ORIGINAL ARTICLE

# Focused Analysis of Exome Sequencing Data for Rare Germline Mutations in Familial and Sporadic Lung Cancer

Yanhong Liu, PhD,[a] Farrah Kheradmand, MD,[a] Caleb F. Davis, PhD,[a] Michael E. Scheurer, PhD,[a] David Wheeler, PhD,[a] Spiridon Tsavachidis, MS,[a] Georgina Armstrong, MPH,[a] Claire Simpson, PhD,[b] Diptasri Mandal, PhD,[c] Elena Kupert, MS,[d] Marshall Anderson, PhD,[e] Ming You, MD, PhD,[e] Donghai Xiong, PhD,[e] Claudio Pikielny, PhD,[f] Ann G. Schwartz, PhD,[g] Joan Bailey-Wilson, PhD,[h] Colette Gaba, MPH,[i] Mariza De Andrade, PhD,[j] Ping Yang, MD, PhD,[j] Susan M. Pinney, PhD,[d] The Genetic Epidemiology of Lung Cancer Consortium, Christopher I. Amos, PhD,[f] Margaret R. Spitz, MD[a,*]

[a]*Baylor College of Medicine, Houston, TX, USA*
[b]*National Institutes of Health, Baltimore, MD, USA*
[c]*Louisiana State University Health Sciences Center, New Orleans, LA, USA*
[d]*University of Cincinnati College of Medicine, Cincinnati, OH, USA*
[e]*Medical College of Wisconsin, Milwaukee, WI, USA*
[f]*Dartmouth College, Lebanon, NH, USA*
[g]*Karmanos Cancer Institute, Wayne State University, Detroit, MI, USA*
[h]*Human Genome Research Institute, Bethesda, MD, USA*
[i]*The University of Toledo College of Medicine, Toledo, OH, USA*
[j]*Mayo Clinic College of Medicine, Rochester, MN, USA*

Received 21 July 2015; revised 21 September 2015; accepted 25 September 2015

## ABSTRACT

**Introduction:** The association between smoking-induced chronic obstructive pulmonary disease (COPD) and lung cancer (LC) is well documented. Recent genome-wide association studies (GWAS) have identified 28 susceptibility loci for LC, 10 for COPD, 32 for smoking behavior, and 63 for pulmonary function, totaling 107 nonoverlapping loci. Given that common variants have been found to be associated with LC in genome-wide association studies, exome sequencing of these high-priority regions has great potential to identify novel rare causal variants.

**Methods:** To search for disease-causing rare germline mutations, we used a variation of the extreme phenotype approach to select 48 patients with sporadic LC who reported histories of heavy smoking—37 of whom also exhibited carefully documented severe COPD (in whom smoking is considered the overwhelming determinant)—and 54 unique familial LC cases from families with at least three first-degree relatives with LC (who are likely enriched for genomic effects).

**Results:** By focusing on exome profiles of the 107 target loci, we identified two key rare mutations. A heterozygous p.Arg696Cys variant in the coiled-coil domain containing 147 (*CCDC147*) gene at 10q25.1 was identified in one sporadic and two familial cases. The minor allele frequency (MAF) of this variant in the 1000 Genomes database is 0.0026. The p.Val26Met variant in the dopamine $\beta$-hydroxylase (*DBH*) gene at 9q34.2 was identified in two sporadic cases; the minor allele frequency of this mutation is 0.0034 according to the 1000 Genomes database. We also observed three suggestive rare mutations on 15q25.1: iron-responsive element binding protein neuronal 2 (*IREB2*); cholinergic receptor, nicotinic, alpha 5 (neuronal) (*CHRNA5*); and cholinergic receptor, nicotinic, beta 4 (*CHRNB4*).

**Conclusions:** Our results demonstrated highly disruptive risk-conferring *CCDC147* and *DBH* mutations.

***Corresponding author.**

***Disclosure:** This work was supported by grants from the National Institutes of Health (R01 CA127219, R01 HL082487, R01 HL110883, K07CA181480, R01 CA060691, R01 CA87895, R01 CA80127, R01 CA84354, R01 CA134682, R01 CA134433, R03 CA77118, P20GM103534, P30CA125123, P30CA023108, P30-ES006096, P30CA022453, N01-HG-65404, U01CA076293, U19CA148127, and HHSN268201 200007C). JEB-W was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. Additional support was provided by the National Library of Medicine T15LM007093 (Davis) and the Population Sciences Biorepository at Baylor College of Medicine (BCM). The authors declare no conflict of interest.*

**Address for correspondence:** Margaret R. Spitz, MD, Department of Molecular and Cellular Biology, Dan.L Duncan Cancer Center, Baylor College of Medicine, Houston, TX 77030. E-mail: spitz@bcm.edu



**ISSN: 1556-0864**

http://dx.doi.org/10.1016/j.jtho.2015.09.015

*Keywords:* Exome sequencing; Single-nucleotide variants; Lung cancer; Chronic obstructive pulmonary disease; Familial; Sporadic

## Introduction

Chronic tobacco-induced airway inflammation provokes a milieu conducive to pulmonary carcinogenesis. We and others have previously shown that tobacco-induced chronic obstructive pulmonary disease (COPD), which is also characterized by a sustained inflammatory reaction in the airways and lung parenchyma, is a significant contributor to risk for the development of lung cancer (LC) in smokers.[1] Likewise, reduced pulmonary function is also reported as an important variable when included in risk prediction models.[2] Recent genome-wide association studies (GWAS) have identified 28 susceptibility loci for LC, 10 loci for COPD, 32 loci for smoking behavior (SM), and 63 loci for abnormal pulmonary function (PF) and related phenotypes, totaling 107 unique GWAS susceptibility loci (as of November 2014, Supplemental Table 1). Interestingly, there is considerable overlap among the susceptibility loci for these phenotypes. For example, 6p21.32 major histocompatibility class III region (advanced glycosylation end product–specific receptor [*AGER*]/mut S homolog 5 [*MSH5i*]), 15q24-25.1 cholinergic receptor, nicotinic, alpha 5 (neuronal) (*CHRNA3*)/cholinergic receptor, nicotinic, alpha 5 (neuronal) (*CHRNA5*)/iron-responsive element binding protein neuronal 2 (*IREB2*), and 19q13.2 cytochrome P450, family 2, subfamily A, polypeptide 6 (*CYP2A6*) are shared by all four phenotypes; 5p15.33 telomerase reverse transcriptase (*TERT*)/CLPTMI-like (*CLPTM1L*)/aryl hydrocarbon receptor repressor (*AHRR*), 10q25.1 glutathione S-transferase omega 2 (*GSTO2*)/vesicle transport through interaction with t-SNAREs 1A (*VTI1A*), and 10q23.31 actin, alpha 2, smooth muscle, aorta (*ACTA2*)/phospholipase C, epsilon 1 (*PLCE1*) are shared by three of these phenotypes; and more than 15 loci are shared by two of the four phenotypes (see Supplemental Table 1). Therefore, LC and COPD are not discrete diseases related only through smoking exposure; they may also share genetic predisposition mechanisms.

Given the common variants that have been found to be associated with LC in GWAS, exome sequencing with a focused analysis provides a cost-effective approach for further investigation of high-priority regions of the genome and has great potential to identify rare causal variants in GWAS loci, as targeted studies of inflammatory bowel disease[3] and hypertriglyceridemia[4] have demonstrated. Rare variants, with minor allele frequencies (MAFs) less than 0.01 and modest to high effect sizes,[5–7] may play a crucial role in the etiology of complex traits and could account for missing heritability that is unexplained by common variants.

Our approach to unveiling these hidden rare variants was to sequence selective cases of LC by adopting a modified extreme phenotype approach. Only approximately 13% of cases of LC are reported as familial[8]; however, individuals with a family history of LC are at an approximately twofold to threefold higher risk for development of the disease.[9,10] Therefore, it could be assumed that patients with LC who are from high-risk families would tend to reflect the genetic component of the etiology of LC more clearly than those who are not from high-risk families. In the present study, to search for the disease-causing rare germline mutations within the target 107 GWAS loci, we selected (1) 48 patients with sporadic LC who reported histories of heavy smoking and 37 of whom exhibited carefully documented severe COPD (in which the environmental factor of smoking is considered overwhelming), and (2) 54 unrelated unique patients with familial LC who were from families with at least three first-degree relatives with LC (and who are likely enriched for genomic signal).

## Methods

### *Study population*

**Study subjects with familial LC.** Phenotype data and biological specimens for 54 patients with LC who had three or more first-degree relatives affected with histologically confirmed LC were provided by the Genetic Epidemiology of Lung Cancer Consortium (GELCC) collection. Only one patient with LC per family was included in the current study. The selection criteria included availability of adequate amounts of good-quality genomic DNA stored at the GELCC biorepository for probands and for whom no DNA samples on other affected family members were available. Samples and data were collected by the familial LC recruitment sites of the GELCC, which included the University of Cincinnati, University of Colorado Health Science Center, Karmanos Cancer Institute at Wayne State University, Louisiana State University Health Sciences Center-New Orleans, Mayo Clinic, University of Toledo, Johns Hopkins University, and Saccomanno Research Institute. The GELCC study population and recruitment scheme have been described in detail previously.[11] COPD phenotype on these patients with familial LC was not available.

**Study subjects with sporadic LC.** Forty-eight patients with sporadic LC were selected from enrollees into an ongoing study of COPD in current or former smokers

that was launched in 2002.[12–14] Ever-smokers older than 40 years were enrolled from three clinics within the Texas Medical Center in Houston, Texas: Ben Taub General Hospital, Houston Methodist Hospital, and Michael E. DeBakey Veterans Affairs Medical Center. The COPD phenotype was carefully defined by irreversible airflow limitation (reduced forced expiratory volume in 1 second <50% predicted and forced expiratory volume in 1 second/forced vital capacity <0.7) assessed by postbronchodilator spirometry. For this analysis, we selected smokers enrolled in this study who had histologically confirmed LC. Information on family history of LC was not available for these patients with sporadic LC.

DNA was isolated from the peripheral blood of the patients with familial LC and those with sporadic LC. The study was approved by the institutional review board of all sites accruing participants and by the institutional review board at the Baylor College of Medicine (BCM) for exome sequencing conducted at the BCM Human Genome Sequencing Center (HGSC).

### *Library preparation and capture enrichment*

DNA samples were constructed into Illumina paired-end precapture libraries according to the manufacturer's protocol (Illumina Multiplexing_SamplePrep_Guide_1005361_D). The complete library and capture protocol, as well as the oligonucleotide sequences, have been described in detail previously.[15] For exome capture, each library pool was hybridized in solution to the BCM-HGSC–designed VCRome 2.1 capture reagent according to the manufacturer's protocol (NimbleGen) with minor revisions.

### *Exome sequencing, alignment, and variant calling*

The sequencing runs were performed in paired-end mode using the Illumina HiSeq 2000 platform. Sequence analysis was performed using the BCM-HGSC Mercury analysis pipeline.[16] All sequence reads were mapped to the GRCh37 human reference genome using the Burrows-Wheeler aligner.[17] The resultant binary alignment/map file was subjected to quality recalibration using Genome Analysis Toolkit.[18] Putative variants, including single-nucleotide variants (SNVs) and insertions or deletions (Indels), were called using the Atlas2 suite.[19] Read qualities were recalibrated with Genome Analysis Toolkit; a minimum quality score of 30 was required, and the variant had to have been present in more than 15% of the reads covering the position.

### *Variant annotation and filtering*

This analysis was restricted to rare mutations mapping to the exons within the 107 selected regions

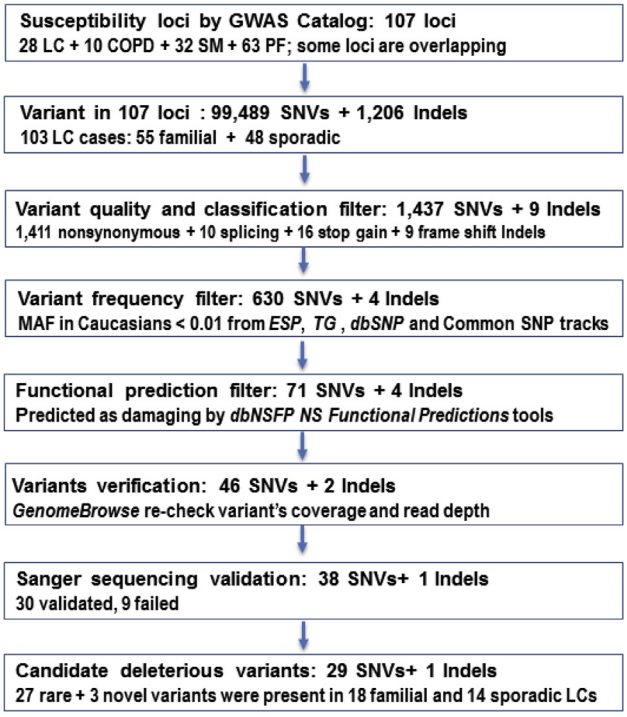


**Figure 1.** Workflow and annotation pipeline for the identification of candidate variants.

described earlier (see Supplemental Table 1 for genomic coordinates). Variants were annotated for effect on the protein and predicted function using the Single-Nucleotide Polymorphism (SNP) & Variation Suite (SVS) software (Golden Helix, Inc.). This suite integrates more than 378 databases for variant information including the following: (1) MAF in the European American population in the reference database (1000 Genomes [TG], Exome Sequencing Project [ESP] 6500) and the University of California, Santa Cruz Common SNPs 135/137/141 tracks, which include all variants with a MAF of at least 0.01 in the general population; (2) experimental evidence from disease variant databases (such as the Catalogue of Somatic Mutations in Cancer [*COSMIC*] and ClinVar); and (3) deleterious prediction of variant function determined either by mutation type (truncating, splicing, frame shift, stop gain/loss, or exonic Indels) or mutation effects predicted by dbNSFP Functional Predictions.

To generate a list of disease-causing candidate variants, we focused on identifying genes with rare and novel variants (never reported in a publicly available database or University of California, Santa Cruz All SNPs 135/137/141 tracks) (Fig. 1). We used scaled C-scores from the combined annotation-dependent depletion (CADD) method[20] for prioritization of causal variants. A C-score of 10, 20, and 30, indicates variants predicted to be in the top 10%, 1%, and 0.1% of the most deleterious in the human genome, respectively.[20] After implementing the

aforementioned filtering schema, we used GenomeBrowse (Golden Helix, Inc.) to visually confirm the potential candidate variants by rechecking the raw binary alignment/map file data. We then tabulated the number of candidate deleterious mutations per gene and within our two study subgroups (familial versus sporadic) and created a Venn diagram for the list of candidate variants that were significantly associated with the four different phenotypes (LC, COPD, PF, and SM) in previous GWAS.

### *Sanger validation*

The potential candidate variants were verified, and segregation was examined by using Sanger capillary bidirectional sequencing in the selected sample sites. Primers specific to the region containing the variant to be tested were designed, polymerase chain reactions (PCRs) were prepared according to the Qiagen Multiplex PCR Kit protocol (Qiagen), and touchdown PCR was performed (all PCR primers and conditions are available upon request). SNVs were identified using SNP Detector and visually displayed in Sequence Scanner v1.0 (Applied Biosystems).

### *Candidate variant protein annotation, structure modeling, and protein-protein interaction*

We used the online databases *Pfam*[21] and *PRINTS*[22] to annotate and classify protein families and domains, the BioGrid and the *STRING*[23] for predicting protein-protein interaction, and the PHYRE2 server[24] for modeling the 3D structure of the candidate variant gene encoded-protein. These resources use sequence-, structure-, and systems biology–based features to predict whether the mutation in the protein is likely to have a functional or phenotypic effect.

## Results

Demographic information, including age, sex, smoking history, and histologic diagnosis, is summarized in Table 1. All 54 unrelated patients with familial LC and 48 with sporadic LC were adult non-Hispanic whites. The mean ages of onset of LC in the patients with familial and sporadic LC were 56.0 and 60.9 years, respectively. More than 85% of those with familial LC and all those with sporadic LC (because of the study design criteria) reported being ever-smokers, with mean pack-years of 52.3, and 60.3, respectively. Overall, non–small cell LC had been diagnosed in 86.0% of those with familial LC and 90.5% of those with sporadic LC. Adenocarcinoma was diagnosed in 40.5% of those in the sporadic group and 30.2% of those in the familial group for whom histologic data were available.

Of 99,489 SNVs and 1206 Indels located in the exons of the target 107 loci, our stepwise filtering strategy identified 39 potential candidate variants (see Fig. 1). Of these 39 variants interrogated by Sanger sequencing, nine mutations failed, and 30 variants (80%) were verified in the original LC samples (Table 2). All the failed mutations were singletons. Of the 30 verified candidate variants, five variants were present in two or more patients, three variants were located in highly likely functional sites (*CHRNA5* g.78880766 splice donor, myozenin 3 [*MYOZ3*] g.150051315 splice acceptor, and chromosome 10 open reading frame 11 [*C10orf11*] p.Ser8 frameshift), and three SNVs were novel (patatin-like phospholipase domain containing 8 [*PNPLA8*] p.Ile479Ser, pantothenate kinase 1 [*PANK1*] p.Phe163Ser, and insulin-degrading enzyme [*IDE*] p.Asp9Asn) (see Table 2).

**Table 1.** Demographic and histologic characteristics of patients with familial and sporadic lung cancer

| Characteristics | Familial LC (n = 54) | Sporadic LC[a] (n = 48) |
|---|---|---|
| Age of diagnosis | | |
| Mean (SD) | 56.0 (10.1) | 60.9 (4.7) |
| Range | 30-70 | 48-65 |
| Sex | | |
| Male (%) | 22 (40.0) | 44 (91.7) |
| Female (%) | 32 (60.0) | 4 (8.3) |
| Smoking history | | |
| Ever-smoker (%) | 46 (85.2) | 48 (100) |
| Nonsmoker (%) | 8 (14.8) | 0 (0) |
| Cigarette pack-years | | |
| Mean (SD) | 52.3 (30.8) | 60.3 (30.9) |
| Range | 0-165 | 14-150 |
| Histologic characteristics[b] | | |
| NSCLC | 37 (86.0) | 37 (90.5) |
| Adenocarcinoma | 13 (30.2) | 17 (40.5) |
| Squamous cell carcinoma | 17 (39.5) | 16 (38.1) |
| Large cell carcinoma | 7 (16.3) | 5 (11.9) |
| SCLC | 6 (14.0) | 4 (9.5) |

[a]Of the patients in the 48 sporadic cases of LC, 37 also had severe COPD.
[b]Numbers do not add up because of missing data.
LC, lung cancer; NSCLC, non-small cell lung cancer; SCLC, small cell lung cancer.

Overall, the total number and proportion of patients with LC (N = 32) who carried these 30 candidate variants were only slightly higher in the group with familial cases (18 cases, 18/54 = 33.3%) than in the group with sporadic cases (14 cases, 14/48 = 29.2%, with 11 of these 14 patients also having severe COPD). The mean ages of the familial and sporadic candidate mutation carriers were not different from the overall means. In terms of smoking intensity, however, carriers of familial mutations reported fewer pack-years than their mean (43 versus 52), whereas there was no difference in smoking intensity among the carriers of sporadic mutations.

We identified two highly deleterious mutations occurring in more than three patients with LC (see Table 2

**Table 2.** List of 30 candidate deleterious germline mutations in familial and sporadic cases of lung cancer

| Region | Disease association | Gene | Marker[a] | SNV/Indels | Ref./ Alt. | RS ID | MAF in TG/ESP | CADD C-score[b] | N mutated familial vs. sporadic | Total N mutated LC cases |
|---|---|---|---|---|---|---|---|---|---|---|
| 10q25.1 | LC+SM+PF | *CCDC147* | 10:106163533-SNV | p.Arg696Cys | C/T | rs41291850 | 0.0026/0.0072 | 16.2 | 2 : 1 | 3 |
| 9q34.2 | SM | *DBH* | 9:136501569-SNV | p.Val26Met | G/A | rs76856960 | 0.0034/0.0045 | 17.3 | 0 : 2[c] | 3 |
| | | | 9:136522317-SNV | p.Met563Thr | T/C | rs201973877 | 0.0002/0.0002 | 20.3 | 1 : 0 | |
| 15q25.1 | LC+SM+PF+COPD | *IREB2* | 15:78783019-SNV | p.Gly747Glu | G/A | rs139092247 | 0.0014/0.0034 | 35 | 1 : 0 | 3 |
| | | *CHRNA5* | 15:78880766-SNV | g. splice donor | G/A | rs200616965 | NA | 22.7 | 0 : 1[c,d] | |
| | | *CHRNB4* | 15:78921343-SNV | p.Ala435Val | G/A | rs56317523 | 0.0008/0.0028 | 27.2 | 1 : 0 | |
| 16q23.1 | PF | *KARS* | 16:75665388-SNV | p.Arg421Gln | C/T | rs149772470 | 0.0002/0.0018 | 26.1 | 0 : 1[c] | 3 |
| | | | 16:75665146-SNV | p.Arg448Cys | G/A | rs77573084 | 0.0006/0.0030 | 19.6 | 0 : 1[c,d] | |
| | | *WWOX* | 16:78466521-SNV | p.Arg310Cys | C/T | rs193001955 | 0.0006/0.0006 | 24.1 | 0 : 1[c] | |
| 1q44 | SM | *C1orf100* | 1:244541827-SNV | p.Asp71His | G/C | rs41269385 | 0.0022/0.0065 | *14.2* | 2 : 0 | 2 |
| 2q35 | PF | *TNS1* | 2:218686643-SNV | p.Glu1027Val | T/A | rs112371945 | 0.0006/0.0013 | 22.7 | 1 : 0 | 2 |
| | | | 2:218669288-SNV | p.Thr1701Met | G/A | rs61740054 | 0.0010/0.0034 | 27.9 | 0 : 1[c] | |
| 5q32 | LC+PF | *FBXO38* | 5:147817940-SNV | p.Pro893Arg | C/G | rs141168806 | NA/0.0001 | 22.9 | 1 : 0 | 2 |
| | | | 5:147821690-SNV | p.Val1108Ile | G/A | rs143682696 | 0.0002/0.0008 | 26.4 | 1 : 0 | |
| 7q31.1 | SM | *PNPLA8* | 7:108154659-SNV | p.Cys379Gly | A/C | rs141089628 | 0.0002/0.0033 | 15.1 | 0 : 1[c] | 2 |
| | | | 7:108137944-SNV | p.Ile479Ser | A/C | Novel | NA | 25.2 | 1 : 0 | |
| 10q23.31 | LC+SM+COPD | *PANK1* | 10:91359156-SNV | p.Phe163Ser | A/G | Novel | NA | 26.1 | 1 : 0 | 2 |
| | | *IDE* | 10:94243061-SNV | p.Asp9Asn | C/T | Novel | NA | 36 | 0 : 1[c,e] | |
| 13q12.12 | LC | *MIPEP* | 13:24448998-SNV | p.Leu197Pro | A/G | rs150167906 | 0.0002/0.0023 | 21.4 | 1 : 1 | 2 |
| 14q22.1 | PF | *NID2* | 14:52508948-SNV | p.Thr567Met | G/A | rs150406341 | 0.0006/0.0060 | 19.3 | 1[f] : 1[c,d] | 2 |
| 17q24.2 | LC+PF | *BPTF* | 17:65889520-SNV | p.Arg823Gln | G/A | rs375975293 | NA/0.0001 | 19.8 | 1 : 0 | 2 |
| | | | 17:65936627-SNV | p.Thr2237Met | C/T | rs372551122 | NA | 17.2 | 0 : 1[c] | |
| 5q33.1 | LC+PF | *MYOZ3* | 5:150051315-SNV | g. splice acceptor | A/G | rs143036945 | 0.0002/0.0005 | 12.3 | 0 : 1 | 1 |
| 10q22.2-3 | PF | *C10orf11* | 10:77542754-Deletion | p.Ser8 Frameshift | C/- | rs146123023 | 0.007/0.0013 | NA | 1 : 0 | 1 |
| 12q13.3 | PF | *LRP1* | 12:57577915-SNV | p.Arg1993Trp | C/T | rs141826184 | 0.0004/0.0031 | 21.8 | 0 : 1[c] | 1 |
| 12q21.2 | SM | *NAV3* | 12:78392209-SNV | p.Ser278Ile | G/T | rs755721519 | NA | 31 | 0 : 1[c,e] | 1 |
| 14q24.2 | PF | *SLC8A3* | 14:70515508-SNV | p.Val152Met | C/T | rs144289733 | 0.0004/0.0008 | 27 | 1 : 0 | 1 |
| 15q15.2 | LC | *TGM5* | 15:43527092-SNV | p.Tyr502His | A/G | rs146901531 | 0.0002/0.0006 | 18.8 | 0 : 1[c] | 1 |
| 18p11.3 | LC | *LAMA1* | 18:6965341-SNV | p.Arg2381Cys | G/A | rs142063208 | 0.0028/0.0016 | 25 | 1[f] : 0 | 1 |
| 19q13.2 | LC+SM+PF+COPD | *EGLN2* | 19:41307024-SNV | p.Val183Met | G/A | rs117916638 | 0.0002/0.0004 | 16.9 | 1 : 0 | 1 |

[a]All are heterozygous mutations.
[b]C-score is the overall measure of deleteriousness. C-score $\geq$ 20 indicates top 1% deleterious germline mutations in the human genome.
[c]Sporadic LC patient(s) with severe COPD.
[d]Entries followed by superscript "d" refer to the same patient.
[e]Entries followed by superscript "e" refer to the same patient.
[f]Entries followed by superscript "f" refer to the same patient.
SNV, Single-nucleotide variants; Indels, insertions or deletions; Ref., reference; Alt., Alternative; RS ID, reference SNP ID number; MAF, minor allele frequency; TG/ESP, Egl-9 family hypoxia-inducible factor 2; CADD, Combined Annotation-Dependent Depletion; LC, lung cancer; SM, smoking behavior; PF, pulmonary function; *CCDC147*, coiled-coil domain containing 147; *DBH*, dopamine $\beta$-hydroxylase; COPD, chronic obstructive pulmonary disease; *IREB2*, iron-responsive element binding protein 2; *CHRNA5*, cholinergic receptor, nicotinic, alpha 5 (neuronal); *CHRNB4*, cholinergic receptor, nicotinic, beta 4 (neuronal); *KARS*, lysyl-tRNA synthetase; *WWOX*, WW domain-containing oxidoreductase; *C1ORF100*, chromosome 1 open reading frame 100; *TNS1*, tensin 1; *FBXO38*, F-box protein 38; *PNPLA8*, patatin-like phospholipase domain containing 8; *PANK1*, pantothenate kinase 1; *IDE*, insulin-degrading enzyme; *MIPEP*, mitochondrial intermediate peptidase; *NID2*, nidogen 2 (osteonidogen); *BPTF*, bromodomain PHD transcription factor; *MYOZ3*, myozenin 3; *C10ORF11*, chromosome 10 open reading frame 11; *LRP1*, low-density lipoprotein receptor-related protein 1; *NAV3*, neuron navigator 3; *SLC8A3*, solute carrier family 8 (sodium/calcium exchanger), member 3; *TGM5*, transglutaminase 5; *LAMA1*, lamanin, alpha 1.

and Supplemental Fig. 1). The first was a heterozygous c.2086C>T in the coiled-coil domain-containing 147 gene (*CCDC147,* also called *CFAP58*), resulting in a p.Arg696Cys substitution. This variant was identified in two patients with familial LC (both women, mean age 54.5, mean pack-years 40, and squamous histologic features) and one patient with sporadic LC (a 57-year-old man, pack-years 88, with adenocarcinoma but without COPD). Notably, the MAF of this variant is 0.0026 from the TG, and 0.0072 from the ESP6500 databases. The mutation is predicted to be protein damaging by PolyPhen-2 (score: 1.0) and highly-functional by Mutation taster. This variant has a high scaled CADD C-score of 16.2, which indicates that the Arg696 is predicted to be in the top 10% possible deleterious substitutions in the human genome. The *CCDC147* spans 101kb, contains 18 exons, and has 872 amino acids (AAs) (Fig. 2); the p.Arg696Cys is located in exon 14 and affects a strictly evolutionarily conserved AA residue in the crystal structure of tropomyosin (protein ID: Q5T655). This p.Arg696Cys mutation is predicted to perturb the tertiary structure (folding of the domain and stability of the three-dimensional shape) of the protein because the Cys696 forms a covalent bond disulfide bridge with 697Cys. It is also very close to the p.Arg698Gln, which is a confirmed somatic mutation in patients with cutaneous melanoma shown from *COSMIC* database and the acetylation modification site Lys692.

The second candidates were two missense SNVs, p.Val26Met and p.Met563Thr, in the dopamine beta-hydroxylase *(DBH)* gene (see Table 2 and Supplemental Fig. 1). The p.Val26Met was found in two patients with sporadic LC (both men, age 64 and 65, pack-years 40, histologic diagnosis of adenocarcinoma with severe COPD), and p.Met563Thr found in one patient with familial LC (a man, age 30, pack-years 52, histologic diagnosis not specified). The MAFs of these two variants were 0.0034/0.0045 and 0.0002/0.0002 from the TG/ESP6500 databases, respectively. The two mutations were predicted to be protein damaging by PolyPhen-2 (score 0.93 and 0.87), with CADD C-scores of 17.3 and 20.3, and they exhibited extremely high degrees of sequence conservation (0.96 and 0.99, respectively). The *DBH* gene contains 12 exons, spans 23 kb, and has 617 AAs (protein ID: P09172; see Fig. 2). The p.Val26Met is located within exon 1, lies in the hydrophobic transmembrane region, and possesses a helical structure. The p.Met563Thr located in exon 11, and it lies in a highly conserved region of $\alpha$-helix and the dopamine $\beta$-monoxgenase (DBM) motif IX that may influence the stability of the enzyme. This somatic mutation is also reported in patients with acute myeloid leukemia from the *COSMIC* database. These observations suggest that the two *DBH* mutations are likely to have a detrimental effect on the protein.

The other interesting candidates were three SNVs located in the 15q25.1 loci: *IREB2* p.Gly747Glu; *CHRNA5* g.78880766 splice donor; and cholinergic receptor, nicotinic, beta 4 (*CHRNB4*) p.Ala435Val. The *CHRNA5* splicing variant found in a patient with sporadic LC (a 63-year-old man, 48 pack-years of smoking, non–small cell LC, with severe COPD), who was also a carrier of another two candidate mutations (nidogen 2 (osteonidogen) [*NID2*] p.Thr567Met and lysyl-tRNA synthetase [*KARS*] gene p.Arg448Cys); the *IREB2* and *CHRNB4* SNVs were found in two patients with familial LC (both women, ages 45 and 64, pack-years 45 and 80, with small cell LC and unknown histologic diagnoses).

There were three additional candidate variants—*NID2* p.Thr567Met, mitochondrial intermediate peptidase (*MIPEP*) p.Leu197Pro, and chromosome 1 open reading frame 100 *(C1orf100*) p.Asp71His—that were present in multiple LC cases. Other genes that harbored multiple different mutations in different patients included tensin 1 (*TNS1*), F-box protein 38 (*FBXO38*), *PNPLA8, KARS*, and bromodomain PHD transcription factor (*BPTF*). In addition, a patient with sporadic LC and a history of extremely heavy smoking (a 65-year-old man, 150 pack-years of smoking, adenocarcinoma, and severe COPD) was a carrier of two novel mutations with a CADD C-score higher than 30 (*IDE* p.Asp9Asn and neuron navigator 3 [*NAV3*] p.Ser278Ile) (see Table 2).

Of the 30 candidate variants belonging to 20 loci and 24 genes (Fig. 3A), seven genes (including *CCDC147* and *DBH*) had candidate variants observed both in those with familial LC and in those with sporadic LC. Also (as shown in Fig. 3B), among the candidate genes examined in the current study, the *CCDC147, IREB2/CHRNA5/CHRNB4, PANK1/IDE*, and egl-9 family hypoxia-inducible factor 2 (*EGLN2*) genes were shared by three or more phenotypes (LC, COPD, PF, and SM) from the previously published GWAS (see Table 2 and Fig. 3B).

## Discussion

Despite previous family-based linkage studies, intensive population-based GWAS analyses, and candidate gene screening, a large proportion of the heritability of LC remains unexplained. Using an extreme phenotype design, this report describes the first exome sequencing approach comparing heavy smokers with familial and sporadic LC and evaluating the effects of rare coding variation in the GWAS loci associated with LC, COPD, SM, and PF. Our results showed that the familial mutation carriers reported fewer pack-years than their group's mean (43 versus 52), whereas there was no difference in smoking intensity among the sporadic carriers. Furthermore, we identified two disease-causing rare mutations on 10q25.1 (*CCDC147* p.Arg696Cys) and

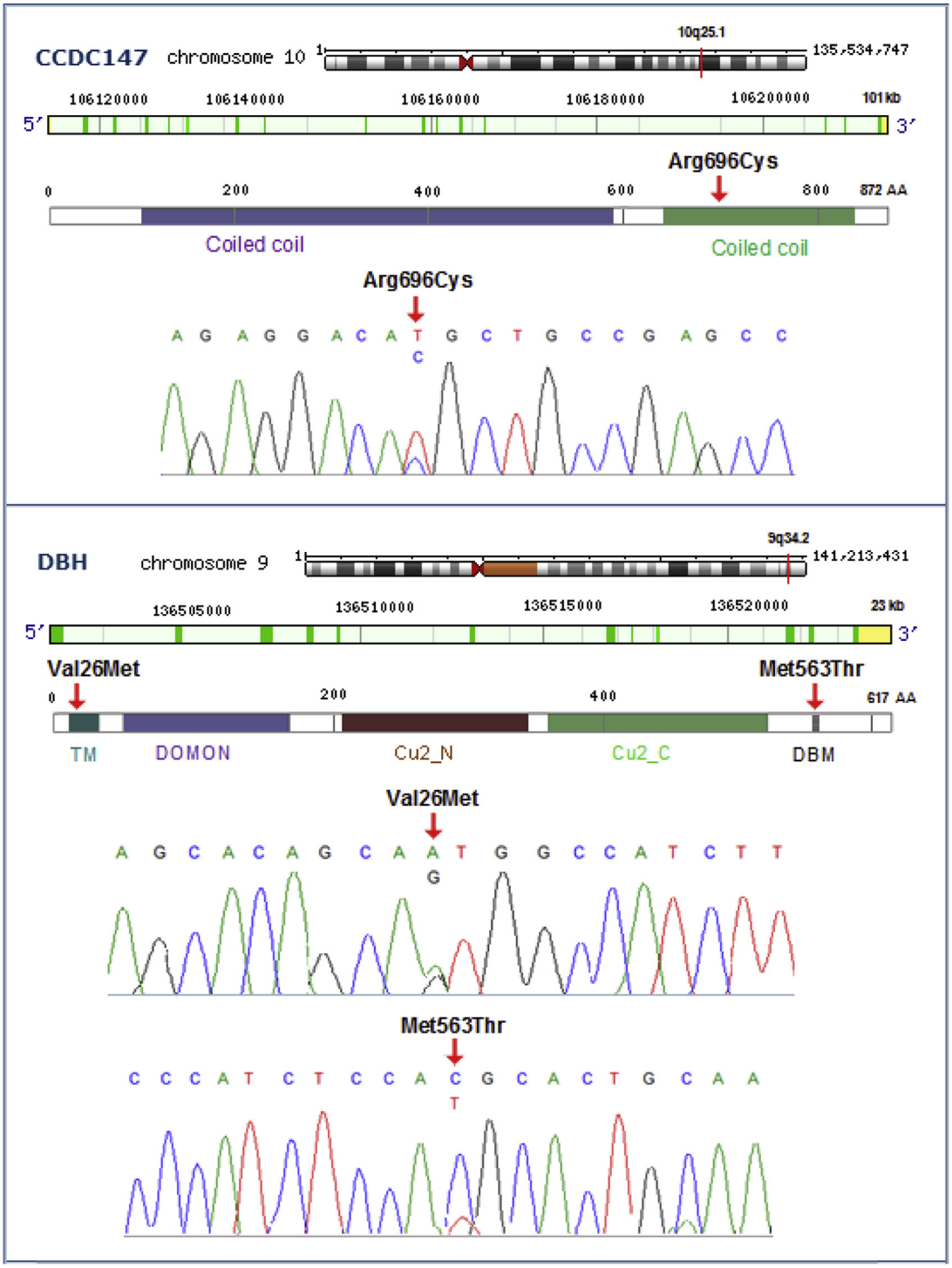


**Figure 2.** Chromosomal position, gene structure, protein domain(s), and sequence of the candidate mutations of coiled-coil domain containing 147 (*CCDC147l*) and dopamine $\beta$-hydroxylase (*DBH*) genes. The mutations were confirmed with Sanger sequencing and are indicated with red arrows. *CCDC147* includes two coiled coil regions: 106-595 AA and 642-839 AA. DBH consist of five domains: transmembrane (20-37 AA), DOMON (dopamine $\beta$-monooxygenase N-terminal; 56-172 AA), Cu2_N (copper type II, N-terminal; 213-341 AA), Cu2_C (copper type II, C-terminal; 360-524 AA), and dopamine $\beta$-monooxygenase motif IX; 552-572 AA).

9q34.2 (*DBH* p.Val26Met and p.Met563Thr), and three suggestive rare mutations on 15q25.1 (*IREB2* p.Gly747-Glu, *CHRNA5* g.78880766 splice donor, and *CHRNB4* p.Ala435Val), although the findings require replication. Patients with familial LC and patients with sporadic LC are indistinguishable at initial clinical examination, and our results demonstrated that the two forms of LC may have both shared determinants and distinct components.

Strong evidence for an LC-conferring deleterious mutation was observed at *CCDC147* p.Arg696Cys in three patients with LC (two with familial LC and one with sporadic LC). Interestingly, the two familial carriers were lighter smokers and had an earlier age of onset than the overall mean for familial cases. The sporadic carrier was a heavier smoker with a history of 88 pack-years and no documented COPD. Although several genes in 10q25.1 loci have been implicated in susceptibility to LC,[25] PF,[26] and SM[27] in GWAS, very little is known about the function of the *CCDC147* gene in humans or mice, although it is thought to produce a functional protein as described in the Proteomics database. CCDC147 protein, which is also known as cilia- and flagella-associated protein 58 (CFAP58), demonstrates high expression in T cells, nasal epithelium, lungs, and alveolar fluids (http://www.genecards.org/cgi-bin/carddisp.pl?gene=CFAP58). It is believed to interact with members of the shelterin complex, the human telomere repeat binding factor 1 (TRF1) and protection of telomeres 1 (POT1), as reported in the BioGRID database and STRING Interaction Network. Interestingly, recent studies have shown that rare mutations in the gene *POT1* are associated with chronic lymphocytic leukemia,[28] familial melanoma,[29] and familial glioma,[30] in which it is thought to result in telomere deprotection and length extension associated with cancer. Furthermore, one of the most important functions of shelterin includes modulation of telomerase activity, which has been detected in approximately 85% of cancers and is linked to genomic instability and tumorigenesis. Although direct evidence regarding the biological function of CCDC147 is lacking, our finding of CCDC147 as a novel telomere-interacting protein underscores the need for future work that could elucidate the role of this gene in LC pathogenesis.

Another main finding was the highly disruptive and deleterious rare mutations on 9q34.2 *DBH,* p.Val26Met and p.Met563Thr, in three patients with LC (one with familial LC and two with sporadic LC). The familial carrier was very young (age 30). Both sporadic carriers had adenocarcinoma and severe COPD. Previous GWAS identified *DBH* rs3025343 as a locus associated with SM.[31] The *DBH* (OMIM 609312) gene contributes primarily to conversion of dopamine to noradrenaline. Dopamine is known to be released from neurons in response to nicotine and plays a well-documented role in determining an individual's predisposition to nicotine dependence through its role in mediating drug reward in the brain.[31–34] The contribution of cigarette smoking to both LC and COPD could invoke a variety of underlying biological processes, including inflammation, epithelial-mesenchymal transition, oxidative stress, DNA repair, and abnormal cellular proliferation.

*NID2* is a known GWAS hit for PF[35] and blood lipid phenotypes[36] and a new biomarker for ovarian cancer,[37]

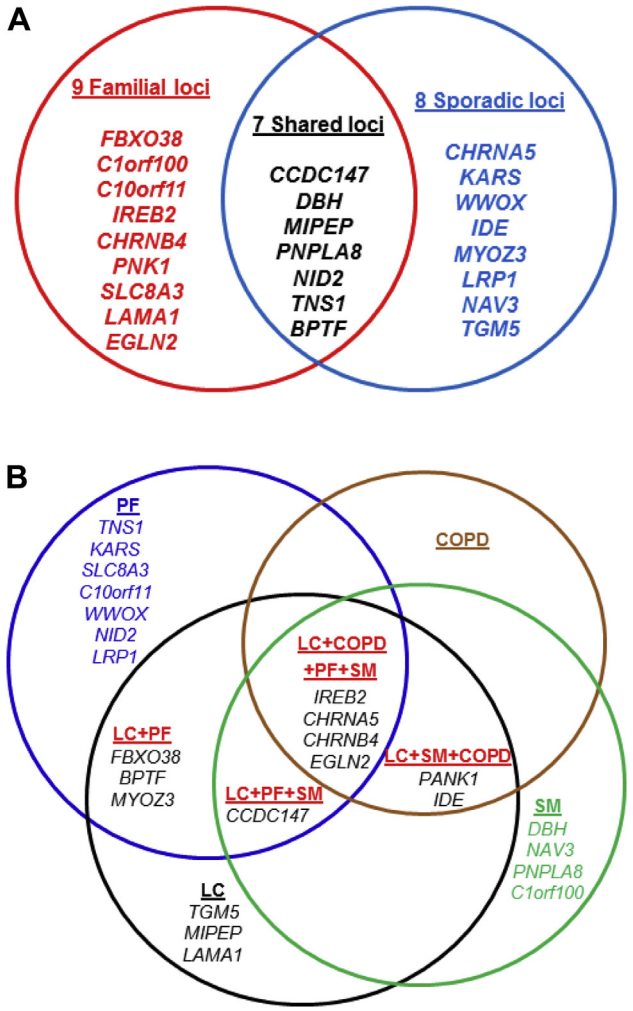


**Figure 3.** Venn diagrams and schematic representations of all genes with candidate mutations in the familial and sporadic lung cancer groups. (*A*) Shared and specific genes with candidate deleterious variants in the familial, sporadic, or both lung cancer groups. (*B*) The list of genes with candidate deleterious variants that were significantly associated with these four different phenotypes in previous genome-wide association studies.

hepatocellular carcinoma,[38] and oral squamous cell carcinoma.[39] *NID2* (OMIM 605399) encodes a member of the highly conserved nidogen family of basement membrane proteins. This protein binds collagens I and IV and laminin, is involved in stabilizing and maintaining the structure of the basement membrane, and plays a key role in the cell-extracellular matrix. Unbalanced proteolysis in the extracellular matrix is a potential mechanism to explain inflammatory processes within the emphysematous lung. *NID2* mutation in patients with LC may be conducive to invasion and metastasis of tumor cells by loosening cell interaction with basal membrane and weakening the strength of the basement membrane itself, and it could be a marker of progression as well.

A main strength of this study is its focus on patients with extreme phenotypes, who are the most likely to be informative. For quantitative traits, one can select individuals with extreme trait values after adjusting for known covariates. Alternatively, in disease-focused studies, selection of individuals with extreme phenotypes can be conducted on the basis of known risk factors. Smoking, family history of LC, and COPD are all well-documented risk factors for LC. Because the frequencies of alleles that contribute to the trait/disease are enriched in phenotype extremes (such as familial LC or patients with both LC and COPD), studying extremes has been shown to provide more than five times the power (only 20% of the subjects compared with in traditional designs).[7] In the present study, the recurrent rare mutations described herein suggest that it may be possible to identify susceptibility genes in a relatively small sample size, although we cannot rule out the possibility that the results have been observed by chance. The small sample size and lack of validation of the identified mutations in a separate large-scale cohort limit the relevance of our findings. Another limitation of this analysis is phenotype misclassification between familial and sporadic LC. For the patients with familial LC, we lacked COPD phenotype data, and for those with sporadic LC, family history of LC was not available. Also, we acknowledge the existence of a sex imbalance between the familial and sporadic cases that could cause bias and limit applicability of the findings to the general population.

In summary, our results demonstrated highly disruptive germline mutations in the genes *CCDC147* and *DBH* in patients with LC that are interesting candidates for LC risk alleles. The overlap in risk loci between familial and sporadic LC, and that between COPD and LC, may be due to genes and mutations involving telomere maintenance, to inflammation, or to the lack of family history in the sporadic cases being the result of no smoking exposure in other carriers of the mutation in their families. Therefore, going forward, comprehensive genomic analyses of whole genomes (from point mutations to large structural variants) and a large number of LC samples from diverse race/ethnic groups for validation, as well as further functional works for the top two candidate genes, will be needed to better understand the underlying molecular genetics and guide screening for mutations in this unique subset of patients to assess their potential risk for LC.

## Acknowledgments

This work was supported by grants from the National Institutes of Health (R01 CA127219, R01 HL082487, R01 HL110883, K07CA181480, R01 CA060691, R01 CA87895, R01 CA80127, R01 CA84354, R01 CA134682, R01 CA134433, R03 CA77118, P20GM103534, P30CA125123, P30CA023108, P30-ES006096, P30CA022453, N01-HG-65404, U01CA076293, U19CA148127, and HHSN268201200007C). Dr. Bailey-Wilson was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. Additional support was provided by the National Library of Medicine T15LM007093 (Davis) and the Population Sciences Biorepository at BCM. 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 HGSC-BCM for performing the exome sequencing for all the samples in this study.

## Appendix

### *Web resources*

GWAS, Genome-wide Association Studies Catalog, www.genome.gov/gwastudies/

TG, Thousand Genomes, http://www.1000genomes.org

ESP, Exome Sequencing Project, http://evs.gs.washington.edu/EVS/

COSMIC, Catalogue of Somatic Mutations in Cancer, http://cancer.sanger.ac.uk/cosmic

OMIM, Online Mendelian Inheritance in Man, http://www.omim.org

CADD, Combined Annotation Dependent Depletion, cadd.gs.washington.edu/

Pfam, http://pfam.janelia.org/

PRINTS, protein motif fingerprint database, http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/

BioGRID, Biological General Repository for Interaction Datasets, http://www.thebiogrid.org

STRING, Search Tool for the Retrieval of Interacting Genes, http://string-db.org/

Phyre2, www.sbg.bio.ic.ac.uk/phyre2/

## Supplementary Data

Note: To access the supplementary material accompanying this article, visit the online version of *Journal of Thoracic Oncology* at www.jto.org and at http://dx.doi.org/10.1016/j.jtho.2015.09.015.

## References

1. Etzel CJ, Kachroo S, Liu M, et al. Development and validation of a lung cancer risk prediction model for African-Americans. *Cancer Prev Res (Phila)*. 2008;1:255–265.
2. Tammemagi CM, Pinsky PF, Caporaso NE, et al. Lung cancer risk prediction: Prostate, Lung, Colorectal And Ovarian Cancer Screening Trial models and validation. *J Natl Cancer Inst*. 2011;103:1058–1068.
3. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of *GWAS* loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet*. 2011;43:1066–1073.

4. Johansen CT, Wang J, Lanktree MB, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet*. 2010;42:684–687.
5. Kang G, Lin D, Hakonarson H, et al. Two-stage extreme phenotype sequencing design for discovering and testing common and rare genetic variants: efficiency and power. *Hum Hered*. 2012;73:139–147.
6. Lamina C. Digging into the extremes: a useful approach for the analysis of rare variants with continuous traits? *BMC Proc*. 2011;5(suppl 9):S105.
7. Li D, Lewinger JP, Gauderman WJ, et al. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genet Epidemiol*. 2011;35:790–799.
8. Bailey-Wilson JE, Amos CI, Pinney SM, et al. A major lung cancer susceptibility locus maps to chromosome 6q23-25. *Am J Hum Genet*. 2004;75:460–474.
9. Ooi WL, Elston RC, Chen VW, et al. Increased familial risk for lung cancer. *J Natl Cancer Inst*. 1986;76:217–222.
10. Jonsson S, Thorsteinsdottir U, Gudbjartsson DF, et al. Familial risk of lung carcinoma in the Icelandic population. *JAMA*. 2004;292:2977–2983.
11. 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.
12. Lee SH, Goswami S, Grudo A, et al. Antielastin autoimmunity in tobacco smoking-induced emphysema. *Nat Med*. 2007;13:567–569.
13. 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.
14. 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.
15. 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.
16. 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.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754–1760.
18. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–498.
19. 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.
20. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46:310–315.
21. Finn RD, Mistry J, Schuster-Bockler B, et al. Pfam: clans, web tools and services. *Nucleic Acids Res*. 2006;34 (database issue):D247–D251.
22. Attwood TK, Bradley P, Flower DR, et al. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*. 2003;31:400–402.
23. Stark C, Breitkreutz BJ, Reguly T, et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. 2006;34(database issue):D535–D539.
24. Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc*. 2009;4:363–371.
25. Lan Q, Hsiung CA, Matsuo K, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet*. 2012;44:1330–1335.
26. 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.
27. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*. 2010;42:441–447.
28. Ramsay AJ, Quesada V, Foronda M, et al. *POT1* mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat Genet*. 2013;45:526–530.
29. Robles-Espinoza CD, Harland M, Ramsay AJ, et al. *POT1* loss-of-function variants predispose to familial melanoma. *Nat Genet*. 2014;46:478–481.
30. Bainbridge MN, Armstrong GN, Gramatges MM, et al. Germline mutations in shelterin complex genes are associated with familial glioma. *J Natl Cancer Inst*. 2015;107:384.
31. Siedlinski M, Cho MH, Bakke P, et al. Genome-wide association study of smoking behaviours in patients with COPD. *Thorax*. 2011;66:894–902.
32. Shiels MS, Huang HY, Hoffman SC, et al. A community-based study of cigarette smoking behavior in relation to variation in three genes involved in dopamine metabolism: catechol-*o*-methyltransferase (COMT), dopamine beta-hydroxylase (DBH) and monoamine oxidase-A (MAO-A). *Prev Med*. 2008;47:116–122.
33. Freire MT, Marques FZ, Hutz MH, et al. Polymorphisms in the *DBH* and *DRD2* gene regions and smoking behavior. *Eur Arch Psychiatry Clin Neurosci*. 2006;256:93–97.
34. Zhang XY, Chen da C, Xiu MH, et al. Association of functional dopamine-beta-hydroxylase (*DBH*) 19 bp insertion/deletion polymorphism with smoking severity in male schizophrenic smokers. *Schizophr Res*. 2012;141:48–53.
35. Wilk JB, Walter RE, Laramie JM, et al. Framingham Heart Study genome-wide association: results for pulmonary function measures. *BMC Med Genet*. 2007;8(suppl 1):S8.
36. Kathiresan S, Manning AK, Demissie S, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet*. 2007;8(suppl 1):S17.
37. Kuk C, Gunawardana CG, Soosaipillai A, et al. Nidogen-2: a new serum biomarker for ovarian cancer. *Clin Biochem*. 2010;43:355–361.
38. Cheng ZX, Huang XH, Wang Q, et al. Clinical significance of decreased nidogen-2 expression in the tumor tissue and serum of patients with hepatocellular carcinoma. *J Surg Oncol*. 2012;105:71–80.
39. Guerrero-Preston R, Soudry E, Acero J, et al. NID2 and HOXA9 promoter hypermethylation as biomarkers for prevention and early detection in oral cavity squamous cell carcinoma tissues and saliva. *Cancer Prev Res (Phila)*. 2011;4:1061–1072.