Genome-wide interaction study of smoking behavior and non-small cell lung cancer risk in Caucasian population

Yafang Li, Xiangjun Xiao, Younghun Han, Olga Gorlova, David Qian, Natasha Leighl, Jakob S. Johansen, Matt Barnett, Chu Chen, Gary Goodman, Angela Cox, Fiona Taylor, Penella Woll, H.-Erich Wichmann, Judith Manz, Thomas Muley, Angela Risch, Albert Rosenberger, Susanne M. Arnold, Eric B. Haura, Ciprian Bolca, Ivana Holcatova, Vladimir Janout, Milica Kontic, Jolanta Lissowska, Anush Mukeria, Simona Ognjanovic, Tadeusz M. Orlowski, Ghislaine Scelo, Beata Swiatkowska, David Zaridze, Per Bakke, Vidar Skaug, Shanbeh Zienolddiny, Eric J. Duell, Lesley M. Butler, Richard Houlston, María Soler Artigas, Kjell Grankvist, Mikael Johansson, Frances A. Shepherd, Michael W. Marcus, Hans Brunnström, Jonas Manjer, Olle Melander, David C. Muller, Kim Overvad, Antonia Trichopoulou, Rosario Tumino, Geoffrey Liu, Stig E. Bojesen, Xifeng Wu, Loic Le Marchand, Demetrios Albanes, Heike Bickeböller, Melinda C. Aldrich, William S. Bush, Adonina Tardon, Gad Rennert, M. Dawn Teare, John K. Field, Lambertus A. Kiemeney, Philip Lazarus, Aage Haugen, Stephen Lam, Matthew B. Schabath, Angeline S. Andrew, Pier Alberto Bertazzi, Angela C. Pesatori, David C. Christiani, Neil Caporaso, Mattias Johansson, James D. McKay, Paul Brennan, Rayjean J. Hung, Christopher I Amos*

*To whom correspondence should be addressed. Tel: 603-650-1972; Email: Christopher.I.Amos@Dartmouth.edu

Abstract

Non-small cell lung cancer (NSCLC) is the most common type of lung cancer. Both environmental and genetic risk factors contribute to lung carcinogenesis. We conducted a genome-wide interaction analysis between SNPs and smoking status (never vs ever smokers) in a European-descent population. We adopted a two-step analysis strategy in the discovery stage: we first conducted a case-only interaction analysis to assess the relationship between SNPs and smoking behavior using 13,336 NSCLC cases. Candidate SNPs with p-value less than 0.001 were further analyzed using a standard case-control interaction analysis including 13970 controls. The significant SNPs with p-value less than $3.5x10^{-5}$ (correcting for multiple tests) from the case-control analysis in the discovery stage were further validated using an independent replication dataset comprising 5377 controls and 3054 NSCLC cases. We further stratified the analysis by histological subtypes. Two novel SNPs, rs6441286 and rs17723637, were identified for overall lung cancer risk. The interaction odds ratio and meta-analysis p-value for these two SNPs were 1.24 with $6.96x10^{-7}$ and 1.37 with $3.49x10^{-7}$, respectively. Additionally, interaction of smoking with rs4751674 was identified in squamous cell lung carcinoma with an odds ratio of 0.58 and p-value of $8.12x10^{-7}$. This study is by far the largest genome-wide SNP-smoking interaction analysis reported for lung cancer. The three identified novel SNPs provide potential candidate biomarkers for lung cancer risk screening and intervention. The results from our study reinforce that gene-smoking interactions play important roles in the etiology of lung cancer and account for part of the missing heritability of this disease.

Summary: We conducted a genome-wide gene-smoking interaction analysis in non-small cell lung cancer using genotype from 35,737 individuals including both discovery and validation datasets. We identified three novel SNPs with significant interactions with tobacco smoking.

Introduction

Lung cancer is one of the most common cancers worldwide and the leading cause of cancer-related death in both men and women in the United States (1). Non-small cell lung cancer (NSCLC) contributes to about 80-85% of lung cancer cases (2). NSCLC has three major subtypes: adenocarcinoma, squamous cell carcinoma and large cell carcinoma. About 40% of NSCLC are adenocarcinoma, while squamous cell carcinoma represents about 25-30% of NSCLC and is strongly related to a history of having ever smoked (3-5).

Genome-wide association studies (GWAS) have been successful in identifying common variants associated with lung cancer in the past decade. The identified susceptibility genes include the *CHRNA5*, *CHRNA3* and *CHRNB4* genes at 15q25, *TERT* at 5p15, the HLA region at 6p21, *TP63* at 3q28, and several additional variants (6-13). Most of the identified common variants have a relatively small genetic effect (Odds Ratio (OR) < 1.5) and together account for a fraction of the heritability of lung cancer. Gene-environment interactions are believed to explain part of the missing heritability (14). Tobacco smoking is the major risk factor associated with lung cancer risk and about 80% to 90% of European-descent lung cancer cases have a history of exposure to cigarette smoke (15). Interactions between genes and smoking behavior play an important role in the development of lung cancer (16-18). An interaction effect manifests itself when the disease risk associated with a genotype varies by smoking behavior. In 2014, Zhang and his colleague detected two SNPs rs1316298 and rs4589502 (OR: 0.71, p value 6.73×10^{-6}; and OR 1.55, p value 3.84×10^{-6}, respectively) in a genome-wide gene-smoking interaction scanning using genotype data from 3865 cases and 4566 controls from a Han Chinese population (16). Studies of gene-smoking interactions are important in deciphering the lung cancer etiology because they will reveal those genes involved in lung tumorigenesis that interacting with tobacco smoking that would not be discovered by main effect association analysis without jointly modeling with smoking status. The

identified genetic variants with heterogeneous effects between subgroups defined by smoking behavior will contribute to lung cancer risk prediction and disease prevention.

However, genome-wide interaction scanning remains a challenge. Most GWAS studies were designed for main effect association analysis and have limited power for interaction analysis. Analyses of power show that a sample size at least a four-fold larger is required for interaction analysis if a standard case-control design is used and the power limitations are more extreme when the effect size is modest or the risk allele has a lower frequency (19). In the absence of gene-environment correlation, a case-only approach has been shown to be much more powerful than a standard case-control design (20-21). If the gene environment independence assumption is not met, then false positives can be introduced when a case-only design is followed. A two-step test strategy was proposed by researchers for gene-environment interaction analysis: step 1, comprises a case-only test to test the association between SNPs and environmental risk factor; step 2, candidate SNPs from step 1 were further submitted to standard case-control logistic interaction analysis (21). There are two advantages using this 2-step study design: first, the step 1 test allows us to filter the SNPs tested in step2 thus reducing the power loss from multiple comparisons in the step 2 test; and second, standard case-control interaction analysis in step 2 is more stringent and is robust to the gene-environment requirement of the case-only design, thus reducing the false discovery rate that may otherwise plague the case-only design.

The current reports on genome-wide gene-smoking interaction analysis in lung cancer are still quite limited (16). To explore gene-smoking interactions in NSCLC lung cancer development in a European-decent population, we conducted a genome-wide interaction analysis based on about 500,000 SNPs genotype data from about 27,000 individuals of European descent. We tested the interactions between each SNP and the smoking status (never-smokers vs ever smokers). The interaction analyses were further categorized by lung cancer histology subtypes including adenocarcinoma and squamous cell

carcinoma. The candidate SNPs were further validated using independent genotype data from another sample of about 8, 400 individuals. As far as we know, this is the largest genome-wide SNP-smoking interaction analysis in lung cancer study up to date.

Materials and methods

Study populations

The discovery genotype data in this study came from OncoArray consortium which was designed to identify genetic variants associated with common cancers including breast, colon, lung, prostate and ovarian cancers (22). We restricted the analysis to individuals with European ancestry and valid information on smoking status and lung cancer histology (23). The smoking status was denoted as never- versus ever-smokers, and ever-smokers included current and former smokers based on self-reported information about smoking status when the samples were recruited. The large sample size (n>25,000) in the discovery phase derives from samples that were collected from 28 individual institutes. To minimize the potential for false positive findings, we randomly grouped the data into three balanced data sets S1-S3 (Table S1). The three subsets serve as internal replication datasets for the associations, and help to reduce the potential for spurious association findings. The sample size from the 28 sites varies from 146 to 3195. We "randomly" distributed the sites to three groups following two criteria: 1, there are sites with sample size > 1000 and sites with sample size < 1000 in each group; 2, the sample size of each group are balanced (within range of average±500). There are 9480, 9059 and 8767 individuals in S1-S3 which sum to 13970 controls and 13336 patients with NSCLC lung cancer (Table 1). The NSCLC lung cancer cases includes 7015 adenocarcinoma patients and 4529 squamous cell carcinoma patients. All the samples were genotyped using the Illumina OncoArray-500K BeadChip (22). The independent replication data includes 5377 controls and 3054 NSCLC cases genotyped on a separate Affymetrix array (24). The smoking statuses in the replication data were recorded following the same

classification as in the discovery data. The percentage of never smokers in the control samples are 32.14% and 29.85% in discovery and replication data; and 10.49% and 11.43% in the disease samples in the discovery and replication data, respectively (Table 1).

Ethics statement

All subjects provided informed consent, and the institutional review boards of each participating institutes approved this collaborative study.

Genotype data quality control

In the discovery stage, we started with genotypes from 43,959 samples on 517820 SNPs. We inferred ancestry information using the FastPop program and individuals with probability of European ancestry greater than 0.8 were inferred as having European-descent population (25). IBD analysis and sex checking were conducted as quality control checks to identify close relatives or possible sample processing issues. Individuals and SNPs with genotype call rate less than 0.95 were excluded from the analysis. IBD analysis was further performed among samples between discovery and replication data sets, and duplicate samples included in discovery study were removed. A total of 27,306 individuals including 13,970 controls and 13,336 patients with NSCLC lung cancer were included in the discovery study. FlashPCA was used for PCA analysis and we adjusted for the first three principal components in the interaction analysis (26). A total of 502,933 SNPs were analyzed in the interaction analysis (23).

In the replication study, a total of 12,651 individuals were genotyped using Affymetrix Array platform on 404740 SNPs (24). IBD analysis was conducted to remove duplicate samples or close relatives within the data set. Individuals with genotype call rate less than 0.95 were excluded from the analysis. The Structure program was run to infer ancestry origin and 0.8 was used as the cutoff for European-descent population inference (27). A total of 8431 samples were included in replication study including 5377

controls and 3054 patients with NSCLC lung cancer. EIGENSTRAT was run for PCA analysis and we adjusted for the first three principal components in the analysis (28).

Statistical analysis

We conducted a genome-wide interaction analysis comprising a discovery stage in which candidate SNPs were identified, and then these SNPs were validated in a subsequent replication study using an independent set of cases and controls (Supplemental Figure S1)> A two-step analysis strategy was adopted in discovery stage: step 1, a genome-wide case-only logistic regression analysis was performed to assess the association between each SNP and smoking status using formula (1) (E denotes smoking status) using all the discovery data; SNPs with case-only p value less than 0.001 were further submitted to step 2 analysis with a standard case-control logistic model as denoted in formula (2) (D denotes disease status).

$$logit\ (E) = \beta_0 + \beta_1 \times snp + \sum \beta_i \times cov_i \quad (1)$$

$$logit\ (D) = \beta_0 + \beta_1 \times snp + \beta_2 \times smoking + \beta_3 \times snp \times smoking + \sum \beta_i \times cov_i \quad (2)$$

The Bonferroni corrected cutoff p-value in the step 2 case-control analysis was set to 0.05 divided by the number of SNPs entering the step 2 analysis. For example, if 500 SNPs had case-only p-value less than 0.001 then the cutoff p-value in case-control analysis was $0.05/500=1x10^{-4}$. The significant candidate SNPs following the step 2 test were chosen for further study based on two additional criteria: 1, the SNPs have case-control interaction p value less than 0.1 from each of three subsets in discovery data; 2, case-control interaction p values less than the Bonferroni corrected p-value from the combined discovery data. The candidate SNPs were further submitted for verification in replication study. In the interaction analysis, SNPs were coded in an additive model (0, 1 or 2). There were three categories of reported smoking status, never smoker, current smoker and ex-smoker, in the phenotype data. And ex-

Commented [CIA1]: Could you move the equation numbers further over to the right because they look like they are part of the equation where they are placed right now.

smoker was defined as time since last smoking greater than 2 years. We grouped the samples into never smokers (0) and ever-smokers (1, including both current smokers and ex-smokers). The first three principal components were adjusted in the interaction analysis.

The interaction analysis was further stratified by histology subtypes including adenocarcinoma and squamous cell carcinoma. For those SNPs validated in replication study (case-control interaction p value < 0.05), we also performed a meta-analysis to combine the information from both discovery and replication data.

Genotype imputation

To increase the density of SNP markers at regions surrounding the significant SNPs verified in replication study, we used IMPUTE2 to impute the flanking SNPs in ~250 kb of the three validated SNPs rs6441286, rs17723637, and rs4751674 in the discovery data. Because of the limited overlap in SNP panels between discovery and replication data, we also conducted imputation to increase the SNP density and overlap in the results for the replication data. The 1000 Genomes Project Phase 3 release was used as the reference dataset (29). The output dosage file from IMPUTE2 was used as input in logistic regression analysis and the first three PCs were adjusted in the imputed genotype analysis.

Deleted: whole genome in

Deleted: D

Results

Discovery study

In discovery study, we first performed the genome-wide interaction analysis to test the association between SNPs and smoking behavior using only the lung cancer patients; the samples with p value < 0.001 were submitted to interaction analysis to test the association between SNP-smoking interaction and lung cancer risk using S1-S3 subset as well as the combined data in the discovery stage. Figure 1A-C display the Manhattan plot of –log10(p) from the case-only studies including 13336 NSCLC cases, 7015

adenocarcinoma cases and 4529 squamous cell carcinoma cases, respectively. The QQ-plots displayed the observed p-values vs expected p-values, and the observed genomic inflation factor (λ, lambda) were 1.13, 1.06 and 1.00 for NSCLC, adenocarcinoma and squamous cell carcinoma, respectively. Since the lambda value scales with sample size, we also computed the inflation factor for an equivalent study of 1000 cases (30). The scaled lambda values were 1.01, 1.01 and 1.00 for interaction analysis in NSCLC, adenocarcinoma and squamous cell carcinoma, respectively (Figure 1A-C). No obvious inflation of type I error rate was detected in the study. In the association analysis between smoking behavior and SNPs using only cases, 1379, 867 and 468 SNPs, including the SNPs at the well-known chr15q24.3—chr15q25.1 region (the *CHRNA5*, *CHRNA3*, *CHRNB4*, *IREB2*, PSMA4 gene cluster) with p-value less than 0.001 were detected in NSCLC, adenocarcinoma and squamous cell carcinoma case-only interaction analysis, respectively (Table S2). And these SNPs entered the step 2 test in discovery stage to test the associations between gene-smoking interactions and lung cancer disease using all the cases and controls data.

In step 2 test, the Bonferroni corrected p-values were computed by dividing 0.05 by the number of SNPs entered the analysis. And we got $3.63x10^{-5}$, $5.77x10^{-5}$, $1.07x10^{-4}$ for NSCLC, adenocarcinoma and squamous cell carcinoma subgroup studies, respectively (Table S2). For consistency, we used $3.5x10^{-5}$ as the cutoff in step 2 case-control interaction analysis across all the three studies by histology. The significant SNPs from step 2 test in discovery stage were chosen based on two criteria: 1, the association between disease status and gene-smoking interaction has a p vale < 0.1 from each of the S1-S3 subset; 2, has a p value < $3.5x10^{-5}$ from the combined data analysis. For example, 438, 766 and 925 SNPs had a case-control interaction p value less than 0.1 from subset 1-3 in NSCLC cohort and 105 of them were common to all the three subsets. Among the 105 SNPs, 52 had a case-control interaction p value less than $3.5x10^{-5}$ using the combined data (Table S2). In adenocarcinoma and squamous cell carcinoma lung cancer cohort, 41 and 10 SNPs were selected as significant markers for further replication analysis (Table

S2, S3). 33 and 26 SNPs at chr15q24.3—chr15q25.1 region had significant interaction p-values in discovery stage from NSCLC and adenocarcinoma interaction analysis, respectively. The CHRNA5 region had been extensively studied in several independent studies (7, 12, 18). However, these p-values were much less significant compared with that from main-effect-only association analysis, which means the association effect between disease status and SNPs were much more significant when only genetic main effect was considered in the model compared with the association effect between disease status and gene-smoking interactions (Table S3). These SNPs on chromosome 15q were not novel SNPs and the interaction effect between smoking and SNPs was not as striking as that found in main effect analysis.

Replication study and meta-analysis

The replication data came from a separate study so the genotype panel was different from that of discovery data. Some of the selected candidate SNPs from discovery study were not available in validation data but we still validated the signals at three novel SNPs using genotypes from replication data. In the cohort including all NSCLC cases, SNP rs6441286 on chromosome 3q25.33 had a p-value of 6.30×10^{-6} in gene and smoking behavior association analysis (step 1) and p-value of 1.16×10^{-5} in gene-smoking and disease status association analysis (step 2) using combined discovery data. The step 2 interaction p-values were 3.39×10^{-3}, 1.67×10^{-2}, and 4.79×10^{-3} for S1-S3 subset. The replication data produced an interaction p-value of 2.02×10^{-2}. The interaction odds ratios (OR) varied from 1.21 to 1.31 across the different subsets and the overall OR was 1.24. This SNP has a p-value of 6.96×10^{-7} in the meta-analysis to combine both the discovery and replication data. rs6441286 was located at the intron of *IL12A-AS1* gene, which is an antisense RNA regulating *IL12* gene, a key regulator in immune response.

Another validated SNP in NSCLC cohort was SNP rs17723637 located in *ZNF462* gene. It had a case-only interaction p value of 4.92×10^{-4} in the gene and smoking behavior association analysis, a gene-smoking and disease status association p value of 1.06×10^{-5} in combined discovery data and p value of 9.76×10^{-3}

in validation analysis. The interaction ORs were 1.40 (95% CI: (1.09, 1.79)), 1.32 (95% CI: (1.06, 1.64), 1.45 (95% CI: (1.12, 1.86)), and 1.43 (95%CI: (1.09, 1.88)) for S1-S3 and replication data, respectively. The overall interaction OR was 1.37 with p-value of 3.49×10^{-7} in the meta-analysis.

In the genome-wide gene-smoking interaction analysis stratified by different tumor subtype, SNP rs4751674 had a case-only interaction p-value of 3.69×10^{-5} and the case-control interaction p-value of 1.07×10^{-5} in discovery stage and 2.62×10^{-2} in replication stage in squamous cell carcinoma lung cancer group (Table 2). The interaction ORs were 0.59, 0.54, 0.68 and 0.58 for S1-S3 and replication data, respectively. The overall OR was 0.58 from the meta-analysis with a p value of 8.12×10^{-7}. We also identified another neighbor rs2244178 with a gene and smoking behavior association p-value of 2.23×10^{-4} and gene-smoking interaction and squamous cell lung cancer disease status association p-value of 3.14×10^{-5} from combined discovery data analysis. The OR of lung cancer risk associated with rs2244178 was 0.58 (95% CI: (0.45, 0.75)). Unfortunately, rs2244178 was not available at the replication data and we couldn't verify it. Both rs2244178 and rs4751674 were located at gene *AFAP1L2 (XB130)*, which is an adaptor that regulated signal transduction in lung.

We further checked the interaction effect of the three SNPs at different lung cancer subtypes. Both SNP rs6441286 and rs17723637 had a marginal interaction effect in adenocarcinoma and squamous cell lung cancer. These interaction effect only achieved genome-wide significance in NSCLC analysis when both adenocarcinoma and squamous cell carcinoma patients were included in the analysis to get a larger sample size (Figure 2A). The interaction effect between SNP rs4751674 and smoking behavior was only detected in squamous cell lung cancer subtype. This effect was not existing in adenocarcinoma subtype although there were 3293 more samples in adenocarcinoma than in squamous cell group.

Imputation analysis

To further verify the replicated interactions between SNPs and smoking behavior, we imputed the ~250kb flanking regions around each of the three significant SNPs using the genotypes from the discovery data to increase the density of markers in the regions harboring the three target SNPs. We plotted the signals from 1000 up- and down-stream SNP markers of the target SNPs. Because the genotype data for imputation analysis went through additional QC procedures, the final sample size was a little smaller than that used in genotype analysis. There were 12624 controls and 12979 NSCLC cases in imputed data analysis. In the imputation analysis, we found another eight imputed SNPs with interaction p-values less than $3.5x10^{-5}$ around rs6441286 and the most significant SNP was rs66785795 with interaction p value $7.63x10^{-6}$ (Table S4 and Figure 1D). All of the most significant SNPs were within *IL12A-AS1* gene. For SNP rs17723637, we found another 6 SNPs with interaction p-values less than $3.5x10^{-5}$ and all of the most significant SNPs were within gene *ZNF462* which encodes a zinc finger protein (Table S4 and Figure 1E). For SNP rs4751674 from interaction analysis in squamous cell lung cancer, we detected 14 more SNPs with interaction p-values less than $3.5x10^{-5}$ harboring both rs2244178 and rs4751674. All of these SNPs were located within the gene *AFAP1L2* and the most significant p-value came from imputed SNP rs2483911 with a p-value of $2.87x10^{-6}$ (Table S4 and Figure 1F). The results from imputed genotype analysis strongly supported the identified SNPs in our gene-smoking interactions analysis.

Because the SNP panels between discovery and replication data were different and only limited number of SNPs were available in both panels, we also imputed the replication genotype data using 1000 Genome as the reference. We identified another three SNPs with case-control interaction p value < 0.05 in the imputed replication data. They are rs10477550 on chromosome 5, rs4557740 on chromosome 8 and rs11544453 on chromosome 22 (Table S3). rs10477550 was located within COMMD10 gene, and it has interaction p value of 9.05x10-2, 7.69x10-4, 4.30x10-2 with adenocarcinoma disease in the three subsets S1-S3 in discovery data and 4.33x10-3 in the imputed replication data. rs11544453 was within

Deleted: Since

Deleted: SNPs

Deleted: And a

Moved (insertion) [1]

Deleted: We

WNT7B gene and it has interaction p value of 5.61x10-4, 7.07x10-2, 2.17x10-3 with squamous cell lung cancer in three subsets in discovery data and 4.05x10-2 in imputed replication data (Table S3). The results from imputed replication data suggest the potential interaction effect with smoking behavior at these two genes but direct genotype data would be more reliable for further validation.

Commented [CIA2]: I think you should present a table like Table 3 for these additional 3 snps as a supplemental table for completeness. I would make this an additional supplemental table.

Risk effect of lung cancer at significant SNPs stratified by smoking status

For the replicated SNPs with significant interactions with smoking status, we further investigated the risk effect of the SNPs in smoking and never smoking groups separately. There were 5899 never-smokers in the NSCLC cohort and 1399 individuals were NSCLC patients in the never-smokers (Table 1). The minor allele at SNP rs6441286 had a protective effect on NSCLC in never smokers and the overall OR was 0.83 (95%CI: (0.77, 0.90)) when we combined both the discovery and replication data set (Figure 2B). However, this protective effect did not exist in samples from only smokers (OR=1.04, 95%CI: (1.00, 1.07)) in the study combining both discovery and replication data. Similarly, SNP rs17723637 had a protective effect on NSCLC in nonsmokers with the overall OR was 0.76 (95% CI: (0.68, 0.85)) in never smoking group. No significant effect was identified in the smoking group.

SNP rs4751674 had a negative interaction with smoking behavior in squamous cell carcinoma cohort. Among the 4649 never-smokers, only 159 of them were squamous cell carcinoma lung cancer patients. SNP rs4751674 had a squamous cell lung cancer risk effect with the overall OR of 1.66 (95%CI: (1.35, 2.05)) in nonsmokers when we combined both discovery and replication data (Figure 2B). This risk effect for lung cancer did not exist in the smoking group (OR: 0.99, 95% CI: (0.94, 1.04)). SNP rs4751674 had a risk allele A, and there was no significant difference between the allele frequencies in never-smokers vs ever-smokers in controls (OR=1, p value=0.98) (Table S5). In patients with squamous cell lung cancer, 63.52% of never smokers have at least one risk allele, compared with 45.73% in ever smoker patients

(OR=2.07, p-value=1.44×10^{-5}). The neighbor SNP rs2244178 had a risk effect with OR of 1.60 (95%CI: (1.25, 2.04)) in never-smoker group and no significant effect in ever-smoker group.

Joint analysis of SNP and smoking behavior in lung cancer

To better understand the interaction effect between the SNP and smoking status we conducted a joint analysis with never smoker without risk allele as the reference group (Table 3). For ever-smokers without risk allele group, the NSCLC risk at SNP rs6441286 was 3.51 (95% CI: (3.18, 3.88)); for never-smokers with risk allele group, the risk was 0.79 (95% CI: (0.70, 0.90)); for ever-smokers with risk allele group, the disease risk was 3.66 (95%CIl (3.30, 4.06)). The interaction effect between the risk genotype and smoking behavior was 1.33 (95% CI: (1.16, 1.52)). A similar pattern was found at SNP rs17723637. For people carrying at least one risk allele, the OR of lung cancer was 0.77 (95% CI: (0.67, 0.88)) in never-smokers and 3.94 (95% CI: (3.61, 4.30)) in ever-smokers. The interaction between smoking and the SNP was 1.38 (95% CI; (1.18, 1.60)). The joint analysis at these two SNPs displayed that for a person who was a carrier of the risk allele the lung cancer risk varied dramatically depending on the smoking behavior of the person. Cigarette smoking had a synergetic effect on the risk genotype and abstinence from smoking among those risk allele carrier population would significantly decrease their risk for NSCLC lung cancer.

For SNP rs4751674 we found that smoking had a very big risk effect for squamous cell lung carcinoma which alone contributed an OR of 19.67 (95% CI: (15.08, 25.65). SNP rs4751674 was located at a potential tumor gene *AFAP1L2* and the risk allele contributed an OR of 1.92 (95% CI: (1.38, 2.67)). The OR was decreased to 19.04 (95%CI: (14.59, 24.85)) when both risk factors occurred which meant that smoking behavior had an antagonistic effect on the risk allele and there was a negative interaction between the SNP and smoking behavior (OR=0.48 (95% CI: (0.34, 0.67)).

SNP rs4751674 is located at gene *AFAP1L2* (alias: *XB130*) on chromosome 10 which encodes an adaptor protein that participates in many cellular functions including cell proliferation and survival process in various cancers (31). *AFAP1L2* is a potential oncogene and the knockdown of *AFAP1L2* by RNAi was associated with induced cell death in human lung cancer cells (32). The results from our statistical analysis of SNP rs4751674 suggested that this gene was involved in squamous cell carcinoma in nonsmokers.

Discussion

Lung cancer has a complicated disease mechanism and both genetic and environmental factors impact the disease development. Tobacco smoking is the most important environmental risk factor associated with lung cancer. In this study, we conducted a genome-wide gene-smoking behavior interaction analysis on NSCLC lung cancer using genotype data from about 36,000 samples including both discovery and validation datasets. As far as we know this is by far the largest genome-wide SNP-smoking interaction analysis in lung cancer. The three subsets at the discovery stage and independent validation data at the replication stage allow us to identify SNPs with consistent effect in interaction analysis across different data sets and provide solid evidence for interactions between reported SNPs and smoking behavior. The GWAS studies are designed for main effect association analysis and have limited power for interaction effect detection. We adopted a two-step test for the interaction analysis in discovery stage. The case-only interaction analysis at step 1 allowed us to filter the SNPs for further case-control interaction analysis at step 2. Based on the number of SNPs in standard interaction analysis at step 2, we computed the genome-wide interaction significant level in a conservative way, i.e., 0.05 divided by the number of SNPs tested in step2 analysis and we chose 3.5×10^{-5} as the significance cutoff for all the analyses in discovery study. And the candidate SNPs were submitted to replication study using independent data set.

Although the SNPs at the chr15q24.3—chr15q25.1 region passed the significance criteria at gene-smoking association analysis in case-only study and had a significant interaction effect with smoking behavior in disease status association analysis. The interaction effect was much less significant compared with the main effect association analysis when no interaction effect was considered. For example, rs7163730 and rs11638372, both at chr15q25 region, had a p value of $5.66x10^{-32}$ and $5.24x10^{-25}$ in the main effect model, respectively. The interaction p values were only $8.78x10^{-6}$ and $1.95x10^{-5}$ when SNP-smoking interaction was considered in the model (Table S3).

52, 41 and 10 SNPs were identified from discovery study in NSCLC, ADE and SQC subgroups, respectively. Because of the limited overlap in SNP panel between discovery and replication genotype data, only 35, 26 and 1 of them were available in the replication genotype data, which limited our ability in replication study (Table S2). Three novel SNPs, rs6441286, rs17723637 and rs4751674, were validated in the replication analysis with significant interactions with smoking behavior in lung cancer development. The gene-smoking interaction and lung cancer disease association p-values from meta-analysis are $6.96x10^{-7}$, $3.49x10^{-7}$ and $8.12x10^{-7}$ for these three SNPs. The overall ORs combing both discovery and replication data are 1.24, 1.37 for SNP rs6441286 and rs17723637 in NSCLC, respectively; and 0.58 for SNP rs4751674 in squamous cell carcinoma lung cancer. The large sample size in this study allow us to identify two SNP-smoking interactions with moderate effect (OR of 1.24 and 1.37). The minor alleles C at SNP rs64412866 and G at rs17723637 both have protective effect for NSCLC in never smokers but these protective effects are not existing in smokers. SNP rs64412866 is located at gene *IL12A-AS1* which encodes an antisense RNA of *IL12A* gene. Antisense RNA is widely transcribed in human genome and is an important regulatory mechanism human gene expression (37-38). Studies have shown that cigarette smoking affects non-coding RNA, such as microRNA and antisense RNA, expression in humans (39-40). One study found that some stress-induced non-coding RNAs were up-regulated by exposure to tobacco carcinogen nicotine-derived nitrosamine betone (NNK) in lung cancer and breast

Deleted: identified

cancer cell lines (40). Xi et al. found that the exposure of human respiratory epithelial cells and lung cancer cells to cigarette smoke increased the expression of micro RNA miR-31 in both these two types of cells and overexpression of miR-31 was associated with increased lung cancer risk (39). *IL12A* encodes the subunit of *IL12* which has been shown to be a potent cytokine with antitumor activity in human (41). Our results suggest smoking behavior interact with *IL12A-AS1* gene and increase the risk for NSCLC lung cancer in smokers. The other SNP rs17723637 is located at gene *ZNF462*, which is a member of zinc finger protein transcription factor family in human. The functions of zinc finger proteins in human tumorigenesis vary in different cancers and the report about *ZNF462* is still quite limited. Studies showed that *ZNF462* could be involved in chronic obstructive pulmonary disease development (42). Our results suggested it had a protective effect for lung cancer in nonsmokers.

SNP rs6441286 and rs17723637 are common variants with minor allele frequency of 0.4 and 0.15, respectively. The lung cancer risk among individuals carrying the risk allele of each of these two SNPs varies drastically by smoking status (OR 0.79 vs. 3.66 at SNP 6441286 and 0.77 vs. 3.94 at SNP 17723637 between never- and ever-smokers, Table 3). The positive interactions between smoking behavior and these two SNPs illustrated the adverse effect of smoking behavior in NSCLC development again. The results at these two SNPs provided us another evidence that smoking is harmful to our health and quitting smoking will greatly reduce the risk for lung cancer in human.

In the interaction analysis stratified by disease subtype, we identified some significant interaction in adenocarcinoma cohort in discovery study but not validated successfully in replication study. In squamous cell carcinoma cohort, we found two SNPs, rs2244178 and rs4751674, with p-values less than $3x10^{-5}$ in case-control interaction analysis in discovery study but only rs4751674 were available and successfully validated in replication study. The minor allele A at SNP rs4751674 has a strong interaction effect with smoking status and the OR is 0.58 in squamous cell carcinoma lung cancer risk evaluation.

The negative interaction effect between gene and smoking behavior, i.e., tobacco smoking decreases the genetic risk for lung cancer disease at a genetic locus, is rare but still existing in lung cancer development. For example, Zhang and his colleagues identified the negative interaction between rs1316298 and smoking behavior. This SNP has an OR of 1.12 (95% CI: 1.01-1.25) in non-smoking group whereas an OR of 0.79 (95% CI: 0.71-0.87) is found among smokers (16). SNP rs1316298 is located within a potential tumor suppressor gene and close to genes with tumor-related functions as well. The SNP rs1316298 was not available in our genotype data so we couldn't validate their findings. In our analysis, rs4751674 is located at gene *AFAP1L2* (alias: *XB130*) which is a member of actin filament associated protein (AFAP) family. *AFAP* genes are adaptor proteins and have been shown to be related with tumorigenesis in prostate, lung and breast cancer (31-32,43). Study showed that XB130 regulated survival, cell cycle, migration and invasion of cancer by interacting with binging proteins (44). Our results support its oncogene function in never smokers. Tobacco smoking has a complicated effect in the genome including impact on the signaling pathways, gene expression and induced methylation at many genes (45-48). A study on cadmium, one of the important toxic chemicals in cigarette, showed that cadmium suppressed *AFAP1L2* gene expression (36). Tobacco smoking may reduce the *AFAP1L2* gene expression thus reduced its tumorigenesis risk effect in smokers.

Squamous cell carcinoma of the lung constitutes about 25% to lung cancers and is closely related with smoking history (49). In our squamous carcinoma cohort, only 3.51% and 3.99% of the patients are never-smokers in the discovery data and replication data, respectively, compared with 15.63% in adenocarcinoma cohort (Table 1). The large sample size in the study enabled us to identify the significant interactions in never-smoking squamous cell carcinoma patients. However, the sample size in squamous cell carcinoma cohort is still limited and there are only 159 and 38 never-smoker patients in the discovery and replication cohort, respectively. We are hoping more never-smoker patients will be available in the future so these results can be further validated.

The three SNPs, rs6441286, rs17723637 and rs4751674, identified in our study stratify lung cancer risk by smoking behavior. The interaction ORs are 1.24, 1.37 and 0.58 for these three SNPs, respectively. rs6441286 and rs17723637 have increased risk effect for lung cancer in ever smokers whereas rs4751674 has a protective effect in ever smokers compared with never smokers. These three SNPs can be potential biomarkers used to improve the precision to which we can categorize an individual's risk of lung cancer disease by smoking behavior. rs6441286 and rs17723637 have interaction effect with smoking behavior in NSCLC development, and rs4751674 only interacts with tobacco smoking in squamous cell carcinoma lung cancer. These lung cancer subtype-specific biomarkers will further help us categorize the disease risk by tumor histology which is helpful for individualized prognosis and prediction of treatment plan.

All the three identified novel SNPs have little evidence for association with lung cancer risk in main-effect-only association analysis (p values vary from 0.29 to 0.94 in main effect analysis), which displays that the gene-environment interaction analysis is an essential approach in exploring the missing heritability of lung cancer disease. There are significant gene-smoking interactions at well-known chr15q24.3—chr15q25.1 region in lung adenocarcinoma. But the interaction p-values are much less significant than that from the main-effect-only association analysis, which suggests the dominant roles of the main effect in lung cancer development (Table S3). The interaction effect at SNP rs4751674 only exists in squamous cell carcinoma also suggests the difference of genomic features between squamous cell carcinoma and adenocarcinoma in lung cancer from perspective of gene-smoking interaction analysis. This reported study was restricted to Caucasian population and the results may not be generalized to other ethnicities because of the different genetic backgrounds. The limited overlap between discovery genotype and replication genotype may have reduced the power in our validation study. We believe as more genotype data becomes available in the future we can discover more important gene-smoking interaction in lung cancer disease.

Moved up [1]: Because the SNP panels between discovery and replication data were different and only limited number of SNPs were available in both panels, We also imputed the replication genotype data using 1000 Genome as the reference. We identified another three SNPs with case-control interaction p value < 0.05 in imputed replication data. They are rs10477550 on chromosome 5, rs4557740 on chromosome 8 and rs11544453 on chromosome 22 (Table S3). rs10477550 was located within *COMMD10* gene, and it has interaction p value of $9.05x10^{-2}$, $7.69x10^{-4}$, $4.30x10^{-2}$ with adenocarcinoma disease in the three subsets S1-S3 in discovery data and $4.33x10^{-3}$ in the imputed replication data. rs11544453 was within WNT7B gene and it has interaction p value of $5.61x10^{-4}$, $7.07x10^{-2}$, $2.17x10^{-3}$ with squamous cell lung cancer in three subsets in discovery data and $4.05x10^{-2}$ in imputed replication data (Table S3). The results from imputed replication data suggest the potential interaction effect with smoking behavior at these two genes but direct genotype data would be more reliable for further validation. ¶

Acknowledgement

This study is supported by grant U19CA148127.

Reference

1. http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-key-statistics
2. http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-what-is-non-small-cell-lung-cancer
3. http://www.cancer.org/acs/groups/cid/documents/webcontent/003115-pdf.pdf
4. The Cancer Genome Atlas Research Network (2014) Comprehensive molecular profiling of lung adenocarcinoma. Nature, 511: 543-550.
5. The Cancer Genome Atlas Research Network (2012) Comprehensive genomic characterization of lung squamous cell lung cancers. Nature, 489: 519-525.
6. Buttitta F, et al. (2006) Mutational analysis of the HER2 gene in lung tumors from Caucasian patients: Mutations are mainly present in adenocarcinomas with bronchioloalveolar features. Int. J. Cancer, 119:2586-2591.
7. Amos CI, et al. (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat. Genet., 40(5):616-622.
8. Le Marchand L, et al. (2008) Smokers with the CHRNA lung cancer-associated variants are exposed to higher levels of nicotine equivalents and a carcinogenic tobacco-specific nitrosamine. Cancer Res., 68(22):9137-9140.

9. McKay JD, et al. (2008) Lung cancer susceptibility locus at 5p15.33. Nat. Genet., 40(12):1404-1406.
10. Landi MT, et al. (2009) A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am. J. Hum. Genet., 85(5):679-691.
11. Wang Y, et al. (2008) Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat. Genet., 40(12):1407-1409.
12. Truong T, et al. (2010) Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the international lung cancer consortium. J. Natl. Cancer Inst., 102(13):959-971.
13. Miki D, et al. (2010) Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. Nat. Genet., 42(10):893-896.
14. Maher B. (2008) Personal genomes: the case of the missing heritability. Nature, 456:18-21.
15. http://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm
16. Zhang R, et al. (2014) A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. Carcinogenesis, 35(7):1528-1535.
17. Thorgeirsson TE, et al. (2010) Commentary: gene-environment interactions and smoking-related cancers. Int. J. Epidemiol., 39(2):577-579.
18. CanderWeele TJ, et al. (2012) Genetic variants on 15q25.1, smoking and lung cancer: an assessment of mediation and interaction. Am. J. Epidemiol., 175(10):1013-1020.
19. Smith PG, et al. (1984) The design of case-control studies: the influence of confounding and interaction effect. Int. J. Epidemiol., 13:356-365.
20. Kraft P, et al. (2007) Exploiting gene-environment interaction to detect genetic associations. Hum. Hered., 63: 111-119.

21. Murcray CE, et al. (2009) Gene-environment interaction in genome-wide association studies. Am J. Epidemiol., 169 (2):219-226.

22. Amos CI, et al. (2016) The OncoArray consortium: a network for understanding the genetic architecture of common cancers. Cancer Epidemiol. Biomarkers Prev., doi: 10.1158/1055-9965

23. McKay JD, et al. (2017) Large scale genetic analysis identifies novel loci and histological variability in susceptibility to lung cancer. Nat genetics, In Press.

24. Kachuri L, et al. (2016) Fine mapping of chromosome 5p15.33 based on a targeted deep sequencing and high density genotyping identifies novel lung cancer susceptibility loci. Carcinogenesis, 37(1): 96-105.

25. Li Y, et al. (2016) FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. BMC Bioinformatics, 17:122, DOI: 10.1186/s12859-016-0965-1.

26. Abraham G, et al. (2014) Fast Principal Component Analysis of Large-Scale Genome-Wide Data. PLoS One, 9(4): doi: 10.1371/journal.pone.0093766.

27. Pitchard JK, et al. (2000) Inference of population structure using multilocus genotype data. Genetics, 155:945-959.

28. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet., 28:904-909.

29. Howie BN, et al. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics, 5(6): e1000529.

30. PIW Bakker, et al. (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet., 17(R2): R122-R128.

31. Shiozaki A, et al. (2011) Roles of XB130, a novel adaptor protein, in cancer. J. Clin. Bioinforma., 1:10.

32. Lodyga M, et al. (2005) P-080 Prognostic expression of a novel adaptor protein XB130 innon-small cell lung cancer. Lung Cancer, 49:S1352005.

33. Charlesworth JC, et al. (2010) Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. BMC Med. Genomics, 3:29.

34. Leeuwen DM, et al. (2007) Cigarette smoke-induced differential gene expression in blood cells from monozygotic twin pairs. Carcinogenesis, 28(3):691-697.

35. Staff J, et al. (2012) Relation between smoking history and gene expression profiles in lung adenocarcinomas. BMC Med. Genomics, 5:22.

36. Benton MA, et al. (2011) Comparative genomic analyses identify common molecular pathways modulated upon exposure to low doses of arsenic and cadmium. BMC Genomics, 12:173.

37. Pelechano V, et al. (2013) Gene regulation by antisense transcription. Nat. Rev. Genet., 14: 880-893.

38. Balbin PA, et al. (2015) The landscape f antisense gene expression in human cancers. Genome Res., 25(7):1068-1079.

39. Xi S, Yang M, Tao Y, Xu H, Shan J, et al. (2010). Cigarette smoke induces C/EBP-beta-mediated activation of miR-31 in normal human respiratory epithelia and lung cancer cells. PLoS ONE 5, e13764.10.1371/journal.pone.0013764

40. Silva M, Perez DS, Pritchett JR, Halling ML, Tang H, Smith DI. (2010). Identification of long stress-induced non-coding transcripts that have altered expression in cancer. Genomics 95, 355–36210.1016/j.ygeno.2010.02.009

41. Tugues S, et al. (2015) New insights into IL-12mediated tumor suppression. Cell Death Differ., 22: 237-246.

42. http://www.atsjournals.org/doi/pdf/10.1164/ajrccm-conference.2016.193.1_MeetingAbstracts.A7480

43. Zeng Z, et al. (2016) AFAP1-AS1, a long noncoding RNA upregulated in lung cancer and promotes invasion and metastasis. Tumor Biology, 37(1): 729-737.

44. Shiozaki A, Liu M. (2011) Roles of XB130, a novel adaptor protein, in cancer. J. Clin. Bioinforma., 1:10.

45. Zeilinger S, et al. (2013) Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. PLoS ONE, 8(5): e63812. doi:10.1371/journal.pone.0063812

46. Huang T, et al. (2015) Meta-analysis of gene methylation and smoking behavior in non-small cell lung cancer patients. Sci. Rep., 5:8897. doi:10.1038/srep08897

47. Vink JM, et al. (2015) Differential gene expression patterns between smokers and non-smokers: cause or consequence? Addict. Biol., doi: 10.1111/adb.12322.

48. Birrell MA, et al. (2008) Impact of tobacco-smoke on key signaling pathways in the innate immune response in lung macrophages. J. Cell Physiol., 214(1): 27-37.

49. Kenfield SA, et al. (2008) Comparison of aspects of smoking among the four histological types of lung cancer. Tobacco Control, 17 (3): 198–204.