang R, Zhao Y, et al. A genome-wide gene-gene interaction analysis identifies an epistatic gene pair for

lung cancer susceptibility in Han Chinese. *Carcinogenesis*. 2014;35:572-577.

11. VanderWeele TJ, Knol MJ. A tutorial on interaction. *Epidemiol Methods*. 2014;3:33-72.
12. Shen S, Wei Y, Li Y, et al. A multiomics study links TNS3 and SEPT7 to long-term former smoking NSCLC survival. *NPJ Precis Oncol*. 2021;5:39.
13. Liu CY, Stücker I, Chen C, et al. Genome-wide gene-asbestos exposure interaction association study identifies a common susceptibility variant on 22q13.31 associated with lung cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2015;24:1564-1573.
14. Malhotra J, Sartori S, Brennan P, et al. Effect of occupational exposures on lung cancer susceptibility: a study of gene-environment interaction analysis. *Cancer Epidemiol Biomarkers Prev*. 2015;24:570-579.
15. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*. 2012;109:1193-1198.
16. Ashworth A, Lord CJ, Reis-Filho JS. Genetic interactions in cancer progression and treatment. *Cell*. 2011;145:30-38.
17. Tao S, Feng J, Webster T, et al. Genome-wide two-locus epistasis scans in prostate cancer using two European populations. *Hum Genet*. 2012;131:1225-1234.
18. Ciampa J, Yeager M, Amundadottir L, et al. Large-scale exploration of gene-gene interactions in prostate cancer using a multistage genome-wide association study. *Cancer Res*. 2011;71:3287-3295.
19. 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*. 2017;140:2075-2084.
20. Jiao S, Hsu L, Berndt S, et al. Genome-wide search for gene-gene interactions in colorectal cancer. *PLoS One*. 2012;7:e52535.
21. Milne RL, Herranz J, Michailidou K, 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*. 2014;23:1934-1946.
22. Su WH, Yao Shugart Y, Chang KP, Tsang NM, Tse KP, Chang YS. How genome-wide SNP-SNP interactions relate to nasopharyngeal carcinoma susceptibility. *PLoS One*. 2013;8:e83034.
23. Ji X, Bossé Y, Landi MT, et al. Identification of susceptibility pathways for the role of chromosome 15q25.1 in modifying lung cancer risk. *Nat Commun*. 2018;9:3221.
24. McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet*. 2017;49:1126-1132.
25. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779.
26. Dai J, Lv J, Zhu M, 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;7:881-891.
27. Amos CI, Dennis J, Wang Z, et al. The OncoArray consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol Biomarkers Prev*. 2017;26:126-135.
28. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48:1443-1448.
29. Gyenesei A, Moody J, Laiho A, Semple CA, Haley CS, Wei WH. BiForce Toolbox: powerful high-throughput computational analysis of gene-gene interactions in genome-wide association studies. *Nucleic Acids Res*. 2012;40:W628-W632.
30. Hung RJ, Spitz MR, Houlston RS, et al. Lung cancer risk in never-smokers of European descent is associated with genetic variation in the 5(p)15.33 tert-CLPTM1Ll region. *J Thorac Oncol*. 2019;14:1360-1369.
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57:289-300.
32. Xu Z, Taylor JA. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res*. 2009;37:W600-W605.
33. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22:1790-1797.
34. Ward LD, Kellis M, HaploReg Kellis M. a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40:D930-D934.
35. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348:648-660.
36. Hung RJ, Warkentin MT, Brhane Y, et al. Assessing lung cancer absolute risk trajectory based on a polygenic risk model. *Cancer Res*. 2021;81:1607-1615.
37. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat Comput*. 2015;25:173-187.
38. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics*. 2012;16:284-287.
39. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*. 2018;50:1335-1341.
40. Dey R, Schmidt EM, Abecasis GR, Lee S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am J Hum Genet*. 2017;101:37-49.
41. Mazzone PJ, Silvestri GA, Souter LH, et al. Screening for lung cancer: CHEST guideline and expert panel report. *Chest*. 2021;160:e427-e494.
42. Field JK, Vulkan D, Davies MPA, Duffy SW, Gabe R. Liverpool Lung Project Lung cancer risk stratification model: calibration and prospective validation. *Thorax*. 2021;76:161-168.
43. Schwartzentruber J, Cooper S, Liu JZ, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat Genet*. 2021;53:392-402.

44. Bansal V, Mitjans M, Burik CAP, et al. Genome-wide association study results for educational attainment aid in identifying genetic heterogeneity of schizophrenia. *Nat Commun*. 2018;9:3078.
45. Barban N, Jansen R, de Vlaming R, et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat Genet*. 2016;48:1462-1472.
46. Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. *Eur Respir J*. 2016;48:889-902.
47. Denholm R, Schüz J, Straif K, et al. Is previous respiratory disease a risk factor for lung cancer? *Am J Respir Crit Care Med*. 2014;190:549-559.
48. Park HY, Kang D, Shin SH, et al. Chronic obstructive pulmonary disease and lung cancer incidence in never smokers: a cohort study. *Thorax*. 2020;75:506-509.
49. Parris BA, O'Farrell HE, Fong KM, Yang IA. Chronic obstructive pulmonary disease (COPD) and lung cancer: common pathways for pathogenesis. *J Thorac Dis*. 2019;11:S2155-S2172.
50. Lambert SA, Gil L, Jupp S, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet*. 2021;53:420-425.
51. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15:2759-2772.
52. Zhang R, Chen C, Dong X, 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*. 2020;158:808-819.
53. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N Engl J Med*. 2020;382:503-513.
54. Oudkerk M, Devaraj A, Vliegenthart R, et al. European position statement on lung cancer screening. *Lancet Oncol*. 2017;18:e754-e766.
55. Oudkerk M, Liu S, Heuvelmans MA, Walter JE, Field JK. Lung cancer LDCT screening and mortality reduction - evidence, pitfalls and future perspectives. *Nat Rev Clin Oncol*. 2021;18:135-151.
56. Force U, Krist AH, Davidson KW, et al. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *JAMA*. 2021;325:962-970.
57. Tammemägi MC. Selecting lung cancer screenees using risk prediction models-where do we go from here. *Transl Lung Cancer Res*. 2018;7:243-253.
58. Balderston S, Taulbee JJ, Celaya E, et al. Discrimination of single-point mutations in unamplified genomic DNA via Cas9 immobilized on a graphene field-effect transistor. *Nat Biomed Eng*. 2021;5:713-725.
59. Zanetti KA, Wang Z, Aldrich M, et al. Genome-wide association study confirms lung cancer susceptibility loci on chromosomes 5p15 and 15q25 in an African-American population. *Lung Cancer*. 2016;98:33-42.
60. Pande M, Spitz MR, Wu X, Gorlov IP, Chen WV, Amos CI. Novel genetic variants in the chromosome 5p15.33 region associate with lung cancer risk. *Carcinogenesis*. 2011;32:1493-1499.
61. Chen Z, Wang J, Bai Y, 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*. 2017;24:20-27.
62. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet*. 2008;40:1404-1406.
63. Liu R, Tan J, Shen X, 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*. 2021;118:e2022779118.
64. James MA, Vikis HG, Tate E, Rymaszewski AL, You M. CRR9/CLPTM1L regulates cell survival signaling and is required for Ras transformation and lung tumorigenesis. *Cancer Res*. 2014;74:1116-1127.
65. Calado RT, Chen J. Telomerase: not just for the elongation of telomeres. *BioEssays*. 2006;28:109-112.
66. Zienolddiny S, Skaug V, Landvik NE, et al. The tert-CLPTM1L lung cancer susceptibility variant associates with higher DNA adduct formation in the lung. *Carcinogenesis*. 2009;30:1368-1371.
67. Gealy R, Zhang L, Siegfried JM, Luketich JD, Keohavong P. Comparison of mutations in the p53 and K-ras genes in lung carcinomas from smoking and nonsmoking women. *Cancer Epidemiol Biomarkers Prev*. 1999;8:297-302.
68. Lan Q, Hsiung CA, Matsuo K, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet*. 2012;44:1330-1335.
69. Wang G, Bai Y, Fu W, 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*. 2019;179:108747.
70. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. *Nat Rev Immunol*. 2018;18:325-339.
71. Ferreiro-Iglesias A, Lesseur C, McKay J, et al. Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. *Nat Commun*. 2018;9:3927.
72. Beatty GL, Gladney WL. Immune escape mechanisms as a guide for cancer immunotherapy. *Clin Cancer Res*. 2015;21:687-692.
73. Chen XL, Hong LL, Wang KL, et al. Deregulation of CSMD1 targeted by microRNA-10b drives gastric cancer progression through the NF-kappaB pathway. *Int J Biol Sci*. 2019;15:2075-2086.
74. Matevossian A, Resh MD. Hedgehog acyltransferase as a target in estrogen receptor positive, HER2 amplified, and tamoxifen resistant breast cancer cells. *Mol Cancer*. 2015;14:72.