Carcinogenesis, 2018, Vol. 39, No. 9, 1135–1140

doi:10.1093/carcin/bgy080
Advance Access publication June 19, 2018
Original Article

ORIGINAL ARTICLE

Genome-wide association study of familial lung cancer

Jinyoung Byun[1,†], Ann G. Schwartz[2,†], Christine Lusk[2], Angela S. Wenzlaff[2], Mariza de Andrade[3], Diptasri Mandal[4], Colette Gaba[5], Ping Yang[3], Ming You[6], Elena Y. Kupert[6], Marshall W. Anderson[6], Younghun Han[1], Yafang Li[1], David Qian[1], Adrienne Stilp[7], Cathy Laurie[7], Sarah Nelson[7], Wenying Zheng[7], Rayjean J. Hung[8], Valerie Gaborieau[9], James Mckay[9], Paul Brennan[9], Neil E. Caporaso[10], Maria Teresa Landi[10], Xifeng Wu[11], John R. McLaughlin[12], Yonathan Brhane[7], Yohan Bossé[13], Susan M. Pinney[14], Joan E. Bailey-Wilson[15] and Christopher I. Amos[1,16,*]

[1]Department of Biomedical Data Science, Dartmouth Geisel School of Medicine, Lebanon, NH 03756, USA, [2]Karmanos Cancer Institute, Wayne State University, Detroit, MI 48201, USA, [3] Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA, [4] Louisiana State University Health Sciences Center, New Orleans, LA 70112, USA, [5]University of Toledo Dana Cancer Center, Toledo, OH 43606, USA, [6]Medical College of Wisconsin, Milwaukee, WI 53226, USA, [7]Genetic Analysis Center, University of Washington, Seattle, WA 98195, USA, [8]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada, [9]Genetic Epidemiology Group, International Agency for Research on Cancer (IARC), 69372 Lyon, France, [10]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA, [11]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, [12]Public Health Ontario, Toronto, Canada, [13]Institut Universitaire de Cardiologie et de Pneumologie de Québec, Department of Molecular Medicine, Laval University, Québec, Canada, [14] University of Cincinnati College of Medicine, Cincinnati, OH 45267, USA, [15]National Human Genome Research Institute, National Institutes of Health, Baltimore, MD 21224, USA and [16]Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX 77030, USA

[*] To whom correspondence should be addressed. Tel: +1 603 650 1972; Fax: +1 603 653 6696; Email:chrisa@bcm.edu
Correspondence may also be addressed to Ann Schwartz Tel: +1 313 578 4201; Fax: +1 313 578 4306; Email:schwarta@karmanos.org
[†] These authors contributed equally to this work.

Abstract

To identify genetic variation associated with lung cancer risk, we performed a genome-wide association analysis of 685 lung cancer cases that had a family history of two or more first or second degree relatives compared with 744 controls without lung cancer that were genotyped on an Illumina Human OmniExpressExome-8v1 array. To ensure robust results, we further evaluated these findings using data from six additional studies that were assembled through the Transdisciplinary Research on Cancer of the Lung Consortium comprising 1993 familial cases and 33 690 controls. We performed a meta-analysis after imputation of all variants using the 1000 Genomes Project Phase 1 (version 3 release date September 2013). Analyses were conducted for 9 327 222 SNPs integrating data from the two sources. A novel variant on chromosome 4p15.31 near the *LCORL* gene and an imputed rare variant intergenic between *CDKN2A* and *IFNA8* on chromosome 9p21.3 were identified at a genome-wide level of significance for squamous cell carcinomas. Additionally, associations of *CHRNA3* and *CHRNA5* on chromosome 15q25.1 in sporadic lung cancer were confirmed at a genome-wide level of significance in familial lung cancer. Previously identified variants in or near *CHRNA2, BRCA2, CYP2A6* for overall lung cancer, *TERT, SECISPB2L* and *RTEL1* for adenocarcinoma and *RAD52* and *MHC* for squamous carcinoma were significantly associated with lung cancer.

Received: December 20, 2017; Revised: April 12, 2018; Accepted: June 18, 2018

Abbreviations

eQTL	expression quantitative trait loci
EAGLE	Environment And Genetics in Lung cancer Etiology
GELCC	The Genetic Epidemiology of Lung Cancer Consortium
GWAS	Genome-wide association studies
ILCCO	International Lung Cancer Consortium
IARC	International Agency for Research on Cancer
MDACC	MD Anderson Cancer Center
PLCO	Prostate, Lung, Colon, and Ovary screening studies
SNP	single nucleotide polymorphism
TRICL	Transdisciplinary Research in Cancer of the Lung

Introduction

Lung cancer is estimated to account for over 13% of all new cancer diagnoses and almost 27% of all cancer deaths (http://SEER.cancer.gov/statfacts/html.lungb.html). Despite diagnostic advances in recent years, lung cancer usually presents at advanced stages and mortality remains high. While most lung cancer is attributable to tobacco smoke exposure, many studies have identified genetic variation that influences lung cancer risk (1–12). Because cancers arise through an accumulation of deleterious somatic mutations, environmental or inherited risk factors due to inherited deleterious germline genetic variants are critical (12).

The Genetic Epidemiology of Lung Cancer Consortium (GELCC) has conducted the only family-based lung cancer linkage study to identify rare, highly penetrant lung cancer susceptibility genes, providing evidence of linkage on chromosomal regions 6q23-25, 18p11.23, 2p22.2, 14q13.1, 16p13 and 20q13.11 (4,5,10–13). Linkage analysis is designed to identify highly penetrant uncommon variants but does not identify more common variants. Genome-wide association studies (GWAS) have provided highly significant and reproducible results associating common genetic variations to lung cancer. For example, a locus on chromosome 15q25, spanning a neuronal nicotinic acetylcholine receptor gene cluster comprising *CHRNA3*, *CHRNA5* and *CHRNB4* subunits, is significantly associated with lung cancer risk (5,7,14). Regions on chromosomes 6p21 (*BAG6, MSH5*), 5p15 (*TERT, CLPTM1L*), 13q (*BRCA2*) and 22q (*CHEK2*) have also been identified from GWAS as being consistently associated with histology-specific lung cancer risk (5,6,11,15).

To further identify genetic variation associated with familial lung cancer risk, we conducted a GWAS meta-analysis focused on familial lung cancer cases. Familial lung cancer cases used in the analysis came from the GELCC family-based study and from the Transdisciplinary Research in Cancer of the Lung (16) and the International Lung Cancer Consortium (TRICL-ILCCO) (11). Lung cancer genome-wide association studies enriched for familial lung cancer cases represent a novel application of GWAS methodology to a population most likely to have a genetic component to their risk. Selecting cases with a family history improves the power to detect uncommon variants that are likely to have a larger inherited impact on lung cancer risk (12,17).

Materials and methods

GELCC study subjects

The GELCC enrolled lung cancer cases with a family history of lung cancer from the following data collection sites: Karmanos Cancer Institute/Wayne State University, University of Cincinnati College of Medicine, Mayo Clinic, Medical College of Toledo, and Louisiana State University Health Sciences Center—New Orleans. Local Institutional Review Boards approved all protocols and each participant provided written informed consent. Familial lung cancer cases (ICD-9 162.0–162.9) were defined as individuals with a histologically confirmed diagnosis of lung cancer and at least one first or second degree relative with lung cancer. Family history information was collected by interview. The diagnosis of lung cancer in both the index case and relatives was verified with medical records, tumor registries and death certificates when possible. For the GWAS, one lung cancer case from each of 685 families of European descent with at least two first or second-degree relatives with lung cancer was selected. 744 controls of European descent were selected from persons related by marriage to lung cancer cases or from the general population in which the cases were identified, and did not have a lung cancer diagnosis and had no first or second-degree relatives with lung cancer (Supplementary Table 1, available at *Carcinogenesis* Online). Controls were frequency matched to cases on age (5-year groups), sex and smoking status as closely as possible. The detailed demographic information on GELCC study is presented in Supplemented Table 2, available at *Carcinogenesis* Online.

Participation included completing a questionnaire (either by phone, in the clinic, or by mail), providing a blood or buccal/saliva sample for DNA, and providing access to medical records for confirmation of diagnosis and histology type. The study questionnaire included basic demographics, health history including history of other lung diseases, smoking history, and family history of cancer. Clinical data about diagnostic procedures, stage of disease, and histology type were obtained from medical records.

GELCC study genotyping

DNA samples were obtained from several sources including blood (68.2%), mouthwash (0.3%), saliva (2.3%) and whole genome amplification from blood (29.2%). Germline DNA was genotyped at the Center for Inherited Disease Research (CIDR) using the Illumina Human OmniExpressExome-8v1 array, designed to human genotype build 37. The GenTrain Version 1.0 calling algorithm in Genome Studio version 2011.1 was used, as was zCall for post-processing rare variant calling. Subsequent quality control steps were conducted collaboratively between the Genetic Analysis Center at the University of Washington in Seattle and the statistical group at Dartmouth College. The array included 18 642 pairs and 51 trios of duplicated SNPs. Where SNPs were duplicated, the SNP with the lowest missing rate was retained for analysis. In the analysis of the array, 49 blind duplicate samples were analyzed of which 31 were study-specific duplicates and 18 were derived from HapMap referent panel. One HapMap control and one internal control were included on each 96-well plate that was analyzed. The median discordance rate for genotypes for these 49 blind duplicates was 1.7×10^{-5}. The genotyped HapMap samples included five trios, and of 936 945 unique markers, there were 1008 markers with one Mendelian error and 75 with two or more errors all of which were excluded from further analysis. A selected set of 907 markers from the whole genome amplified samples analyzed at the Mayo Clinic, one of the participating sites, provided different genotype frequencies from other participating sites. Results from the Mayo Clinic for those specific SNPs were removed from analyses. The median call rate for genotypes on this array was 99.9% and the error rate assessed from 31 pairs of study sample duplicates was 8.5×10^{-6}. All genotyped samples had a call rate of 97% or higher and were retained for further quality control checks. Relatively few SNPs were removed due to quality control steps, but on this array 14.47% of SNPs were monomorphic and were therefore removed from analyses (Supplementary Table 3, available at *Carcinogenesis* Online). Of the 951,117 SNPs assayed, 29% had minor allele frequencies of 1% or less. The detailed QC processes are presented in Supplementary Figure 1, available at *Carcinogenesis* Online.

Imputation in GELCC study

We performed genotype imputation in the familial lung cancer study, using the software package IMPUTE2 v2.3.0 (18) and 1000 Genomes Project Phase 1 as a reference panel, which contains whole genome sequence data from 1092 individuals including approximately 39 million variants,

insertions/deletions, and structural variants. Imputation was done to 1000 Genomes Project Phase 1 version 3 integrated haplotypes produced with SHAPEIT2 for chromosome 1–22; for chromosome X, the previous version was used. The study samples were imputed in a two-stage procedure using SHAPEIT2 (version 2.r644, c2011-2012) to derive phased genotypes from the array SNPs and IMPUTE2 (version 2.3.0) to perform imputation of the phased data. We used the default parameters to derive phased genotypes with SHAPEIT2, increasing the number of conditioning haplotypes to be used during the phasing process to 200 ("--states 200 or –S 200"). The imputation was performed using 5MB segments over the length of each chromosome from the first to last position appearing in the reference panel. The flag "-use_prephased_g" was provided to indicate that prephased haplotypes were being used and the buffer region was increased to 500kb ("-buffer 500"). Imputation target variants were limited to those with at least two copies of the minor allele observed in any of the four reference panel super-populations, yielding 27 728, 264 total variants in the imputed dataset.

TRICL-ILCCO Data

The Transdisciplinary Research in Cancer of the Lung (16) and the International Lung Cancer Consortium (TRICL-ILCCO) is an international consortium that has assembled and analyzed GWAS data from lung cancer cases and controls. Among the studies that have collected data for TRICL-ILCCO, a subset had information about the presence of lung cancer in a first degree relative. From TRICL-ILCCO, 1,308 cases with a family history of lung cancer in a first degree relative from six studies were included (Supplementary Table 2, available at *Carcinogenesis* Online). An additional sample of 32 946 controls was available for analysis from the TRICL data. In all (GELCC plus TRICL-ILCCO), 1993 familial cases and 33 690 controls were included (Supplementary Table 1, available at *Carcinogenesis* Online). The same referent panel and procedures were used for the United States, Canadian and French sites for TRICL-ILCCO as were used for GELCC. The deCODE analysis included a referent set derived from Icelandic individuals who have been sequenced; it therefore provides more accurate imputation for rare variants than imputation analysis used by other sites.

Imputation in TRICL studies

Whole genome imputations for the MD Anderson Cancer Center (MDACC), deCODE, Samuel Lunenfeld Research Institute (SLRI) Study, Environmental and Genetics in Lung Cancer Etiology (EAGLE), and Prostate, Lung, Colon, and Ovary (PLCO) screening studies in TRICL-ILCCO consortium were conducted using IMPUTE2 v2.5 based on the 1000 Genomes Project Phase 1. For the International Agency for Research on Cancer (IARC) study imputation was performed using minimac (version 2012.10.3) based on 1000 Genome Project Phase 1 For MDACC, SLRI, EAGLE and PLCO studies, we used the default parameters to derive phased genotypes with SHAPEIT, increasing: the number of burn-in iterations used by the algorithm to reach a good starting point to 10 ("--burn 10"), the number of pruning iterations to find a parsimonious graph for each individual to 10 ("--prune 10"), and the number of iterations to compute transition probabilities in haplotype graphs to 50 ("--main 50"). We conducted imputation with IMPUTE2 using 5MB non-overlapping intervals. The flag "-use_prephased_g" was applied to indicate the prephased haplotypes were being used. The missing genotypes were replaced with imputed genotypes using the option "-pgs_miss" and the buffer region was increased to 500KB ("-buffer 500"). Supplementary Table 4, available at *Carcinogenesis* Online, provides the details for the number of SNPs that passed quality control on each study.

Statistical analysis

Using genotyped data from 1429 subjects of European descent in the GELCC study, we applied the EIGENSTRAT method to carry out principal components analysis. To assure robustness of our analytical approach, we performed association analyses adjusting for the top three eigenvectors based on the principal components analysis. Logistic regression was applied to evaluate the odds ratio per allele of each SNP, adjusting for age, sex and eigenvectors. Analyses were conducted using PLINK.

Meta-Analysis

Meta-analysis of the GELCC and TRICL-ILCCO studies was conducted using METAL software (http://csg.sph.umich.edu/abecasis/metal/index.html) that combined test statistics and standard errors across studies (Supplementary Table 5, available at *Carcinogenesis* Online). In addition, we determined whether there were differences between case groups according to the most common histologic subtypes of lung cancer: squamous (n = 367) and adenocarcinoma (n = 775). For the overall meta-analysis, GELCC and 6 TRICL-ILCCO studies were used (Supplementary Table 6A, available at *Carcinogenesis* Online). Histology specific meta-analysis was performed on GELCC and 5 TRICL-ILCCO studies (excluding the IARC study) (Supplementary Tables 6B and C, available at *Carcinogenesis* Online). For some histology specific analyses, some populations contributed small numbers of cases, a situation where rare variants can create unstable estimates of the effect size. Therefore, we excluded markers with imputation quality info score less than 0.5 and removed results when Odd Ratios (OR) were greater than 100 or less than 0.01 for each study to reduce false positive findings. Common SNPs that were present in at least five studies including GELCC plus at least four studies out of five TRICL-ILCCO studies were used for histology specific analyses. We also did not consider SNPs when the Mantel–Haenszel heterogeneity P-value test was ≤ 0.05. The lambda values showed minimal inflation except for squamous carcinoma, 1.015 and 1.004, respectively, before and after adjustment to reflect a standardized sample size of 1000 cases and 1000 controls (since lambda increases with sample size). Lambda values for histology-specific analyses were 1.040 and 1.026 (adenocarcinoma) and 1.079 and 1.108 (squamous cell) before and after correction based on 1000 cases and 1000 controls, respectively. Q-Q plots are presented in Supplementary Figure 2, available at *Carcinogenesis* Online. We used GCTA to evaluate joint and conditional models with chromosome 15q21.1 region markers (19,20).

Lung expression QTL analysis

Whole-genome gene expression profiling in 409 adjacent normal lung samples obtained from patients undergoing resection for lung cancer therapy was performed on a custom Affymetrix array (GPL10379). Microarray pre-processing and quality controls were described previously (21,22). Genotyping was carried out on the Illumina Human 1M-Duo BeadChip array. Expression traits were adjusted for age, sex and smoking status. *Cis*-eQTL were evaluated for transcripts (probe sets) located 1 Megabase up- and downstream of the sentinel lung cancer associated SNPs. Association tests were carried out using PLINK version 1.9. For the chromosome 9 region transcripts that were evaluated included *IFNA* gene cluster and the *CDKN2A/CDKN2B* region.

Results

In the meta-analysis of the GELCC and 6 TRICL-ILCCO studies, 44 and 41 variants (Supplementary Tables 7A and B, available at *Carcinogenesis* Online) were identified as associated with overall lung cancer and squamous cell carcinoma lung cancer (367 cases) in family history positive lung cancer cases, respectively, at a genome-wide level of significance ($P < 5 \times 10^{-8}$) (Supplementary Figure 3, available at *Carcinogenesis* Online). No variants were identified for adenocarcinoma (775 cases) at a genome-wide significant level. In the analysis of overall lung cancers, the top hits ($P < 5 \times 10^{-8}$) fall primarily within the *CHRNA3/CHRNA5/CHRNB4* nicotinic acetylcholine receptor gene region on chromosome 15 (Supplementary Figure 4A, available at *Carcinogenesis* Online) (23). The most significant SNP on chromosome15, rs8040868, is in the region of *CHRNA3*. As presented in Table 1, the estimate of risk associated with this relatively common SNP (MAF = 0.42) in familial lung cancer is OR = 1.33 ($P = 4.38 \times 10^{-13}$). Even though this is a synonymous change, statistically significant changes in gene expression for *CHRNA3*, as well as *CHRNA5* and *PSMA4*, are associated with this polymorphism, based on expression quantitative trait analysis (eQTL) from 409 noncancerous lung tissues removed for lung cancer resections, providing additional

Table 1. Meta-analysis association test results for top SNPs, by histology type

Strata	SNP	Chr	Position	Nearest Gene	Function	Allele[a]	EAF	OR (95% CI)	*P*-value	Mean of QC Info Score
Overall	rs8040868	15q25.1	78911181	*CHRNA3*	Synonymous (V-V)	T_C	0.42	1.33 (1.23, 1.43)	4.38×10^{-13}	0.99
SQC	rs71603396	4p15.31	18012035	*LCORL*	Intron	G_A	0.14	1.92 (1.55, 2.37)	1.73×10^{-9}	0.88
	rs12686364	9p21.3	21670087	*CDKN2A*	Intergenic	A_G	0.004	8.49 (4.13, 17.46)	6.21×10^{-9}	0.96

OR, odds (log additive) ratio; 95% CI, 95% confidence interval; SQC, squamous cell carcinoma.
[a]Denotes reference_effect(coded) allele; Genome positions relative to GRCh37; EAF, effective allele frequency.

support for the role of this polymorphism or another variant in linkage disequilibrium with it in influencing lung cancer risk (Table 2; Supplementary Figure 5, available at *Carcinogenesis* Online).

In histology-specific analyses, two of the most significant associations were with familial lung squamous cell carcinoma. The first, rs71603396 on chromosome 4p15.31 is in the region of the ligand dependent nuclear receptor corepressor (*LCORL*) gene (MAF = 0.14, OR = 1.92, $P = 1.73 \times 10^{-9}$) (Supplementary Figure 4B, available at *Carcinogenesis* Online). This SNP rs71603396 in the *LCORL* gene was associated with gene expression in eQTL analysis (Table 2). The second significant SNP, rs12686364, is located in an intergenic region (Supplementary Figure 4C, available at *Carcinogenesis* Online, on chromosome 9p21.3 (MAF = 0.004, OR = 8.49, $P = 6.21 \times 10^{-9}$) that is upstream of *IFNA8* and downstream of *CDKN2A*). eQTL analysis demonstrated an association between this SNP and expression of interferon-α8 (*IFNA8*) (Table 2). Results by study center are presented in Supplementary Table 8, available at *Carcinogenesis* Online.

Table 3 presents results for previously identified SNPs that influence lung cancer risk in European descent individuals and that had previously reached genome-wide levels of significance. Of the previously identified SNPs, only the rs55781567 variant on chromosome 15q25.1 reached a genome-wide level of significance. The nicotinic acetylcholine region of chromosome 15q25.1 has consistently been the most strongly associated region for lung cancer risk among unselected cases of European descent lung cancer. We further examined the two SNPs on 15q25.1; rs8040868 from familial lung cancer and rs55781567 from unselected lung cancer, which are 53,195 bp apart. The linkage disequilibrium between the two SNPs is D' = 0.95 (R^2 = 0.73), indicated that these SNPs are in high linkage disequilibrium. We, therefore, performed the conditional analysis to evaluate whether rs8040868 ($P = 4.38 \times 10^{-13}$) has association signal conditioning on rs55781567 ($P = 1.69 \times 10^{-11}$) (22). The conditional test of rs8040868 after adjusting for rs55781567 yielded $P = 3.4 \times 10^{-3}$. This indicates that rs8040868 still has association signal for familial lung cancer risk. While rs8040868 provided an independent signal, many SNPs in this region are strongly correlated and the signal from rs804868 could reflect effects from another causal variant in the region such as rs16969968, a causal SNP in CHRNA5. Conditional analysis of rs804868 allowing for effects from rs16969968 also showed a residual effect with $P = 0.002$. On chromosome 9p21.3, the effects between the previously identified SNP rs885518 and the newly identified SNP rs12686364 was very low (R^2 = 0.007, D' = 0.452), with the very low R^2 value reflecting very different allele frequencies of the two variants. After conditioning on the previously identified SNP rs885518, which predisposes to adenocarcinoma, the association of rs12686364 remained highly significant ($P = 5.93 \times 10^{-9}$), supporting completely independent effects of these two SNPs on lung cancer risk.

Supplementary Tables 6A–C, available at *Carcinogenesis* Online, present results stratified by analysis in GELCC alone, TRICL alone or combined. Results for the GELCC study only identified several additional variants that were not replicated on analysis of the data from the TRICL studies. Variants in *BCHE, PDGFRA* and *CFTR* genes were highly significant ($P < 5 \times 10^{-7}$) for overall lung cancer, but all protective of developing lung cancer and did not replicate in the TRICL analysis. Similarly, highly significant associations ($P < 5 \times 10^{-7}$) were noted for adenocarcinoma including the *GSG1L* and *ZC3H12D* genes and for squamous associations were noted for RP11-335M9.1, EPDR1, CNTN4 and a gene desert on chromosome 14, but none of these associations were supported when combined with the TRICL data.

Discussion

Prior studies of familial lung cancer have identified rare variants that substantially increase lung cancer risk (24,25). Common genetic variants have been associated with sporadic lung cancer risk in GWAS studies (26). The study presented is the largest study to use a GWAS approach to more broadly identify genetic risk in familial lung cancer. Selecting cases with a family history of disease increases the power to detect rare variants that increase cancer risk (27). This study of familial lung cancer confirmed genetic variation on *CHRNA3* and *CHRNA5* on chromosome 15q25.1 previously reported in sporadic lung cancer. For all lung cancers, the top hits ($P < 5 \times 10^{-8}$) fall primarily within the *CHRNA3/CHRNA5/CHRNB4* nicotinic acetylcholine receptor regions on chromosome 15. The most significant SNP on chromosome15, rs8040868, has shown to be associated with lung cancer risk regardless of family history (16) (Supplementary Figure 4A, available at *Carcinogenesis* Online). In European descent individuals, rs8040868 is very strongly associated (R^2 = 0.78) with the causal variant rs16969968 that changes signal transduction of *CHRNA5* (28,29) The variant rs55781567 that was previously identified in unselected lung cancer cases (11) as most strongly associated with lung cancer SNP was also highly significant in our study (MAF = 0.37; OR = 1.31, $P = 1.69 \times 10^{-11}$). This variant was previously shown to tag an insertion/deletion polymorphism in the promoter of *CHRNA5* that affects its expression (30,31). The estimate of risk we report in familial lung cancer is similar to risk estimates for SNPs in the chromosome 15q25.1 region reported for sporadic lung cancer (5,7).

Two of the most significant associations were with familial lung squamous cell carcinoma. The first SNP, rs71603396 on chromosome 4p15.31, is in the region of the *LCORL* gene. The association between familial lung cancer risk and this SNP in the *LCORL* gene was also supported by the eQTL analysis. *LCORL* regulates transcription from RNA polymerase II and mutations in *LCORL* have frequently been noted in all histologies of lung cancer. In TCGA data, deletions of *LCORL* were observed in 0.5% of squamous lung cancers (32) and *LCORL* is often deleted

Table 2. Gene expression QTL analysis in the lung for the three top SNPs

SNP	Chr	Major allele	Minor allele	MAF	# probes tested	ProbeSet	Gene	BETA	SE	*P*-value[a]	% variance
rs8040868	15	T	C	0.474	45	100154936_IF147302_TGI_at	*AF147302*	−0.235	0.023	7.84×10^{-22}	20.3
						100154936_IREB2_TGI_at	*IREB2*	−0.244	0.027	4.23×10^{-18}	16.9
						100156434_TGI_at	*CHRNA5*	−0.121	0.019	3.43×10^{-10}	9.3
						100148404_TGI_at	*PSMA4*	0.059	0.014	1.9×10^{-5}	4.4
						100123649_TGI_at	*CHRNA3*	−0.151	0.039	1.5×10^{-4}	3.5
						100312267_TGI_at	*AJ584709*	−0.109	0.032	0.001	2.7
						100143446_TGI_at	*IREB2*	−0.066	0.023	0.005	1.9
						100145993_TGI_at	*CHRNA3*	0.037	0.015	0.012	1.5
						100154605_TGI_at	*CRABP1*	0.078	0.035	0.025	1.3
						100142282_TGI_at	*TBC1D2B*	−0.027	0.013	0.031	1.2
						100135897_TGI_at	Not known	0.071	0.033	0.032	1.1
						100126596_TGI_at	*DNAJA4*	0.057	0.028	0.044	1.0
rs71603396	4	G	A	0.113	21	100160526_TGI_at	*MED28*	−0.063	0.026	0.01498	1.5
						100142682_TGI_at	*LF383968*	−0.128	0.054	0.01769	1.4
						100158710_TGI_at	*BX648830*	0.082	0.039	0.03589	1.1
rs12686364	9	A	G	0.0049	38	100144400_TGI_at	*IFNA8*	−0.390	0.191	0.04117	1.0

[a]Only genes with statistically significant changes in gene expression associated with SNP genotypes are presented.

Table 3. Comparison to familial lung cancer study using the recently published susceptible loci in OncoArray lung cancer

Strata	Chromosome	SNP	Position	Gene	Allele[a]	EAF	OR (95%CI)	*P*-value
Lung	1p31.1	rs71658797	77967507	*FUBP1*	T_A	0.099	1.12 (0.99–1.27)	7.54×10^{-2}
Lung	6q27	rs6920364	167376466	*RNASET2*	G_C	0.449	1.08 (1.00–1.17)	5.25×10^{-2}
Lung	**8p21.1**	**rs11780471**	**27344719**	*EPHX2, CHRNA2*	**G_A**	**0.063**	**0.80 (0.67–0.95)**	$\mathbf{1.00 \times 10^{-2}}$
Lung	**13q13.1**	**rs11571833**	**32972626**	*BRCA2*	**A_T**	**0.010**	**1.66 (1.14–2.41)**	$\mathbf{7.78 \times 10^{-3}}$
Lung	15q21.1	rs66759488	47577451	*SEMA6D*	G_A	0.362	1.05 (0.97–1.14)	1.91×10^{-1}
Lung	**15q25.1**	**rs55781567**	**78857986**	*CHRNA5*	**C_G**	**0.366**	**1.31 (1.21–1.41)**	$\mathbf{1.69 \times 10^{-11}}$
Lung	**19q13.2**	**rs56113850**	**41353107**	*CYP2A6*	**C_T**	**0.429**	**0.91 (0.83–0.99)**	$\mathbf{2.86 \times 10^{-2}}$
Adeno	3q28	rs13080835	189357199	*TP63*	G_T	0.489	0.96 (0.86–1.08)	5.28×10^{-1}
Adeno	**5p15.33**	**rs7705526**	**1285974**	*TERT*	**C_A**	**0.343**	**1.28 (1.13–1.46)**	$\mathbf{1.45 \times 10^{-4}}$
Adeno	8p12	rs4236709	32410110	*NRG1*	A_G	0.225	1.10 (0.96–1.26)	1.56×10^{-1}
Adeno	**9p21.3**	**rs885518**	**21830157**	*MTAP/CDNK2A*	**A_G**	**0.110**	**1.37 (1.16–1.62)**	$\mathbf{2.20 \times 10^{-4}}$
Adeno	10q24.3	rs11591710	1.06E+08	*OBFC1*	A_C	0.143	1.14 (0.97–1.33)	1.21×10^{-1}
Adeno	11q23.3	rs1056562	1.18E+08	*MPZL3/AMICA1*	C_T	0.471	1.08 (0.96–1.21)	1.89×10^{-1}
Adeno	**15q21.1**	**rs77468143**	**49376624**	*SECISBP2L*	**T_G**	**0.249**	**0.87 (0.76–0.99)**	$\mathbf{4.00 \times 10^{-2}}$
Adeno	**20q13.33**	**rs41309931**	**62326579**	*RTEL1*	**G_T**	**0.119**	**1.35 (1.14–1.60)**	$\mathbf{6.14 \times 10^{-4}}$
SQC	**6p21.33**	**rs116822326**	**31434111**	*MHC*	**A_G**	**0.161**	**1.26 (1.03–1.54)**	$\mathbf{2.36 \times 10^{-2}}$
SQC	**12p13.33**	**rs7953330**	**998819**	*RAD52*	**G_C**	**0.313**	**0.79 (0.66–0.94)**	$\mathbf{7.05 \times 10^{-3}}$
SQC	22q12.1	rs17879961	29121087	*CHEK2*	A_G	0.004	2.50 (0.76–8.15)	1.29×10^{-1}

EAF, effective allele frequency; OR, odds (log additive) ratio; 95% CI, 95% confidence interval; Adeno, adenocarcinoma; SQC, squamous cell carcinoma.
[a]Denotes reference_effect(coded) allele. Variants in bold are associated with familial lung cancer at nominal significance level of 0.05. Genome positions relative to GRCh37.

in pancreatic cancers (5.5% of cases) (33). The second significant SNP associated with familial squamous cell lung cancer, rs12686364, is located in an intergenic region on chromosome 9p21.3, upstream of *IFNA8* and downstream of *CDKN2A*. eQTL analysis again provided support for this association. Published results show that *IFNA8* was deleted in 30% of glioblastomas (34), 24.8% of pancreatic cancers (33), and deleted or mutated in 12.9% of squamous lung cancers (32). Deletions of *IFNA8* are associated with improved survival in glioma patients (35,36).

Our results were also compared to the recently published variants associated with risk of sporadic lung cancer from a very large study (Table 3) (11). There is some heterogeneity in findings for familial lung cancer compared with those previously reported for the largest study to date of unselected lung cancer cases. In particular, variants in *TP63* have been reproducibly associated with lung adenocarcinoma risk, but showed no association in this study. The rare variant rs17879961 of *CHEK2* in sporadic lung cancer is highly protective from lung cancer development but in these familial cancers, the association analysis yielded strongly positive association (odds ratio of 2.5). Given the small sample size, this association was not significant and larger studies are needed to evaluate this finding further. While this is the largest study to date of a GWAS for familial lung cancer, the sample size is still quite limited, which may largely explain why many known associations for sporadic lung cancer were not associated with significant findings in this study. Another limitation in this study is the difference in family history between the Genetic Epidemiology of Lung Cancer Consortium, which recruited families with two or more first or second degree relatives and the restriction to families with one or more first degree relatives for the other studies. However, in conducting this study we sought to identify reliable findings

that met genome wide significance and to be able to evaluate the findings in multiple independent studies.

In summary, a novel variant rs71603396 on chromosome 4p15.31 and an imputed rare variant rs12686364 on 9p21.3 were identified at a genome-wide level of significance for familial squamous cell carcinomas. Also, the associations of *CHRNA3* and *CHRNA5* on chromosome 15q25.1 were confirmed in this familial lung cancer data. The identification of high risk variants in individuals selected through a positive family history provides insights into the etiology of lung cancer and may improve our ability to target selected individuals for lung cancer screening.

Supplementary material

Supplementary materials can be found at *Carcinogenesis* online.

Funding

National Institute of Health (U01CA76293, U19CA148127, P30CA22453, HHSN26820100007C); The Intramural Research Program of the National Human Genome Research Institute; National Institute of Health (U19CA148127, CA148127S1) (to the Transdisciplinary Research for Cancer in Lung (TRICL)); Cancer Care Ontario Research Chair of Population Studies (to RH); Canadian Cancer Society Research Institute (no. 020214) (to the SLRI GWAS); Ontario Institute for Cancer Research (to RH); National Institute of Health (R01CA060691, R01CA87895) (to Detroit GELCC study). National Institute of Health -National Cancer Institute (CA77118, CA80127, CA84353) (to Mayo Clinic Lung Study); National Institute of Health (P30ES006096) (to Cincinnati study); Genotyping services for the GELCC study were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract numbers HHSN268200782096C and HHSN268201100011.

Acknowledgements

The authors would like to thank the staff at the Respiratory Health Network Tissue Bank of the FRQS for their valuable assistance with the lung eQTL dataset at Laval University.
Conflict of Interest Statement: None declared.

References

1. Sellers, T.A. *et al.* (1990) Evidence for mendelian inheritance in the pathogenesis of lung cancer. J. Natl. Cancer Inst., 82, 1272–1279.
2. Schwartz, A.G. *et al.* (1996) Familial risk of lung cancer among nonsmokers and their relatives. Am. J. Epidemiol., 144, 554–562.
3. Sellers, T.A. *et al.* (1994) Segregation analysis of smoking-associated malignancies: evidence for Mendelian inheritance. Am. J. Med. Genet., 52, 308–314.
4. Bailey-Wilson, J.E. *et al.* (2004) A major lung cancer susceptibility locus maps to chromosome 6q23-25. Am. J. Hum. Genet., 75, 460–474.
5. Amos, C.I. *et al.* (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat. Genet., 40, 616–622.
6. Wang, Y. *et al.* (2014) Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat. Genet., 46, 736–741.
7. Hung, R.J. *et al.* (2008) A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature, 452, 633–637.
8. Landi, M.T. *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, 679–691.
9. Shi, J. *et al.* (2012) Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma. Cancer Discov., 2, 131–139.
10. Amos, C.I. *et al.* (2010) A susceptibility locus on chromosome 6q greatly increases lung cancer risk among light and never smokers. Cancer Res., 70, 2359–2367.
11. McKay, J.D. *et al.*; SpiroMeta Consortium. (2017) Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat. Genet., 49, 1126–1132.
12. Musolf, A.M. *et al.* (2017) Familial lung cancer: a brief history from the earliest work to the most recent studies. Genes (Basel), 8, E36.
13. Musolf, A.M. *et al.* (2016) Parametric linkage analysis identifies five novel genome-wide significant loci for familial lung cancer. Hum. Hered., 82, 64–74.
14. Thorgeirsson, T.E. *et al.* (2008) A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature, 452, 638–642.
15. McKay, J.D. *et al.*; EPIC Study. (2008) Lung cancer susceptibility locus at 5p15.33. Nat. Genet., 40, 1404–1406.
16. Timofeeva, M.N. *et al.*; Transdisciplinary Research in Cancer of the Lung (TRICL) Research Team. (2012) Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. Hum. Mol. Genet., 21, 4980–4995.
17. Peng, G. *et al.* (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. Eur. J. Hum. Genet., 18, 111–117.
18. Howie, B. *et al.* (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet., 44, 955–959.
19. Yang, J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet., 88, 76–82.
20. Yang, J. *et al.*; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet., 44, 369–75, S1.
21. Hao, K. *et al.* (2012) Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet., 8, e1003029.
22. Nguyen, J.D. *et al.* (2014) Susceptibility loci for lung cancer are associated with mRNA levels of nearby genes in the lung. Carcinogenesis, 35, 2653–2659.
23. Johnson, A.D. *et al.* (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics, 24, 2938–2939.
24. Xiong, D. *et al.* (2015) A recurrent mutation in PARK2 is associated with familial lung cancer. Am. J. Hum. Genet., 96, 301–308.
25. Hwang, S.J. *et al.* (2003) Lung cancer risk in germline p53 mutation carriers: association between an inherited cancer predisposition, cigarette smoking, and cancer risk. Hum. Genet., 113, 238–243.
26. Bossé, Y. *et al.* (2018) A decade of GWAS results in lung cancer. Cancer Epidemiol. Biomarkers Prev., 27, 363–379.
27. Peng, B. *et al.* (2010) Power analysis for case-control association studies of samples with known family histories. Hum. Genet., 127, 699–704.
28. Doyle, G.A. *et al.* (2011) *In vitro* and ex vivo analysis of CHRNA3 and CHRNA5 haplotype expression. PLoS One, 6, e23373.
29. Bierut, L.J. *et al.* (2008) Variants in nicotinic receptors and risk for nicotine dependence. Am. J. Psychiatry, 165, 1163–1171.
30. Falvella, F.S. *et al.* (2013) Multiple isoforms and differential allelic expression of CHRNA5 in lung tissue and lung adenocarcinoma. Carcinogenesis, 34, 1281–1285.
31. Chen, D. *et al.* (2011) A sex-specific association between a 15q25 variant and upper aerodigestive tract cancers. Cancer Epidemiol. Biomarkers Prev., 20, 658–664.
32. Cancer Genome Atlas Network (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature, 517, 576–582.
33. Witkiewicz, A.K. *et al.* (2015) Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nat. Commun., 6, 6744.
34. Brennan, C.W. *et al.*; TCGA Research Network. (2013) The somatic genomic landscape of glioblastoma. Cell, 155, 462–477.
35. Kohanbash, G. *et al.* (2012) Differential activity of interferon-α8 promoter is regulated by Oct-1 and a SNP that dictates prognosis of glioma. Oncoimmunology, 1, 487–492.
36. Okada, H. *et al.* (2013) Integration of epidemiology, immunobiology, and translational research for brain tumors. Ann. N. Y. Acad. Sci., 1284, 17–23.