ORIGINAL ARTICLE

Rare Variants in Known Susceptibility Loci and Their Contribution to Risk of Lung Cancer

Yanhong Liu, PhD,[a,*] Christine M. Lusk, PhD,[b] Michael H. Cho, PhD,[c,†] Edwin K. Silverman, PhD,[c,†] Dandi Qiao, PhD,[c,†] Ruyang Zhang, PhD,[d] Michael E. Scheurer, PhD,[e] Farrah Kheradmand, PhD,[a,f] David A. Wheeler, PhD,[g] Spiridon Tsavachidis, MS,[a] Georgina Armstrong, MPH,[a] Dakai Zhu, MS,[a,h] Ignacio I. Wistuba, MD,[i] Chi-Wan B. Chow, MD,[i] Carmen Behrens, MD,[j] Claudio W. Pikielny, PhD,[k] Christine Neslund-Dudas, PhD,[l] Susan M. Pinney, PhD,[m] Marshall Anderson, PhD,[m] Elena Kupert, MS,[m] Joan Bailey-Wilson, PhD,[n] Colette Gaba, MPH,[o] Diptasri Mandal, PhD,[p] Ming You, MD,[q] Mariza de Andrade, PhD,[r] Ping Yang, PhD, MD,[r] John K. Field, PhD,[s] Triantafillos Liloglou, PhD,[s] Michael Davies, PhD,[s] Jolanta Lissowska, PhD,[t] Beata Swiatkowska, PhD,[u] David Zaridze, PhD,[v] Anush Mukeriya, PhD,[v] Vladimir Janout, PhD,[w] Ivana Holcatova, PhD, MD,[x] Dana Mates, PhD, MD,[y] Sasa Milosavljevic, PhD,[z] Ghislaine Scelo, PhD,[aa] Paul Brennan, PhD,[aa] James McKay, PhD,[aa] Geoffrey Liu, PhD,[bb] Rayjean J. Hung, PhD,[cc] David C. Christiani, MD,[d] Ann G. Schwartz, PhD,[b] Christopher I. Amos, PhD,[a,h] Margaret R. Spitz, MD[a]

[a]*Dan L. Duncan Comprehensive Cancer Center, Department of Medicine, Baylor College of Medicine, Houston, Texas*
[b]*Karmanos Cancer Institute, Wayne State University, Detroit, Michigan*
[c]*Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts*
[d]*Harvard University School of Public Health, Boston, Massachusetts*
[e]*Department of Pediatrics, Baylor College of Medicine, Houston, Texas*
[f]*Michael E. DeBakey Veterans Affairs Medical Center; Houston, Texas*
[g]*Department of Molecular and Human Genetics, Human Genome Sequence Center, Baylor College of Medicine, Houston, Texas*
[h]*Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, Texas*
[i]*Department of Translational Molecular Pathology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas*
[j]*Department of Thoracic/Head and Neck Medical Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas*
[k]*Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire*
[l]*Department of Public Health Sciences, Henry Ford health System, Detroit, Michigan*
[m]*University of Cincinnati College of Medicine, Cincinnati, Ohio*
[n]*National Human Genome Research Institute, Bethesda, Maryland*
[o]*The University of Toledo College of Medicine, Toledo, Ohio*
[p]*Louisiana State University Health Sciences Center, New Orleans, Louisiana*
[q]*Medical College of Wisconsin, Milwaukee, Wisconsin*
[r]*Mayo Clinic College of Medicine, Rochester, Minnesota*
[s]*Roy Castle Lung Cancer Research Programme, The University of Liverpool, Department of Molecular and Clinical Cancer Medicine, Liverpool, United Kingdom*
[t]*Maria Sklodowska-Curie Institute of Oncology Center, Warsaw, Poland*
[u]*Nofer Institute of Occupational Medicine, Department of Environmental Epidemiology, Lodz, Poland*
[v]*Russian N. N. Blokhin Cancer Research Centre, Moscow, Russian Federation*

*Corresponding author.

†Writing on behalf of The COPDGene Investigators.

Disclosures: The authors declare no conflict of interest.

Address for correspondence: Yanhong Liu, PhD, Baylor College of Medicine, 1 Baylor Plaza, TX 77030. E-mail: yl10@bcm.edu

ISSN: 1556-0864

https://doi.org/10.1016/j.jtho.2018.06.016

[w]Faculty of Health Sciences, Palacky University, Olomouc, Czech Republic
[x]Institute of Public Health and Preventive Medicine, Charles University, Second Faculty of Medicine, Prague, Czech Republic
[y]National Institute of Public Health, Bucharest, Romania
[z]International Organization for Cancer Prevention and Research, Belgrade, Serbia
[aa]International Agency for Research on Cancer, Lyon, France
[bb]Princess Margaret Cancer Center, Toronto, Ontario, Canada
[cc]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada

Received 12 April 2018; revised 6 June 2018; accepted 17 June 2018
Available online - 4 July 2018

ABSTRACT

Background: Genome-wide association studies are widely used to map genomic regions contributing to lung cancer (LC) susceptibility, but they typically do not identify the precise disease-causing genes/variants. To unveil the inherited genetic variants that cause LC, we performed focused exome-sequencing analyses on genes located in 121 genome-wide association study–identified loci previously implicated in the risk of LC, chronic obstructive pulmonary disease, pulmonary function level, and smoking behavior.

Methods: Germline DNA from 260 case patients with LC and 318 controls were sequenced by utilizing VCRome 2.1 exome capture. Filtering was based on enrichment of rare and potential deleterious variants in cases (risk alleles) or controls (protective alleles). Allelic association analyses of single-variant and gene-based burden tests of multiple variants were performed. Promising candidates were tested in two independent validation studies with a total of 1773 case patients and 1123 controls.

Results: We identified 48 rare variants with deleterious effects in the discovery analysis and validated 12 of the 43 candidates that were covered in the validation platforms. The top validated candidates included one well-established truncating variant, namely, BRCA2, DNA repair associated gene (*BRCA2*) K3326X (OR = 2.36, 95% confidence interval [CI]: 1.38–3.99), and three newly identified variations, namely, lymphotoxin beta gene (*LTB*) p.Leu87Phe (OR = 7.52, 95% CI: 1.01–16.56), prolyl 3-hydroxylase 2 gene (*P3H2*) p.Gln185His (OR = 5.39, 95% CI: 0.75–15.43), and dishevelled associated activator of morphogenesis 2 gene (*DAAM2*) p.Asp762Gly (OR = 0.25, 95% CI: 0.10–0.79). Burden tests revealed strong associations between zinc finger protein 93 gene (*ZNF93*), *DAAM2*, bromodomain containing 9 gene (*BRD9*), and the gene *LTB* and LC susceptibility.

Conclusion: Our results extend the catalogue of regions associated with LC and highlight the importance of germline rare coding variants in LC susceptibility.

Keywords: Exome sequencing; Rare variants; Lung cancer; Susceptibility loci

Introduction

Although almost 80% of lung cancers (LCs) are attributed to smoking, LC develops in only about 15% of smokers. Therefore, it remains of great importance to understand the genetic factors that contribute to LC risk. It is well recognized that tobacco-induced chronic obstructive pulmonary disease (COPD) is an important predictor of LC risk. Genome-wide association studies (GWASs) have identified 45 genome-wide significant loci for LC, 22 loci for COPD, 32 loci for smoking behavior (SM), and 63 loci for pulmonary function (PF) levels, totaling 121 unique susceptibility loci (Supplementary Table 1). Interestingly, there is considerable overlap among these susceptibility loci and genes for these phenotypes (LC, COPD, SM, and PF levels). For example, the 6p21-22, 15q24-25.1, and 19q13.2 regions are shared by all four phenotypes; 5p15.33, 6p21.32, 10q23.31, and 10q25 are shared by three phenotypes; and 15 loci shared by two phenotypes (see Supplementary Table 1).

Although GWASs have been successful in identifying common (minor allele frequency [MAF] >5%) variants of small effect, the overall amount of LC heritability explained by these known common variants remains small. Further, because the tag single-nucleotide polymorphisms (SNPs) utilized in GWASs are used to identify genomic regions of interest rather than being selected for causality, identification of the functional variant at a specific locus generally poses a significant challenge. For example, of the 93 common LC GWAS top hits from the 45 reported susceptibility loci,[1] only two are protein coding (cholinergic receptor nicotinic alpha 3 subunit gene (*CHRNA3*) p.Tyr215 and cholinergic receptor nicotinic alpha 5 subunit gene (*CHRNA5*) p.Asp398Asn), with the remaining 91 variants falling in noncoding regions (four in untranslated regions, seven in flanking regions, 70 in intron regions, and 10 in intergenic regions). Alleles that are functionally deleterious will tend to be underrepresented at high frequencies, which is an assertion that is supported by the observation of a relationship between putative functionality and MAF. Recent studies suggest that multiple low-frequency (1% < MAF < 5%) or rare (MAF <1%) variants exhibit stronger effect sizes (ORs) than common variants do and contribute to the missing heritability.[2] Supporting this hypothesis is the observation that several genes containing known low-frequency or rare variants with moderate to large effect are associated with LC (e.g., poly(A)-specific ribonuclease gene [*PRKN*] p.Arg275Trp,[3] BRCA2, DNA repair associated gene [*BRCA2*] p.Lys3326X, checkpoint kinase 2 gene [*CHEK2*] p.Ile157Thr,[4] cilia and flagella associated protein 58 gene [*CCDC147*] p.Arg696Cys, and dopamine beta-hydroxylase gene [*DBH*] p.Val26Met).[5]

To unveil the inherited germline rare variants, we used whole exome sequencing with a focused analysis of the known 121 high-priority GWAS susceptibility loci (260 potential target genes). Smoking, family history of LC, and COPD are all well-documented risk factors for LC. To efficiently identify the most probable causative variants and genes, we have sequenced selected case patients with extreme phenotypes (high-risk patients with familial LC and sporadic patients reporting a history of heavy smoking and or severe COPD) and controls who report a history of heavy smoking but have normal spirometry results and are considered resistant to the effects of smoking.

Methods

Study Population in Discovery

Case patients with LC were derived from four independent case series, including sporadic cases from Baylor College of Medicine (BCM) (n = 68),[6–8] Harvard School of Public Health (n = 101), and M. D. Anderson Cancer Center (n = 37) and familial cases from the Genetic Epidemiology of LC Consortium (n = 54) (each familial case was chosen from one family considered at high risk of LC and having three or more affected first-degree members).[5,9] All case patients were white and had histologically confirmed NSCLC. Patient clinical and demographic information, such as smoking history (status and pack-years [PY]), was obtained by using self-administered questionnaires. For sporadic patients with LC, moderate-to-severe COPD phenotype was carefully defined by PF tests (reduced forced expiratory volume in 1 second [FEV_1] <80% predicted, and ratio of FEV_1 to forced vital capacity [FEV_1/FVC] <0.7). COPD phenotyping data were not available for familial cases.

The smoking controls were selected from two independent studies: (1) **the** Genetic Epidemiology of COPD Study (n = 298), which included 10,192 current or former smokers and which was a multicenter investigation to examine the genetic epidemiology of COPD and smoking-related lung diseases,[10] and (2) the BCM COPD and LC study (n = 20), which enrolled current or former smokers and was launched in 2002 within the Texas Medical Center in Houston, Texas.[6–8] All subjects underwent study-related testing that included spirometry, computed tomography scans of the chest, and blood collection. The controls were white, resistant smokers with normal PF data (defined as postbronchodilator FEV_1 ≥80% predicted and FEV_1/FVC ≥0.7) and at least a 10-PY history of smoking cigarettes.

DNA was isolated from peripheral blood or saliva both from patients with LC and from controls. The study was approved by the institutional review boards of all sites accruing participants and by the institutional review board at BCM for exome sequencing conducted at the Human Genome Sequencing Center (HGSC).

Library Preparation, Capture Enrichment, and Exome Sequencing

DNA samples were constructed into Illumina paired-end precapture libraries (Illumina, San Diego, CA) according to the manufacturer's protocol. The complete library and capture protocol, as well as the oligonucleotide sequences, have been described in detail previously.[11,12] For exome capture, each library pool was hybridized in solution to the BCM-HGSC–designed VCRome 2.1 probe set (Roche NimbleGen, Madison, WI) according to the manufacturer's protocol. This exome capture probe set targets the Vertebrate Genome Annotation, Consensus Coding Sequence Project, and RefSeq gene models, with 45.2 megabase capture targeting 23,585 genes. Exome sequencing was performed in paired-end mode with use of the Illumina HiSeq 2000 platform (Illumina). Sequencing runs generated approximately 300 to 400 million successful reads on each lane of a flow cell, yielding 7 to 13 gigabases per sample. For exome sequencing yields, samples achieved an average depth of coverage of 200× over exonic regions. Sequence analysis was performed by using the BCM-HGSC Mercury analysis pipeline.[13] All sequence reads were mapped to the GRCh37 human reference genome by using the Burrows-Wheeler aligner.[14] Putative variants, including single-nucleotide variants (SNVs) and insertions or deletions (indels), were called by using the Atlas2 suite.[15] Read qualities were recalibrated with the Genome Analysis Toolkit, and a minimum quality score of 30 was required; also, the variant must have been present in more than 15% of the reads that cover the position.

Single-Rare Variant Filtering and Functional Annotation

Our analysis was restricted to mapping of rare variants within the exonic regions of the 121 known GWAS loci (see Supplementary Table 1 for genomic coordinates and 260 target genes). Variants were annotated for effect on the protein and predicted function by using the SNP and Variation Suite (SVS, Golden Helix, Inc, Bozeman, MT). To identify pathogenic variants, a three-step filtering protocol was designed; it used automated filtering followed by manual review (Fig. 1). In step 1, automated filtering identified variants that fulfilled the following criteria: (1) mutation type, including missense and disruptive (defined as nonsense, stop-gain/loss, splice site destructions, and frameshift indels, which severely disrupt protein structure); (2) mutation effects (i.e., the variant was predicted to result in truncation of the protein or was predicted to be damaging/deleterious [not to be benign/tolerated] to the protein with use of the sorting intolerant from tolerant, PolyPhen-2, Mutation taster and scaled C-scores from the Combined Annotation-Dependent Depletion [CADD] method,[16] which strongly correlates with both molecular functionality and pathogenicity); and (3) MAF less than 1% in the Europeans in the reference databases, including the Exome Aggregation Consortium (ExAC), 1000 Genomes Project (1KGP), UK10K project, and UCSC Common SNPs tracks. Novel variants were defined as never having been reported in a publicly available database and the University of California Santa Cruz All SNPs 135/137/141 tracks. In step 2, after implementation of the aforementioned automated filtering schema, manual review of the raw BAM files were performed by using the GenomeBrowse (Golden Helix, Inc). This filter was used to remove the false-positive events that result from mapping errors and mutations found in a "noisy" background (multiple mismatches or indels in flanking sequences). These highly rare and predicted deleterious mutations were used to perform the gene-based burden analysis. In step 3, we further prioritized candidate variants that were highly enriched in the case patient group (risk alleles) or the control group (protective alleles).

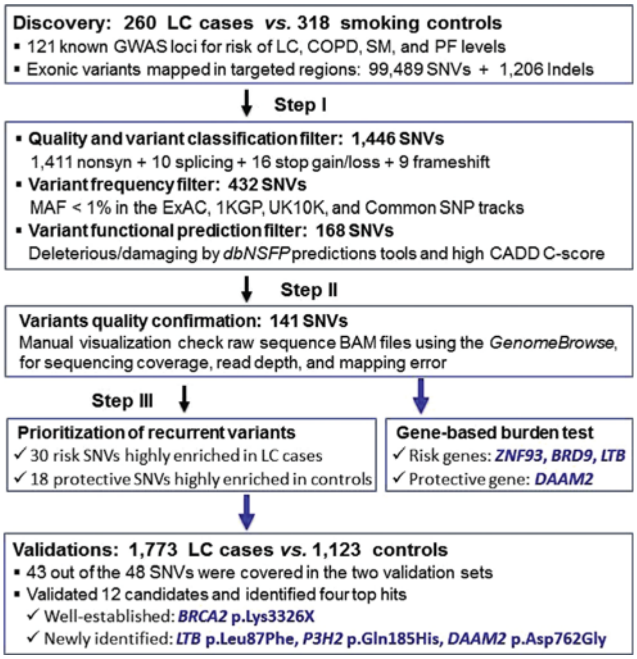
Discovery: 260 LC cases vs. 318 smoking controls
▪ 121 known GWAS loci for risk of LC, COPD, SM, and PF levels
▪ Exonic variants mapped in targeted regions: 99,489 SNVs + 1,206 Indels
Step I
▪ Quality and variant classification filter: 1,446 SNVs
1,411 nonsyn + 10 splicing + 16 stop gain/loss + 9 frameshift
▪ Variant frequency filter: 432 SNVs
MAF < 1% in the ExAC, 1KGP, UK10K, and Common SNP tracks
▪ Variant functional prediction filter: 168 SNVs
Deleterious/damaging by dbNSFP predictions tools and high CADD C-score
Step II
Variants quality confirmation: 141 SNVs
Manual visualization check raw sequence BAM files using the GenomeBrowse, for sequencing coverage, read depth, and mapping error
Step III
Prioritization of recurrent variants
✓ 30 risk SNVs highly enriched in LC cases
✓ 18 protective SNVs highly enriched in controls
Gene-based burden test
✓ Risk genes: ZNF93, BRD9, LTB
✓ Protective gene: DAAM2
Validations: 1,773 LC cases vs. 1,123 controls
▪ 43 out of the 48 SNVs were covered in the two validation sets
▪ Validated 12 candidates and identified four top hits
✓ Well-established: BRCA2 p.Lys3326X
✓ Newly identified: LTB p.Leu87Phe, P3H2 p.Gln185His, DAAM2 p.Asp762Gly

Figure 1. Workflow and annotation pipeline for the identification of candidate variants. LC, lung cancer; GWAS, genome-wide association studies; COPD, chronic obstructive pulmonary disease; SM, smoking behavior; PF, pulmonary function; SNV, single nucleotide variants; Indels, insertions or deletions; ExAC, Exome Aggregation Consortium; MAF, minor allele frequency; 1KGP, 1000 Genomes Project; SNP, single-nucleotide polymorphism; dbNSFP, database for non-synonymous SNPs functional predictions; CADD, Combined Annotation Dependent Depletion; *BAM*, binary alignment map file; *ZNF93*, zinc finger protein 93 gene; *BRD9*, bromodomain containing 9 gene; *LTB*, lymphotoxin beta gene; *DAAM2*, dishevelled associated activator of morphogenesis 2 gene.

Gene-Based Burden Analysis of Multiple Rare Deleterious Variants

To have greater power to detect significant associations with rare variants, we performed gene-based collapsing tests for those genes that included two or more rare and predicted deleterious variants (from filtering steps 1 and 2), including the combined multivariate and collapsing (CMC) test and the kernel-based adaptive cluster (KBAC) test.[17,18] To measure their cumulative effect, the CMC first bins variants according to MAF criterion (thresholds at 1%, 0.1%, and 0.01%) based on the observed data, then collapses the multiple variants within each bin, and finally uses Hotelling's T^2 to perform multivariate testing on the counts across the various bins. The KBAC test first counts multimarker genotypes within a given gene based on the variant data and then performs a special case-control test based on the weighted sum of these allele counts. To account for multiple comparisons, we calculated the false discovery rate (FDR)-adjusted *p* values.[19]

Study Population in Validation

To discover robust associations and validate the promising candidates, we analyzed two independent sets. The first was the Wayne State University (WSU) study, which enrolled participants at Karmanos Cancer Institute or Henry Ford Health System. Study participants either underwent spirometry or had PF test data abstracted from their medical records.[20] We carefully selected case patients with LC and a smoking history of at least 10 PY and smoking controls with normal PF (FEV_1 ≥80% predicted, and FEV_1/FVC ≥0.7) and a smoking history of at least 10 PY. Genotyping was performed by using the Illumina MEGA panel (Illumina), which includes more than 1.7 million variants. The second of the two independent sets was the Transdisciplinary Research in Cancer of the Lung Team of the

International Lung Cancer Consortium (TRICL-ILCCO) study. Subjects were selected from four sites in the TRICL-ILCCO study: Harvard School of Public Health, International Agency for Research on Cancer, University of Liverpool, and Mount Sinai Hospital and Princess Margaret Hospital in Toronto. Exome capture (with the Agilent SureSelect XT Custom ELID and Whole Exome v5, Agilent Technologies, Santa Clara, CA) and sequencing were performed at the Center for Inherited Disease Research. Both validation studies were approved by the institute ethics review committees, and all participants provided written informed consent.

Allelic Association Analysis in the Combined Data Sets

We then tabulated the minor allele and reference allele counts per candidate in the combined discovery and validation data sets and performed allelic association analysis, which compares frequencies of alleles in case patients with LC cases controls, and in case patients with LC versus in the ExAC reference population (non-Finnish Europeans [N = 33,370]). We note that the individuals in the reference set are not necessarily healthy—many have adult-onset diseases, such as type 2 diabetes and schizophrenia. Because the numbers of mutation carriers are small (fewer than 5), Fisher's exact tests were used for the allelic association analysis. ORs, 95% confidence intervals (CIs) and FDR-adjusted *p* values were calculated.

Results

Demographic and clinical information including age, sex, smoking history, histologic type, and PF data are summarized in Table 1. The discovery set included 260 cases of patients with LC (54 familial and 206 sporadic cases); in 75 of the 206 sporadic cases, patients also had moderate-to-severe COPD. In all, there were 318 smoking controls with a smoking history of 15 or more PY and normal PF data. The validation populations (WSU and TRICL-ILCCO populations) included 1773 case patients and 1123 controls. Among the combined 2033 case patients and 1441 controls of European descent, 129 case

Table 1. Basic Characteristics of LC Cases and Controls in the Discovery and Validations

	Discovery			Validation: WSU Study			Validation: TRICL-ILCCO Study		
Characteristics	Case Patients (n = 260)[a]	Smoking Controls (n = 318)[b]	*p* Value[c]	Case Patients (n = 831)	Smoking Controls (n = 266)[b]	*p* Value[c]	Case Patients (n = 942)	Controls (n = 857)[b]	*p* Value[c]
Age, y									
Mean (SD)	64 (6.3)	62.6 (4.9)	0.258	63.6 (9.8)	59.5 (9.1)	<0.001	62.4 (12.3)	60.8 (11.8)	0.006
Range	30-87	55-80		31-88	35-86		23.8-91	19.8-90	
Sex									
Male (%)	165 (63.4)	172 (54.1)	0.039	398 (47.9)	129 (48.5)	0.920	515 (54.7)	498 (58.1)	0.15
Female (%)	95 (39.7)	146 (45.9)		433 (52.1)	137 (51.5)		427 (45.3)	359 (41.9)	
Smoking status									
Never (%)	21 (8.1)	–	<0.001	–	–	<0.001	108 (11.5)	303 (35.6)	<0.001
Former (%)	163 (62.7)	156 (49.1)		372 (44.8)	153 (57.5)		380 (40.4)	357 (41.9)	
Current (%)	76 (29.2)	162 (50.9)		459 (55.2)	113 (42.5)		450 (47.9)	194 (22.5)	
Smoking pack-years									<0.001
Mean (SD)	45.5 (32.9)	53.6 (18.4)	<0.001	51.8 (30.0)	35.3 (20.5)	<0.001	42.6 (28.0)	22.8 (18.6)	–
Range	0-165	10-97		10-216	10-124		0-196	0-105	
FEV1 (% pred) #									
Mean (SD)	69.1	93.9 (10.5)	<0.001	68.6 (20.4)	94.5 (10.3)	<0.001	–	–	–
Range	22-124	80-129.1		15-135.1	80-123.2		–	–	
FEV1/FVC #									
Mean (SD)	0.59	0.77 (0.05)	<0.001	0.66 (0.12)	0.79 (0.04)	<0.001	–	–	–
Range	0.27-0.94	0.70-0.9		0.27-0.97	0.70-0.94		–	–	
Histologic type									
Adenocarcinoma	136 (52.3)	–	–	415 (51.3)	–	–	325 (44.3)	–	–
Squamous	82 (31.5)	–		176 (21.8)	–		248 (33.8)	–	
Other	42 (16.2)	–		218 (26.9)	–		161 (21.9)	–	

[a] Of the 260 LC cases, 54 were unrelated familial cases; in 75 of the 206 sporadic LC cases, patients also had severe chronic obstructive pulmonary disease.
[b] Controls with normal pulmonary function are defined as having an FEV$_1$ value greater than 80% and FEV1/FVC ratio greater than 0.7 predicted. These data are not available for familial cases in the discovery and the TRICL-ILCCO study subjects.
[c] *p* Value from the two-sided chi-square test (for categorical variables) and Student's *t* test (for continuous variables).
LC, lung cancer; WSU, Wayne State University; FEV1, forced expiratory volume in 1 second; pred, predicted; FEV1/FVC, ratio of forced expiratory volume in 1 second to forced vital capacity; TRICL-ILCCO, Transdisciplinary Research in Cancer of the Lung of the International Lung Cancer Consortium.

Table 2. Candidate Rare Deleterious Variants Identified in the Discovery and Tested in the Validations

						MAF%		No. of Carriers among Case Patients/Controls[c]		
Known Association (25 Loci)	Gene (33 Genes)	Variant (48 SNVs)	Identifier RS ID	Ref/Alt	CADD C-Score[a]	ExAC[b] (N = 33,370)	Case/Control (n = 2033/1441)	Discovery[d] (n = 260/318)	WSU Study (n = 831/266)	TRICL-ILCCO Study (n = 942/857)
1q23.2 (LC)	*DUSP23*	Cys95Ser	rs147728803	G/C	31	0.07	0.10/0.31	0/4 ^{C1}	0/1	4/4
2q36.3 (PF)	*COL4A4*	Pro1587Arg	rs190148408	G/C	20	0.27	0.32/0.38	0/5	6/1	7/5
	COL4A3	Pro1109Ser	rs55816283	C/T	17	0.56	0.52/0.48	0/6 ^{C2,7}	8/3	13/5
3p24.1 (LC_PF)	*ZCWPW2*	Asn23Ser	rs148504648	A/G	13	0.33	0.19/0.45	0/4 ^{C3}	1/2	7/7
3q13.13 (COPD_SM)	*DPPA2*	Ala157Ser	rs144052288	C/A	17	0.22	0.37/0.34	5 ^{S8}/1	—	4/7
	DZIP3	Gly67Cys	rs745923043	G/T	29	0.003	0.08/0.04	2 ^{S1,2}/0	—	0/1
		Pro990Leu	rs140068430	C/T	26	0.03	0.07/0	2 ^{S3}/0	0/0	1/0
3q28 (LC_SM)	*P3H2*	Gln185His	rs117688924	C/A	27	0.06	0.19/0.03	3/0	2/1	3/0
4p16.1 (LC)	*DRD5*	Met75Thr	rs151282040	T/C	19	0.18	0.12/0.14	4 ^{S5}/1 ^{C4}	0/0	1/3
		Cys335X	rs145497708	C/A	36	0.23	0.23/0.09	4/1	1/0	—
5p15.33 (LC_COPD_SM_PF)	*BRD9*	Splice 3′	rs201402002	T/C	16	0.17	0.33/0.09	5 ^{S1}/0	—	3/2
	SLC12A7	Splice 5′	rs150315797	G/A	13	0.06	0.12/0	3 ^{S6}/0	—	0/0
6p21.2 (PF)	*DAAM2*	Arg172His	rs200589550	G/A	31	0.31	0.22/0.45	1 ^{S3}/6 ^{C1}	4/3	4/4
		Pro555Leu	rs201570348	C/T	25	0.25	0.15/0.14	0/3 ^{C4}	3/0	3/1
		Asp762Gly	rs200287086	A/G	24	0.28	0.10/0.38	0/6 ^{C2,3,5,6}	3/1	1/4
6p21.33 (LC_COPD_SM_PF)	*LTB*	Leu87Phe	rs4647187	G/A	23	0.09	0.27/0.03	3/0	2/0	6/1
	SAPCD1	Gln76X	rs139815351	C/T	35	0.15	0.24/0.14	4 ^{S5}/1	3/1	3/2
8q11.21 (PF)	*SNTG1*	Splice 5′	rs201831443	G/T	14	0.11	0.08/0.13	2 ^{S1,3}/0	—	0/3
		Val121Leu	rs138262840	G/C	27	0.20	0.42/0.24	3 ^{S6}/0	5/4	9/3
9q22.32 (PF)	*PTCH1*	Asp436Asn	rs142274954	C/T	23	0.11	0.15/0.03	2/0	2/0	2/1
9q34.2 (SM)	*DBH*	Val195Met	rs145059403	G/A	28	0.09	0.05/0.10	0/3	1/0	1/0
10q23 (LC_COPD_SM)	*IFIT3*	Leu390Arg	rs116926108	T/G	24	0.12	0.17/0.34	3 ^{S2,7}/0	1/1	3/4
	CEP55	Arg191Gln	rs368583889	G/A	34	0.002	0.38/0	2/0	—	—
		Glu321Lys	rs146992036	G/A	28	0.26	0.15/0.03	3 ^{S11}/0	1/1	2/0
	PLCE1	Thr467Ile	rs192219615	C/T	25	0.16	0/0.34	0 /4 ^{C5}	0/1	0/0
10q25.1 (LC_SM_PF)	*CCDC147*	Arg696Cys	rs41291850	C/T	28	1.09	0.84/0.59	3 ^{S7}/0	14/1	17/16
	ITPRIP	Arg181Trp	rs151176986	G/A	27	0.15	0.12/0.55	2/0	1/1	2/7
		Asp236Tyr	rs372849615	C/A	23	0.003	0.12/0	2 ^{S11}/0	—	1/0
11p14.1 (SM)	*BDNF*	Thr2Ile	rs8192466	G/A	24	0.15	0.15/0.10	3 ^{F1}/0	1/0	2/3
12p13.33 (LC)	*RAD52*	Arg396Cys	rs112677599	G/A	15	0.15	0.17/0.55	3/0	1/1	3/7
13q12.12 (LC)	*MIPEP*	Leu197Pro	rs150167906	A/G	32	0.11	0.27/0.10	3/1	3/1	5/1
13q13.1 (LC)	*BRCA2*	Phe12Ser	rs587782872	T/C	23	novel	0/0.31	0/2	—	—
		Tyr42Cys	rs4987046	A/G	15	0.22	0.12/0.55	4 ^{F2}/1 ^{C6}	1/2	0/5
		Gly1433Trp	rs1036091086	G/T	24	0.003	0/0.31	0/2	—	—
		Lys3326X	rs11571833	A/T	38	0.90	1.47/0.62	7/4	22/3	31/11
16q21 (PF)	*CCDC113*	Lys100Asn	rs144246110	A/T	21	0.34	0.27/0.24	1/5 ^{C8}	1/0	9/2

(continued)

Table 2. Continued

						MAF%		No. of Carriers among Case Patients/Controls[c]		
Known Association (25 Loci)	Gene (33 Genes)	Variant (48 SNVs)	Identifier RS ID	Ref/Alt	CADD C-Score[a]	ExAC[b] (N = 33,370)	Case/Control (n = 2033/1441)	Discovery[d] (n = 260/318)	WSU Study (n = 831/266)	TRICL-ILCCO Study (n = 942/857)
16q23.1 (PF)	*ADAMTS18*	Gln146His	rs151326659	C/G	23	0.17	0.25/0.34	0/3	—	6/5
		Arg1053Trp	rs148703569	G/A	28	0.23	0.22/0.42	0/4 ^{C5,8}	3/2	6/6
	CLEC3A	Ala66Pro	rs150149068	G/C	24	0.29	0.54/0.13	3/0	—	10/3
18p11.3 (LC)	*LAMA1*	Gly967Asp	rs141851670	C/T	26	0.25	0.24/0.14	2/0	5/0	3/4
		Gly1227Arg	rs776158943	C/T	25	0.008	0.38/0	2 ^{F2}/0	—	—
19p12 (SM)	*ZNF93*	Gly337Glu	rs145491369	G/A	26	0.61	1.08/0.42	6 $^{S4,8,9,10/}$1	—	20/9
		Lys388Asn	rs140935689	G/C	24	0.08	0.17/0.10	3 $^{S8,9,10/}$0	0/1	4/2
20q13.33 (LC)	*RTEL1*	Gln397Glu	rs150285674	C/G	24	0.06	0.17/0.07	2 ^{F1,S8}/0	3/1	2/1
		Met652Thr	rs148080505	T/C	16	0.03	0.10/0.10	0/3 ^{C7}	0/0	4/0
22q12.1 (LC)	*CHEK2*	Ile157Thr	rs17879961	A/G	21	0.47	0.49/0.62	1/4	5/0	14/14
		Met424Val	rs375130261	T/C	26	0.005	0/0.31	0/2	—	—
22q12.2 (LC)	*MTMR3*	Pro1192His	rs773098171	C/A	29	0.006	0.08/0	2/0	—	0/0

[a] The CADD C-score is the overall measure of deleteriousness; a score of 20 or higher indicates the top 1%, and a score of 30 or higher indicates the top 0.1% in the human genome.

[b] MAF% were reported for the non-Finnish Europeans in ExAC database (N = 33,370).

[c] Of the 48 SNVs, 15 were not covered in the WSU Study, six were not covered in TRICL-ILCCO Study, and five were not covered by either validation sets (shown by an em dash). The MAF% of these SNVs was based on the available cases and controls.

[d] In the discovery set, 13 LC case patients and eight controls were carrying multiple candidates (see details in Supplementary Table 2); in column 9, a superscript *C* refer to the same control subject, a superscript *F* refers to the same familial cases, and a superscript *S* refers to the same sporadic cases.

SNV, single-nucleotide variant; RS ID, reference single-nucleotide polymorphism identifie; Ref, reference; Alt, alternative; CADD, combined annotation-dependent depletion; MAF, minor allele frequency; ExAC, Exome Aggregation Consortium; WSU, Wayne State University; TRICL-ILCCO, Transdisciplinary Research in Cancer of the Lung team of the International Lung Cancer Consortium; LC, lung cancer; PF, pulmonary function; COPD, chronic obstructive pulmonary disease; SM, smoking behavior; *DUSP23*, dual specificity phosphatase 23 gene; *COL4A4*, collagen type IV alpha 4 chain gene; *COL4A3*, collagen type IV alpha 3 chain gene; *ZCWPW2*, zinc finger CW-type and PWWP domain containing 2 gene; *DPPA2*, developmental pluripotency associated 2 gene; *DZIP3*, DAZ interacting zinc finger protein 3 gene; *P3H2*, prolyl 3-hydroxylase 2 gene; *DRD5*, dopamine receptor D5 gene; *BRD9*, bromodomain containing 9 gene; *SLC12A7*, solute carrier family 12 member 1 gene; *DAAM2*, dishevelled associated activator of morphogenesis 2 gene; *LTB*, lymphotoxin beta gene; *SAPCD1*, suppressor APC domain containing 1 gene; *SNTG1*, syntrophin gamma 1 gene; *PTCH1*, patch 1; *DBH*, dopamine beta-hydroxylase gene; *IFIT3*, interferon induced protein with tetratricopeptide repeats 3 gene; *CEP55*, centrosomal protein 55 gene; *PLCE1*, phospholipase C epsilon 1 gene; *CCDC147*, cilia and flagella associated protein 58 gene; *ITPRIP*, inositol 1,4,5-triphosphate receptor interacting protein gene; *BDNF*, brain derived neurotrophic factor gene; *RAD52*, RAD52 homolog, DNA repair protein gene; *MIPEP*, mitochondrial intermediate peptidase gene; *RTEL1*, regulator of telemere elongation helicase gene; *BRCA2*, BRCA2, DNA repair associated gene; *CCDC113*, coiled-coil domain containing 113; *ADAMTS18*, ADAM metallopeptidase with thrombospondin type 1 motif 18 gene; *CLEC3A*, C-type lectin domain family 3 member A gene; *LAMA1*, laminin subunit alpha 1 gene; *ZNF93*, zinc finger protein 93 gene; *RTEL1*, regulator of telemere elongation helicase gene; *CHEK2*, checkpoint kinase 2 gene; *MTMR3*, myotubularin related protein 3 gene.

patients (6%) and 303 controls (21%) were nonsmokers, mostly from the TRICL-ILCCO study. In terms of smoking intensity (mean PY value), in the discovery, lower PY values were reported in the case patients with LC than in the controls (mean 46 versus 54 [$P < 0.001$]), whereas much higher PY values were reported in case patients than in controls in the two validation studies (mean 52 versus 35 in the WSU study, and 43 versus 23 in the TRICL-ILCCO study, respectively [both $P < 0.001$]). Regarding LC histologic type, adenocarcinoma was the most common type across the three data sets, with 52% in the discovery set and 51% and 44% in the two validation sets, respectively.

Analysis of Recurrent Rare and Deleterious Variants

In the discovery set, of 99,489 SNVs and 1206 indels mapped in the exons of the target 121 known loci (260 genes), 1446 were functional mutation types (1411 nonsynonymous SNVs, 10 splice sites, 16 stop gain/loss, and nine frameshifts), of which 432 were rare and 168 were further predicted to be potential deleterious. Our stepwise filtering strategy (Fig. 1 and Table 2) identified 48 recurrent candidate variants of which 30 were highly enriched in patients with LC (risk-conferring) and 18 enriched in controls (protective), including three stop-gains, three splice-sites, and 42 nonsynonymous SNVs. These 48 candidates were located in 33 genes at 25 of the risk loci and presented in a total of 68 smoking controls and 85 patients (17 familial and 68 sporadic cases; eight sporadic carriers had severe COPD, of whom 70% had the adenocarcinoma histologic type). Among the candidates carriers, 13 (two with familial and 11 with sporadic cases, none of whom had severe COPD) and eight controls were multicarriers who had carried at least two candidates (Supplementary Table 2). It is interesting to note that four of the 13 multicarrier patients had p.Gly337Glu in zinc finger protein 93 gene (*ZNF93*) (Online Mendelian Inheritance in Man [OMIM] No. 603975). In particular, one patient with adenocarcinoma (age 60 years, male sex, and smoking history of 30 PY) was a carrier of four candidates, including two candidates from *ZNF93.* For the controls, six of eight multicarriers carried one or two candidates from disheveled-associated activator of morphogenesis 2 gene (*DAAM2*) (OMIM No. 606627).

In the validation sets, of the 48 candidates, five (10%) were not covered by both validation studies. Specifically, 15 (31%) were not covered by the WSU study, whereas six (13%) were not covered in the TRICL-ILCCO study. As shown in Table 3, the top most risk-conferring variant from the allelic association analysis is a known stop codon, p.Lys3326X (K3326X) in *BRCA2* (OMIM No. 600185). This stop gain results from an A > T transversion in the 27th exon that leads to the loss of the final 93 amino acids of the BRCA2, DNA repair associated gene protein (UniProt No. P51587). The MAF of K3326X in the case patients was significantly higher than in the

Table 3. Top Hits from Allelic Association Analysis of Combined Discovery and Validation Sets

	Case/Control/ExAC[a] (N = 2033/1441/33,370)		Allelic OR (95% CI) and FDR-Adjusted *p* Value[b]			
Candidate Variants	No. of Minor Alleles	MAF%	Case vs. Control		Case vs. ExAC	
Strong association						
BRCA2 Lys3326X	60/18/602	1.47/0.62/0.90	2.36 (1.38-3.99)	**0.0004**	1.68 (1.29-2.20)	**0.0002**
LTB Leu87Phe	11/1/57	0.27/0.03/0.09	7.52 (1.01-16.56)	**0.008**	3.07 (1.61-5.85)	**0.001**
P3H2 Gln185His	8/1/40	0.19/0.03/0.06	5.39 (0.75-15.43)	**0.032**	3.33 (1.56-7.12)	**0.003**
DAAM2 Asp762Gly	4/11/34	0.10/0.38/0.27	0.25 (0.10-0.79)	**0.007**	0.34 (0.11-0.94)	**0.011**
ZNF93 Gly337Glu[c]	26/10/404	1.08/0.42/0.61	2.51 (1.22-5.20)	**0.005**	1.82 (1.22-2.72)	**0.003**
CLEC3A Ala66Pro[c]	13/3/195	0.54/0.13/0.29	4.18 (1.19-11.69)	**0.008**	1.89 (1.08-3.31)	**0.021**
BRD9 splice acceptor[c]	8/2/14	0.33/0.09/0.17	3.77 (0.95-10.41)	**0.041**	2.28 (0.97-4.93)	**0.039**
Suggestive signal						
PLCE1 Thr467Ile[d]	0/5/109	0/0.34/0.16	0.15 (0.01-1.02)	0.053	0.15 (0.10-0.75)	**0.013**
MIPEP Leu197Pro	11/3/76	0.27/0.10/0.11	2.58 (0.87-9.22)	0.059	2.41 (1.28-4.53)	**0.007**
RTEL1 Gln397Glu	7/2/41	0.17/0.07/0.06	2.45 (0.52-11.76)	0.065	2.83 (1.27-6.27)	**0.009**
*SNTG1*Val121Leu	17/7/126	0.42/0.24/0.20	1.72 (0.71-4.11)	0.118	2.13 (1.28-3.54)	**0.004**
ZNF93 Lys388Asn	7/3/50	0.17/0.10/0.08	1.63 (0.43-6.27)	0.152	2.34 (1.06-5.16)	**0.028**

[a] The non-Finnish Europeans in the ExAC database (N = 33,370).

[b] *p* Values were calculated by Fisher's exact test and bolded if significant after FDR adjustment.

[c] These variants were not covered in the Wayne State University Study, the allele counts were based on the discovery and Transdisciplinary Research in Cancer of the Lung Team of the International Lung Cancer Consortium validation sets, including 1202 LC cases and 1175 controls.

[d] This variant was absent in the LC case; thus, we added 0.5 to each cell in the analysis.

ExAC, Exome Aggregation Consortium; CI, confidence inteval; FDR, false discovery rate; MAF, minor allele frequency; *BRCA2*, BRCA2, DNA repair associated gene; *LTB*, lymphotoxin beta gene; *P3H2*, prolyl 3-hydroxylase 2 gene; *DAAM2*, dishevelled associated activator of morphogenesis 2 gene; *ZNF93*, zinc finger protein 93 gene; *CLEC3A*, C-type lectin domain family 3 member A gene; *BRD9*, bromodomain containing 9 gene; *PLCE1*, phospholipase C epsilon 1 gene; *MIPEP*, mitochondrial intermediate peptidase gene; *RTEL1*, regulator of telemere elongation helicase gene; *SNTG1*, syntrophin gamma 1 gene.

controls (MAF 1.47% versus 0.62%; OR = 2.36, 95% CI: 1.38–3.99) and ExAC population (MAF 1.47% versus 0.9%; OR = 1.68, 95% CI: 1.29–2.20). This truncating variant had the highest scaled CADD C-score of 38 and was predicted to be in the top 0.1% most deleterious substitutions in the human genome. The K3326X occurred in the highly conserved COOH-terminal domain, which plays a critical role in the homology-directed repair of DNA double-strand breaks.

Another two top candidates were missense variants, p.Leu87Phe in lymphotoxin beta gene (*LTB*) (OMIM No. 600978) and p.Gln185His in prolyl 3-hydroxylase 2 gene (*P3H2*) (also known as *LEPREL1*) (OMIM No. 610341). Both variants were carried by only one control (MAF 0.034%) but occurred in 11 and eight cases (MAF 0.27% and 0.19%), respectively, with an effect size of 7.52 (95% CI: 1.01–16.56) and 5.39 (95% CI: 0.75–15.43), respectively. Likewise, these two variants were exceedingly rare in the ExAC reference population (MAF 0.09% and 0.06, respectively), and they were predicted to be among the 1% most deleterious (CADD scores of 23 and 27, respectively). The *LTB* p.Leu87Phe occurred at the third exon of the gene and a remarkably conserved β-strand structure that links the transmembrane and tumor necrosis factor domains of the protein (UniProt No. Q06643 [Fig. 2*A*]). The *P3H2* p.Gln185His is located in the second exon of the gene and between the second and third tetratricopeptide-like helical repeats of the protein (UniProt No. Q8IVL5 [Fig. 2*B*]).

In contrast to the aforementioned three risk-conferring variants, we also identified one protective variant, p.Asp762Gly in *DAAM2*. The LC risk for carriers

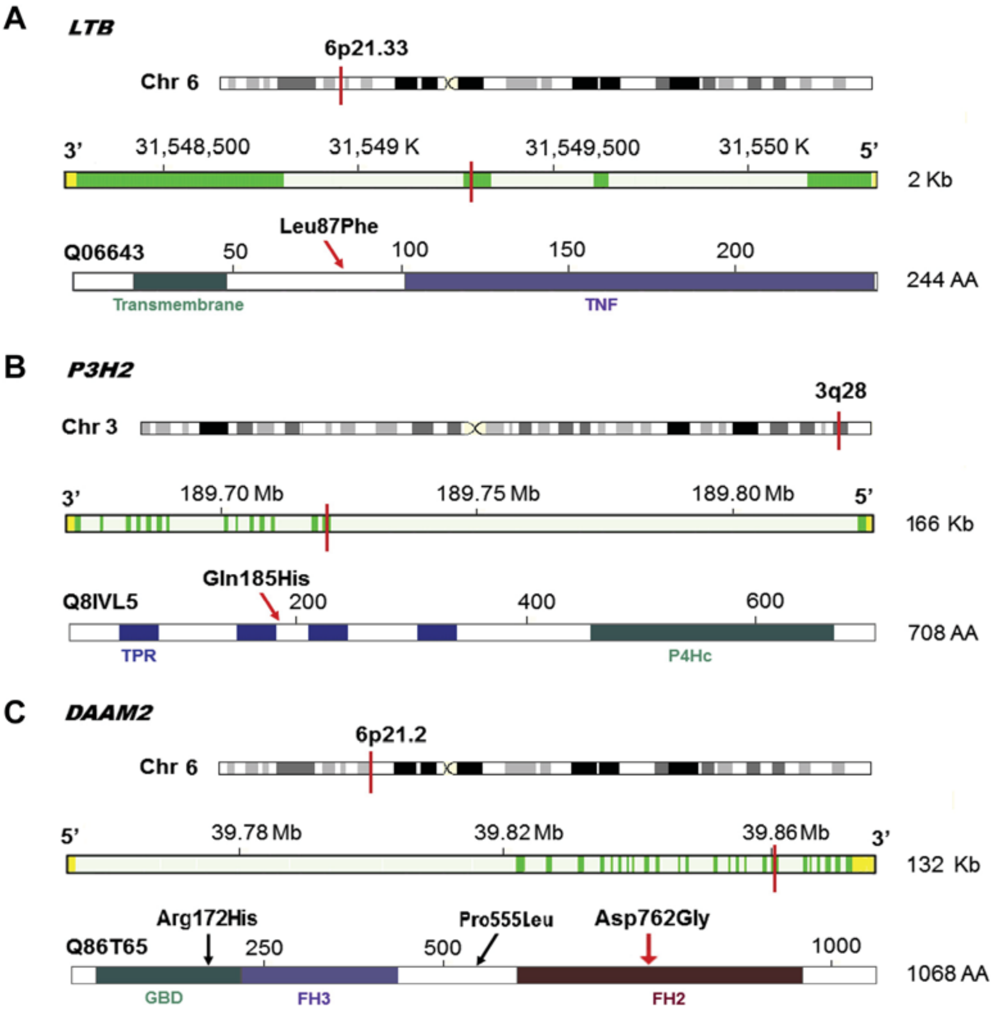

Figure 2. Chromosomal position, gene exon, protein domain(s), and the top candidates. (*A*) Lymphotoxin beta gene (*LTB*) p.Leu87Phe located in the third exon, the β-strand (which links the transmembrane and tumor necrosis factor (TNF) domains. (*B*) Prolyl 3-hydroxylase 2 gene (*P3H2*) p.Gln185His located in the second exon, between the second and third tetratricopeptide-like helical repeat (TPR) domains. (*C*) Dishevelled associated activator of morphogenesis 2 gene (*DAMM2*) p.Asp762Gly located the 18th exon, the second formin homology (FH) domain. The top candidate mutations are indicated with red lines in the chromosome and gene exons (genomic location, assembly GRCh37) and with red arrows in the protein. The gene annotation also shows forward (*DAMM2*) or reverse (*LTB* and *P3H2*) strands of the chromosome.

of *DAAM2* p.Asp762Gly decreased fourfold compared with that for the study controls (OR = 0.25, 95% CI: 0.10–0.79) and threefold compared with that for the ExAC population (OR = 0.34, 95% CI: 0.11–0.94). The p.Asp762Gly is located in the 18th exon of the gene, which is close to two acetylation sites, Lys765 and Lys766, and lies in the second formin homology domain of the protein (UniProt No. Q86T65 [Fig. 2*C*]). In the same gene, another candidate, p.Arg172His, although not statistically significant, was also enriched in controls, with an MAF of 0.45% in smoker controls and 0.31% in the ExAC reference population, compared with 0.22% for case patients.

Other promising candidates with consistent allelic associations include p.Gly337Glu in *ZNF93*, p.Ala66Pro in C-type lectin family 3 member A gene (*CLEC3A*) (OMIM No. 613588), and a splice acceptor (rs201402002) in bromodomain containing 9 gene (*BDR9*). Unfortunately, these three variants were not covered in the WSU study. In addition, five SNVs showed suggestive evidence (only significant in LC case versus in the ExAC population): mitochondrial intermediate peptidase gene (*MIPEP*) p.Leu197Pro, regulator of telemere elongation helicase (*RTEL1*) p.Gln397Glu, phospholipase C epsilon 1 gene (*PLCE1*) p.Thr467Ile, syntrophin gamma 1 gene [(*SNTG1*) p.Val121Leu, and *ZNF93* p.Lys388Asn (see Table 3).

Gene-Based Burden Analysis of Rare Variants

Table 4 summarizes the burden test results from the gene-based multiple rare and predicted deleterious SNVs. Among the 21 candidate genes with multiple rare deleterious SNVs, four genes, namely, *ZNF93*, *DAAM2*, *BRD9*, and *LTB*, showed strong association, with FDR adjusted *p* values less than 0.05 in both the CMC and KBAC tests.

Table 4. Gene-Based Association Collapsing Tests in the Discovery Data

		No. of SNVs MAF% Distribution				FDR-Adjusted *p* Value[c]	
Genes[a] (21 Genes)	No. of Rare Deleterious SNVs per Gene[b]	Bin 1: 0.1-1	Bin 2: 0.01-0.1	Bin 3: <0.01	No. of Carriers in Cases/ Control (n = 260/318)	CMC Test	KBAC Test
Risk genes							
ZNF93	4	1	1	2	10/2	**0.011**	**0.009**
BRD9	2	0	1	1	6/1	**0.046**	**0.039**
LTB	4	0	2	2	6/1	**0.048**	**0.039**
DRD5	3	2	1	0	8/3	0.090	0.077
SNTG1	3	2	1	0	5/1	0.094	0.079
LAMA1	4	1	1	2	5/1	0.095	0.079
SLC12A7	4	1	2	1	5/1	0.095	0.079
IFIT3	3	1	1	1	6/2	0.140	0.115
CEP55	4	1	2	1	6/2	0.141	0.115
BDNF	2	1	0	1	4/1	0.204	0.152
RAD52	3	1	1	1	4/1	0.205	0.154
ITPRIP	3	1	1	1	4/1	0.205	0.154
P3H2	4	1	1	2	5/2	0.266	0.189
DZIP3	4	0	2	2	5/2	0.267	0.189
BRCA2	6	2	1	3	13/9	0.275	0.193
CCDC147	2	1	1	0	4/2	0.415	0.329
RTEL1	3	0	2	1	3/3	0.992	0.998
Protective genes							
DAAM2	4	3	1	0	3/15	**0.019**	**0.013**
ADAMTS18	4	2	1	1	2/8	0.186	0.125
CHEK2	4	1	2	1	2/7	0.265	0.188
DBH	2	0	1	1	1/3	0.692	0.486

[a] Only genes with two or more rare deleterious variants are included in the analysis from the discovery set.

[b] Number of rare deleterious SNVs (after filtering steps 1 and 2) within the genes.

[c] Significant *p* values are bolded.

SNV, single-nucleotide variant; MAF, minor allele frequency; FDR, false discovery rate; CMC, combined multivariate and collapsing; KBAC, kernel-based adaptive cluster; *ZNF93*, zinc finger protein 93 gene; *BRD9*, bromodomain containing 9 gene; *LTB*, lymphotoxin beta gene; *DRD5*, dopamine receptor D5 gene; *ITPRIP*, inositol 1,4,5-triphosphate receptor interacting protein gene; *SNTG1*, syntrophin gamma 1 gene; *LAMA1*, laminin subunit alpha 1 gene; *SLC12A7*, solute carrier family 12 member 1 gene; *IFIT3*, interferon induced protein with tetratricopeptide repeats 3 gene; *CEP55*, centrosomal protein 55 gene; *BDNF*, brain derived neurotrophic factor gene; *RAD52*, RAD52 homolog, DNA repair protein gene; *P3H2*, prolyl 3-hydroxylase 2 gene; *DZIP3*, DAZ interacting zinc finger protein 3 gene; *BRCA2*, BRCA2, DNA repair associated gene; *CCDC147*, cilia and flagella associated protein 58 gene; *RTEL1*, regulator of telemere elongation helicase gene; *DAAM2*, dishevelled associated activator of morphogenesis 2 gene; *ADAMTS18*, ADAM metallopeptidase with thrombospondin type 1 motif 18 gene; *CHEK2*, checkpoint kinase 2 gene; *DBH*, dopamine beta-hydroxylase gene.

Discussion

Despite previous family-based linkage studies and intensive population-based GWAS analyses and candidate gene screening, a large proportion of the heritability of LC remains unexplained. Our focused analyses led to identification of four rare and deleterious inherited variants associated with LC susceptibility, including one well-established truncating variant (*BRCA2* K3326X) and three newly identified missense variations (*LTB* p.Leu87Phe, *P3H2* p.Gln185His, and *DAAM2* p.Asp762Gly). It should be noted that none of the candidate rare variants that we have identified in the present study were in linkage disequilibrium with the known LC-GWAS common SNPs. The limits of linkage disequilibrium between common and rare variants were quantified by Wray[21] and supported by previous studies.[22,23]

This study confirms a robust association between a known rare truncating variant, K3326X in 13q13.1 *BRCA2*, and risk of LC. The effect size in the current study (OR = 2.36) is nearly identical to the previously identified association with risk of squamous cell LC (OR = 2.47)[4] and upper aerodigestive tract cancer (OR = 2.53),[24] and it far exceeds the small increase in risk of breast and ovarian cancer (OR = 1.26).[25,26] The molecular mechanisms that underpin this finding are unknown. In relation to the effect on cellular and biochemical properties of this variant, cancer cell lines show that mice and cells with the exon 27 truncated protein are hypersensitive to ionizing radiation[27] and cross-linking agents[28,29] and exhibited increased susceptibility to various types of solid tumors.[30] It has been demonstrated that patients who have ovarian cancer and *BRCA1/BRCA2* germline mutations respond favorably to poly(ADP-ribose) polymerase 1 inhibitors in clinical trials.[31–33] Therefore, it is possible that patients who have LC with *BRCA2* K3326X may also similarly respond favorably to poly(ADP-ribose) polymerase inhibition and benefit from treatment.

An interesting finding is the association with the immunity-related gene *LTB,* which is localized to the 6p21.33 major histocompatibility complex region (which has been previously implicated in risk of LC, COPD, SM, and PF levels).[34–37] Functional studies in gene knockout and transgenic mouse systems have shown that *LTB* is of fundamental importance in fibrogenesis and carcinogenesis because of its action through a distinct receptor lymphotoxin beta receptor and nuclear factor kappa B.[38] The lymphotoxin beta protein plays a key role in innate immunity and inflammation, which have been the focus of intensive basic science and translational research.[39] Another finding in the chromosome 6p region was the protective effects of *DAAM2.* This 6p21.2 region is a known susceptibility loci to multiple diseases including PF levels,[34] smoking cessation,[40] renal cancer,[41] schizophrenia,[42] and hypospadias.[43] The *DAAM2* gene is involved in the regulation of actin cytoskeleton in several different tissues, including the tissue of the tracheal airways system,[44] and it is abnormally regulated in nasopharyngeal carcinoma[45] and COPD.[46] The dishevelled associated activator of morphogenesis 2 protein is one of the key WNT/plantar cell polarity signaling pathway proteins and has been documented to promote oncogenesis, stem cell renewal, and tumor proliferation.[47,48]

The 3q28 *P3H2* plays a critical role in collagen metabolic processes and oxidation reduction, and it inhibits cell proliferation.[49] Previous work has shown that *P3H2* genes are novel targets for epigenetic silencing in breast cancer.[50] The pathogenic mutation p.Gly508Val was associated with high degree myopia.[51–53] Moreover, from the TISSUES the database, *P3H2* expression shows the highest levels in lung tissue.[54]

A main strengths of this study are the accurate PF data and smoking exposure data. In the discovery set, in which only a small number of individuals were exome sequenced, the inclusion of even a small proportion of misclassified individuals could have affected the analysis. On the other hand, extreme phenotypes increase statistical power. Our discovery set (260 cases) included the 102 cases from our previous study[5], 48 sporadic cases in which patients reported histories of heavy smoking and severe COPD, and 54 familial cases in which patients were likely enriched for disease-associated genetic signals. In the WSU validation study, the smoking controls were strictly selected in terms of normal PF despite a history of heavy smoking, and thus, they were considered to be resistant to the effects of smoking; in contrast, for the LC cases, there were only 75 sporadic LC cases in which patients had severe COPD and in which the tobacco exposure would be considered quite substantial. It should be noted however that our study is still underpowered for the association analysis of each rare variant in a limited number of 2033 cases and 1441 controls. Although the association of *CCDC147* p.Arg696Cys and *DBH* p.Val26Met between the case patients with LC and the ExAC reference population did not attain statistical significance, our result is in line with the results of our previous work.[5] Another limitation is that we have focused on missense, nonsense, stop-gain/loss, splice sites, and frameshift variants in our variant filtration strategies and we have not evaluated certain classes of variants, such as large gene-disrupting duplications and noncoding variants (such as flanking intronic and untranslated regions that may disrupt gene expression). Although larger studies and/or whole genome sequencing analysis might identify more rare variants with deleterious effects, the paucity of findings of recurrent rare variants impacting LC risk is intriguing.

In conclusion, our results provide evidence that the rare deleterious germline variants *BRCA2* p.Lys3326X, *LTB* p.Leu87Phe, *P3H2* p.Gln185His, and *DAAM2* p.Asp762Gly contribute to LC susceptibility. However, further in-depth functional follow-up studies are still needed to evaluate the pathogenicity of each of the strong candidates reported in this study.

Acknowledgments

This work was supported by grants from the National Institutes of Health (R01CA127219, R01CA141769, R01CA060691, R01CA87895, R01CA80127, R01CA 84354, R01CA134682, R01CA134433, R01CA074386, R01CA092824, R01HL089856, R01HL089897, R01 HL113264, R01HL082487, R01HL110883, R03CA77 118, P20GM103534, P30CA125123, P30CA023108, P30CA022453, P30ES006096, P50CA090578, U01CA 76293, U19CA148127, K07CA181480, N01-HG-65404, HHSN261201 300011I, and HHSN268201 200007C), the Intramural Research Program of the National Human Genome Research Institute (to Dr. Bailey-Wilson), and the Herrick Foundation. The Mount Sinai Hospital and Princess Margaret Hospital study is supported by The Canadian Cancer Society Research Institute (grant 020214 to Dr. Hung), Ontario Institute of Cancer and the Alan Brown Chair (to Dr. Liu), and Lusi Wong Programs at the Princess Margaret Hospital Foundation. We would like to thank the patients and their families for participating in this research. We thank Dr. Richard Gibbs, Donna Muzny, Xiaoyun Liao, Van Le, Sandra Lee, and Margi Sheth from the Human Genome Sequencing Center at Baylor for performing the exome sequencing for all the samples in the discovery phase.

Supplementary Data

Note: To access the supplementary material accompanying this article, visit the online version of the *Journal of Thoracic Oncology* at www.jto.org and at https://doi.org/10.1016/j.jtho.2018.06.016.

References

1. Bosse Y, Amos CI. A decade of GWAS results in lung cancer. *Cancer Epidemiol Biomarkers Prev.* 2018;27:363-379.
2. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet*. 2008;82:100-112.
3. Xiong D, Wang Y, Kupert E, et al. A recurrent mutation in PARK2 is associated with familial lung cancer. *Am J Hum Genet*. 2015;96:301-308.
4. Wang Y, McKay JD, Rafnar T, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet*. 2014;46:736-741.
5. Liu Y, Kheradmand F, Davis CF, et al. Focused analysis of exome sequencing data for rare germline mutations in familial and sporadic lung cancer. *J Thorac Oncol*. 2016;11:52-61.
6. Lee SH, Goswami S, Grudo A, et al. Antielastin autoimmunity in tobacco smoking-induced emphysema. *Nat Med*. 2007;13:567-569.
7. Grumelli S, Corry DB, Song LZ, et al. An immune basis for lung parenchymal destruction in chronic obstructive pulmonary disease and emphysema. *PLoS Med*. 2004;1:e8.
8. Shan M, Cheng HF, Song LZ, et al. Lung myeloid dendritic cells coordinately induce TH1 and TH17 responses in human emphysema. *Sci Transl Med*. 2009;1:4ra10.
9. Liu P, Vikis HG, Wang D, et al. Familial aggregation of common sequence variants on 15q24-25.1 in lung cancer. *J Natl Cancer Inst*. 2008;100:1326-1330.
10. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD*. 2010;7:32-43.
11. Bainbridge MN, Wang M, Wu Y, et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol*. 2011;12:R68.
12. Lupski JR, Gonzaga-Jauregui C, Yang Y, et al. Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy. *Genome Med*. 2013;5:57.
13. Reid JG, Carroll A, Veeraraghavan N, et al. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics*. 2014;15:30.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754-1760.
15. Challis D, Yu J, Evani US, et al. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics*. 2012;13:8.
16. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46:310-315.
17. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008;83:311-321.
18. Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet*. 2010;6:e1001156.
19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B*. 1995;57:289-300.
20. Schwartz AG, Lusk CM, Wenzlaff AS, et al. Risk of lung cancer associated with COPD phenotype based on quantitative image analysis. *Cancer Epidemiol Biomarkers Prev*. 2016;25:1341-1347.
21. Wray NR. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and

interpretation of association studies. *Twin Res Hum Genet*. 2005;8:87–94.

22. Lopez de Maturana E, Ibanez-Escriche N, Gonzalez-Recio O, et al. Next generation modeling in GWAS: comparing different genetic architectures. *Hum Genet*. 2014;133:1235–1253.
23. de Los Campos G, Sorensen D, Gianola D. Genomic heritability: what is it? *PLoS Genet*. 2015;11:e1005048.
24. Delahaye-Sourdeix M, Anantharaman D, Timofeeva MN, et al. A rare truncating BRCA2 variant and genetic susceptibility to upper aerodigestive tract cancer. *J Natl Cancer Inst*. 2015;107:djv037.
25. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet*. 2013;45:353–361, 361e351–e352.
26. Meeks HD, Song H, Michailidou K, et al. BRCA2 Polymorphic stop codon K3326X and the risk of breast, prostate, and ovarian cancers. *J Natl Cancer Inst*. 2016;108:dijv315.
27. Morimatsu M, Donoho G, Hasty P. Cells deleted for Brca2 COOH terminus exhibit hypersensitivity to gamma-radiation and premature senescence. *Cancer Resh*. 1998;58:3441–3447.
28. Atanassov BS, Barrett JC, Davis BJ. Homozygous germ line mutation in exon 27 of murine Brca2 disrupts the Fancd2-Brca2 pathway in the homologous recombination-mediated DNA interstrand cross-links' repair but does not affect meiosis. *Genes Chromosomes Cancer*. 2005;44:429–437.
29. Wang X, Andreassen PR, D'Andrea AD. Functional interaction of monoubiquitinated FANCD2 and BRCA2/FANCD1 in chromatin. *Mol Cell Biol*. 2004;24:5850–5862.
30. McAllister KA, Bennett LM, Houle CD, et al. Cancer susceptibility of mice with a homozygous deletion in the COOH-terminal domain of the Brca2 gene. *Cancer Res*. 2002;62:990–994.
31. Audeh MW, Carmichael J, Penson RT, et al. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet*. 2010;376:245–251.
32. Fong PC, Yap TA, Boss DS, et al. Poly(ADP)-ribose polymerase inhibition: frequent durable responses in BRCA carrier ovarian cancer correlating with platinum-free interval. *J Clin Oncol*. 2010;28:2512–2519.
33. Ledermann JA, Harter P, Gourley C, et al. Overall survival in patients with platinum-sensitive recurrent serous ovarian cancer receiving olaparib maintenance monotherapy: an updated analysis from a randomised, placebo-controlled, double-blind, phase 2 trial. *Lancet Oncol*. 2016;17:1579–1589.
34. Repapi E, Sayers I, Wain LV, et al. Genome-wide association study identifies five loci associated with lung function. *Nat Genet*. 2010;42:36–44.
35. 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.
36. Broderick P, Wang Y, Vijayakrishnan J, et al. Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res*. 2009;69:6633–6641.
37. Hancock DB, Eijgelsheim M, Wilk JB, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet*. 2010;42:45–52.
38. Drutskaya MS, Efimov GA, Kruglov AA, Kuprash DV, Nedospasov SA. Tumor necrosis factor, lymphotoxin and cancer. *IUBMB Life*. 2010;62:283–289.
39. Aggarwal BB. Signalling pathways of the TNF superfamily: a double-edged sword. *Nat Rev Immunol*. 2003;3:745–756.
40. Uhl GR, Liu QR, Drgon T, et al. Molecular genetics of successful smoking cessation: convergent genome-wide association study results. *Arch Gen Psychiatry*. 2008;65:683–693.
41. Hirata H, Hinoda Y, Nakajima K, et al. Wnt antagonist gene polymorphisms and renal cancer. *Cancer*. 2009;115:4488–4503.
42. Meda SA, Ruano G, Windemuth A, et al. Multivariate analysis reveals genetic associations of the resting default mode network in psychotic bipolar disorder and schizophrenia. *Proc Natl Acad Sci U S A*. 2014;111:E2066–E2075.
43. Geller F, Feenstra B, Carstensen L, et al. Genome-wide association analyses identify variants in developmental genes associated with hypospadias. *Nat Genet*. 2014;46:957–963.
44. Matusek T, Djiane A, Jankovics F, Brunner D, Mlodzik M, Mihaly J. The Drosophila formin DAAM regulates the tracheal cuticle pattern through organizing the actin cytoskeleton. *Development*. 2006;133:957–966.
45. Zeng ZY, Zhou YH, Zhang WL, et al. Gene expression profiling of nasopharyngeal carcinoma reveals the abnormally regulated Wnt signaling pathway. *Hum Pathol*. 2007;38:120–133.
46. Wu X, Sun X, Chen C, Bai C, Wang X. Dynamic gene expressions of peripheral blood mononuclear cells in patients with acute exacerbation of chronic obstructive pulmonary disease: a preliminary study. *Crit Care*. 2014;18:508.
47. Barrow JR. Wnt/PCP signaling: a veritable polar star in establishing patterns of polarity in embryonic tissues. *Semin Cell Dev Biol*. 2006;17:185–193.
48. Tanaka K. Formin family proteins in cytoskeletal control. *Biochem Biophys Res Commun*. 2000;267:479–481.
49. Pokidysheva E, Boudko S, Vranka J, et al. Biological role of prolyl 3-hydroxylation in type IV collagen. *Proc Natl Acad Sci U S A*. 2014;111:161–166.
50. Shah R, Smith P, Purdie C, et al. The prolyl 3-hydroxylases P3H2 and P3H3 are novel targets for epigenetic silencing in breast cancer. *Br J Cancer*. 2009;100:1687–1696.
51. Mordechai S, Gradstein L, Pasanen A, et al. High myopia caused by a mutation in LEPREL1, encoding prolyl 3-hydroxylase 2. *Am J Hum Genet*. 2011;89:438–445.
52. Guo H, Tong P, Peng Y, et al. Homozygous loss-of-function mutation of the LEPREL1 gene causes severe non-syndromic high myopia with early-onset cataract. *Clin Genet*. 2014;86:575–579.
53. Feng CY, Huang XQ, Cheng XW, Wu RH, Lu F, Jin ZB. Mutational screening of SLC39A5, LEPREL1 and LRPAP1 in a cohort of 187 high myopia patients. *Sci Rep*. 2017;7:1120.
54. Santos A, Tsafou K, Stolte C, Pletscher-Frankild S, O'Donoghue SI, Jensen LJ. Comprehensive comparison of large-scale tissue expression datasets. *PeerJ*. 2015;3:e1054.