Association analysis of driver-gene related genetic variants identified novel lung cancer susceptibility loci with 20,871 lung cancer cases and 15,971 controls

Yuzhuo Wang [1,2†], Olga Y. Gorlova [3†], Ivan P. Gorlov [3], Meng Zhu [1,2,4], Juncheng Dai [1,4], Demetrius Albanes [5], Stephen Lam [6], Adonina Tardon [7], Chu Chen [8], Gary Goodman [9], Stig E. Bojesen [10,11], Maria Teresa Landi [12], Mattias Johansson [13], Angela Risch [14,15,16], H-Erich Wichmann [17,18,19], Heike Bickeboller [20], David C. Christiani [21], Gadi Rennert [22], Susanne Arnold [23], Paul Brennan [13], John K. Field [24], Sanjay Shete [25], Loic Le Marchand [26], Olle Melander [27,28], Hans Brunnstrom [27], Geoffrey Liu [29], Rayjean J. Hung [30], Angeline Andrew [31], Lambertus A. Kiemeney [32], Shanbeh Zienolddiny [33], Kjell Grankvist [34], Mikael Johansson [35], Neil Caporaso [36], Penella Woll [37], Philip Lazarus [38], Matthew B. Schabath [39], Melinda C. Aldrich [40], Victoria L. Stevens [41], Hongxia Ma [1,4], Guangfu Jin [1,4], Zhibin Hu [1,4], Christopher I. Amos [42]*, Hongbing Shen [1,4]*

1 Department of Epidemiology, International Joint Research Center on Environment and Human Health, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China.

2 Department of Thoracic Surgery, Jiangsu Key Laboratory of Molecular and Translational Cancer Research, Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research & Nanjing Medical University Affiliated Cancer Hospital, Nanjing, China.

3 Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, United States of America.

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

5 Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States of America.

6 Department of Integrative Oncology, British Columbia Cancer Agency, Vancouver, British Columbia, Canada.

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

8 Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America.

9 Public Health Sciences Division, Swedish Cancer Institute, Seattle, Washington, United States of America.

10 Department of Clinical Biochemistry, Copenhagen University Hospital, Copenhagen, Denmark.

11 Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

12 National Cancer Institute, Bethesda, Maryland, United States of America.

13 Genetic Epidemiology Group, International Agency for Research on Cancer, Lyon, France.

14 Cancer Center Cluster Salzburg at PLUS, Department of Molecular Biology, University of Salzburg, Salzburg, Austria.

15 Biobank and Tumor Documentation, Thoraxklinik at University Hospital Heidelberg, Heidelberg, Germany.

16 Translational Lung Research Center Heidelberg (TLRC-H), German Center for Lung Research (DZL), Heidelberg, Germany.

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

18 Helmholtz Zentrum Munchen, German Research Center for Environmental Health (GmbH), Institute of Epidemiology, Neuherberg, Germany.

19 Institute of Medical Statistics and Epidemiology, Technical University Munich, Munich, Germany.

20 Department of Genetic Epidemiology, University Medical Center Goettingen, Goettingen, Germany.

21 Departments of Environmental Health and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America.

22 Technion Faculty of Medicine, Carmel Medical Center, Haifa, Israel.

23 Markey Cancer Center, University of Kentucky, Lexington, Kentucky, United States of America.

24 Molecular and Clinical Cancer Medicine, Roy Castle Lung Cancer Research Programme, The University of Liverpool Institute of Translational Medicine, Liverpool, United Kingdom.

25 Department of Epidemiology, The University of Texas, MD Anderson Cancer Center, Houston, Texas, United States of America.

26 Epidemiology Program, University of Hawai'i Cancer Center, Honolulu, Hawai'i, United States of America.

27 Clinical Sciences, Lund University, Lund, Sweden.

28 Department of Internal Medicine, Skåne University Hospital, Malmö, Sweden.

29 Princess Margaret Cancer Centre, Toronto, Ontario, Canada.

30 Prosseman Centre for Population Health Research, Lunenfeld-Tanenbuaum Research Institute, Sinai Health System, Toronto, Ontario, Canada.

31 Department of Neurology, Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire, United States of America.

32 Department of Health Evidence, Radboud University Medical Center, Nijmegen, Germany.

33 National Institute of Occupational Health (STAMI), Oslo, Norway.

34 Department of Medical Biosciences, Umeå University, Umea, Sweden.

35 Department of Radiation Sciences, Umeå University, Umea, Sweden.

36 Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America.

37 Academic Unit of Clinical Oncology, University of Sheffield, Sheffield, United Kingdom.

38 College of Pharmacy, Washington State University, Spokane, Washington, United States of America.

39 Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, United States of America.

40 Department of Thoracic Surgery, Division of Epidemiology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America.

41 Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, United States of America.

42 Department of Medicine, Epidemiology Section, Institute for Clinical and Translational Research, Baylor Medical College, Houston, Texas, United States of America

* **Correspondence to:** Hongbing Shen, Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing 211166, China, Tel/fax: +86 25

868 68437, E-mail: hbshen@njmu.edu.cn. Christopher I. Amos, Department of Medicine, Epidemiology Section, Institute for Clinical and Translational Research, Baylor Medical College, Houston, 77030, TX, USA, Tel: +1 603 650 1972, Fax: +1 603 653 6696, Email: chris.amos@bcm.edu.

† These authors contributed equally to this work.

Running title: Genetic variants in lung CDGs influence lung cancer risk

Keywords: genetic variation, lung cancer driver gene, lung cancer risk, genome-wide association study, somatic alteration

Financial support

This work was supported by National Natural Science of China (81820108028, 81521004, 81922061, 81703295). The TRILC-ILCCO OncoArray and phenotype data was supported by NIH U19CA203654 and U19CA148127, and RR170048 Cancer Prevention and Research Institute of Texas.

Disclosure of Potential Conflicts of Interest

The authors declare no potential conflicts of interest.

Abstract

Background: A substantial proportion of cancer driver genes (CDGs) are also cancer predisposition genes. However, the associations between genetic variants in lung CDGs and the susceptibility to lung cancer have rarely been investigated.

Methods: We selected expression-related single nucleotide polymorphisms (eSNPs) and nonsynonymous variants of lung CDGs, and tested their associations with lung cancer risk in two large-scale genome-wide association studies (20,871 cases and 15,971 controls of European descent). Conditional and joint association analysis was performed to identify independent risk variants. The associations of independent risk variants with somatic alterations in lung CDGs or recurrently altered pathways were investigated using data from The Cancer Genome Atlas (TCGA) project.

Results: We identified seven independent SNPs in five lung CDGs that were consistently associated with lung cancer risk in discovery ($P < 0.001$) and validation ($P < 0.05$) stages. Among these loci, rs78062588 in *TPM3* (1q21.3) was a new lung cancer susceptibility locus (OR = 0.86, $P = 1.65\times10^{-6}$). Subgroup analysis by histological types further identified nine lung CDGs. Analysis of somatic alterations found that, in lung adenocarcinomas, rs78062588[C] allele (*TPM3* in 1q21.3) was associated with elevated somatic copy number of *TPM3* (OR = 1.16, $P = 0.02$). In lung adenocarcinomas, rs1611182 (*HLA-A* in 6p22.1) was associated with truncation mutations of the transcriptional misregulation in cancer pathway (OR = 0.66, $P = 1.76\times10^{-3}$).

Conclusions: Genetic variants can regulate functions of lung CDGs and influence lung cancer susceptibility.

Impact: Our findings might help unravel biological mechanisms underlying lung cancer susceptibility.

Introduction

Lung cancer has been one of the most commonly diagnosed malignancies and the leading cause of cancer death worldwide (1). The development of lung cancer is a multi-step process that involves both genetic and environmental factors (2-4). Genome wide association studies (GWASs) have been proven to be a powerful approach to dissect genetic architectures of complex diseases. To date, GWASs have identified 51 lung cancer susceptibility loci in various populations (5,6). However, the information provided by GWAS remains inadequate. The heritability of lung cancer was estimated to be 20.6% in European populations (7), while only a small proportion of lung cancer heritability could be explained by risk loci that were identified in previous lung cancer GWASs (8). Therefore, more risk loci for lung cancer are needed to be identified.

Several waves of technology have facilitated the identification of lung cancer driver genes (lung CDGs), which are improving our understanding of oncogenic process for lung cancer. Based on The Cancer Genome Atlas (TCGA) research on lung cancer, the most commonly mutated oncogenes in lung adenocarcinoma (lung ADC) included *KRAS*, *EGFR*, *BRAF*, *PIK3CA*, and *MET*; mutations in tumor suppressors such as *TP53*, *STK11*, *KEAP1*, *NF1*, *RB1*, and *CDKN2A* were also frequently detected in lung ADC (9-11). Although *TP53*, *RB1*, *ARID1A*, *CDKN2A*, *PIK3CA*, and *NF1* were significantly mutated in both lung ADC and lung squamous cell carcinoma (lung SqCC), significantly mutated genes like *NOTCH1* and *HRAS* were only identified in lung SqCC (10-12). In addition to somatic mutations, somatic copy number alterations (SCNAs) and rearrangements also play important roles in lung cancer development. Amplification of *TERT* and *EGFR*, as well as fusions involving *ALK* and *ROS1* were commonly identified in lung ADC. Deletions of

CDKN2A have been identified in both lung ADC and SqCC (9,10,12).

Emerging evidence has shown that a substantial proportion of cancer driver genes are also cancer predisposition genes (13). The TCGA PanCanAtlas Germline Working Group identified 44 genes that showed co-clustering or co-localization of pathogenic germline variants with recurrent somatic mutations, implying shared oncogenic processes in germline and somatic genomes (14). In addition, susceptibility variants could regulate the functions of nearby cancer driver genes. For example, rs2736100, a risk variant of lung cancer, is located in the first intron of driver gene *TERT*, and was associated with increased expression of *TERT* in lung tumors (15). However, the associations between common genetic variants in lung CDGs and lung cancer risk have rarely been explored. Therefore, we integrated lung CDGs, genetics of gene expression, and functional annotation databases with large-scale lung cancer GWAS datasets to systematically investigate the associations between lung CDG-related genetic variants and lung cancer risk.

Materials and methods

GWAS datasets

The present study utilized data from two existing GWASs of European descent: the OncoArray dataset (16) and Division of Cancer Epidemiology and Genetics (DCEG) Lung Cancer Study (17). The OncoArray dataset was derived from the Transdisciplinary Research of Cancer in Lung of the International Lung Cancer Consortium (TRICL-ILCCO) and the Lung Cancer Cohort Consortium (LC3). Quality control and imputation processes were described previously (16), resulting in 18,444 cases and 14,027 controls remained. The DCEG Lung Cancer GWAS data were obtained from dbGap phs000336.v1.p1 (17). Detailed quality control and imputation processes have been described previously (18). We further excluded

individuals in the DCEG Lung Cancer Study that overlapped with or were related to individuals from the OncoArray dataset based on identity by descent (IBD) analysis (IBD > 0.45). As a result, a total of 2,427 cases and 1,944 controls from the DCEG Lung Cancer Study remained. All participants signed informed consents and study protocols were approved by the ethical review boards of each institution.

Selection of lung CDG-related genetic variants

Genes were annotated as lung CDGs if they fulfilled any of the following criteria: (1) lung cancer related genes in the COSMIC Cancer Gene Census (v78) (19); (2) mutational-drivers, somatic copy number alteration (SCNA)-drivers, and fusion-drivers detected by the IntOGen pipeline in lung tumors (20); (3) significantly mutated genes (SMGs) and candidate cancer driver genes with significant SCNAs that were identified in lung ADC and/or lung SqCC by the TCGA projects (10).

To investigate functional variants in lung CDGs, we included single nucleotide polymorphisms (SNPs) if they satisfied either of the following criteria: (1) SNPs that were associated with expressions of lung CDGs (expression-related SNPs, or eSNPs) in normal lung tissues based on the Genotype-Tissue Expression Project (GTEx, v6p release) ($P < 0.05$) (21); (2) nonsynonymous variants of lung CDGs identified using Variant Effect Predictor (22). The selected eSNPs and nonsynonymous variants were extracted from the two GWAS datasets. SNP with imputation INFO < 0.8, minor allele frequency (MAF) in controls <0.005, Hardy-Weinberg equilibrium test *P* in controls $< 1\times10^{-7}$, or HWE test *P* in cases $< 1\times10^{-12}$ was excluded from the analysis.

Statistical analyses

Association analysis

We performed logistic regression to generate odds ratios and confidence intervals (CIs) for each SNP. The OncoArray dataset was used in the discovery stage with age,

gender, and the first three principal components (PCs) adjusted (16). Variations with association $P < 0.001$ were further tested in the DCEG Lung Cancer Study (the validation stage), and we adjusted age, gender, and the first PC in logistic regression model (23). SNPTEST v2.5 was used for the association analysis, taking dosage format of imputed genotypes. For variations with $P < 0.05$ in the validation stage, meta-analysis that combined effect estimates from the two datasets was performed using GWAMA v2.0.2 (24). The index of heterogeneity (I^2) and *P* value based on Cochran's Q test were calculated to assess the heterogeneity between studies. Fixed-effect model was used for absent of heterogeneity between studies (*P* value for heterogeneity > 0.05); otherwise random-effect model was adopted. Variations with the same direction of effect in both GWAS datasets and $P < 1\times10^{-5}$ in the meta-analysis were considered as suggestive risk SNPs (**Supplementary Fig. S1**).

In addition to the overall lung cancer, we also investigated the associations of lung CDG-related SNPs with risk of lung ADC and lung SqCC. As the DCEG Lung Cancer Study lacked information of histological types, we performed association analysis using logistic regression model in the OncoArray dataset. To control the false discovery rate (FDR), we used Benjamini-Hochberg step-down method to calculate FDR for each variation. Variations with FDR < 0.01 were considered as suggestive risk SNPs.

We mapped suggestive risk SNPs to lung CDGs based on the GTEx v6p release, and performed functional prediction for significant nonsynonymous variants using SIFT (25) and PolyPhen2 (26), which were implemented in ANNOVAR (27). For lung CDGs with multiple risk SNPs, conditional and joint association analysis were performed to identify independent signals using genome-wide complex trait analysis (GCTA) (28). During the model selection process, the testing SNP was not selected if

its regression R^2 on the selected SNPs was greater than 0.1. The threshold *P*-value of 0.0001 was adopted to identify significant independent hits. SNPs that were significant after the multiple testing correction and that were not in linkage disequilibrium (LD, $r^2 < 0.1$) with and were located at least 500 kilobases apart from known risk variants were considered as novel susceptibility SNPs.

Co-expression and pathway enrichment analysis

Expression data on 56,238 genes for 320 normal lung tissues were downloaded from the GTEx website (21). Genome-wide expression correlation analysis was performed using a linear regression model to identify genes co-expressed with significant lung CDGs. Significant co-expressed genes that satisfied the Bonferroni correction (0.05/(56318 × 17 significant genes)) were used for pathway enrichment analysis. We downloaded pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database from MSigDB (29-31), and performed pathway enrichment analysis using "PHYPER" function as implemented in R software (version 3.4.1), which computes a p-value for each pathway based on hypergeometric distribution.

Associations between independent SNPs and somatic alterations

TCGA datasets of lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) were used to model the association between independent SNPs and somatic alterations in lung CDGs (9,12). Germline genotype data generated using Affymetrix Genome-Wide Human SNP Array 6.0 were applied for and approved in Feb, 2015. Standard quality control and genotype imputation process have been described previously (32).

We downloaded Mutation Annotation Format derived from whole-exome sequencing, as well as somatic copy numbers calculated using GISTIC2 from the

Broad Institute Genome Data Analysis Center (GDAC) Firehose portal (stamp analyses_2016_01_28) (33). For each patient, a lung CDG was considered mutated if one or more somatic mutations mapped to this gene. We also assessed truncation mutations (frame shift insertion/deletion, nonsense, nonstop, and splice site mutations) (34) in pathways that are recurrently altered in lung cancer, including cell cycle, spliceosome, Notch signaling pathway, transcriptional misregulation in cancer, Ras signaling pathway, and PI3K-Akt signaling pathway (11). A pathway was considered as mutated if one or more truncation mutations were observed in this pathway. We used logistic regression models to evaluate the association between independent SNPs and mutational status of lung CDGs or pathways. In the analysis of SCNAs, somatic copy number of lung CDG was used as outcome, and we used linear regression to model the association between independent SNPs and SCNAs. Age, gender, smoking status, clinical stage, and the first ten PCs were adjusted as covariates. The association analysis between independent SNPs and somatic alterations were performed in lung ADC and lung SqCC, separately. Benjamini-Hochberg step-down method was used to calculate FDR for each SNP-lung CDG (or SNP-pathway) pair in order to control the false discovery rate. Association analysis was conducted using the R software (version 3.4.1).

Results

The OncoArray dataset included 18,444 cases and 14,027 controls. The mean (± standard error) age of the subjects was 63.79 ± 10.44 for cases and 61.77 ± 10.29 for controls. For the DCEG Lung Cancer Study, a total of 2,427 cases and 1,944 controls were included. Among participants across both studies with known histological types, there were 6,819 lung ADCs and 4,490 lung SqCCs. Detailed characteristics and clinical features of participants in each data set were shown in **Supplementary Table**

S1.

Genetic variants associated with lung cancer risk

A total of 348 protein-coding lung CDGs were included from published data (**Supplementary Table S2**). We identified 139,666 eSNPs and 2,041 nonsynonymous variants of lung CDGs. Among SNPs that passed the quality control process, a total of 234 SNPs were identified in the OncoArray dataset ($P < 0.001$) and validated in the DCEG Lung Cancer Study ($P < 0.05$), which were mapped to five lung CDGs (**Supplementary Table S3**). After conditional analysis, seven independent signals were identified. Among these loci, rs78062588, which was mapped to *TPM3* in chromosome 1q21.3, was a new lung cancer susceptibility locus (OR = 0.87, 95%CI: 0.81-0.92, $P = 1.55\times10^{-5}$ in the OncoArray dataset; OR = 0.82, 95%CI: 0.68-0.98, $P = 3.11\times10^{-2}$ in the DCEG Lung Cancer Study; and OR = 0.86, 95%CI: 0.81-0.91, $P = 1.65\times10^{-6}$ in the meta-analysis) (**Table 1**, **Table 2**, **Supplementary Table S3**). In addition, rs71658797 in *FUBP1* (1p31.1), rs1655931 and rs2517586 in *HLA-A* (6p22.1), rs2887532 in *KDM5A* (12p13.33), rs7359276 and rs7161774 in *IREB2* (15q25.1) had been reported by previous GWASs as lung cancer susceptibility loci (**Table 1**, **Table 2**, **Supplementary Table S3**) (5,6).

Stratified analyses in lung ADC and lung SqCC found another nine susceptibility genes, including seven genes that were identified in lung ADC and two genes that were identified only in lung SqCC (**Fig. 1A and 1B**, **Supplementary Table S4**). Independent variants derived from conditional analysis are shown in **Supplementary Table S5**. Of these loci, rs2700389 in *KALRN* (3q21.1), rs79518818 in *MGA* (15q15.1), and rs62054832 in *EFTUD2* (17q21.31) were first identified as risk loci for lung ADC, while rs148797791 in *IRF6* (1q32.2) was found as a novel risk locus for lung SqCC. SNPs rs7823498 in *NRG1* (8p12), rs10757256 and rs1011970 in

CDKN2A (9p21.3), rs79040073 in *COPS2* (15q21.1), rs2281925 in *ARFGAP1* (20q13.33), and rs17879961 in *CHEK2* (22q12.1) had been reported by previous GWASs as lung cancer susceptibility loci (5,6).

Functional evaluation for significant SNPs

Among 234 significant SNPs in overall lung cancer, three were nonsynonymous variants. Two additional nonsynonymous variants (rs1136688 in *HLA-A* and rs17879961 in *CHEK2*) were identified in lung SqCC (**Supplementary Table S6**). We predicted functional consequence of nonsynonymous variants using SIFT and Polyphen-2 (25,26). Notably, risk variant rs707910 in *HLA-A* (NM_001242758, c.G203A) was predicted as deleterious by SIFT and possibly damaging by Polyphen-2. SNP rs17879961 in *CHEK2* (NM_007194, c.T470C) was predicted as tolerated by SIFT and possibly damaging by Polyphen-2.

To explore biological processes underlying significant lung CDGs, we performed genome-wide co-expression and KEGG pathway enrichment analysis. We identified essential pathways in lung carcinogenesis such as apoptosis, MAPK signaling pathway, spliceosome, cell cycle, and nucleotide excision repair (**Supplementary Table S7**) (11).

Associations between independent risk SNPs and somatic alterations

We investigated the associations between independent SNPs and somatic alterations in lung CDGs. The protective rs78062588[C] allele (*TPM3* in 1q21.3) was associated with increased expression of *TPM3* in normal lung tissues (OR = 1.14, P = 0.04) and elevated somatic copy number of *TPM3* in TCGA lung adenocarcinomas (OR = 1.16, P = 0.02) (**Supplementary Fig. S2**). However, the analysis of somatic mutations in lung CDGs did not identify any association with $P < 0.05$. As the mutational frequencies of lung CDGs are relatively low, we further analyzed the

associations between independent risk SNPs and truncation mutations at the pathway level. Among patients with lung ADC, we found that rs1611182 (*HLA-A* in 6p22.1), a risk SNP for lung ADCs, was associated with decreased frequency of truncation mutations in the transcriptional misregulation in cancer pathway (OR = 0.66, 95%CI: 0.50-0.85, $P = 1.76\times10^{-3}$, FDR < 0.25) (**Table 3**, **Supplementary Table S8**, **Supplementary Fig. S3**).

Discussion

The present study comprehensively incorporated lung cancer GWASs, lung CDGs, genetics of gene expression, somatic alterations in lung tumors, and functional annotation databases to investigate the associations of cancer driver gene-related genetic variants with lung cancer risk. We identified five lung CDGs in overall lung cancer. Subgroup analysis by histological types further identified seven and two genes in lung ADC and lung SqCC, respectively. Genes co-expressed with the identified lung CDGs were involved in essential pathways including cell cycle, MAPK signaling, and nucleotide excision repair pathways. Incorporation of somatic alterations identified lung cancer risk variants that were associated with somatic alterations in lung CDGs or recurrently mutated pathways.

TPM3 is included in the COSMIC Cancer Gene Census. Translocation of *TPM3* could form oncogenic fusion proteins, such as TPM3-ROS1 observed in advanced lung adenocarcinoma (35). Previously conducted functional assessment in NIH3T3 cells showed that TPM3-ALK fusion protein can interact with endogenous tropomyosin, which may induce changes in cell morphology and cytoskeleton organization and further bestowed higher metastatic capacities (36). Our results found that the protective allele of rs78062588 was associated with increased *TPM3* expression as well as increased somatic copy number alterations of *TPM3* in lung

adenocarcinomas. However, reaching a better understanding of the functional impact of *TPM3* on lung cancer development warrants further investigation.

CDKN2A in 9p21.3 encodes several alternatively spliced transcripts, among which are p16 and ARF. p16 is a tumor suppressor that functions as an inhibitor of CDK4 and CDK6 (37). Another tumor suppressor protein, ARF, functions as a stabilizer of the tumor suppressor protein p53. Both p16 and ARF have functionality in cell cycle G1 control. *CDKN2A* is recognized as an important tumor suppressor gene. Deletion of *CDKN2A* was frequently identified in lung tumors (10). In addition, *CDKN2A* has been identified as susceptibility gene for lung adenocarcinoma (16). We validated this locus and identified a second signal within *CDKN2A*. Consistently, the risk alleles of independent SNPs were associated with decreased expression of *CDKN2A* in normal lung tissues.

The transcription factor interferon regulatory factor 6 (*IRF6*) was identified as significantly mutated gene in TCGA lung squamous cell carcinomas (10). IRF6 has essential role in epidermal development. It is induced in differentiation through a Notch-dependent mechanism. Down-regulation of *IRF6* in epithelial squamous cell carcinomas promotes ras-induced tumor formation and reintroduction of *IRF6* strongly inhibits cell growth (38,39). The tumor suppressor role of *IRF6* has also been demonstrated in vulvar squamous cell carcinoma (40). In addition, elevated *IRF6* expression in nasopharyngeal carcinomas suppressed cell proliferation and growth (41). We identified *IRF6* as a susceptibility gene for lung squamous cell carcinoma. Consistent with the tumor suppressor role of *IRF6*, the risk allele of rs148797791 was associated with decreased expression of *IRF6* in normal lung tissues. These results indicate that germline variant might contribute to lung cancer risk by down-regulation of *IRF6*.

Genes co-expressed with the identified lung CDGs were enriched in essential pathways such as apoptosis, MAPK signaling pathway, spliceosome, cell cycle, and nucleotide excision repair. A comprehensive molecular profiling of lung ADC demonstrated recurrent somatic alterations in cell cycle and MAPK signaling pathway (9,42). In addition, deregulated RNA Splicing is involved in lung ADC, and cell cycle pathway is involved in both lung ADC and lung SqCC (11,42).

We comprehensively collected 348 lung CDGs from three databases, and tested associations between functional SNPs of lung CDGs and risk of lung cancer in large-scale lung cancer GWASs of Europeans. We identified five novel susceptibility loci of lung cancer, and validated nine loci that had been reported by previous lung cancer GWASs. These results showed that genetic variants in lung CDGs contribute to lung cancer susceptibility. Our findings might help to unravel biological functions of lung cancer susceptibility loci.

Authors' Contributions

Conception and design: Y.Z. Wang, O.Y. Gorlova, I.P. Gorlov, C.I. Amos, H.B. Shen

Development of methodology: Y.Z. Wang, M. Zhu, J.C. Dai

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): D. Albanes, S. Lam, A. Tardon, C. Chen, G. Goodman, S.E. Bojesen, M.T. Landi, M. Johansson, A. Risch, H.E. Wichmann, H. Bickeboller, D.C. Christiani, G. Rennert, S. Arnold, P. Brennan, J.K. Field, S. Shete, L.L. Marchand, O. Melander, H. Brunnstrom, G. Liu, R.J. Hung, A. Andrew, L.A. Kiemeney, S. Zienolddiny, K. Grankvist, M. Johansson, N. Caporaso, P. Woll, P. Lazarus, M.B. Schabath, M.C. Aldrich, V.L. Stevens, C.I. Amos, H.B. Shen

Analysis and interpretation of data (e.g., statistical analysis, biostatistics,

computational analysis): Y.Z. Wang, M. Zhu, J.C. Dai

Writing, review, and/or revision of the manuscript: Y.Z. Wang, O.Y. Gorlova, I.P. Gorlov, M. Zhu, J.C. Dai, R.J. Hung, H. Brunnstrom, K. Grankvist, M. Johansson, D.C. Christiani, H.X. Ma, G.F. Jin, Z.B. Hu, C.I. Amos, H.B. Shen

Study supervision: J.C. Dai, C.I. Amos, H.B. Shen

Acknowledgments

We thank the study participants and research staff for their contributions and commitment to this study.

References

1. 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 Cancer J Clin **2018**;68:394-424.
2. Tokuhata GK, Lilienfeld AM. Familial aggregation of lung cancer in humans. J Natl Cancer Inst **1963**;30:289-312.
3. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M*, et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med **2000**;343:78-85.
4. Matakidou A, Eisen T, Houlston RS. Systematic review of the relationship between family history and lung cancer risk. Br J Cancer **2005**;93:825-33.
5. Bosse Y, Amos CI. A Decade of GWAS Results in Lung Cancer. Cancer Epidemiol Biomarkers Prev **2018**;27:363-79.
6. Dai J, Lv J, Zhu M, Wang Y, Qin N, Ma H*, 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. Sampson JN, Wheeler WA, Yeager M, Panagiotou O, Wang Z, Berndt SI*, et al.* Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. J Natl Cancer Inst **2015**;107:djv279.
8. Dai J, Shen W, Wen W, Chang J, Wang T, Chen H*, et al.* Estimation of heritability for nine common cancers using data from genome-wide association studies in Chinese population. Int J Cancer **2017**;140:329-36.
9. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature **2014**;511:543-50.
10. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS*, et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet **2016**;48:607-16.
11. Swanton C, Govindan R. Clinical Implications of Genomic Discoveries in Lung Cancer. N Engl J Med **2016**;374:1864-73.
12. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature **2012**;489:519-25.
13. Rahman N. Realizing the promise of cancer predisposition genes. Nature **2014**;505:302-8.
14. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C*, et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. Cell **2018**;173:355-70 e14.
15. Wei R, Cao L, Pu H, Wang H, Zheng Y, Niu X*, et al.* TERT Polymorphism rs2736100-C Is Associated with EGFR Mutation-Positive Non-Small Cell Lung Cancer. Clin Cancer Res **2015**;21:5173-80.
16. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC*, et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet **2017**;49:1126-32.
17. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M*, et al.* A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet **2009**;85:679-91.
18. Wang Y, Wu W, Zhu M, Wang C, Shen W, Cheng Y*, et al.* Integrating expression-related SNPs

into genome-wide gene- and pathway-based analyses identified novel lung cancer susceptibility genes. Int J Cancer **2018**;142:1602-10.

19. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R*, et al.* A census of human cancer genes. Nat Rev Cancer **2004**;4:177-83.
20. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A*, et al.* IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods **2013**;10:1081-2.
21. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet **2013**;45:580-5.
22. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A*, et al.* The Ensembl Variant Effect Predictor. Genome Biol **2016**;17:122.
23. Dai J, Li Z, Amos CI, Hung RJ, Tardon A, Andrew AS*, et al.* Systematic analyses of regulatory variants in DNase I hypersensitive sites identified two novel lung cancer susceptibility loci. Carcinogenesis **2019**;40:432-40.
24. Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics **2010**;11:288.
25. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res **2012**;40:W452-7.
26. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P*, et al.* A method and server for predicting damaging missense mutations. Nat Methods **2010**;7:248-9.
27. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res **2010**;38:e164.
28. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet **2011**;88:76-82.
29. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res **1999**;27:29-34.
30. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A **2005**;102:15545-50.
31. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst **2015**;1:417-25.
32. Wang Y, Wang C, Zhang J, Zhu M, Zhang X, Li Z*, et al.* Interaction analysis between germline susceptibility loci and somatic alterations in lung cancer. Int J Cancer **2018**;143:878-85.
33. Marx V. Drilling into big cancer-genome data. Nat Methods **2013**;10:293-7.
34. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MD, Wendl MC*, et al.* Integrated analysis of germline and somatic variants in ovarian cancer. Nat Commun **2014**;5:3156.
35. Zhu YC, Liao XH, Wang WX, Xu CW, Zhuang W, Wei JG*, et al.* Dual drive coexistence of EML4-ALK and TPM3-ROS1 fusion in advanced lung adenocarcinoma. Thorac Cancer **2018**;9:324-7.
36. Armstrong F, Lamant L, Hieblot C, Delsol G, Touriol C. TPM3-ALK expression induces changes in cytoskeleton organisation and confers higher metastatic capacities than other ALK fusion proteins. Eur J Cancer **2007**;43:640-6.
37. Ohtani N, Yamakoshi K, Takahashi A, Hara E. The p16INK4a-RB pathway: molecular link between cellular senescence and tumor suppression. J Med Invest **2004**;51:146-53.
38. Botti E, Spallone G, Moretti F, Marinari B, Pinetti V, Galanti S*, et al.* Developmental factor IRF6

exhibits tumor suppressor activity in squamous cell carcinomas. Proc Natl Acad Sci U S A **2011**;108:13710-5.

39. Restivo G, Nguyen BC, Dziunycz P, Ristorcelli E, Ryan RJ, Ozuysal OY*, et al.* IRF6 is a mediator of Notch pro-differentiation and tumour suppressive function in keratinocytes. EMBO J **2011**;30:4571-85.
40. Rotondo JC, Borghi A, Selvatici R, Magri E, Bianchini E, Montinari E*, et al.* Hypermethylation-Induced Inactivation of the IRF6 Gene as a Possible Early Event in Progression of Vulvar Squamous Cell Carcinoma Associated With Lichen Sclerosus. JAMA Dermatol **2016**;152:928-33.
41. Xu L, Huang TJ, Hu H, Wang MY, Shi SM, Yang Q*, et al.* The developmental transcription factor IRF6 attenuates ABCG2 gene expression and distinctively reverses stemness phenotype in nasopharyngeal carcinoma. Cancer Lett **2017**
42. Ji Y, Zheng MF, Ye SG, Chen JY, Chen YJ. PTEN and Ki67 expression is associated with clinicopathologic features of non-small cell lung cancer. J Biomed Res. **2014**; 28: 462–467.

Tables

Table 1. The associations between independent variants representing each lung cancer locus and overall lung cancer risk in the OncoArray dataset.

Cytoband [a]	Location (bp) [b]	SNP	Gene	Effect allele	Reference allele	INFO	EAF in case	EAF in control	OR (95%CI)	*P*
1p31.1	77967507	rs71658797	*FUBP1*	A	T	1.00	0.11	0.10	1.14 (1.08-1.20)	1.04E-06
1q21.3	154566225	rs78062588*	*TPM3*	C	T	0.95	0.06	0.07	0.87 (0.81-0.92)	1.55E-05
6p22.1	29897438	rs1655931	*HLA-A*	A	G	0.96	0.17	0.15	1.15 (1.10-1.20)	3.79E-10
6p22.1	30205174	rs2517586	*HLA-A*	T	C	0.99	0.33	0.35	0.92 (0.89-0.95)	8.84E-07
12p13.33	1051495	rs2887532	*KDM5A*	T	C	1.00	0.17	0.18	0.93 (0.89-0.97)	3.90E-04
15q25.1	78892661	rs7359276	*IREB2*	T	C	1.00	0.80	0.76	1.27 (1.22-1.32)	9.74E-35
15q25.1	79069734	rs7161774	*IREB2*	T	G	0.96	0.57	0.60	0.85 (0.82-0.88)	9.39E-23

Abbreviations: EAF: effect allele frequency; OR: odds ratio; 95%CI: 95% confidence interval.

[a] Cytogenetic band;

[b] SNP position, build 37;

* SNPs (or loci) that were first identified as potential lung cancer susceptibility loci in the present study.

Table 2. The associations between independent variants representing each lung cancer locus and overall lung cancer risk in the DCEG Lung Cancer Study.

Cytoband [a]	Location (bp) [b]	SNP	Gene	Effect allele	Reference allele	INFO	EAF in case	EAF in control	OR (95%CI)	*P*
1p31.1	77967507	rs71658797	*FUBP1*	A	T	0.98	0.13	0.11	1.18 (1.04-1.35)	1.22E-02
1q21.3	154566225	rs78062588*	*TPM3*	C	T	0.97	0.05	0.07	0.82 (0.68-0.98)	3.11E-02
6p22.1	29897438	rs1655931	*HLA-A*	A	G	0.97	0.14	0.13	1.15 (1.01-1.30)	3.37E-02
6p22.1	30205174	rs2517586	*HLA-A*	T	C	0.98	0.35	0.37	0.89 (0.82-0.98)	1.34E-02
12p13.33	1051495	rs2887532	*KDM5A*	T	C	1.00	0.20	0.21	0.88 (0.79-0.98)	2.10E-02
15q25.1	78892661	rs7359276	*IREB2*	T	C	1.00	0.78	0.74	1.31 (1.18-1.45)	1.57E-07
15q25.1	79069734	rs7161774	*IREB2*	T	G	0.96	0.63	0.66	0.87 (0.79-0.95)	2.71E-03

Abbreviations: EAF: effect allele frequency; OR: odds ratio; 95%CI: 95% confidence interval.

[a] Cytogenetic band;

[b] SNP position, build 37;

* SNPs (or loci) that were first identified as potential lung cancer susceptibility loci in the present study.

Table 3. Associations between rs1611182 and truncation mutations in the transcriptional misregulation in cancer pathway.

SNP	Allele [a]	Histological types	Cases [b]	Controls [b]	EAF		OR (95%CI) [c]	*P* [c]
					Cases	Controls		
rs1611182	G / T	Lung ADC	30/93/81	71/144/86	0.38	0.48	0.66 (0.50-0.85)	1.76E-03
		Lung SqCC	50/105/74	56/131/66	0.45	0.48	0.91 (0.70-1.18)	4.66E-01

Abbreviations: Lung ADC, lung adenocarcinoma; Lung SqCC, lung squamous cell carcinoma; EAF, effect allele frequency; OR, odds ratio; 95%CI, 95% confidence interval.

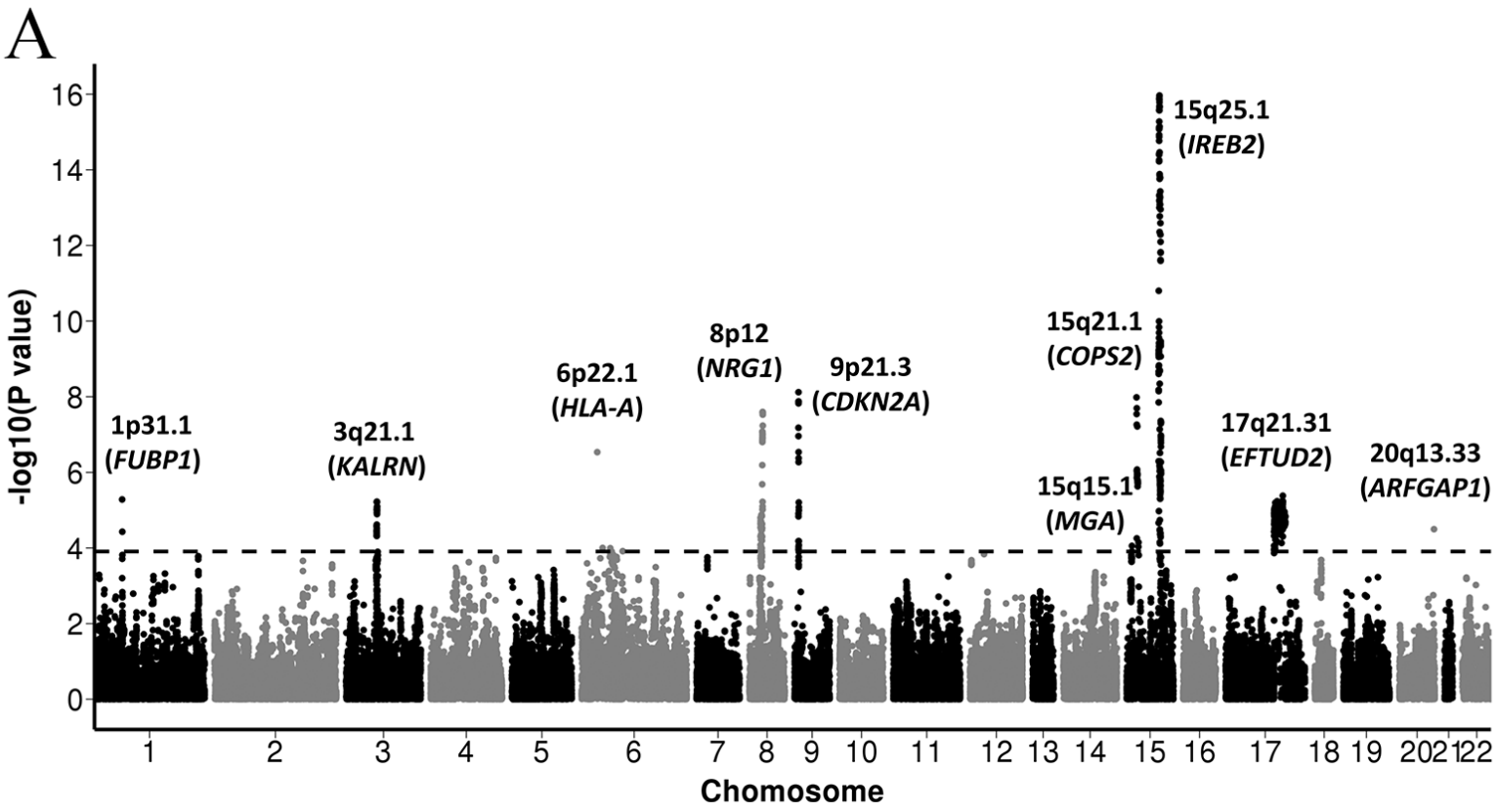
[a] Reference / effect allele;

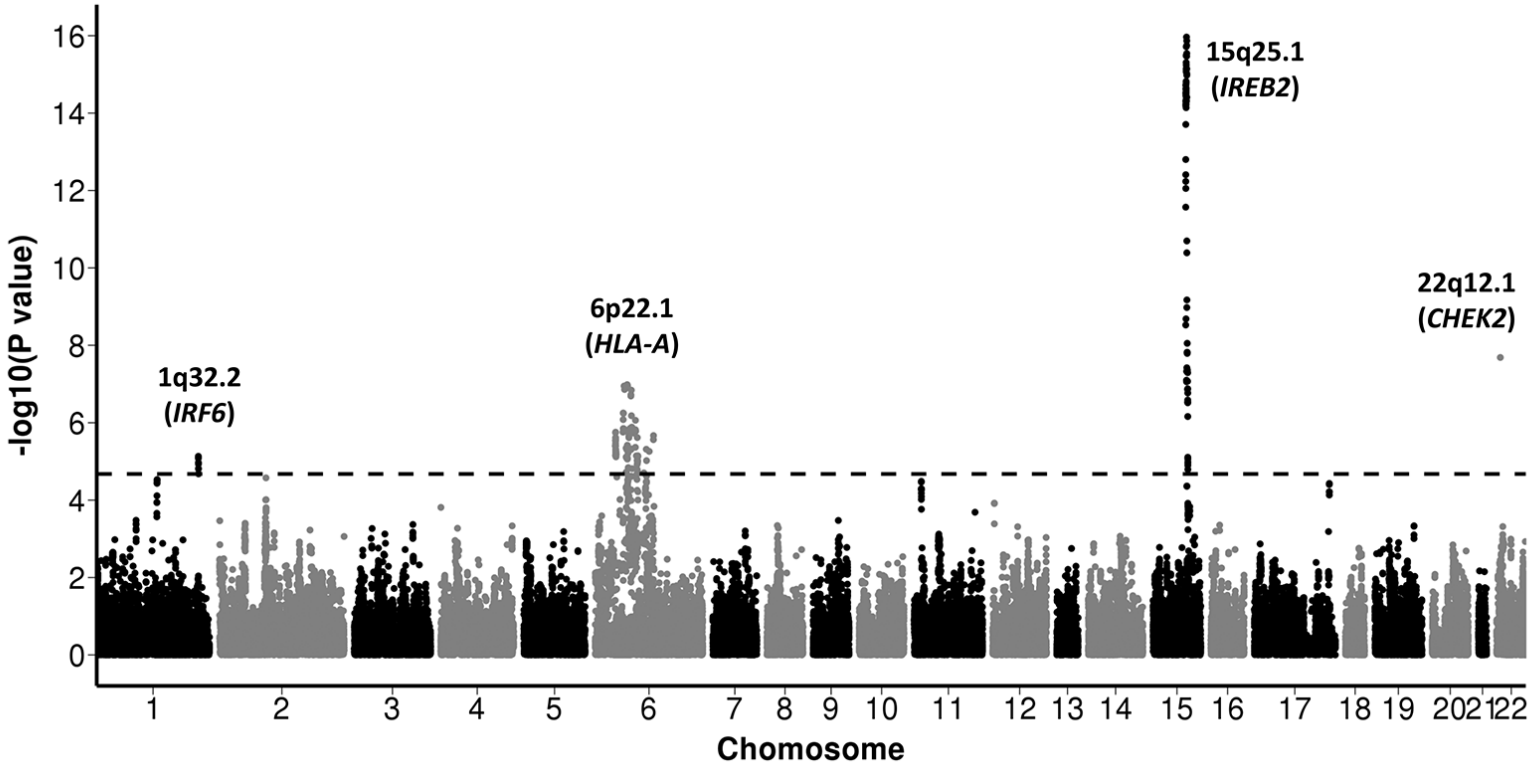
[b] Variant homozygote/heterozygote/wild-type homozygote. Patients with one or more truncation mutations in corresponding pathway were cases. Otherwise the patients were defined as controls.

[c] Adjusted by age, gender, smoking status, clinical stage, and the first ten principals.

Figure legends

Figure 1. Manhattan plot showing $-\log_{10}$(*P* values) for SNP associations with risk of lung adenocarcinoma and squamous cell carcinoma. (A) Lung adenocarcinoma (6,819 cases and 14,027 controls); (B) Lung squamous cell carcinoma (4,490 cases and 14,027 controls). Each locus is annotated by its cytoband location and corresponding lung cancer driver genes. The x axis represents chromosomal location, and the y axis represents −log10 (*P* value). The horizontal line denotes false discovery rate (FDR) < 0.01.


A
-log10(P value)
Chomosome
1p31.1 (FUBP1)
3q21.1 (KALRN)
6p22.1 (HLA-A)
8p12 (NRG1)
9p21.3 (CDKN2A)
15q21.1 (COPS2)
15q15.1 (MGA)
15q25.1 (IREB2)
17q21.31 (EFTUD2)
20q13.33 (ARFGAP1)


B
-log10(P value)
Chomosome
1q32.2 (IRF6)
6p22.1 (HLA-A)
15q25.1 (IREB2)
22q12.1 (CHEK2)