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

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. Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, United States of America.
3. Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, Nanjing Medical University, Nanjing, China.
4. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States of America.
5. Department of Integrative Oncology, British Columbia Cancer Agency, Vancouver, British Columbia, Canada.
6. Faculty of Medicine, University of Oviedo and CIBERESP, Oviedo, Spain.
7. Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America.
8. Public Health Sciences Division, Swedish Cancer Institute, Seattle, Washington, United States of America.
9. Department of Clinical Biochemistry, Copenhagen University Hospital, Copenhagen, Denmark.
10. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
11. National Cancer Institute, Bethesda, Maryland, United States of America.
12. Genetic Epidemiology Group, International Agency for Research on Cancer, Lyon, France.
13. Cancer Center Cluster Salzburg at PLUS, Department of Molecular Biology, University of Salzburg, Salzburg, Austria.
14. Biobank and Tumor Documentation, Thoraxklinik at University Hospital Heidelberg, Heidelberg, Germany.
15. Translational Lung Research Center Heidelberg (TLRC-H), German Center for Lung Research (DZL), Heidelberg, Germany.
16. Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig Maximilians University, Munich, Bavaria, Germany.
17. Helmholtz Zentrum Munchen, German Research Center for Environmental Health (GmbH), Institute of Epidemiology, Neuherberg, Germany.
18. Institute of Medical Statistics and Epidemiology, Technical University Munich, Munich, Germany.
19. Department of Genetic Epidemiology, University Medical Center Goettingen, Goettingen, Germany.
20. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America.
21. Technion Faculty of Medicine, Carmel Medical Center, Haifa, Israel.
22. Markey Cancer Center, University of Kentucky, Lexington, Kentucky, United States of America.
23. Molecular and Clinical Cancer Medicine, Roy Castle Lung Cancer Research Programme, The University of Liverpool Institute of Translational Medicine, Liverpool, United Kingdom.
24. Department of Epidemiology, The University of Texas, MD Anderson Cancer Center, Houston, Texas, United States of America.
25. Epidemiology Program, University of Hawai'i Cancer Center, Honolulu, Hawai'i, United States of America.
26. Clinical Sciences, Lund University, Lund, Sweden.
27. Department of Internal Medicine, Skåne University Hospital, Malmö, Sweden.
28. Epidemiology Division, Princess Margaret Cancer Centre, Toronto, Ontario, Canada.
29. Epidemiology Division, Lunenfeld-Tanenbuaum Research Institute, Sinai Health System, Toronto, Ontario, Canada.
30. Department of Neurology, Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire, United States of America.
31. Department of Health Evidence, Radboud University Medical Center, Nijmegen, Germany.
32. National Institute of Occupational Health (STAMI), Oslo, Norway.
33. Unit of Clinical Chemistry, Department of Medical Biosciences, Umeå University, Umea, Sweden.
34. Department of Radiation Sciences, Umeå University, Umea, Sweden.
35. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, United States of America.
36. Academic Unit of Clinical Oncology, University of Sheffield, Sheffield, United Kingdom.
37. College of Pharmacy, Washington State University, Spokane, Washington, United States of America.
38. Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, United States of America.
39. Department of Thoracic Surgery, Division of Epidemiology, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America.
40. Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, United States of America.
41. Department of Medicine, Epidemiology Section, Institute for Clinical and Translational Research, Baylor Medical College, Houston, Texas, United States of America

† These authors contributed equally to this work.

**\* 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](mailto: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.

**Abstract**

Emerging evidence has shown that a substantial proportion of cancer driver genes are also cancer predisposition genes. However, the associations between genetic variants in lung cancer driver genes (lung CDGs) and the susceptibility to lung cancer have rarely been investigated. We selected 141,631 potentially functional variants that are mapped to lung CDGs, and tested their associations with lung cancer risk in large-scale case-control studies (20,871 cases and 15,971 controls of European descent). Associations between independent risk variants and somatic alterations in lung CDGs or recurrently mutated pathways (cell cycle, spliceosome, Notch signaling pathway, transcriptional misregulation, Ras signaling pathway, and PI3K-Akt signaling pathway) were analyzed using data from the Cancer Genome Atlas (TCGA) project. We identified 17 independent susceptibility loci achieving statistical significance, either in the overall or histology-stratified analyses. Seven of the 17 loci were novel loci, with *TPM3* (1q21.3), *ANK3* (10q21.2), and *ARHGEF12* (11q23.3) associated with overall lung cancer, *KALRN* (3q21.1), *MGA* (15q15.1), and *EFTUD2* (17q21.31) associated with lung adenocarcinoma (lung ADC), and *IRF6* (1q32.2) associated with lung squamous cell carcinoma (lung SqCC). We identified four independent risk variants that were associated with somatic alterations in lung CDGs or pathways. SNP rs78062588 (*TPM3* in 1q21.3) was associated with elevated somatic copy number of *TPM3* in TCGA lung ADCs (OR = 1.16, *P* = 0.02). SNP rs6090387 (*ARFGAP1* in 20q13.33) was associated with increased frequency of truncation mutations in the Ras signaling pathway (OR = 1.51, *P* = 5.74×10-4 in the meta-analysis that combined association results from histological subtypes). In lung ADCs, rs1611182 (*HLA-A* in 6p22.1) was associated with decreased frequency of truncation mutations in the transcriptional misregulation pathway (OR = 0.66, *P* = 1.76×10-3). In lung SqCCs, rs145433737 (*ARHGEF12* in 11q23.3) was associated with increased frequency of truncation mutations in the Notch signaling pathway (OR = 9.19, *P* = 1.22×10-3). Our findings might help to 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](#_ENREF_1" \o "Bray, 2018 #51). The development of lung cancer is a multi-step process that involves both genetic and environmental factors [2-4](#_ENREF_2" \o "Lichtenstein, 2000 #52). Genome wide association studies (GWAS) have been proven to be a powerful approach to dissect genetic architectures of complex diseases. To date, GWASs have identified 45 lung cancer susceptibility loci in various populations [5](#_ENREF_5" \o "Bosse, 2018 #4). However, the information provided by GWAS remains inadequate. The heritability of lung cancer was estimated to be 20.6% in European populations [6](#_ENREF_6), while only a small proportion of lung cancer heritability could be explained by risk loci identified in previous lung cancer GWAS [7](#_ENREF_7). Therefore, more risk loci for lung cancer are warranted 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 frequently detected in lung ADC [8-10](#_ENREF_8" \o ", 2014 #26). 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 [9-11](#_ENREF_9). 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 [8](#_ENREF_8), [10](#_ENREF_10), [11](#_ENREF_11).

Emerging evidence has shown that a substantial proportion of cancer driver genes are also cancer predisposition genes [12](#_ENREF_12). Recently, 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 [13](#_ENREF_13" \o "Huang, 2018 #50). 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 [14](#_ENREF_14" \o "Wei, 2015 #12). 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 of cancer driver gene-related genetic variants with lung cancer risk. A flow chart of study design is shown in **Figure S1**.

**Materials and methods**

***GWAS datasets***

The present study utilized data from two existing GWASs of European descent: the OncoArray dataset [15](#_ENREF_15) and Division of Cancer Epidemiology and Genetics (DCEG) Lung Cancer Study [16](#_ENREF_16" \o "Landi, 2009 #14). 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 [15](#_ENREF_15" \o "McKay, 2017 #13), resulting in 18,444 cases and 14,027 remained. The DCEG Lung Cancer GWAS data were obtained from dbGap phs000336.v1.p1 [16](#_ENREF_16" \o "Landi, 2009 #14). Detailed quality control and imputation processes have been described previously [17](#_ENREF_17" \o "Wang, 2018 #59). We further excluded individuals in the DCEG Lung Cancer GWAS that overlapped with or were related to individuals from the OncoArray dataset based on IBD analysis (identity by descent, IBD > 0.45). As a result, a total of 2,427 cases and 1,944 controls from the DCEG Lung Cancer GWAS remained. All participants signed informed consents and study protocols were approved by the ethical review boards of each institution.

***Selection of genetic variants within lung cancer driver genes***

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) [18](#_ENREF_18); (2) mutational-drivers, somatic copy number alteration (SCNA)-drivers, and fusion-drivers detected by the IntOGen pipeline in lung tumors [19](#_ENREF_19); (3) significantly mutated genes (SMGs) and candidate cancer driver genes with significant SCNAs identified by the TCGA LUAD and/or LUSC projects [10](#_ENREF_10).

To investigate functional variants in lung CDGs, we included 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 (GTEx) database (v6p release) (*P* < 0.05) [20](#_ENREF_20); (2) nonsynonymous variants of lung CDGs identified using Variant Effect Predictor [21](#_ENREF_21). The selected eSNPs and nonsynonymous variants were extracted from the two GWAS datasets. SNPs with imputation INFO < 0.8, MAF in controls <0.005, HWE test *P* in controls < 1×10-7, or HWE test *P* in cases < 1×10-12 were excluded from the analysis.

***Association analysis***

We performed logistic regression to generate odds ratios (ORs) and confidence intervals (CIs) for each SNP with adjustment of age, gender and principal components (PCs). In the OncoArray dataset, the first three PCs were adjusted [15](#_ENREF_15" \o "McKay, 2017 #13). In the DCEG Lung Cancer GWAS, we adjusted the first PC in the logistic regression model [22](#_ENREF_22" \o "Dai, 2019 #64). SNPTEST v2.5 was used for the association analysis, taking dosage format of imputed genotypes. Then meta-analysis of the two datasets was performed using GWAMA v2.0.2 [23](#_ENREF_23). The index of heterogeneity (I2) 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. To control the false discovery rate (FDR), we used the Benjamini-Hochberg step-down method to generate an FDR-adjusted p-value (FDR) for each variation. Variations with FDR < 0.01 were considered to be significantly associated with susceptibility to lung cancer. We mapped significant eSNPs to lung CDGs based on the GTEx v6p database, and performed functional prediction for significant nonsynonymous variants using SIFT and PolyPhen2 implemented in ANNOVAR [24](#_ENREF_24).

For lung CDGs with multiple significant SNPs, conditional and joint association analysis were performed to identify independent signals using genome-wide complex trait analysis (GCTA) [25](#_ENREF_25" \o "Yang, 2011 #23). During the model selection process, the testing SNP was not selected if its regression R2 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, r2 < 0.1) with and were located at least 500 kilobases apart from known risk variants were considered as novel susceptibility SNPs.

***Co-expression and KEGG enrichment analysis***

Expression data on 56,238 genes for 320 normal lung tissues were downloaded from the GTEx website (v6p release) [20](#_ENREF_20). 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 KEGG enrichment analysis, which is implemented in R package ‘clusterProfiler’ [26](#_ENREF_26).

***Associations between independent SNPs and somatic alterations***

Data from TCGA LUAD and LUSC projects were used to model the association between independent SNPs and somatic alterations in lung CDGs [8](#_ENREF_8), [11](#_ENREF_11). 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 [27](#_ENREF_27).

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) [28](#_ENREF_28). 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) [29](#_ENREF_29) in pathways that are recurrently altered in lung cancer, including cell cycle, spliceosome, Notch signaling pathway, transcriptional misregulation, Ras signaling pathway, and PI3K-Akt signaling pathway [9](#_ENREF_9). 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 principal components were adjusted as covariates. The association analysis between independent SNPs and somatic alterations were performed in lung ADC and lung SqCC, separately. We then combined association results of the two histological subtypes using meta-analysis. Cochran’s Q test was used to assess the heterogeneity between lung ADC and lung SqCC. We used fixed-effect meta-analysis for absent of heterogeneity (*P* value for heterogeneity > 0.05); otherwise random-effect meta-analysis was adopted. Benjamini-Hochberg step-down method was used to generate an FDR-adjusted p-value (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), and the meta-analysis were performed using R package “meta”.

**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 **Table S1**.

***Genetic variants associated with lung cancer risk***

A total of 348 protein-coding lung CDGs were included from published data (**Table S2**). We identified 139,666 eSNPs and 2,041 nonsynonymous variants of lung CDGs. Among 121,861 SNPs that passed the quality control process, 705 SNPs were found significantly associated with overall lung cancer risk (FDR<0.01), which were mapped to 10 driver genes (**Figure 1**, **Table S3**). After conditional analysis, 12 independent signals were identified. Among these loci, rs78062588 in *TPM3* (1q21.3), rs4948409 in *ANK3* (10q21.2), and rs145433737 in *ARHGEF12* (11q23.3) were first identified as lung cancer susceptibility SNPs (**Table 1**, **Table S4**). In addition, rs71658797 in *FUBP1* (1p31.1), rs1655931 and rs2517586 in *HLA-A* (6p22.1), rs2887532 in *KDM5A* (12p13.33), rs79040073 in *COPS2* (15q21.1), rs7359276 and rs7161774 in *IREB2* (15q25.1), and rs6090387 in *ARFGAP1* (20q13.33) had been reported as lung cancer susceptibility loci [15](#_ENREF_15) (**Table 1**, **Table S4**). And rs3769823 in *CASP8* (2q33.1) was identified as a risk locus by a recent GWAS of non-small cell lung cancer [30](#_ENREF_30).

Stratified analyses in lung ADC and lung SqCC found another seven susceptibility genes, including five genes that were identified in lung ADC and two genes that were significant only in lung SqCC (**Figure 1**, **Table S3**). Independent variants derived from conditional analysis are shown in **Table 2**. 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.

***Functional evaluation for significant SNPs***

Among 705 significant SNPs in overall lung cancer, 17 were nonsynonymous variants. An additional nonsynonymous variant in *CHEK2* was identified in lung SqCC (**Table S5**). We predicted functional consequence of nonsynonymous variants using SIFT and Polyphen-2 [31](#_ENREF_31), [32](#_ENREF_32). Notably, risk variant rs3179982 in *HLA-A* (NM\_001242758, c.C952T) was predicted as deleterious by SIFT and probably damaging by Polyphen-2. SNP rs17879961 in *CHEK2* (NM\_007194, c.T470C) was predicted as tolerated by SIFT and possibly damaging by Polyphen-2. SNP rs3769823 in *CASP8* (NM\_001080125, c.A41G) was predicted as tolerated by SIFT and benign by Polyphen-2, but the G allele of rs3769823 was associated with higher expression of *CASP8* in normal lung tissues (effect on expression = 0.43 and *P* = 1.68×10-22 in the GTEx v6p release, **Table 1**).

To explore biological processes underlying significant lung CDGs, we performed genome-wide co-expression and KEGG enrichment analysis. Results showed that genes co-expressed with identified lung CDGs were enriched in pathways such as cell cycle, MAPK signaling pathway, spliceosome, p53 signaling pathway, and nucleotide excision repair (**Table S6**). Some of these pathways play key roles in lung carcinogenesis [9](#_ENREF_9" \o "Swanton, 2016 #7).

***Associations between germline risk SNPs and somatic alterations***

We investigated the associations between independent SNPs and somatic alterations in lung CDGs. SNP rs78062588 (*TPM3* in 1q21.3) was associated with increased expression of *TPM3* in normal lung tissues (*P* = 0.04) and elevated somatic copy number of *TPM3* in TCGA lung adenocarcinomas (*P* = 0.02) (**Figure 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 SNPs and truncation mutations at the pathway level. We identified three associations with FDR < 0.2 (**Table S7**). In the meta-analysis that combined association results from lung ADC and lung SqCC, the risk allele (G) of rs6090387 (*ARFGAP1* in 20q13.33) was associated with increased frequency of truncation mutations in the Ras signaling pathway (OR = 1.51, 95%CI: 1.19-1.90, *P* = 5.74×10-4) (**Figure S3**, **Table S7**). In lung ADCs, the risk allele (T) of rs1611182 (*HLA-A* in 6p22.1) was associated with decreased frequency of truncation mutations in the transcriptional misregulation pathway (OR = 0.66, 95%CI: 0.50-0.85, *P* = 1.76×10-3) (**Figure S3**, **Table S7**). In lung SqCCs, the risk allele (A) of rs145433737 (*ARHGEF12* in 11q23.3) was associated with increased frequency of truncation mutations in the Notch signaling pathway (OR = 9.19, 95%CI: 2.40-35.27, *P* = 1.22×10-3) (**Figure S3**, **Table S7**).

**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 17 significant lung CDGs in overall lung cancer, lung ADC, and lung SqCC. Of these lung CDGs, seven were identified as susceptibility genes of lung cancer for the first time. Genes co-expressed with significant lung CDGs were involved in essential pathways including cell cycle, MAPK signaling, and nucleotide excision repair pathways. Incorporation of somatic alterations identified four independent 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 [33](#_ENREF_33" \o "Zhu, 2018 #31). 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 [34](#_ENREF_34" \o "Armstrong, 2007 #32). 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.

The IntOGen pipeline identified *CASP8* as a mutational-driver with mutational frequencies of 1.15% in lung SqCC and 0.77% in lung ADC [19](#_ENREF_19). *CASP8* encodes a member of the cysteine-aspartic acid protease (caspase) family. Sequential activation of caspases plays a central role in the execution-phase of cell apoptosis. The role of *CASP8* in carcinogenesis has been extensively investigated. Somatic mutations in *CASP8* could lead to decreased cell death and had been detected in colorectal carcinomas, hepatocellular carcinomas, and gastric cancers [35](#_ENREF_35). In addition, loss of caspase-8 renders neuroblastomas resistant to apoptosis and potentiates neuroblastoma metastasis [36](#_ENREF_36), [37](#_ENREF_37). Consistently, the protective allele (G) of rs3769823 was associated with increased *CASP8* expression in normal lung tissues. Population studies showed that a variant in medium LD with this SNP (rs700635 [C], r2 = 0.74, corresponding to risk allele A of rs3769823) was associated with increased risk of basal-cell carcinoma [38](#_ENREF_38" \o "Stacey, 2015 #56). SNP rs700635 tags an SVA-E retrotransposon that is inserted into intron 8 of *CASP8* and was associated with transcriptional and splicing anomalies involving *CASP8* [38](#_ENREF_38), [39](#_ENREF_39). We speculate that risk variant rs3769823 might modulate lung cancer risk by regulating *CASP8* transcription.

*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 [40](#_ENREF_40). 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](#_ENREF_10" \o "Campbell, 2016 #8). In addition, *CDKN2A* has been identified as susceptibility gene for lung adenocarcinoma [15](#_ENREF_15" \o "McKay, 2017 #13). 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](#_ENREF_10" \o "Campbell, 2016 #8). 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 [41](#_ENREF_41), [42](#_ENREF_42). The tumor suppressor role of *IRF6* has also been demonstrated in vulvar squamous cell carcinoma [43](#_ENREF_43" \o "Rotondo, 2016 #40). In addition, elevated *IRF6* expression in nasopharyngeal carcinomas suppressed cell proliferation and growth [44](#_ENREF_44" \o "Xu, 2017 #41). We identified *IRF6* as lung cancer susceptibility gene. 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*.

To uncover biological mechanisms underlying significant lung CDGs, we performed KEGG enrichment analysis, and found that genes co-expressed with significant lung CDGs were enriched in important pathways such as cell cycle, MAPK signaling pathway, spliceosome, mTOR signaling pathway, p53 signaling pathway, and nucleotide excision repair. A comprehensive molecular profiling of lung ADC demonstrated recurrent somatic alterations in cell cycle, mTOR signaling pathway and MAPK signaling pathway [8](#_ENREF_8). In addition, p53 signaling pathway is involved in lung ADC, and cell cycle pathway is involved in both lung ADC and lung SqCC [9](#_ENREF_9" \o "Swanton, 2016 #7).

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 seven novel susceptibility loci of lung cancer, and validated ten 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.

**Conflicts of Interest**

The authors declare no competing financial interest.

**Acknowledgments**

This work was supported by National Natural Science of China (81521004), the Outstanding Young Fund of Jiangsu Province (BK20160046), and the Priority Academic Program for the Development of Jiangsu Higher Education Institutions [Public Health and Preventive Medicine] and Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (PPZY2015A067). 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: a cancer journal for clinicians 2018;68:394-424.

2. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K. 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;343:78-85.

3. Matakidou A, Eisen T, Houlston RS. Systematic review of the relationship between family history and lung cancer risk. British journal of cancer 2005;93:825-33.

4. Tokuhata GK, Lilienfeld AM. Familial aggregation of lung cancer in humans. Journal of the National Cancer Institute 1963;30:289-312.

5. 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;27:363-79.

6. Sampson JN, Wheeler WA, Yeager M, Panagiotou O, Wang Z, Berndt SI, Lan Q, Abnet CC, Amundadottir LT, Figueroa JD, Landi MT, Mirabello L, 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:djv279.

7. Dai J, Shen W, Wen W, Chang J, Wang T, Chen H, Jin G, Ma H, Wu C, Li L, Song F, Zeng Y, et al. Estimation of heritability for nine common cancers using data from genome-wide association studies in Chinese population. International journal of cancer 2017;140:329-36.

8. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511:543-50.

9. Swanton C, Govindan R. Clinical Implications of Genomic Discoveries in Lung Cancer. The New England journal of medicine 2016;374:1864-73.

10. Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, Shukla SA, Guo G, Brooks AN, Murray BA, Imielinski M, Hu X, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nature genetics 2016;48:607-16.

11. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489:519-25.

12. Rahman N. Realizing the promise of cancer predisposition genes. Nature 2014;505:302-8.

13. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, Scott AD, Krassowski M, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. Cell 2018;173:355-70 e14.

14. Wei R, Cao L, Pu H, Wang H, Zheng Y, Niu X, Weng X, Zhang H, Favus M, Zhang L, Jia W, Zeng Y, et al. TERT Polymorphism rs2736100-C Is Associated with EGFR Mutation-Positive Non-Small Cell Lung Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research 2015;21:5173-80.

15. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, Caporaso NE, Johansson M, Xiao X, Li Y, Byun J, Dunning A, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nature genetics 2017;49:1126-32.

16. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, Mirabello L, Jacobs K, Wheeler W, Yeager M, Bergen AW, Li Q, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. American journal of human genetics 2009;85:679-91.

17. Wang Y, Wu W, Zhu M, Wang C, Shen W, Cheng Y, Geng L, Li Z, Zhang J, Dai J, Ma H, Chen L, et al. Integrating expression-related SNPs into genome-wide gene- and pathway-based analyses identified novel lung cancer susceptibility genes. International journal of cancer 2018;142:1602-10.

18. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nature reviews. Cancer 2004;4:177-83.

19. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. Nature methods 2013;10:1081-2.

20. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nature genetics 2013;45:580-5.

21. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome biology 2016;17:122.

22. Dai J, Li Z, Amos CI, Hung RJ, Tardon A, Andrew AS, Chen C, Christiani DC, Albanes D, van der Heijden E, Duell EJ, Rennert G, et al. Systematic analyses of regulatory variants in DNase I hypersensitive sites identified two novel lung cancer susceptibility loci. Carcinogenesis 2019;40:432-40.

23. Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC bioinformatics 2010;11:288.

24. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 2010;38:e164.

25. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. American journal of human genetics 2011;88:76-82.

26. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics : a journal of integrative biology 2012;16:284-7.

27. Wang Y, Wang C, Zhang J, Zhu M, Zhang X, Li Z, Dai J, Ma H, Hu Z, Jin G, Shen H. Interaction analysis between germline susceptibility loci and somatic alterations in lung cancer. International journal of cancer 2018;143:878-85.

28. Marx V. Drilling into big cancer-genome data. Nature methods 2013;10:293-7.

29. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MD, Wendl MC, Zhang Q, Koboldt DC, Xie M, Kandoth C, McMichael JF, Wyczalkowski MA, et al. Integrated analysis of germline and somatic variants in ovarian cancer. Nature communications 2014;5:3156.

30. Dai J, Lv J, Zhu M, Wang Y, Qin N, Ma H, He Y, Zhang R, Tan W, Fan J, Wang T, Zheng H, et al. Risk loci identification and polygenic risk score in prediction of lung cancer risk: a large-scale prospective cohort study in Chinese population. The Lancet Respiratory Medicine 2019;(Accepted).

31. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 2009;4:1073-81.

32. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nature methods 2010;7:248-9.

33. Zhu YC, Liao XH, Wang WX, Xu CW, Zhuang W, Wei JG, Du KQ. Dual drive coexistence of EML4-ALK and TPM3-ROS1 fusion in advanced lung adenocarcinoma. Thoracic cancer 2018;9:324-27.

34. 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.

35. Shalini S, Dorstyn L, Dawar S, Kumar S. Old, new and emerging functions of caspases. Cell death and differentiation 2015;22:526-39.

36. Teitz T, Wei T, Valentine MB, Vanin EF, Grenet J, Valentine VA, Behm FG, Look AT, Lahti JM, Kidd VJ. Caspase 8 is deleted or silenced preferentially in childhood neuroblastomas with amplification of MYCN. Nature medicine 2000;6:529-35.

37. Stupack DG, Teitz T, Potter MD, Mikolon D, Houghton PJ, Kidd VJ, Lahti JM, Cheresh DA. Potentiation of neuroblastoma metastasis by loss of caspase-8. Nature 2006;439:95-9.

38. Stacey SN, Helgason H, Gudjonsson SA, Thorleifsson G, Zink F, Sigurdsson A, Kehr B, Gudmundsson J, Sulem P, Sigurgeirsson B, Benediktsdottir KR, Thorisdottir K, et al. New basal cell carcinoma susceptibility loci. Nature communications 2015;6:6825.

39. Stacey SN, Kehr B, Gudmundsson J, Zink F, Jonasdottir A, Gudjonsson SA, Sigurdsson A, Halldorsson BV, Agnarsson BA, Benediktsdottir KR, Aben KK, Vermeulen SH, et al. Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer. Human molecular genetics 2016;25:1008-18.

40. Ohtani N, Yamakoshi K, Takahashi A, Hara E. The p16INK4a-RB pathway: molecular link between cellular senescence and tumor suppression. The journal of medical investigation : JMI 2004;51:146-53.

41. Botti E, Spallone G, Moretti F, Marinari B, Pinetti V, Galanti S, De Meo PD, De Nicola F, Ganci F, Castrignano T, Pesole G, Chimenti S, et al. Developmental factor IRF6 exhibits tumor suppressor activity in squamous cell carcinomas. Proceedings of the National Academy of Sciences of the United States of America 2011;108:13710-5.

42. Restivo G, Nguyen BC, Dziunycz P, Ristorcelli E, Ryan RJ, Ozuysal OY, Di Piazza M, Radtke F, Dixon MJ, Hofbauer GF, Lefort K, Dotto GP. IRF6 is a mediator of Notch pro-differentiation and tumour suppressive function in keratinocytes. The EMBO journal 2011;30:4571-85.

43. Rotondo JC, Borghi A, Selvatici R, Magri E, Bianchini E, Montinari E, Corazza M, Virgili A, Tognon M, Martini F. Hypermethylation-Induced Inactivation of the IRF6 Gene as a Possible Early Event in Progression of Vulvar Squamous Cell Carcinoma Associated With Lichen Sclerosus. JAMA dermatology 2016;152:928-33.

44. Xu L, Huang TJ, Hu H, Wang MY, Shi SM, Yang Q, Lin F, Qiang YY, Mei Y, Lang YH, Li CZ, Peng LX, et al. The developmental transcription factor IRF6 attenuates ABCG2 gene expression and distinctively reverses stemness phenotype in nasopharyngeal carcinoma. Cancer letters 2017.

**Figure legends**

**Figure 1**. Manhattan plot showing −log10(*P* values) for SNP associations with lung cancer risk overall and by histological subtypes. (A) Lung cancer risk overall. (B) Lung adenocarcinoma. (C) Lung squamous cell carcinoma. Each locus is annotated by its cytoband location and nearby genes. The x axis represents chromosomal location, and the y axis represents −log10 (*P* value). The green horizontal line denotes False Discovery Rate (FDR) < 0.01.