Elsevier Editorial System(tm) for The Lancet
Manuscript Draft

Manuscript Number:

Title: Risk loci identification and polygenic risk score in prediction of lung cancer: a large-scale prospective cohort study in Chinese

Article Type: Article

Corresponding Author: Professor Hongbing Shen, Ph.D.

Corresponding Author's Institution: Nanjing Medical University,School of Public Health

First Author: Juncheng Dai

Order of Authors: Juncheng Dai; Jun Lv; Wen Tan; Jingyi Fan; Tianpei Wang; Qi Sun; Lijung Wang; Mingtao Huang; Zijun Ge; Canqing Yu; Yu Guo; Tong-min Wang; Robin G Walters; Yongqiao He; Iona Millwood; Zhengming Chen; Bian Zheng; Jie Wang; Lin Xu; Meng Zhu; Weibing Wu; Xi-Zhao Li; Xin Wang; Rayjean Hung; Haiquan Chen; Mengyun Wang; Chen Wang; Yue Jiang; Kexin Chen; Guangfu Jin; Tangchun Wu; Dongxin Lin; Zhibin Hu; Chen Wu; Christopher I Amos; Qingyi Wei; Wei-Hua Jia; Liming Li; Yuzhuo Wang; Na Qin; Hongxia Ma; Ruoxin Zhang; Hong Zheng; Liang Chen; Hongbing Shen

Abstract: Background Genetic variation plays an important role in the development of non-small cell lung cancer (NSCLC). However, major genetic factors for lung cancer have not been fully identified, especially in Chinese populations, which deters us from using a polygenic risk score (PRS) to identify sub-populations at high-risk of lung cancer for prevention.

Methods To systematically identify genetic variants for NSCLC risk, we newly genotyped 19,546 samples and conducted a meta-analysis of genome-wide association studies (GWASs) of 27,120 cases and 27,355 controls. We then built a PRS for Chinese populations and evaluated its utility and effectiveness in predicting high-risk populations of lung cancer in an independent prospective cohort of 95,408 individuals from China Kadoorie Biobank (CKB).

Findings We identified 19 susceptibility loci to be significantly associated with NSCLC risk at 5*0×10-8, including six novel ones. When applied to the CKB cohort, the PRS of the risk loci successfully predicted lung cancer incidence in a dose-response manner (Ptrend=2*02×10-9). Specially, we observed apparently separate predictive morbidity curves for low, intermediate, and high genetic risk populations respectively, and PRS was an independent effective risk stratification indicator beyond age and pack-years during a 10-year follow-up time of the cohort.

Interpretation Based on the systematic identification of the risk loci for NSCLC, we have proved for the first time that GWAS-derived PRS can be effectively used in screening for high-risk populations of lung cancer, potentially leading to a feasible PRS-based lung cancer screening program for individualized prevention in Chinese populations.

Funding National Natural Science Foundation of China, the Priority Academic Program for the Development of Jiangsu Higher Education Institutions, National Key R&D Program of China, and China's Thousand Talents Program.

Risk loci identification and polygenic risk score in prediction of lung cancer: a large-scale prospective cohort study in Chinese

Juncheng Dai, Ph.D. [1,2 †], Jun Lv, Ph.D. [3 †], Meng Zhu, Ph.D. [1,2 †], Yuzhuo Wang, M.Sc. [1 †], Na Qin, M.Sc. [1 †], Hongxia Ma, Ph.D. [1,4 †], Yong-Qiao He, M.Sc. [5 †], Ruoxin Zhang, Ph.D. [6 †], Wen Tan, Ph.D. [7], Jingyi Fan, M.Sc. [1], Tianpei Wang, M.Sc [1], Hong Zheng, M.D. [8], Qi Sun, B.S. [1], Lijuan Wang, B.S. [1], Mingtao Huang, M.Sc. [1], Zijun Ge, B.S. [1], Canqing Yu, Ph.D. [3], Yu Guo, M.D. [9], Tong-Min Wang, Ph.D. [5], Jie Wang, M.D. [10], Lin Xu, M.D. [10], Weibing Wu, M.D. [11], Liang Chen, M.D. [11], Zheng Bian, M.D. [9], Robin Walters, Ph.D. [12], Iona Millwood, Ph.D. [12], Xi-Zhao Li, Ph.D. [5], Xin Wang, M.D. [13], Rayjean J. Hung, Ph.D. [14], Haiquan Chen, M.D. [15], Mengyun Wang, M.D. [6], Cheng Wang , Ph.D. [1,16], Yue Jiang, M.Sc. [1,2], Kexin Chen, M.D. [8], Zhengming Chen, Ph.D. [12], Guangfu Jin, Ph.D. [1,4], Tangchun Wu, Ph.D. [17], Dongxin Lin, Ph.D. [7], Zhibin Hu, Ph.D. [1,4 *], Christopher I. Amos, Ph.D. [18 *], Chen Wu, Ph.D. [7 *], Qingyi Wei, Ph.D. [6,19,20 *], Wei-Hua Jia, Ph.D. [5 *], Liming Li, Ph.D. [3 *], Hongbing Shen, Ph.D. [1,2 *]

1 Department of Epidemiology and Biostatistics, International Joint Research Center on Environment and Human Health, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China
2 Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, Nanjing Medical University, Nanjing, China
3 Department of Epidemiology and Biostatistics, School of Public Health, Peking University Health Science Center, Beijing, China
4 State Key Laboratory of Reproductive Medicine, Center for Global Health, Nanjing Medical University, Nanjing, China
5 State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China
6 Cancer Institute, Fudan University Shanghai Cancer Center; Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China
7 Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
8 Department of Epidemiology and Biostatistics, Key Laboratory of Cancer Prevention and Therapy, Tianjin Key Laboratory of Breast Cancer Prevention and Therapy, Ministry of Education, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China
9 Chinese Academy of Medical Sciences, Beijing, China
10 Jiangsu Key Laboratory of Molecular and Translational Cancer Research; Department of Thoracic Surgery, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research, Nanjing Medical University Affiliated Cancer Hospital, Nanjing, China
11 Department of Thoracic Surgery, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China
12 Clinical Trial Service Unit & Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, United Kingdom

13 Department of Thoracic Surgery, Sun Yat-sen University Cancer Center, Guangzhou, China
14 Lunenfeld-Tanenbaum Research Institute of Sinai Health System, University of Toronto, Toronto, Canada
15 Department of Thoracic Surgery, Fudan University Shanghai Cancer Center; Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China
16 Department of Bioinformatics, School of Basic Medical Sciences, Nanjing Medical University, Nanjing, China
17 Department of Occupational and Environmental Health and Ministry of Education Key Lab for Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
18 Baylor College of Medicine, Institute for Clinical and Translational Research, Houston, Texas, United States of America
19 Duke Cancer Institute, Duke University Medical Center, Durham, NC, United States of America
20 Department of Population Health Sciences, Duke University School of Medicine, Durham, NC, United States of America

† These authors contributed equally to this work.

* Correspondence to: Hongbing Shen, Department of Epidemiology and Biostatistics, Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, Center for Global Health, School of Public Health, Nanjing Medical University, 101 Longmian Avenue, Nanjing 21116, China, Phone: 86-25-86868439, Email: hbshen@njmu.edu.cn;
or Liming Li, Email: lmlee@vip.163.com; or Wei-Hua Jia, Email: jiawh@sysucc.org.cn; or Qingyi Wei, Email: qingyi.wei@duke.edu; or Christopher I. Amos, Email: Chris.Amos@bcm.edu; or Chen Wu, Email: wuc@cicams.ac.cn; or Zhibin Hu, Email: zhibin_hu@njmu.edu.cn.

Summary

Background Genetic variation plays an important role in the development of non-small cell lung cancer (NSCLC). However, major genetic factors for lung cancer have not been fully identified, especially in Chinese populations, which deters us from using a polygenic risk score (PRS) to identify sub-populations at high-risk of lung cancer for prevention.

Methods To systematically identify genetic variants for NSCLC risk, we newly genotyped 19,546 samples and conducted a meta-analysis of genome-wide association studies (GWASs) of 27,120 cases and 27,355 controls. We then built a PRS for Chinese populations and evaluated its utility and effectiveness in predicting high-risk populations of lung cancer in an independent prospective cohort of 95,408 individuals from China Kadoorie Biobank (CKB).

Findings We identified 19 susceptibility loci to be significantly associated with NSCLC risk at $5{\cdot}0\times10^{-8}$, including six novel ones. When applied to the CKB cohort, the PRS of the risk loci successfully predicted lung cancer incidence in a dose-response manner ($P_{trend}=2{\cdot}02\times10^{-9}$). Specially, we observed apparently separate predictive morbidity curves for low, intermediate, and high genetic risk populations respectively, and PRS was an independent effective risk stratification indicator beyond age and pack-years during a 10-year follow-up time of the cohort.

Interpretation Based on the systematic identification of the risk loci for NSCLC, we have proved for the first time that GWAS-derived PRS can be effectively used in screening for high-risk populations of lung cancer, potentially leading to a feasible PRS-based lung cancer screening program for individualized prevention in Chinese populations.

Funding National Natural Science Foundation of China, the Priority Academic Program for the Development of Jiangsu Higher Education Institutions, National Key R&D Program of China, and China's Thousand Talents Program.

Research in context

Evidence before this study

We systematically searched PubMed and reviewed for research articles published in English before Aug 23, 2018, with search terms including “polygenetic risk score” and “GWAS”. In the past decades, hundreds of susceptibility loci have been identified for complex diseases using GWAS strategy. Based on previous findings, genetic risk loci from previous GWASs could be applied as a PRS strategy to identify individuals at an increased risk of diseases for prevention, such as Alzheimer’s disease, cardiovascular diseases, and breast cancer, etc. For lung cancer, 45 risk loci were identified among different populations, and a few case-control studies have been conducted to evaluate PRS, but such a PRS has poor discriminating ability in predicting lung cancer risk, largely because the lack of subsequent prospective studies to validate its efficiency and significance. Furthermore, most of the studies were performed among populations of European descent, and few studies included other ethnic groups.

Added value of this study

We identified six novel susceptibility loci to be significantly associated with NSCLC. We generated a PRS score by using the NSCLC related risk loci, and then we evaluated, for the first time, its utility and effectiveness in prediction of lung cancer risk in an independent large-scale prospective cohort. The PRS of the risk loci successfully predicted lung cancer incidence in Chinese populations. Specially, we observed apparently separate predictive morbidity curves and PRS was an independent effective risk stratification indicator beyond age and pack-years.

Implications of all the available evidence

The risk loci identified in the present study provided additional insights into genetic basis of lung cancer. Based on the findings of PRS prediction, we propose that all smokers over the age of 55 should have a genetic risk test to formulate an individualized lung cancer screening plan according to individual risk. PRS can be effectively used in screening for high-risk populations of lung cancer, potentially leading to a feasible PRS-based lung cancer screening program for individualized prevention in Chinese populations.

Introduction

Lung cancer is the leading type of cancer with the highest incidence and mortality in China [1] and the world. [2] It is estimated that more than 2·09 million new lung cancer cases will occur in 2018, accounting for approximately 11·6% of the total cancer diagnoses. [2] Non-small cell lung cancer (NSCLC) accounts for ~ 85% of total lung cancer cases and presents severe threats to the health of the populations. [3] Although environmental risk factors (e.g. smoking) contribute the most to risk of NSCLC, [4] genetic variants can explain approximately 12% ~ 21% of the heritability of lung cancer. [5,6] In the past decade, genome-wide association studies (GWASs) have identified 45 risk loci of lung cancer in different ethnic populations; [7] however, genetic factors associated with lung cancer risk have not been fully identified, especially in Chinese populations.

Recent large-scale population studies have suggested that the combined effect of common genetic variants might serve as an efficient screening tool for some complex diseases. [8] The findings of genetic risk loci from previous GWASs could be applied as a polygenic risk score (PRS) strategy to identify individuals at an increased risk of diseases for prevention, such as Alzheimer's disease, [9] cardiovascular diseases, [10,11] and breast cancer. [12] For lung cancer, a few case-control studies have been conducted to use a PRS derived from genetic risk loci, but such a PRS has poor discriminating ability in predicting lung cancer risk, [13,14] largely because the lack of subsequent prospective studies to validate its efficiency and significance. Furthermore, most of the studies were performed among populations of European descent, and few studies included other ethnic groups.

To systematically identify genetic factors for NSCLC risk, we newly genotyped additional 19,546 samples in Chinese populations, and then performed a meta-analysis of GWASs from both Chinese and European populations, with a total of 27,120 cases and 27,355 controls. We generated a PRS score by using the NSCLC related risk loci, and then we evaluated, for the first time, its utility and effectiveness in prediction of lung cancer risk in a subset of 95,408 individuals randomly selected from an independent large-scale prospective cohort from China Kadoorie Biobank (CKB) with more than ten years' follow-up.

Methods

Study design and subjects

All participants in the present study signed an informed consent form which was approved by the local internal review boards or ethics committees. The workflow chart for the study design is illustrated in Fig. S1 in the Supplementary Appendix.

Individuals for NSCLC risk-loci identification. For the present study, we newly performed genome-wide scan for 19,546 samples (NJMU GSA project: 10,248 cases and 9,298 controls) using Illumina Global Screening Array (GSA), which included three studies (Nanjing GSA study: 4,149 cases and 3,198 controls from Nanjing and Shanghai; Beijing GSA study: 2,155 cases and 2,035 controls from Beijing and Tianjin; and Guangzhou GSA study: 3,944 cases and 4,065 controls from Guangdong). We then performed a meta-analysis of GWASs with 27,120 NSCLC cases and 27,355 controls from two populations (13,327 cases and 13,328 controls of Chinese descent [15] as well as 13,793 cases and 14,027 controls of European descent). [16] For populations of Chinese descent, except for 19,546 samples that were newly genotyped (NJMU GSA), the remaining samples were derived from our previous lung cancer GWAS data (2,126 cases and 3,077 controls) [15] and our unpublished lung cancer OncoArray data (953 cases and 953 controls). All these lung cancer cases were histologically confirmed as incident NSCLC by at least two pathologists, who had not received chemo- or radio-therapy before diagnosis. Cancer-free controls were selected from those participants in the community screening of non-communicable diseases, frequency-matched to cases by age, sex, and geographic regions. All participants provided core data on age, sex, smoking status, smoking pack-years, and histological types. Demographic characteristics of all participants are displayed in Table S1-1 and Table S1-2 in the Supplementary Appendix. For populations of European descent, all data were integrated from the TRICL-ILCCO OncoArray project (13,793 cases and 14,027 controls). [16] Detailed information is provided in Appendix 1 in the Supplementary Appendix.

The prospective cohort study for PRS application. This prospective cohort study included 95,793 individuals with both phenotype and genotype data who were randomly selected from 512,891 Chinese adults of CKB, [17,18] a nationwide prospective cohort. Participants in CKB were followed up for cancer events mainly through the linkage with official death certificates, chronic disease registries, and the national health insurance system, which are annually supplemented with local residential records. [19] Detailed description for CKB is provided in Appendix 1 in the Supplementary Appendix and the demographic characteristics of the participants are provided in Table S2 in the Supplementary Appendix.

Genotyping and quality control

Genotyping array. For risk-loci identification, additional genotyping was performed using Illumina Infinium® Global Screening Array (GSA) v1.0, which contains 700,078 markers. The genotyping was completed by using the Illumina iScan System according to the manufacturer's protocols. Information on the other GWAS datasets included in this study is provided in Appendix 2 in the Supplementary Appendix. The 95,408 CKB samples used in the PRS application were genotyped using a customized Affymetrix Axiom® CKB array (optimized for usage of Han Chinese subjects) by the BGI company, which consists of ~ 700,000 markers.

Quality control and imputation for genotyping samples and variations. For all datasets used in the present study, we performed standard quality control at both sample-level and variant-level according to the literature. [15,16,20] Qualified samples and genotypes in each dataset were phased and imputed with SHAPEIT v2 [20,21] and IMPUTE v2, [22] respectively by using default parameters and the 1000 Genomes Project Phase III database (released in October, 2014) as the reference. Details concerning the processes of quality control and imputation are described in Appendix 2 in the Supplementary Appendix.

Utility and effectiveness of polygenic risk score (PRS)

SNPs selection for PRS. We derived a PRS specific for Chinese populations from all reported SNPs (i.e., 81 SNPs in 40 loci) that have been reported to be associated with lung cancer risk at genome-wide significance level in both current and previous studies. [7] To ensure the efficiency of the model, several criteria were applied to filter out redundant variants: (1) MAF < 0·01 in Chinese populations; (2) SNPs in linkage disequilibrium (LD, r^2 ≥0·1); (3) *P* value ≥ 0·00125 (after multiple comparison at 0·05/40, 40 independent loci remained according to the above two criteria) in Chinese populations included in the present study. As a result, 19 risk variants were kept for the PRS calculation, 16 of which with a *P* value < $5{\cdot}0 \times 10^{-08}$ in the current study. The detailed process is illustrated in Fig. S1 in the Supplementary Appendix.

PRS calculation. In the present study, PRS was generated by multiplying the genotype dosage of each risk allele for each variant by its respective weight (lnOR) in Chinese population, summing all variants together. Effect sizes for all variants were derived from the association of NSCLC patients of Chinese descent in the present study, which were all flipped into risk alleles, where appropriate. PRS was calculated by statisticians (J.D. and M.Z.) who were blinded to the end points in the CKB cohort, and then the association between PRS and lung cancer risk was evaluated independently.

Statistical analyses

For the GWAS array datasets, per-allele odds ratios (ORs) and standard errors (SEs) were calculated using logistic regression analysis with the SNPTEST (v2.5.4) software based on a probabilistic dosage model. [23] A fixed-effects meta-analysis was performed to combine individual association estimates from each GWAS dataset using the METAL software. [24] Other methods used in the present study are illustrated in Appendix 2 in the Supplementary Appendix. For the CKB cohort study, we used the Cox proportional-hazards regression model to test the association between genetic risk and incident lung cancer events with the adjustments of age, sex, source of region, and smoking status. Participants were classified into ten equal parts according to the distribution of PRS, and we compared hazard ratios (HR) for each part with those at

the lowest tenth. Individuals within the top 5%, 5%-95%, and the bottom 5% of PRS were considered as populations at high, intermediate, and low genetic risk respectively. We used Cox regression to calculate cumulative event rates by the end of 2016 in each of the three subgroups, which were standardized to the mean of the above-mentioned adjustments. All the analyses were performed using the R software (version 3.5.1, R Project for Statistical Computing).

Role of the funding source

The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Characteristics of the subjects for risk loci identification

For the NSCLC risk loci identification, we included 27,120 NSCLC cases and 27,355 controls in the analysis. No obvious population stratification was observed either for overall samples or specific subgroup samples (Fig. S2 in the Supplementary Appendix). The individuals' age and sex were well balanced between the case-control groups, in which 57·5% cases were classified as lung adenocarcinoma (LUAD) and 30·8% as lung squamous cell carcinoma (LUSC) (Table S1-1 in the Supplementary Appendix).

Identification of NSCLC risk loci

We identified 19 risk loci at the GWAS-significant level (i.e., $5{\cdot}0\times10^{-8}$) for the overall NSCLC or histological association analysis. Among the 19 loci, six were novel ones, including three loci for overall NSCLC: 2q33.1 (rs3769821: odds ratios (OR) = 1·08, $P = 4{\cdot}45\times10^{-8}$), 3q26.2 (rs2293607: OR = 1·10, $P = 1{\cdot}82\times10^{-10}$), 14q13.1 (rs1200399: OR = 1·11, $P = 3{\cdot}05\times10^{-9}$); two for LUAD: 2p14 (rs17038564: OR = 1·15, $P = 1{\cdot}87\times10^{-8}$), 9p13.3 (rs35201538: OR = 1·10, $P = 1{\cdot}99\times10^{-8}$); and one for LUSC: 9q33.2 (rs4573350: OR = 1·13, $P = 3{\cdot}23\times10^{-9}$) (Figure 1, Table 1). We also replicated

13 previously reported loci at the GWAS-significant level, including 12 loci for overall NSCLC: 3q28, 5p15.33, 6p22.1, etc.; one for LUAD: 15q21.1 (Figure 1, Table 1). It is noteworthy that three risk loci (8p12: rs4236709; 9p21.3: rs10429489 and 11q23.3: rs55768116) previously reported in European populations were also identified in Chinese populations for the first time. Regional plots for these 19 risk loci were used to illustrate the LD pattern and related genes in each region (Fig. S3 in the Supplementary Appendix).

As shown in Figure 1 A and 1 B, we identified four novel loci for LUAD: 2p14 and 9p13.3 were specifically significant in LUAD subgroup, while 3q26.2 and 14q13.1 were detected in both overall NSCLC and LUAD. As shown in Figure 1 B and 1 C, the genetic architecture of lung cancer varied markedly between histological subtypes, with striking differences between LUAD and LUSC. Fewer signals were found for LUSC, and only one signal was specific for LUSC (9q33.2).

To compare genetic architectures between Chinese and European populations, forest plots with bubbles were used to indicate the effect estimates and effect allele frequencies (EAFs) for each lead SNP across overall and histological datasets (Figure 2). As shown in Figure 2, 13 out of 19 identified risk loci showed consistent effect estimates across ancestries, which means that these risk-associated loci are shared across ethnic populations. On the other hand, we also observed the heterogeneity of the risk-associated loci and their effects across different populations (e.g., 3q28, 5p15.33, 6p21.1, and 9q33.2). We also performed subgroup analysis for the six novel risk loci by sex, smoking, and histology status (Table S3 and Fig. S4 in the Supplementary Appendix). Compared with the overall samples, most of the signals had similar effect sizes in specific subgroups.

PRS construction and evaluation

For the PRS application, a total of 95,408 Chinese participants from CKB were included in the prediction study. During the follow-up, 1,316 lung cancer patients were registered and confirmed with a median follow-up time of 10·44 years (Table S2 in the Supplementary Appendix). We derived a PRS calculation model specific for

Chinese populations by using 19 SNPs, which were selected from 81 previously reported SNPs associated with lung cancer risk (Fig. S1 and Table S4 in the Supplementary Appendix). The process of PRS calculation is provided in Table S5 in the Supplementary Appendix.

The GWAS-derived PRS could significantly predict lung cancer risk in the CKB cohort. A risk gradient across each grade of genetic risk was observed such that the participants at a high genetic risk (i.e., in the top tenth of the PRS) were at significantly higher risk of lung cancer than those at a low genetic risk (i.e., in the lowest tenth genetic risk), with an adjusted HR of 1·96 (95% CI, 1·53 to 2·51) (P_{trend}=2·02×10^{-9}) (Figure 3 A). If participants within the top 5% and bottom 5% of PRS were defined as high and low genetic risk populations, the apparently separate predictive morbidity curves were observed in the CKB cohort during the follow-up, with a relative risk of lung cancer 137% higher among participants at high genetic risk than those at low genetic risk (HR=2·37, 95% CI: 1·64-3·44) (Figure 3 B). Similar prediction results were observed in nonsmokers and smokers, females and males, even if we extended the definition of high-risk group to the top 10% population (Fig. S5 in the Supplementary Appendix).

A cumulative effect of the PRS and smoking was observed for incident lung cancer events according to the results from CKB. Figure 3 C showed that the cancer risk in light smokers (pack-year, PY<30) at low genetic risk was similar to nonsmokers (HR=1·17, 95% CI: 0·64-2·15), while a gradually increased cancer risk was observed for light smokers (pack-year, PY<30) at intermediate genetic risk (HR=1·79, 95% CI: 1·49-2·14), heavy smokers (PY≥30) at low genetic risk (HR=2·08, 95% CI: 1·18-3·67), light smokers at high genetic risk (HR=2·93, 95%CI: 1·92-4·49), heavy smokers at intermediate genetic risk (HR=3·27, 95% CI: 2·71-3·94), and heavy smokers at high genetic risk (HR=3·98, 95% CI: 2·64-5·99) after adjustment for age, sex, and region (Figure 3 C). Although smoking was the most important risk factor of lung cancer, light smokers at high genetic risk showed a higher risk compared with heavy smokers at low genetic risk in our cohort. Within each category of genetic risk, the degrees of smoking were associated with an increase in relative risk of lung cancer (Figure 3 D). Participants who smoked heavily at low genetic risk had a lower rate (267·5 per 100,000 person-years) of lung cancer compared to that (339·9 per 100,000

person-years) of light smokers at high genetic risk. These results suggest that the PRS has the ability to predict risk of lung cancer and that it potentially optimizes the definition of sub-populations at high-risk in individualized lung cancer prevention beyond pack-years and other predictors. Besides, a low genetic risk could be largely offset by smoking. A much higher standardized cumulative lung cancer risk was observed among heavy smokers at low genetic risk compared with never smokers at high genetic risk (267·5 *vs.* 156·1 per 100,000 person-years), which support public health efforts that emphasize a healthy lifestyle of non-smoking for everyone, even for those individuals who are at a low genetic risk.

Discussion

In the present study, we performed a large-scale GWAS of lung cancer, and we found six novel variants and confirmed 13 previously reported variants that were associated with NSCLC risk. In the independent prospective cohort study, the GWASs derived PRS significantly predicted lung cancer incidence in a dose-response manner, and the apparently separate predictive morbidity curves were observed during the follow-up for low, intermediate, and high genetic risk populations respectively. We proved, for the first time, that PRS can be an effective tool to predict lung cancer risk and potentially applied in individualized cancer prevention.

Based on the results of subgroup analysis by histology of lung cancer, we observed a strikingly histological heterogeneity between LUAD and LUSC. Among the 19 loci we identified, 16 loci (e.g., 2p14, 3q26.2, 9p13.3 14q13.1, and 15q21.1) were significant in the LUAD subgroup, while only one signal was specific for LUSC (9q33.2). Previous studies revealed that LUAD and LUSC were commonly thought to be different diseases because of their distinct biology and genomic abnormalities in terms of either germline variations or somatic alterations. [16,25-28] It has been reported that smoking is more strongly associated with LUSC than LUAD, indicating different mechanisms of carcinogenesis for these histological subtypes. [29] As for the comparisons of different ethnic groups, most of the risk-associated loci are shared across ethnic populations. Larger sample sizes and more complete ascertainment of variants (particularly in Asian studies) would better assess genetic architecture of lung cancer across divergent populations. [7] Although we observed genetic heterogeneity

between populations, our findings extended the knowledge that most of the risk loci of NSCLC were consistent and shared among populations of both Chinese and European descent.

The effectiveness of the PRS is often evaluated by determining whether it can help separate the population into categories with distinct degrees of absolute risk to drive clinical or personal decision-making. [8] Although several articles have evaluated the predictive performance of PRS in lung cancer, the limited sample size with case-control design in previous studies often led to discouraging results. [13,14] In the present study, we, for the first time, proved that GWAS-derived PRS significantly predicted risk of lung cancer in a nationwide prospective cohort study. The distinct incidence rates for populations within different categories of genetic risk as defined by the PRS provide a strong evidence that GWAS findings can be used in lung cancer screening and individualized prevention. Compared with previous studies on PRS, [13,14] the strength of the present study is that samples from the CKB cohort were absolutely independent from those used in the GWAS, which has avoided overestimation of the PRS effect.

Clinical utility of PRS can be roughly categorized into three major classes of interventions: PRS-informed disease screening, PRS-informed life planning, and PRS-informed therapeutic intervention. [8] In the present study, we observed that GWAS-based PRS was a potential predictor for lung cancer risk beyond age and pack-years of smoking, the main screening eligibility criteria used in the current guidelines by the United States Preventive Services Task Force and the NCCN Clinical Practice Guidelines. [30-32] By demonstrating that different cumulative event rates of lung cancer were observed for heavy smokers at low, intermediate, and high genetic risk during follow-up in the CKB cohort, the present study provides the necessity of individualized screening plans, such as changing the interval between lung cancer screening according to individual risk to reduce potential hazards from the radiation of CT scans. Besides, the present findings revealed that light smokers with a high genetic risk showed a high incident rate of lung cancer close to heavy smokers in the CKB cohort, indicating that these participants should be included in the screening program. Given the overall findings, we propose that all smokers over the age of 55 should have a genetic risk test to formulate an individualized lung

cancer screening plan according to individual risk. Moreover, the PRS may have its utility even in the absence of, or a personal desire to avoid, preventive screening. Similar to recent studies on coronary disease, [10] the population with a genetic score of the top one-twentieth, even the top one-tenth, has the ability to offset much of this risk by sticking to not smoking throughout their lifetime, leading to reduction of their overall risk of lung cancer by nearly 60%.

In conclusion, the risk loci identified in the present study provided additional insights into genetic basis of lung cancer, which would further advance our understanding of lung cancer susceptibility and carcinogenesis. More importantly, we proved in the present study, for the first time, that PRS could be effectively used for lung cancer risk prediction and potentially applied in lung cancer screening program for individualized prevention.

Authors' contributions

H.S., L.L., W.J., Q.W., C.W., A.C., Z.H., and J.L. contributed to the study design and sample collection and supervise the whole project. H.S., J.D., and M.Z. contributed to the data interpretation, data analysis, and writing of the manuscript. Y.W., N.Q., H.M., Y.H., and R.Z. contributed to the study design, sample collections, and data interpretation of the present analysis. W.T., J.F., T.W., H.Z., Q.S., L.W., M.H., Z.G., C.Y., W.W., L.C., Z.B, R.W., I.M., X.L., X.W., R.H., H.C., M.W., C.W., Y.J., K.C., Z.C., G.J., T.W., and D.L. contributed to the study design, sample collection and experiment. All of the authors reviewed or revised the manuscript.

Declaration of interests

The authors declare no competing interests.

Acknowledgments

We thank International Lung and Cancer Consortium (ILCCO) and Carotene and Retinol Efficacy Trial (CARET) project for lung cancer, who shared the GWAS data from European populations. We thank all the study participants and research staff for their contributions and commitment to the present study. This work was supported by National Natural Science Foundation of China (81521004 and 81820108028), and the Priority Academic Program for the Development of Jiangsu Higher Education Institutions [Public Health and Preventive Medicine], National Key R&D Program of China (2016YFC0900500, 2016YFC0900501, 2016YFC0900504, 2016YFC302700) and China's Thousand Talents Program, Fudan University, Shanghai, People's Republic of China.

References

1. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. *CA: a cancer journal for clinicians* 2016; **66**(2): 115-32.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 2018.
3. Alberg AJ, Brock MV, Ford JG, Samet JM, Spivack SD. Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* 2013; **143**(5 Suppl): e1S-e29S.
4. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine* 2000; **343**(2): 78-85.
5. Dai J, Shen W, Wen W, et al. Estimation of heritability for nine common cancers using data from genome-wide association studies in Chinese population. *International journal of cancer* 2017; **140**(2): 329-36.
6. Sampson JN, Wheeler WA, Yeager M, et al. Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. *Journal of the National Cancer Institute* 2015; **107**(12): djv279.
7. Bosse Y, Amos CI. 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* 2018; **27**(4): 363-79.
8. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nature reviews Genetics* 2018; **19**(9): 581-90.
9. Desikan RS, Fan CC, Wang Y, et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS medicine* 2017; **14**(3): e1002258.
10. Khera AV, Emdin CA, Drake I, et al. Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *The New England journal of medicine* 2016; **375**(24): 2349-58.
11. Natarajan P, Young R, Stitziel NO, et al. Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* 2017; **135**(22): 2091-101.
12. Maas P, Barrdahl M, Joshi AD, et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA oncology* 2016; **2**(10): 1295-302.
13. Weissfeld JL, Lin Y, Lin HM, et al. Lung Cancer Risk Prediction Using Common SNPs Located in GWAS-Identified Susceptibility Regions. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* 2015; **10**(11): 1538-45.
14. Qian DC, Han Y, Byun J, et al. A Novel Pathway-Based Approach Improves Lung Cancer Risk Prediction Using Germline Genetic Variations. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2016; **25**(8): 1208-15.
15. Hu Z, Wu C, Shi Y, et al. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nature genetics* 2011; **43**(8): 792-6.
16. 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. *Nature*

genetics 2017; **49**(7): 1126-32.
17. Chen Z, Lee L, Chen J, et al. Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). *International journal of epidemiology* 2005; **34**(6): 1243-9.
18. Chen Z, Chen J, Collins R, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International journal of epidemiology* 2011; **40**(6): 1652-66.
19. Pan R, Zhu M, Yu C, et al. Cancer incidence and mortality: A cohort study in China, 2008-2013. *International journal of cancer* 2017; **141**(7): 1315-23.
20. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nature methods* 2011; **9**(2): 179-81.
21. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* 2013; **10**(1): 5-6.
22. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* 2009; **5**(6): e1000529.
23. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* 2007; **39**(7): 906-13.
24. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**(17): 2190-1.
25. Borczuk AC, Toonkel RL, Powell CA. Genomics of lung cancer. *Proc Am Thorac Soc* 2009; **6**(2): 152-8.
26. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014; **511**(7511): 543-50.
27. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012; **489**(7417): 519-25.
28. Wang C, Yin R, Dai J, et al. Whole-genome sequencing reveals genomic signatures associated with the inflammatory microenvironments in Chinese NSCLC patients. *Nature communications* 2018; **9**(1): 2054.
29. Shiraishi K, Kunitoh H, Daigo Y, et al. A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nature genetics* 2012; **44**(8): 900-3.
30. Aberle DR, Adams AM, Berg CD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England journal of medicine* 2011; **365**(5): 395-409.
31. de Koning HJ, Meza R, Plevritis SK, et al. Benefits and harms of computed tomography lung cancer screening strategies: a comparative modeling study for the U.S. Preventive Services Task Force. *Annals of internal medicine* 2014; **160**(5): 311-20.
32. Wood DE, Kazerooni EA, Baum SL, et al. Lung Cancer Screening, Version 3.2018, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network : JNCCN* 2018; **16**(4): 412-41.

Figure1

Click here to download high resolution image

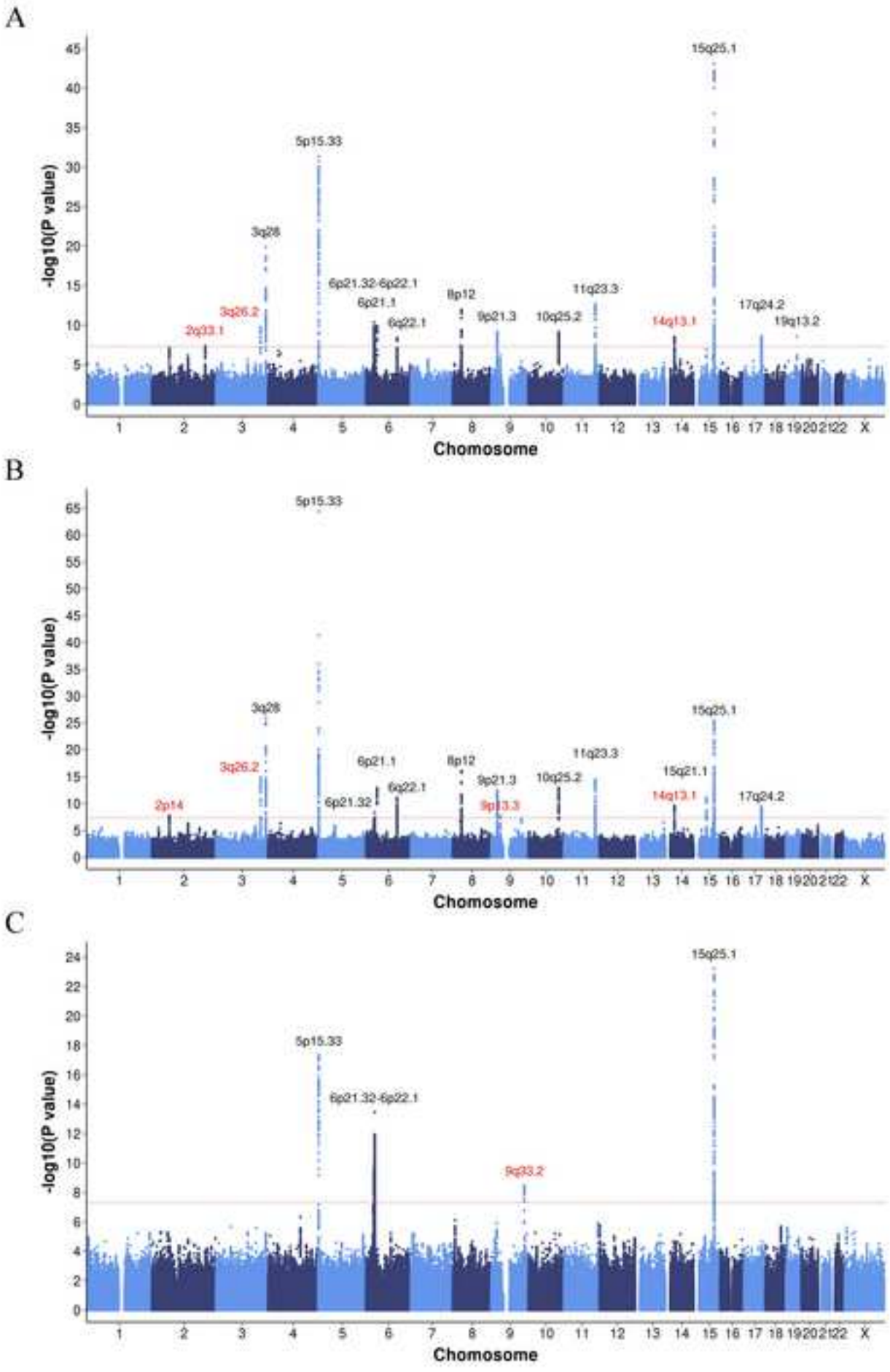

Figure2

Click here to download high resolution image

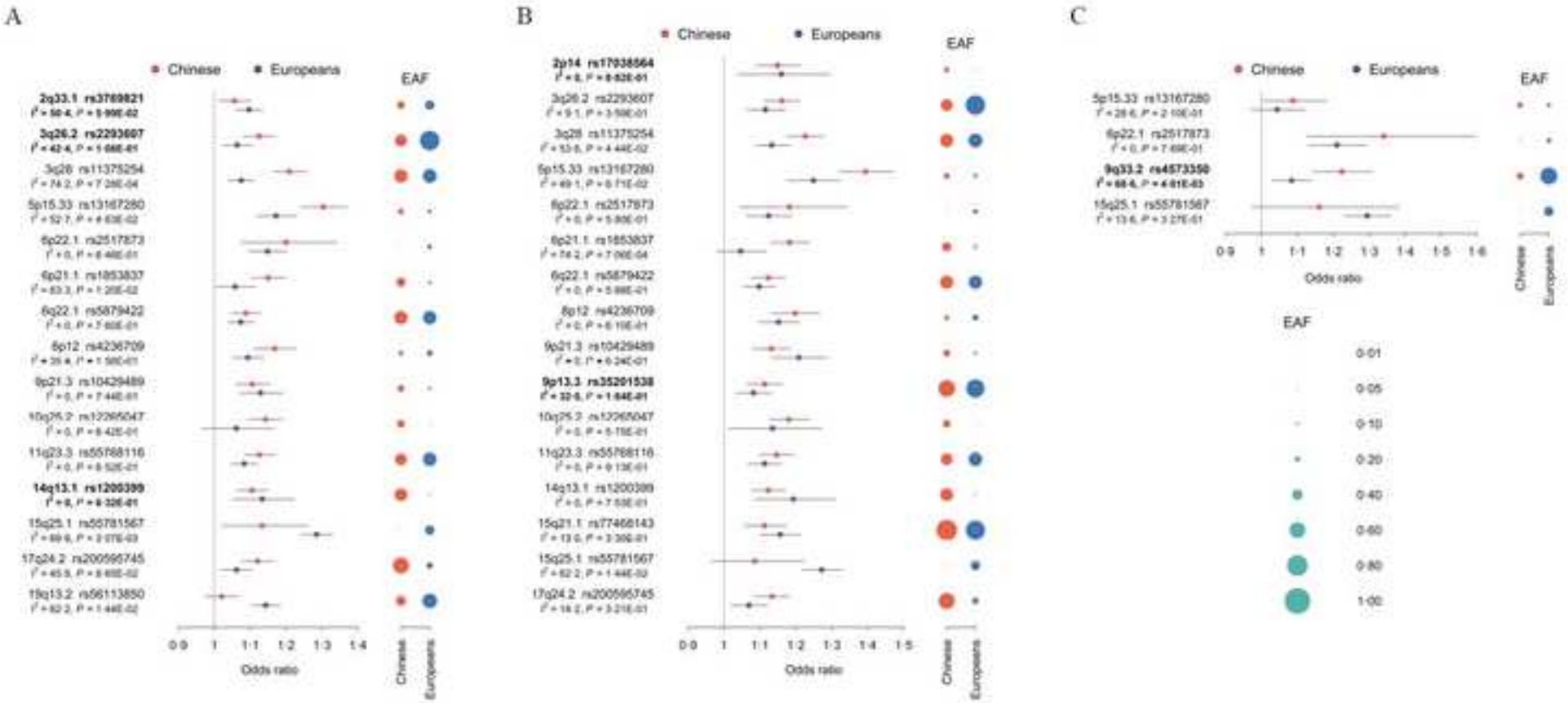

Figure3

Click here to download high resolution image

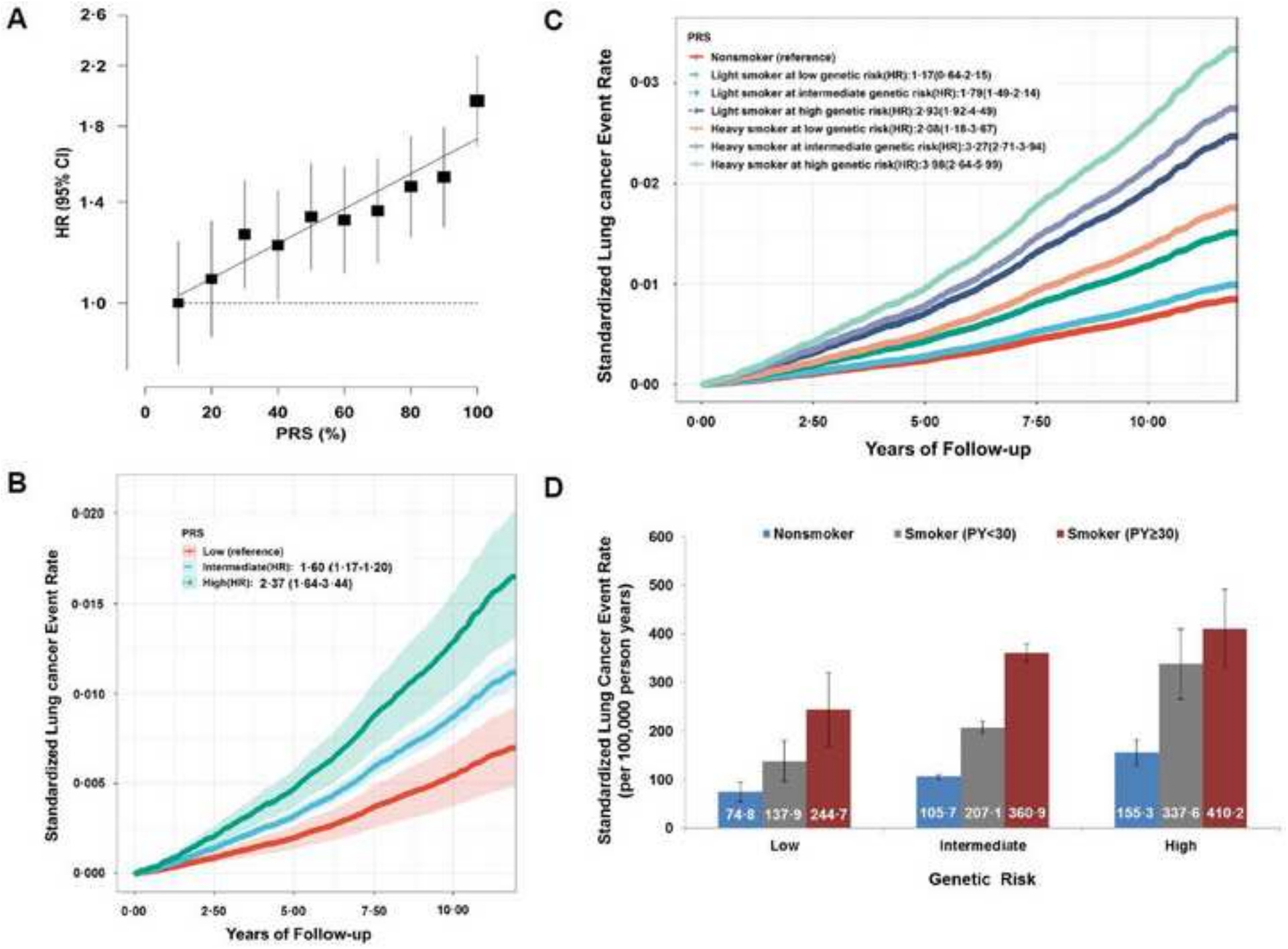

Table1

Table 1. Results for 19 identified NSCLC risk loci from the GWAS meta-analysis

Variants	CytoBand [a]	Pos. [b]	Gene	Phenotype	Effect allele	Reference allele	EAF	OR (95%CI)	*P*	Het *P*-value
Novel associations										
rs3769821	2q33.1	202123430	*CASP8*	NSCLC	C	T	0·321	1·08 (1·05-1·11)	4·45E-08	5·99E-02
rs2293607	3q26.2	169482335	*MYNN*	NSCLC	T	C	0·596	1·10 (1·06-1·13)	1·82E-10	1·08E-01
rs1200399	14q13.1	35293185	*B1Z1A*	NSCLC	C	T	0·405	1·11 (1·07-1·15)	3·05E-09	6·32E-01
rs17038564	2p14	65496058	*AFTPH*	LUAD	G	A	0·161	1·15 (1·10-1·21)	1·87E-08	8·83E-01
rs35201538	9p13.3	33422488	*AQP3*	LUAD	C	CT	0·676	1·10 (1·06-1·13)	1·99E-08	1·84E-01
rs4573350	9q33.2	124955115	*DAB2IP*	LUSC	T	C	0·505	1·13 (1·09-1·18)	3·23E-09	4·01E-03
Previously reported associations										
rs11375254	3q28	189343242	*TP63*	NSCLC	T	TA	0·525	1·13 (1·10-1·16)	1·35E-20	7·28E-04
rs13167280	5p15.33	1280477	*TERT*	NSCLC	A	G	0·170	1·24 (1·19-1·28)	4·59E-32	4·83E-02
rs1853837	6p21.1	41497035	*FOXP4*	NSCLC	A	C	0·258	1·12 (1·08-1·15)	1·21E-10	1·20E-02
rs2517873	6p22.1	29875992	*MHC*	NSCLC	A	G	0·145	1·16 (1·11-1·21)	4·60E-11	6·46E-01
rs5879422	6q22.1	117784658	*DCBLD1*	NSCLC	T	TTG	0·510	1·08 (1·05-1·11)	4·36E-09	7·60E-01
rs4236709	8p12	32410110	*NRG1*	NSCLC	G	A	0·198	1·12 (1·09-1·16)	1·11E-12	1·58E-01
rs10429489	9p21.3	21787521	*CDKN2A*	NSCLC	A	G	0·195	1·11 (1·08-1·15)	6·92E-10	7·44E-01
rs12265047	10q25.2	114487925	*VTI1A*	NSCLC	G	A	0·243	1·13 (1·09-1·17)	7·50E-10	6·42E-01
rs55768116	11q23.3	118108331	*MPZL3*	NSCLC	C	A	0·489	1·10 (1·07-1·13)	2·23E-13	6·52E-01
rs77468143	15q21.1	49376624	*SECISBP2L*	LUAD	T	G	0·767	1·14 (1·10-1·18)	7·48E-12	3·30E-01
rs55781567	15q25.1	78857986	*CHRNA5*	NSCLC	G	C	0·320	1·27 (1·23-1·31)	8·44E-44	3·07E-03
rs200595745	17q24.2	65915289	*BPTF*	NSCLC	A	AAATAATAAT	0·430	1·09 (1·06-1·12)	2·37E-09	8·65E-02
rs56113850	19q13.2	41353107	*CYP2A6*	NSCLC	C	T	0·487	1·09 (1·06-1·12)	2·69E-09	1·44E-02

NSCLC: non-small cell lung cancer; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; EAF, effect allele frequency; OR, odds (log-additive) ratio; 95% CI, 95% confidence interval. [a] Cytogenic band where the variant is positioned; [b] Chromosome position, hg19/GRCh37 build. Cochran's Q test was used to test for heterogeneity in SNP effect sizes across studies (Het *P*-value)

Supplementary Material

Click here to download Supplementary Material: GSA-Supplementary_appendix.pdf