Published in final edited form as:

Lancet Oncol. 2010 April ; 11(4): 321–330. doi:10.1016/S1470-2045(10)70042-5.

Genetic variants and risk of lung cancer in never smokers: a genome-wide association study

Yafei Li^{*}, Chau-Chyun Sheu^{*}, Yuanqing Ye^{*}, Mariza de Andrade^{*}, Liang Wang^{*}, Shen-Chih Chang^{*}, Marie C Aubry, Jeremiah A Aakre, Mark S Allen, Feng Chen, Julie M Cunningham, Claude Deschamps, Ruoxiang Jiang, Jie Lin, Randolph S Marks, V Shane Pankratz, Li Su, Yan Li, Zhifu Sun, Hui Tang, George Vasmatazis, Curtis C Harris, Margaret R Spitz, Jin Jen, Renyi Wang, Zuo-Feng Zhang[†], David C Christiani[†], Xifeng Wu[†], and Ping Yang[†]
(Y Li PhD, Prof M de Andrade PhD, J A Aakre BS, R Jiang BS, V S Pankratz PhD, Y Li MD, Z Sun MD, H Tang PhD, Prof P Yang PhD), **Department of Laboratory Medicine and Pathology** (L Wang PhD, M C Aubry MD, J M Cunningham PhD, G Vasmatazis PhD), **Department of Surgery** (Prof M S Allen MD, Prof C Deschamps MD), **Division of Medical Oncology** (R S Marks MD), and **Department of Pulmonary and Critical Care Medicine** (J Jen PhD), **Mayo Clinic College of Medicine, Rochester, MN, USA**; **Department of Environmental Health, Harvard School of Public Health, Boston, MA, USA** (C-C Sheu MD, F Chen PhD, L Su BS, Prof D C Christiani MD); **Department of Epidemiology, University of Texas MD, Anderson Cancer Center, Houston, TX, USA** (Y Ye PhD, J Lin PhD, Prof M R Spitz MD, Prof X Wu MD); **Department of Epidemiology, School of Public Health, and Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA, USA** (S-C Chang PhD, R Wang MPH, Prof Z-F Zhang MD); **Laboratory of Human Carcinogenesis, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA** (C C Harris MD); and **Pulmonary and Critical Care Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA** (D C Christiani).

Summary

Correspondence to: Dr Ping Yang, Department of Health Sciences Research, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905, USA, yang.ping@mayo.edu.

^{*}These authors contributed equally to this work.

[†]These authors jointly directed the project.

See **Online** for webappendix

For the **Oncomine microarray databases** see www.oncomine.org

Contributors

PY led the study by designing and conducting it, interpreting results, writing the manuscript, and obtaining funding; XW led the MDACC study team and participated in study design, study conduct, participant recruitment, data collection, results interpretation, manuscript writing, and funding support; MRS contributed as the founder of the parent study of the MDACC study, where cases and controls were enrolled, and manuscript preparation; DCC led the Harvard study team and participated in study design, study conduct, participant recruitment, data collection, results interpretation, manuscript preparation, and funding support. ZFZ led the UCLA study team and participated in study design, study conduct, participant recruitment, data collection, results interpretation, manuscript preparation, and funding support. Yafei L, C-CS, YY, MdA, LW, S-CC, Yan L, and RW coordinated the study design, data collection and analysis, results interpretation, and writing of the manuscript. MdA, JAA, VSP, and HT did the statistical analyses. MCA undertook pathology verification. MSA, RSM, and CD coordinated participant enrolment and data collection. JMC led the genotyping of the Mayo Clinic samples. ZS and GV coordinated bioinformatics analysis and gene-expression data quality control. RJ, FC, and JL participated in data collection, preparation, analysis, and results interpretation. LS participated in data collection and laboratory work. JJ directed gene-expression analyses, results interpretation, and manuscript preparation. CCH contributed to results interpretation and discussions. All authors contributed to the final paper.

Conflicts of interest

The authors declared no conflicts of interest.

Background—Lung cancer in individuals who have never smoked tobacco products is an increasing medical and public-health issue. We aimed to unravel the genetic basis of lung cancer in never smokers.

Methods—We did a four-stage investigation. First, a genome-wide association study of single nucleotide polymorphisms (SNPs) was done with 754 never smokers (377 matched case-control pairs at Mayo Clinic, Rochester, MN, USA). Second, the top candidate SNPs from the first study were validated in two independent studies among 735 (MD Anderson Cancer Center, Houston, TX, USA) and 253 (Harvard University, Boston, MA, USA) never smokers. Third, further replication of the top SNP was done in 530 never smokers (UCLA, Los Angeles, CA, USA). Fourth, expression quantitative trait loci (eQTL) and gene-expression differences were analysed to further elucidate the causal relation between the validated SNPs and the risk of lung cancer in never smokers.

Findings—44 top candidate SNPs were identified that might alter the risk of lung cancer in never smokers. rs2352028 at chromosome 13q31.3 was subsequently replicated with an additive genetic model in the four independent studies, with a combined odds ratio of 1.46 (95% CI 1.26–1.70, $p=5.94 \times 10^{-6}$). A *cis* eQTL analysis showed there was a strong correlation between genotypes of the replicated SNPs and the transcription level of the gene *GPC5* in normal lung tissues ($p=1.96 \times 10^{-4}$), with the high-risk allele linked with lower expression. Additionally, the transcription level of *GPC5* in normal lung tissue was twice that detected in matched lung adenocarcinoma tissue ($p=6.75 \times 10^{-11}$).

Interpretation—Genetic variants at 13q31.3 alter the expression of *GPC5*, and are associated with susceptibility to lung cancer in never smokers. Downregulation of *GPC5* might contribute to the development of lung cancer in never smokers.

Introduction

Tobacco smoking remains the principal cause of lung cancer. However, 15% of men and 53% of women (25% of all cases worldwide) who develop lung cancer do so without any history of having smoked tobacco products (never smokers).¹ In Europe and North America, about 10–15% of lung cancers occur in never smokers. By contrast, about 30–40% of lung cancers occur in never smokers in Asian countries.² Many studies have shown that the aetiology, clinical characteristics, and prognosis of lung cancer in never smokers are substantially different to those in smokers, and lung cancer in never smokers is increasingly recognised as a distinct disease entity.^{3,4} Although the causes of lung cancer in never smokers are poorly understood, one of the established risk factors in European and North American countries is exposure to second-hand smoking.⁵ Other—though inconsistently reported—risk factors include environmental factors, hormones, and viral infections.^{3,4} Individual susceptibility to lung cancer has been studied in an attempt to identify and characterise both inherited genetic and acquired somatic changes.^{6,7} However, the specific genetic mechanisms that increase the risk of lung cancer remain to be elucidated.

Recently, genome-wide association studies have identified several candidate genes and genomic loci that have a moderate effect on the risk of lung cancer. Current candidates include nicotinic acetylcholine receptor subunit genes, 5p15.33, 15q25.1, and 6p21.33, with estimated odds ratios ranging from 1.14 to 1.32.^{8–11} A recent study also identified *RGS17* on 6q23–25 as a gene associated with familial lung cancer.¹² To date, no genome-wide association studies have been done with never smokers alone, and the top candidate single nucleotide polymorphisms (SNPs) from previous genome-wide association studies have not been consistently replicated in never smokers.^{13,14} To identify genetic loci and candidate genes that increase the risk of lung cancer in never smokers, we did a genome-wide association study in never smokers with lung cancer and matched controls.

Methods

Patients

Never smokers were defined as individuals who had smoked less than 100 cigarettes during their lifetime. Written informed consent was obtained from all participants at each of the participating institutions. Research protocols were approved by the institutional review boards of Mayo Clinic (Rochester, MN, USA), MD Anderson Cancer Center (MDACC) and Kelsey-Seybold Clinic (Houston, TX, USA), Harvard School of Public Health and Massachusetts General Hospital (Boston, MA, USA), and University of California in Los Angeles (CA, USA).

In the Mayo genome-wide association study, patients with lung cancer who were classed as never smokers were identified and recruited between January, 1997, and September, 2008. A detailed explanation of the recruitment process has been reported previously.^{15,16} Briefly, community residents who were never smokers were selected as controls and matched to patients according to age, sex, and ethnic background. Detailed information on family history of cancer and exposure to second-hand smoke was collected for both the cases and controls through a structured questionnaire and medical records, including information on the source of the exposure (spouses, parents, and co-workers), amount (number of packs, with 20 cigarettes per pack), and duration of exposure (years). Co-worker exposure included exposure in social settings (eg, clubs, bars, and theatres). The source of second-hand smoke was classed as either childhood exposure (mother and/or father), adulthood exposure (spouse and/or co-worker), and lifetime exposure (at least one source from childhood and one from adulthood). The history of chronic obstructive pulmonary disease (COPD) was determined by a review of participants' medical history.

In the MDACC study, the cases were patients with newly diagnosed, histologically confirmed lung cancer recruited from MDACC. Controls were healthy individuals recruited from the Kelsey-Seybold Clinic, which is the largest multispecialty physician group practice in the greater Houston area.¹⁷ Demographic data were obtained in face-to-face interviews by trained MD Anderson staff interviewers. Clinical data were taken from medical records.

Details of participant recruitment for the Harvard study have been described previously.¹⁸ Cases were patients with newly diagnosed, histologically confirmed primary lung cancer; controls were healthy non-blood-related family members and friends of patients with cancer or patients with cardi thoracic conditions undergoing surgery. Interviewer-administered questionnaires were used to collect information on demographics, occupational exposures, and detailed smoking histories from each participant.

In the UCLA study, cases were patients with histologically confirmed lung cancer, and were identified through the Los Angeles County Cancer Registry administered by the cancer surveillance program at the University of Southern California (Los Angeles, CA, USA). The population controls (without history of lung or upper aerodigestive tract cancers) were frequency matched according to sex and age, and were selected by use of an algorithm to identify eligible controls from a census of each case's neighbourhood. Demographic data, information on exposure to second-hand smoke, and detailed smoking histories were collected by in-person interviews through standardised questionnaires administered by trained interviewers.

Procedures

In the Mayo genome-wide association study, genotyping was done with the Illumina HumanHap 370k and 610k BeadChips (Illumina, San Diego, CA, USA), which contain 373397 and 592532 tag SNPs, respectively. CEPH DNA samples (a family trio) were included in each 96-well plate to monitor genotyping calling in this study. Concordance between the replicates

was above 99.5%. For SNP quality control, we excluded SNPs with more than 5% of samples that failed ($n=33549$), minor allele frequencies (MAF) less than 0.005 ($n=7878$), and genotype distributions that were not in Hardy-Weinberg equilibrium among controls ($p < 1 \times 10^{-7}$, $n=52$). For sample quality control, we excluded samples with genotyping call rates less than 95% ($n=3$) and with inconsistencies between genotype and self-reported sex data ($n=1$). A total of 331918 SNPs (common in both the Illumina HumanHap 370k and 610k BeadChips) of 377 case-control pairs were used in the analysis. Population stratification analysis was done before the analysis with EIGENSOFT version 2.0.¹⁹ None of the first ten eigenvectors were significant (p values varying from 0.11 to 0.98) using ANOVA for population differences along each eigenvector. There were no statistically significant differences between case and control samples for all eigenvectors ($p=0.29$); thus, in further SNP-association analyses, the population stratification was not adjusted.

Genotyping data for the top 44 SNPs were obtained from the Illumina HumanHap 660k and 610k BeadChips in the MDACC study. After quality control, 735 never smokers (328 cases and 407 controls) were analysed. All 44 SNPs had call rates above 95%, and all the samples included in this study had genotyping call rates above 95%. The genome control inflation factor, $\lambda(\text{GC})$, for overall population in the additive model was 1.016. We considered the adjustment for potential population stratification by adjusting the first principal component, which resulted in a $\lambda(\text{GC})$ of 1.013, showing no evidence of genome-wide inflation of the observed test statistics, and the effect of population substructure was minimal. Therefore, no principal component was adjusted in the subsequent association analyses.

In the Harvard study, genotyping was done with the Illumina Human610-Quad BeadChip. After quality control, all 44 SNPs had call rates above 95%, and all the samples included in the replication studies had genotyping call rates above 95%. Estimates of pairwise genome-wide identity-by-descent were used to detect cryptic relatedness. Pairs of samples showing PI-HAT greater than 0.05 were inspected individually, and 15 samples were removed. To detect population stratification, we used EIGENSTRAT version 2.0 to do a principal-component analysis.¹⁹ Eight samples were excluded as population outliers. To control for potential confounding from population stratification, we selected the first four principal components, on the basis of significant ($p < 0.05$) Tracy-Wisdom tests²⁰ and $\lambda(\text{GC})$, as covariates for multivariate analyses. The top 44 SNPs were analysed in 253 never smokers (92 cases and 161 controls).

Because the two top SNPs, rs2352028 and rs2352029, were in complete linkage disequilibrium in the control data from Mayo, MDACC, and Harvard, as well as in the CEU HapMap data, only rs2352028 was genotyped in the UCLA study, by use of the TaqMan allelic discrimination method on the TaqMan ABI 7900HT sequence detection system platform (Applied Biosystems, Foster City, CA, USA). The conditions were: 10 ng of dried genomic DNA in a 5 μL reaction mix containing TaqMan SNP genotyping universal master mix and a probe and a primer set (Applied Biosystems) at 95°C for 10 min, followed by 60 cycles of amplification (92°C for 15 s and 60°C for 1 min). After amplification, SDS 2.3 software was used to determine the fluorescent signal from the VIC or FAM-labelled probe. Around 10% of the samples were randomly selected and repeated for quality control, with 100% concordance. A successful genotyping rate of 98.7% was obtained. All SNPs passed the Hardy-Weinburg equilibrium test. The UCLA study did not have genome-wide association study data, and population stratification was not done; therefore, adjustments in the analysis used only demographics and potential confounding variables.

Genome microarray analysis was done with the Illumina Human WG DASL array (Illumina, Inc, San Diego, CA, USA). Raw intensity data were generated in three batches. For each batch, samples were loaded into BeadStudio 3.1 (gene expression module 3.4) for quality control and

calculation of gene or probe level intensity for each individual sample. Average signal intensity, numbers of detected genes, and signal-to-noise ratio (ratio of 95% vs 5% percentile signal intensity) for each sample were assessed and compared with the batch. Samples with a signal-to-noise ratio less than 10 or an extremely low average signal intensity were excluded or repeated. Correlation between the replicates was assessed to ensure the reproducibility of the experiment. Once the samples passed quality control, they were merged and normalised together by use of the R faster cyclic loess function (Fastlo).²¹ After the normalisation, further quality control was done to assess a potential batch effect, and adequacy of normalisation with principal component analysis, unsupervised clustering, and sample replicates. No noticeable batch effect was seen, and the correlation among the replicates across batches was generally high ($r^2 > 0.95$). 143 pairs of normal and tumour samples from never smoking cases were analysed, of which 70 overlapped with the Mayo genome-wide association study; the lung cancer histological types included 77 adenocarcinoma, 29 carcinoid, and 37 other types (sarcomatoid, squamous, large cell, and unspecified non-small cell).

Statistical analysis

In the Mayo study, potential confounders were assessed to determine whether they should be included as variables for adjustment in the subsequent analyses. A stepwise selection process in a conditional logistic regression model was used to select the adjusted variables from all potential confounders, including COPD, exposure to second-hand smoke in adulthood, exposure to second-hand smoke in childhood, lifetime exposure to second-hand smoke, family history of lung cancer, and family history of any other cancers. We selected history of COPD, exposure to second-hand smoke in adulthood, and family history of lung cancer as variables for adjustment. The association between SNPs and lung-cancer risk was analysed by use of multivariate conditional logistic regression assuming additive, dominant, and recessive genetic models. Webappendix p 17 shows the quantile–quantile (Q–Q) plot of the distribution of test statistics. In the MDACC and Harvard replication studies, odds ratios (OR), 95% CI, and p-values were estimated by use of a multivariate unconditional logistic regression adjusting for age, sex, and the first four principal components (Harvard study). All analyses above were done with PLINK version 1.06.²² In the UCLA study, OR and p-values were calculated with a multivariate unconditional logistic regression adjusting for age, sex, ethnic background, education, lifetime exposure to second-hand smoke, and family history of lung cancer in first-degree relatives; all analyses were done with SAS version 9.1.3.

To obtain combined estimates of risk for replicated SNPs from the Mayo genome-wide association study and the three replication studies, a meta-analysis was done to derive the summary estimate of ORs and p values by use of a fixed-effects model. A Cochran's Q test was used to test heterogeneity among studies. A moment-based estimate method was used to assess the variance among studies. The analysis was done with STATA version 8.0. The population-attributable risk was calculated as $PAR\% = 100\% \times P \times (OR - 1) / [P \times (OR - 1) + 1]$, where P is the frequency of the risk allele associated with lung cancer in the control group, and the OR is the combined OR using a fixed effects meta-analysis model.

A linear regression model adjusted for age and sex was used to assess the correlation between genotypes (independent variable, coded as 0, 1, or 2) and gene transcript expression levels (eQTL, dependent variable) in 70 samples of normal lung tissue. Bonferroni correction was used to adjust the multiple tests. The analysis was done with SAS version 9.0.

A paired *t* test was done to identify genes that were expressed differently in tumour samples and samples of adjacent normal tissue. Gene expression was $\log_2$ transformed. The fold change was obtained by raising 2 to the power of the mean difference in expression between tumour and normal tissues. We analysed all histological tumour types. We also analysed separately 77 adenocarcinoma samples and matched normal tissue samples and 29 carcinoid tumour samples

and matched normal tissue. The analysis was done with Partek version 6.4. Bonferroni's method was used to adjust the multiple tests.

Role of the funding source

The funders had no role in the design of the study, data collection, data analysis, data interpretation, writing of the report, or the decision to submit for publication. PY, XW, DCC, and ZFZ had access to the raw data. The corresponding author had full access to all data and had final responsibility for the decision to submit the manuscript for publication.

Results

We did a four-stage study to systematically investigate common genetic variations associated with the risk of lung cancer in never smokers (figure 1). In the first stage (the Mayo genome-wide association study), we analysed 331 918 SNPs in 377 case-control pairs matched according to age, sex, and ethnic origin (webappendix p 2). The strongest association was detected at two intergenic SNPs on chromosome 12, rs11183940 ($p=1.5 \times 10^{-6}$) and rs10880785 ($p=7.1 \times 10^{-6}$; figure 2A). To validate the initial findings, we selected the 44 most significant SNPs (webappendix pp 3–4) by considering potential confounding effects from exposure to second-hand smoke, family history of lung cancer, and a history of COPD. The criteria were p values less than 0.001 in all logistic regression models, under an additive genetic model, with and without adjustment for the three potential confounders. Because most of the patients in the Mayo study had either adenocarcinomas (67.6%) or non-small-cell lung cancer (NSCLC; 81.4%; webappendix p 2), we further investigated the association of the top 44 SNPs with adenocarcinoma-only and NSCLC; all top 44 SNPs remained significant in the two subgroup analyses (webappendix p 5).

In stage 2, we tested the 44 top SNPs in two external datasets from the MDACC and Harvard studies. The MDACC study analysed 735 never smokers (328 cases and 407 controls, webappendix p 6); two SNPs, rs2352028 and rs2352029 (linkage disequilibrium $r^2=1$) at chromosome 13q31.3 (Figure 2B and webappendix p 18) were replicated in both the additive (OR 1.54, 95% CI 1.19–2.00; $p=9.76 \times 10^{-4}$) and dominant genetic models (1.83, 1.32–2.53; $p=2.73 \times 10^{-4}$, webappendix pp 3–4, 7). When the analysis was restricted to adenocarcinoma, the two SNPs remained significant (1.41, 1.06–1.87; $p=0.017$ in the additive model; webappendix p 7). The Harvard study included 253 never smokers (92 cases and 161 controls; webappendix p 6). Although the same two SNPs were not replicated in the Harvard study (0.92, 0.60–1.41; $p=0.70$ in the additive model; webappendix p 7), the combination of the Harvard data with the Mayo and MDACC studies did strengthen the significance of the two SNPs (combined OR 1.48, 95% CI 1.26–1.75; $p=2.2 \times 10^{-5}$).

In stage 3, because rs2352028 and rs2352029 were in complete linkage disequilibrium, we tested rs2352028 in the UCLA study (webappendix p 8). rs2352028 was significantly associated with risk for lung cancer (OR 1.69, 95% CI 1.00–2.84; $p=0.048$) in the dominant model, but not in the additive model (1.37, 0.94–1.98; $p=0.099$; webappendix p 9). When the analysis was restricted to white patients only (data not shown) or adenocarcinoma only (webappendix p 9), the association was not statistically significant.

For all four studies, the combined p -value for the association between rs2352028 and lung cancer in never smokers was 5.94×10^{-6} (OR 1.46, 95% CI 1.26–1.70) under the additive model (figure 3A); when restricted to adenocarcinoma only, the combined p -value was 3.00×10^{-4} (1.39; 1.16–1.66; figure 3B). The estimated percentage of population attributable risk (PAR %) ranged from 10.68% in the additive model to 13.88% in the dominant model (figure 3A), indicating more than 10% of cases of lung cancer in never smokers could be attributed to genetic variation of the SNP rs2352028.

It is well known that SNPs in regulatory elements can affect expression levels of the target genes, and hence alter the susceptibility to diseases.²³ In recent years, expression quantitative trait loci (eQTL) analysis has emerged as a powerful tool for identifying genetic variants that affect gene regulation.²⁴ Thus in stage 4, we did a *cis* eQTL analysis to investigate the association between genotypes of the top 44 SNPs and the gene-expression levels in normal lung tissues from 70 of the 377 Mayo cases. Among all residing and nearby genes (within 50 kb) of the 44 SNPs, we identified 36 genes (webappendix p 10). The eQTL analysis showed that the genotypes of the two replicated SNPs in stage 2, rs2352028 and rs2352029, were strongly associated with expression levels of *GPC5* ($p=1.96\times10^{-4}$ and 1.88×10^{-4} , respectively; webappendix pp 11–12). Individuals with the high-risk allele (A at rs2352028) have a lower *GPC5* expression level than individuals with the common allele (G; figure 4). An identical result was seen for rs2352029, with risk allele C and common allele A (figure 4). No other candidate SNPs were significantly correlated with their residing or nearby genes' expression levels.

We further analysed whole-genome transcript levels in tumour and adjacent normal lung tissues (fresh-frozen) from never-smoking cases. We compared the expression of the 36 genes in all histological tumour types, and separately for 77 adenocarcinomas and matched normal tissue and 29 carcinoid tumours and matched normal tissue. We noted that *GPC5* expression levels were 50% lower in adenocarcinoma than in matched normal lung tissue ($p=6.75\times10^{-11}$; webappendix pp 13–14), consistent with the finding that high-risk alleles of the two validated SNPs were correlated with decreased *GPC5* gene expression. No significant differences in expression were seen when all histological types and carcinoid tumour-type only were compared with matched normal tissues (with either multiple tests adjusted $p>0.05$ or fold-change <2.0). Although seven other genes (*AGTR1*, *ASTN2*, *KIAA1217*, *SGPL1*, *SERPINF1*, *SDK2*, and *CACNA2D1*) were also noted to be differentially expressed between tumour and matched normal tissues, our eQTL analysis did not show any candidate SNPs that were associated with the expression levels of these genes.

Discussion

From the first stage of our genome-wide association study searching for common genetic variations responsible for increasing the risk of lung cancer in never smokers, we identified 44 candidate SNPs. Two of these candidate SNPs, rs2352028 and rs2352029, were replicated in stage 2 of our study. These two SNPs are in complete linkage disequilibrium ($r^2=1$) and located at intron 5 of *GPC5*. rs2352028 was further replicated in stage 3. Subsequent functional analyses, eQTL, and analysis of gene-expression levels, strongly indicate that rs2352028 and rs2352029, or variants tagged by these two SNPs, are associated with risk of lung cancer in never smokers through their regulation of *GPC5* expression.

Our results showed that *GPC5* expression is significantly lower in adenocarcinoma than in matched normal lung tissue, but there was no significant difference in expression levels in lung carcinoid tumours. To further confirm these results, we assessed the Oncomine microarray databases with nine studies on lung cancer, seven of which included lung adenocarcinoma. Two datasets^{25,26} showed significant downregulation of *GPC5* in adenocarcinoma tumours compared with normal lung tissue. Importantly, the two studies^{25,27} included smoking status information, and both showed lower expression in never smokers than in smokers (webappendix pp 15–16). Four studies reported the *GPC5* expression information from other histological types, including carcinoid, squamous, small-cell carcinoma, and large-cell carcinoma, and showed no significant differences in *GPC5* expression between these and normal tissue types (webappendix pp 15–16). Thus, reduced *GPC5* expression could be specific for adenocarcinoma in never smokers. However, owing to sample sizes and the different characteristics of study samples from Oncomine, this conclusion needs to be further

validated. The absence of a significant difference between *GPC5* expression in carcinoid tumours and normal lung tissue in our study and the Oncomine databases might be the result of small sample sizes or the different aetiology of adenocarcinoma—while adenocarcinomas are derived from cells located in the epithelium lining the bronchi, carcinoid tumours arise from neuroendocrine cells.

Many studies have reported consistently that lung cancer risk declines among former smokers with increasing years of abstinence; however, the duration after which lung cancer risk in former smokers reaches the same level as that found in never smokers is unclear.^{28,29} Several studies showed that the risk of lung cancer in former smokers approaches the same risk level as never smokers after 10 years of smoking cessation.^{28,30,31} Therefore, we also analysed the association of the top 44 SNPs, including individuals who had quit smoking for 10 years or more (long-term quitters) in the Harvard study. Interestingly, the two replicated SNPs were also significantly associated with lung cancer risk in former smokers ($p=0.018$ for rs2352028 and 0.015 for rs2352029; webappendix p 19), suggesting that never-smokers and long-term quitters might share common genetic mechanisms in developing lung cancers.

GPC5 is a member of the glypican gene family, has eight exons encoding 572 aminoacids, and spans a large genomic region of 1.47 Mb at 13q31.3.³² Glypicans are a family of heparan sulphate proteoglycans (HSPGs) that are linked to the exocyttoplasmic surface of the plasma membrane by a glycosyl-phosphatidylinositol (GPI) anchor. HSPGs are widely distributed in mammalian tissues and interact with many proteins including growth factors, chemokines, and structural proteins of the extracellular matrix to influence cell growth, differentiation, and the cellular response to the environment.³³ Evidence to date suggests that the main function of the membrane-attached glypicans is to regulate the signalling pathway of Wnt, hedgehog, fibroblast growth factors, and bone morphogenetic proteins.³⁴ Depending on the context, glypicans might have a stimulatory or inhibitory activity on these pathways, which are important in regulating cell proliferation and division,^{35,36} and have been previously shown to be involved in developmental processes and the oncogenesis of various types of human cancer.^{37,38}

Alterations at the *GPC5* locus are a common event in various human tumours. Amplifications at 13q31–32 are frequently seen across several tumour types, including lymphomas, breast cancers, and neurological tumours.^{39–41} For lung cancer, an array CGH-based study reported a homozygous deletion at 13q31.3 in a non-small-cell lung cancer cell line.⁴² Another array CGH-based study recently analysed a series of 14 patients with 13q partial deletion syndrome and noted lung hypoplasia as one of the common phenotypes.⁴³ Among the 14 patients, two had lung hypoplasia. These studies suggest that *GPC5* might be a crucial gene in lung development, and genetic variations of this gene might contribute substantially to an increased risk of lung cancer. For non-cancer association, a recent genome-wide association study identified a role of *GPC5* in multiple sclerosis risk,⁴⁴ while a genome-wide pharmacogenomics analysis indicated the association of *GPC5* polymorphisms with response to interferon- β therapy in patients with multiple sclerosis.⁴⁵ Another genome-wide association study identified *GPC5* as a candidate gene in serum docosahexaonic acid metabolite profiles.⁴⁶ This evidence suggests that *GPC5* has different roles depending on the tissue type and the stage of disease development and progression.

We recognise that our sample sizes (both discovery and replication) do not afford enough power, at a less than 10^{-7} genome-wide level of significance for an OR of 1.6. However, it is important to note that recruiting “pure” never smokers at multiple study sites from countries and regions where cigarette smoking was the cause of 90% of all lung cancer cases in the population has proven to be a difficult task. Nonetheless, despite the small sample size, we were able to identify new genetic variants for never smokers, with the corroboration of

functional data at the specific gene-expression level. Our study design, using a multi-level genomic analytical approach (ie, integrating data from germline SNPs, germline normal-lung-tissue gene expression, and tumour-tissue gene expression) has increased the statistical reliability and biological plausibility of our replicated SNPs.

Although we did not see significant statistical heterogeneity among the four datasets under the three genetic models, there might be unobserved heterogeneity caused by the samples and the study design. Specifically, the control set in the Harvard study was comprised of healthy non-blood-related family members and friends of patients with cancer or patients with cardiothoracic conditions undergoing surgery, while the controls in the three other data sets were enrolled from general community populations. However, it is important to note that despite the potential heterogeneity in the Harvard study, the inclusion of this particular study did not affect our final results and conclusions.

Most of the top 44 SNPs were not validated in the replication studies, and this could be explained by various factors. First, the sample size; and second, the inconsistent adjustment for potential confounders, such as exposure to second-hand smoke, previous diagnosis of COPD, and family history of lung cancer, which were adjusted for in the Mayo study but were not adjusted for in the MDACC and Harvard studies because of a lack of data. Because of the difficulty in collecting large samples with all potential risk factors, our study results will hopefully send out an important message to other investigators in the same field. Our careful selection of the top 44 SNPs based on multiple scenarios of the available covariates was prepared for additional replication studies with varied information that were or are collected by other investigators. Furthermore, three of the study populations used here—the UCLA dataset being the exception—included almost exclusively white participants, which could limit the ability to generalise our results to populations with other ethnic backgrounds. Although replication in the UCLA population, of mixed ethnic origin, confirmed the association of rs2352028 at *GPC5* with lung cancer in never smokers after adjusting for ethnic origin, well-designed genome-wide association studies in other populations are needed to determine whether our findings are population-specific.

To our knowledge, our study presents the first and largest effort so far to characterise comprehensively the genomic alterations in lung cancer in never smokers from germline to somatic level. We used a staged genome-wide association study design and analysis, which was aimed at identifying robust associations and reducing type I errors. Our study was strengthened by incorporating functional analyses of eQTL and differential gene expression. Importantly, our eQTL analysis was done on normal lung tissues from never smokers, while most studies that use eQTL analysis have been based on lymphoblastoid cell lines (LCLs). It has recently been shown that genetic variations might regulate gene transcript expression in either a tissue-specific or tissue-independent manner, and LCLs and primary tissue type cells share only a minority of *cis* eQTL.⁴⁷ Therefore, LCLs might not be truly representative of the regulatory landscape in the affected tissue of interest. Our lung tissue-based eQTL analysis overcomes certain limitations of those LCL-based studies. Finally, our analysis of differential gene expression between tumour and normal tissues further confirmed the germline-based findings (from GWAS and eQTL) at a somatic level, and identified *GPC5* as a candidate lung-cancer-susceptibility gene.

In summary, we have identified a genetic locus at 13q31.3 that regulates the expression of *GPC5*, which might contribute to the development of lung cancer in never smokers. Future studies are needed to investigate the regulatory effect of these SNPs (or tagged variants) and the functional role of *GPC5* in lung tumorigenesis.

Acknowledgments

We thank Susan Ernst for her technical assistance with the manuscript. The Microarray Shared Resource and the Genotyping Shared Resource of the Mayo Clinic Advanced Genomic Technology Center, respectively, did all the genome-wide association and gene-expression analyses described in this study. This study was supported by US National Institutes of Health research grants R01-CA80127 (to PY), R01-CA84354 (to PY), CA111646 (to XW), CA127615 (to XW), CA055769 (to MRS), CA092824 (to DCC), CA074386 (to DCC), CA090578 (to DCC), P50 CA90833 (to ZFZ), T32 CA09142 (to ZFZ), and Mayo Foundation funds (to PY and JJ).

Funding US National Institutes of Health; Mayo Foundation.

References

1. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin* 2005;55:74–108. [PubMed: 15761078]
2. Toh CK, Lim WT. Lung cancer in never-smokers. *J Clin Pathol* 2007;60:337–340. [PubMed: 16916998]
3. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers—a different disease. *Nat Rev Cancer* 2007;7:778–790. [PubMed: 17882278]
4. Subramanian J, Govindan R. Lung cancer in never smokers: a review. *J Clin Oncol* 2007;25:561–570. [PubMed: 17290066]
5. Stayner L, Bena J, Sasco AJ, et al. Lung cancer risk and workplace exposure to environmental tobacco smoke. *Am J Public Health* 2007;97:545–551. [PubMed: 17267733]
6. Risch A, Plass C. Lung cancer epigenetics and genetics. *Int J Cancer* 2008;123:1–7. [PubMed: 18425819]
7. Ding L, Getz G, Wheeler DA, et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 2008;455:1069–1075. [PubMed: 18948947]
8. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* 2008;40:1407–1409. [PubMed: 18978787]
9. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008;452:633–637. [PubMed: 18385738]
10. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet* 2008;40:1404–1406. [PubMed: 18978790]
11. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008;40:616–622. [PubMed: 18385676]
12. You M, Wang D, Liu P, et al. Fine mapping of chromosome 6q23–25 region in familial lung cancer families reveals RGS17 as a likely candidate gene. *Clin Cancer Res* 2009;15:2666–2674. [PubMed: 19351763]
13. Yang P, Li Y, Jiang R, Cunningham JM, Zhang F, de Andrade M. A rigorous and comprehensive validation: common genetic variations and lung cancer. *Cancer Epidemiol Biomarkers Prev* 2010;19:240–244. [PubMed: 20056643]
14. Wang Y, Broderick P, Matakidou A, Eisen T, Houlston RS. Role of 5p15.33 (TERT-CLPTM1L), 6p21.33 and 15q25.1 (CHRNA5-CHRNA3) variation and lung cancer risk in never-smokers. *Carcinogenesis* 2010;31:234–238. [PubMed: 19955392]
15. Yang P, Allen MS, Aubry MC, et al. Clinical features of 5,628 primary lung cancer patients: experience at Mayo Clinic from 1997 to 2003. *Chest* 2005;128:452–462. [PubMed: 16002972]
16. Yang P, Sun Z, Krowka MJ, et al. Alpha1-antitrypsin deficiency carriers, tobacco smoke, chronic obstructive pulmonary disease, and lung cancer risk. *Arch Intern Med* 2008;168:1097–1103. [PubMed: 18504338]
17. Hudmon KS, Honn SE, Jiang H, et al. Identifying and recruiting healthy control subjects from a managed care organization: a methodology for molecular epidemiological case-control studies of cancer. *Cancer Epidemiol Biomarkers Prev* 1997;6:565–571. [PubMed: 9264268]
18. Asomaning K, Miller DP, Liu G, et al. Second hand smoke, age of exposure and lung cancer risk. *Lung Cancer* 2008;61:13–20. [PubMed: 18191495]

19. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904–909. [PubMed: 16862161]
20. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;2:e190. [PubMed: 17194218]
21. Ballman KV, Grill DE, Oberg AL, Therneau TM. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics* 2004;20:2778–2786. [PubMed: 15166021]
22. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–575. [PubMed: 17701901]
23. Morley M, Molony CM, Weber TM, et al. Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004;430:743–747. [PubMed: 15269782]
24. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* 2008;24:408–415. [PubMed: 18597885]
25. Landi MT, Dracheva T, Rotunno M, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One* 2008;3:e1651. [PubMed: 18297132]
26. Su LJ, Chang CW, Wu YC, et al. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics* 2007;8:140. [PubMed: 17540040]
27. Powell CA, Spira A, Derti A, et al. Gene expression in lung adenocarcinomas of smokers and nonsmokers. *Am J Respir Cell Mol Biol* 2003;29:157–162. [PubMed: 12600827]
28. Bjartveit K, Tverdal A. Health consequences of sustained smoking cessation. *Tob Control* 2009;18:197–205. [PubMed: 19228666]
29. Ebbert JO, Yang P, Vachon CM, et al. Lung cancer risk reduction after smoking cessation: observations from a prospective cohort of women. *J Clin Oncol* 2003;21:921–926. [PubMed: 12610194]
30. Speizer FE, Colditz GA, Hunter DJ, Rosner B, Hennekens C. Prospective study of smoking, antioxidant intake, and lung cancer in middle-aged women (USA). *Cancer Causes Control* 1999;10:475–482. [PubMed: 10530619]
31. Lubin JH, Blot WJ, Berrino F, et al. Modifying risk of developing lung cancer by changing habits of cigarette smoking. *Br Med J (Clin Res Ed)* 1984;288:1953–1956.
32. Veugeliers M, Vermeesch J, Reekmans G, Steinfeld R, Marynen P, David G. Characterization of glypican-5 and chromosomal localization of human GPC5, a new member of the glypican gene family. *Genomics* 1997;40:24–30. [PubMed: 9070915]
33. Blackhall FH, Merry CL, Davies EJ, Jayson GC. Heparan sulfate proteoglycans and cancer. *Br J Cancer* 2001;85:1094–1098. [PubMed: 11710818]
34. Filmus J, Capurro M, Rast J. Glypicans. *Genome Biol* 2008;9:224. [PubMed: 18505598]
35. De Cat B, David G. Developmental roles of the glypicans. *Semin Cell Dev Biol* 2001;12:117–125. [PubMed: 11292377]
36. Filmus J. Glypicans in growth control and cancer. *Glycobiology* 2001;11:19R–23R.
37. Klaus A, Birchmeier W. Wnt signalling and its impact on development and cancer. *Nat Rev Cancer* 2008;8:387–398. [PubMed: 18432252]
38. Powers S, Mu D. Genetic similarities between organogenesis and tumorigenesis of the lung. *Cell Cycle* 2008;7:200–204. [PubMed: 18256532]
39. Reardon DA, Jenkins JJ, Sublett JE, Burger PC, Kun LK. Multiple genomic alterations including N-myc amplification in a primary large cell medulloblastoma. *Pediatr Neurosurg* 2000;32:187–191. [PubMed: 10940769]
40. Neat MJ, Foot N, Jenner M, et al. Localisation of a novel region of recurrent amplification in follicular lymphoma to an approximately 6.8 Mb region of 13q32–33. *Genes Chromosomes Cancer* 2001;32:236–243. [PubMed: 11579463]
41. Ojopi EP, Rogatto SR, Caldeira JR, Barbieri-Neto J, Squire JA. Comparative genomic hybridization detects novel amplifications in fibroadenomas of the breast. *Genes Chromosomes Cancer* 2001;30:25–31. [PubMed: 11107172]

42. Imoto I, Izumi H, Yokoi S, et al. Frequent silencing of the candidate tumor suppressor PCDH20 by epigenetic mechanism in non-small-cell lung cancers. *Cancer Res* 2006;66:4617–4626. [PubMed: 16651412]
43. Kirchhoff M, Bisgaard AM, Stoeva R, et al. Phenotype and 244k array-CGH characterization of chromosome 13q deletions: an update of the phenotypic map of 13q21.1-qter. *Am J Med Genet A* 2009;149:894–905. [PubMed: 19363806]
44. Baranzini SE, Wang J, Gibson RA, et al. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum Mol Genet* 2009;18:767–778. [PubMed: 19010793]
45. Byun E, Caillier SJ, Montalban X, et al. Genome-wide pharmacogenomic analysis of the response to interferon beta therapy in multiple sclerosis. *Arch Neurol* 2008;65:337–344. [PubMed: 18195134]
46. Gieger C, Geistlinger L, Altmaier E, et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet* 2008;4:e1000282. [PubMed: 19043545]
47. Dimas AS, Deutsch S, Stranger BE, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 2009;325:1246–1250. [PubMed: 19644074]

Figure 1. Study design and key findings

eQTL=expression quantitative trait loci. SNP=single nucleotide polymorphism.

MDACC=MD Anderson Cancer Center. UCLA=University College Los Angeles.

Figure 2. Scatter plot of chromosome position (x axis) against $-\log_{10}(\text{p value})$ (y axis) from conditional logistic regression analyses adjusted for history of chronic obstructive pulmonary disease, exposure to second-hand smoke at adulthood, and family history of lung cancer
 (A) Black arrows indicate two SNPs: rs11183940 ($p=1.5 \times 10^{-6}$) and rs10880785 ($p=7.1 \times 10^{-6}$) on chromosome 12 with strongest significant association in the Mayo genome-wide association study.
 (B) The arrow indicates two SNPs: rs2352028 ($p=4.3 \times 10^{-5}$) and rs2352029 ($p=5.3 \times 10^{-5}$) on chromosome 13 replicated in the MD Anderson Cancer Center study.

Figure 3. Association of rs2352028 with lung cancer risk overall (A) and in adenocarcinoma only (B) in never smokers in four independent studies

Black squares indicate the odds ratios, with the size of the square inversely proportional to its variance. Horizontal lines represent 95% CI. The combined results under each genetic model are indicated by the black diamond. Heterogeneity test p values were calculated using a Cochran's Q test. PAR%=percentage of population attributable risk. MAF=minor allele frequency (case/control). MDACC=MD Anderson Cancer Center. UCLA=University College Los Angeles.

Figure 4. Expression levels of *GPC5* in normal lung tissues based on the genotype of rs2352028 and rs2352029

Gene expression is in normalised units. Centre lines are the ranges of expression levels; boxes indicate IQR. The p value was calculated with a linear regression model adjusted for age and sex.