SCIENTIFIC REPORTS

OPEN

Genetic variants of *PTPN2* are associated with lung cancer risk: a re-analysis of eight GWASs in the TRICL-ILCCO consortium

Received: 22 November 2016
Accepted: 15 March 2017
Published online: 11 April 2017

Yun Feng[1,2,3], Yanru Wang[2,3], Hongliang Liu[2,3], Zhensheng Liu[2,3], Coleman Mills[2,3], Younghun Han[4], Rayjean J. Hung[5], Yonathan Brhane[5], John McLaughlin[6], Paul Brennan[7], Heike Bickeboeller[8], Albert Rosenberger[8], Richard S. Houlston[9], Neil E. Caporaso[10], Maria Teresa Landi[10], Irene Brueske[11], Angela Risch[12], Yuanqing Ye[13], Xifeng Wu[13], David C. Christiani[14], Christopher I. Amos[4] & Qingyi Wei[2,3]

The T-cell protein tyrosine phosphatase (TCPTP) pathway consists of signaling events mediated by TCPTP. Mutations and genetic variants of some genes in the TCPTP pathway are associated with lung cancer risk and survival. In the present study, we first investigated associations of 5,162 single nucleotide polymorphisms (SNPs) in 43 genes of this TCPTP pathway with lung cancer risk by using summary data of six published genome-wide association studies (GWAS) of 12,160 cases and 16,838 controls. We identified 11 independent SNPs in eight genes after correction for multiple comparisons by a false discovery rate <0.20. Then, we performed *in silico* functional analyses for these 11 SNPs by eQTL analysis, two of which, *PTPN2* SNPs rs2847297 and rs2847282, were chosen as tagSNPs. We further included two additional GWAS datasets of Harvard University (984 cases and 970 controls) and deCODE (1,319 cases and 26,380 controls), and the overall effects of these two SNPs among all eight GWAS studies remained significant (OR = 0.95, 95% CI = 0.92–0.98, and *P* = 0.004 for rs2847297; OR = 0.95, 95% CI = 0.92–0.99, and *P* = 0.009 for rs2847282). In conclusion, the *PTPN2* rs2847297 and rs2847282 may be potential susceptible loci for lung cancer risk.

Lung cancer is one of the most common human malignancies and the leading cause of cancer-related deaths in both men and women[1]. It is estimated that 224,390 new lung cancer cases will be diagnosed in the United States in 2016[2]. Lung cancer risk likely results from joint effects and interactions of environmental and genetic factors.

Single nucleotide polymorphisms (SNPs) are the most common genetic variants and have been shown to be associated with lung cancer risk[3]. Genome-wide association studies (GWAS) have identified 30 loci in 13 genomic regions to be associated with lung cancer risk[4–15]. However, most of the SNPs identified to date have not been

[1]Department of Respiration, Ruijin Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China. [2]Duke Cancer Institute, Duke University Medical Center, Durham, NC, 27710, USA. [3]Department of Medicine, Duke University School of Medicine, Durham, NC, 27710, USA. [4]Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Hanover, NH, 03755, USA. [5]Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada. [6]Public Health Ontario, Toronto, Ontario, M5T 3L9, Canada. [7]Genetic Epidemiology Group, International Agency for Research on Cancer (IARC), 69372, Lyon, France. [8]Department of Genetic Epidemiology, University Medical Center, Georg-August-University Göttingen, 37073, Göttingen, Germany. [9]Division of Genetics and Epidemiology, the Institute of Cancer Research, London, SW7 3RP, UK. [10]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, 20892, USA. [11]Helmholtz Centre Munich, German Research Centre for Environmental Health, Institute of Epidemiology I, 85764, Neuherberg, Germany. [12]Department of Molecular Biology, University of Salzburg, 5020, Salzburg, Austria. [13]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA. [14]Massachusetts General Hospital, Boston, MA 02114, USA, Department of Environmental Health, Harvard School of Public Health, Boston, MA, 02115, USA. Yun Feng and Yanru Wang contributed equally to this work. Correspondence and requests for materials should be addressed to Q.W. (email: qingyi.wei@duke.edu)

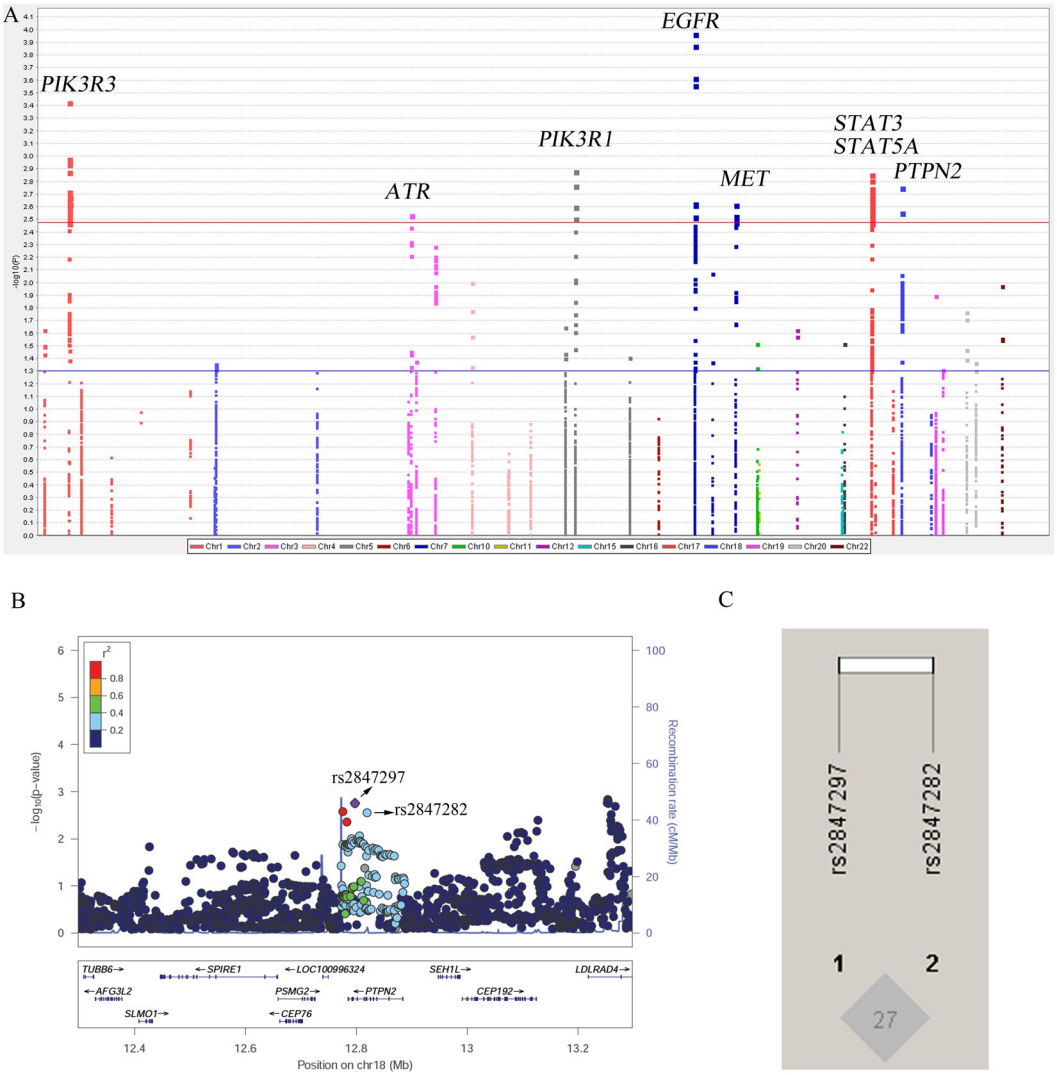

Figure 1. Screening of SNPs in the TCPTP pathway. (**A**) Manhattan plot of genome-wide association results of 5,162 SNPs in 43 TCPTP pathway genes and lung cancer risk in the TRICL-ILCCO Consortium. SNPs are plotted on the X-axis according to their positions on each chromosome. The association *P* values with lung cancer risk are shown on the Y-axis (as $-\log_{10}(P)$ values). The horizontal red line represents FDR threshold 0.20. The horizontal blue line represents *P* value of 0.05; (**B**) SNPs in *PTPN2* with 500 kb up- and downstream of the gene region and (**C**) LD plots of the SNPs in *PTPN2* with FDR <0.20. In **B**, the left-hand y-axis shows the association *P* value of each SNP, which is plotted as $-\log_{10}(P)$ against chromosomal base pair position; the right-hand y-axis shows the recombination rate estimated from the hg19/1000 Genomes European population.

shown to be functional. Other approaches to GWAS including pathway-based analysis with reduced dimension or multiple testing have been emerged to identify possible functional SNPs associated with lung cancer risk.

The T-cell protein tyrosine phosphatase (TCPTP/PTPN2) is an important member of the protein-tyrosine phosphatase (PTP) family. Activating and deactivating mutations in PTP genes often result in enzymes that can either promote or suppress oncogenesis. The TCPTP pathway consists of signaling events mediated by TCPTP through negative regulation of several receptor tyrosine kinases such as epidermal growth factor receptor (EGFR)[16], vascular endothelial growth factor receptor-2 (VEGFR2)[17], platelet-derived growth factor receptor beta (PDGFRβ)[18], signal transducer and activator of transcription subtypes 1 (STAT1)[19], 3 (STAT3)[20], and 6 (STAT6)[21], and the insulin receptor[22].

Studies have shown that mutations and genetic variants of some genes in the TCPTP pathway are associated with lung cancer risk and survival[23,24]. However, SNPs in many candidate genes in the pathway have not been studied and reported. In the present study, we systematically investigated all potentially functional SNPs in TCPTP pathway genes by assessing their associations of lung cancer risk using eight published lung cancer GWAS datasets.

SNP	Gene	Chr.	Allele[a]	SNPinfo	Regulome DB Score	HaploReg	*P*[b]		
							P (additive model)	*P* (dominant model)	*P* (recessive model)
rs7538978	*PIK3R3*	1	A/G	—	1 f	Enhancer histone marks: 9 tissues; DNAse: CRVX; Motifs changed: 6 altered motifs	0.722	0.249	0.175
rs11707731	*ATR*	3	G/T	—	4	Promoter histone marks: 4 tissues; Enhancer histone marks: 4 tissues; DNAse: ESC; Motifs changed: 4 altered motifs	0.579	0.719	0.420
rs706714	*PIK3R1*	5	A/C	TFBS	5	Enhancer histone marks: 7 tissues; DNAse: GI; Motifs changed: GATA,Nkx2,Nkx3	0.281	0.137	0.714
rs2740762	*EGFR*	7	C/A	TFBS	5	Enhancer histone marks: 13 tissues; DNAse: IPSC,MUS,PANC; Motifs changed: Foxo,NF-AT,Pax-4	0.053	0.008	0.338
rs845553	*EGFR*	7	G/A	—	4	Promoter histone marks: 4 tissues; Enhancer histone marks: 17 tissues; DNAse: 17 tissues; Proteins bound: 7; Motifs changed: 5 altered motifs	0.396	0.683	0.280
rs17172432	*EGFR*	7	T/C	—	4	Promoter histone marks: 7 tissues; Enhancer histone marks: 18 tissues; DNAse: 6 tissues; Motifs changed: 4 altered motifs	0.359	0.614	0.280
rs34280975	*MET*	7	A/G	—	2c	Enhancer histone marks: SKIN; DNAse: 4 tissues; Proteins bound: CEBPB; Motifs changed: 7 altered motifs	0.421	0.197	0.538
rs3744483	*STAT3*	17	T/C	miRNA	4	bound: 7; DNAse: 11 tissues; Motifs changed: Foxa,p300	0.907	0.856	0.467
rs1135669	*STAT5A*	17	C/T	Splicing	4	Enhancer histone marks: BLD, THYM; DNAse: OVRY,BRST; Motifs changed: BATF, Pbx3, STAT	0.062	0.079	0.255
rs2847297	*PTPN2*	18	A/G	—	—	DNAse: BLD; Motifs changed: Nkx2,Pax-5	0.005	0.017	0.005
rs2847282	*PTPN2*	18	T/G	—	5	Promoter histone marks: STRM, LIV, BLD; Enhancer histone marks: 9 tissues; DNAse: 4 tissues; Motifs changed: 26 altered motifs	0.029	0.001	0.029

Table 1. Summary of the functional prediction and eQTL analysis results of the 11 selected SNPs in the TCPTP pathways *in silico*. [a]Reference allele/effect allele. [b]*P* value of eQTL analysis results TFBS = transcription factor binding site.

Results

Analysis of six GWAS datasets. Overall, 5162 SNPs from 43 TCPTP pathway genes in the six GWAS datasets from the Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium (TRICL-ILCCO) Consortium were identified, and their associations with lung cancer risk are shown in the Manhattan plot (Fig. 1A). After multiple-testing correction, 112 SNPs in eight genes (*ATR, EGFR, MET, PIK3R1, PIK3R3, PTPN2, STAT3*, and *STAT5A*) remained significantly associated with lung cancer risk with FDR <0.20. The results of associations with lung cancer risk are summarized in Supplementary Table S2. Based on LD analysis ($r^2 > 0.30$) and online functional prediction analyses by using SNPinfo, RegulomeDB, and HaploReg, we selected to perform additional analyses for 11 SNPs: rs11707731 in *ATR;* rs845553, rs1140762 and rs17172432 in *EGFR*; rs34280975 in *MET*; rs706714 in *PIK3R1;* rs7538978 in *PIK3R3;* rs2847297 and rs2847282 in *PTPN2;* rs3744483 in *STAT3;* rs1135669 in *STAT5A* for further study (Supplementary Figure S1 and Supplementary Table S3).

Functional validation by eQTL analysis 21. We assessed associations between the 11 SNPs and mRNA expression levels by using the genotyping and expression data available from the lymphoblastoid cell lines derived from 373 individuals of European descent (http://www.1000genomes.org/), and we found that only rs2847297 and rs2847282 were associated with expression levels of *PTPN2* in additive, dominant and recessive models

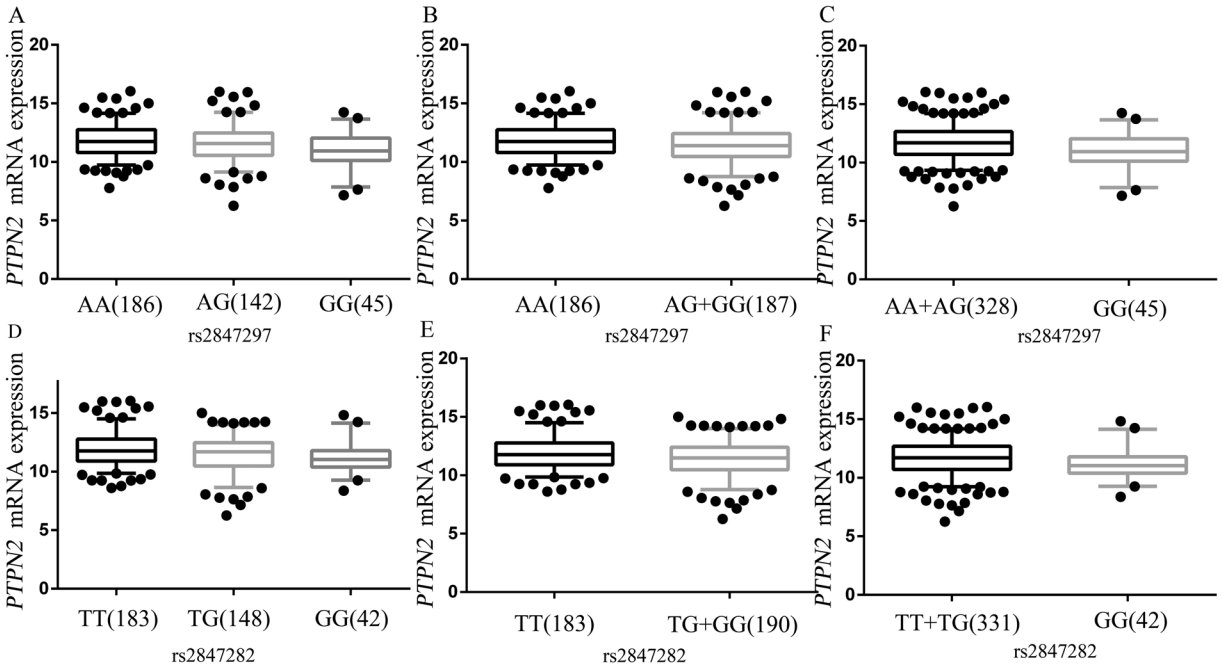

Figure 2. The correlations between identified SNPs and *PTPN2* mRNA expression. rs2847297 in *PTPN2* (**A**) additive model, $P = 0.002$; (**B**) dominant model, $P = 0.017$; (**C**) recessive model, $P = 0.005$) and rs2847282 in *PTPN2* (**D**), additive model, $P = 0.0006$; (**E**) dominant model, $P = 0.001$; (**F**) recessive model, $P = 0.029$).

(Table 1). Regional association plots for rs2847297 and rs2847282 in 500 kb up- and downstream region were shown in Fig. 1B. The SNP rs2847297 was in a low LD with rs2847282 (Fig. 1C). *PTPN2* mRNA expression levels were significantly decreased with an increased number of the rs2847297 G allele in additive ($P = 0.002$) (Fig. 2A), dominant ($P = 0.017$) (Fig. 2B) and recessive model ($P = 0.005$) (Fig. 2C). The eQTL analysis results of rs2847282 were also significant (Fig. 2D,E,F). In addition, we compared mRNA expression levels of *PTPN2* in 109 paired target tissue samples from The Cancer Genome Atlas (TCGA) and found that *PTPN2* mRNA expression levels were significantly increased in tumor tissues than normal tissues ($P = 3.01E-05$) (Supplementary Figure S2). The two SNPs rs2847297 and rs2847282 were chosen as tagSNPs, because they were significantly associated with lung cancer risk as assessed in the overall association analysis and had potential functions according to the eQTL analysis.

Expanded analysis by including additional two GWAS studies. We expanded our analysis by including two additional independent lung cancer GWAS studies, Harvard Lung Cancer Study and Icelandic Lung Cancer Study (deCODE). We performed an overall meta-analysis to evaluate associations between the two *PTPN2* SNPs and lung cancer risk. We found that the overall effects among all eight GWAS studies remained significant (OR = 0.95, 95% CI = 0.92–0.98, *P*het = 0.476, and $P = 0.004$ for rs2847297; OR = 0.95, 95% CI = 0.92–0.99, *P*het = 0.523, and $P = 0.009$ for rs2847282) (Table 2 and Fig. 3A,B).

In subgroup analysis by histology (Table 2, Fig. 3), we found that the rs2847297 G allele was borderline associated with lung adenocarcinoma (AD) risk (OR = 0.95, 95% CI = 0.91–1.00, $P = 0.052$) and significantly associated with squamous cell lung carcinoma (SQ) risk (OR = 0.92, 95% CI = 0.87–0.97, $P = 0.002$, Fig. 3A). We also found the rs2847282 G allele was associated with SQ risk (OR = 0.93, 95% CI = 0.88–0.99, $P = 0.016$), while there was no statistical association with AD risk (OR = 0.96, 95% CI = 0.91–1.01, $P = 0.114$, Fig. 3B). In subgroup analysis by smoking status, there was a marginal significant decrease in lung cancer risk for the rs2847297 G allele among ever smokers (OR = 0.96, 95% CI = 0.91–1.00, $P = 0.042$), but not among never smokers (OR = 0.95, 95% CI = 0.83–1.09, $P = 0.465$, Fig. 3A). However, there was no association with the *PTPN2* rs2847282 G allele and lung cancer risk among ever smokers (OR = 0.96, 95% CI = 0.91–1.00, $P = 0.066$ and never smokers (OR = 1.00, 95% CI = 0.86–1.16, $P = 0.960$, Fig. 3B).

Discussion

In the present study, we sought to investigate associations between genetic variants in the TCPTP pathway genes and lung cancer risk using eight published GWAS studies of 14,463 cases and 44,188 controls. The principal findings included two novel, potentially functional SNPs, rs2847297 and rs2847282 of *PTPN2*, that were both associated with a decreased lung cancer risk and a decreased mRNA expression level of *PTPN2*, particularly in subgroups of ever smokers and squamous cell lung carcinoma. Four articles about pathway-based analysis and lung cancer risk (Centrosome, DNA repair, lncRNA and RNA degradation) have been accepted or published in our laboratory. We found that the loci of two SNPs in *PTPN2* were different from previous studies in our lab and GWAS studies.

Study population	Sample size		Imp. Quality	*PTPN2* rs2847297 A > G		Imp. Quality	*PTPN2* rs2847282 T > G	
	Cases	Controls		OR (95% CI)	*P*		OR (95% CI)	*P*
ICR[1]	1952	5200	1.00	0.97 (0.89–1.04)	0.379	0.88	0.94 (0.87–1.03)	0.180
AD	465	5200	1.00	0.95(0.82–1.09)	0.459	0.87	0.92(0.79–1.07)	0.281
SQ	611	5200	1.00	0.95 (0.84–1.08)	0.425	0.87	0.96 (0.83–1.10)	0.521
MDACC[2]	1150	1134	1.00	0.85 (0.75–0.97)	0.014	0.81	0.85 (0.74–0.99)	0.030
AD	619	1134	1.00	0.87 (0.75–1.01)	0.070	0.81	0.86 (0.72–1.01)	0.073
SQ	306	1134	1.00	0.73 (0.60–0.89)	0.002	0.81	0.88 (0.71–1.09)	0.246
Ever smoking	1150	1134	1.00	0.85 (0.75–0.97)	0.014	0.81	0.85 (0.74–0.99)	0.030
IARC[3]	2533	3791	1.00	0.97 (0.90–1.05)	0.475	0.77	0.94 (0.86–1.03)	0.188
AD	517	2824	1.00	1.03 (0.90–1.19)	0.641	0.77	1.00 (0.85–1.17)	0.961
SQ	911	2968	1.00	0.91 (0.81–1.02)	0.104	0.77	0.89 (0.78–1.02)	0.084
Ever smoking	2367	2508	1.00	0.97 (0.89–1.05)	0.446	0.77	0.95 (0.86–1.04)	0.273
Never smoking	159	1253	1.00	1.06 (0.83–1.36)	0.623	0.77	0.95 (0.71–1.27)	0.735
NCI[4]	5713	5736	1.00	0.94 (0.88–0.99)	0.022	0.87	0.95 (0.89–1.01)	0.116
AD	1841	5736	1.00	0.95 (0.87–1.03)	0.225	0.87	0.95 (0.87–1.04)	0.257
SQ	1447	5736	1.00	0.92 (0.84–1.00)	0.060	0.88	0.95 (0.86–1.04)	0.258
Ever smoking	5342	4336	1.00	0.97(0.91–1.03)	0.297	0.88	0.98 (0.92–1.06)	0.649
Never smoking	350	1379	1.00	0.91(0.74–1.12)	0.376	0.88	0.94 (0.75–1.19)	0.622
Toronto[5]	331	499	1.00	0.86 (0.68–1.07)	0.182	0.85	0.84 (0.65–1.09)	0.180
AD	90	499	1.00	0.84 (0.59–1.19)	0.326	0.85	0.89 (0.60–1.32)	0.566
SQ	50	499	1.00	0.96 (0.61–1.52)	0.870	0.85	0.96 (0.57–1.62)	0.871
Ever smoking	236	272	1.00	0.95(0.71–1.27)	0.735	0.87	0.82 (0.59–1.14)	0.231
Never smoking	95	217	1.00	0.69(0.47–1.03)	0.065	0.87	0.89 (0.57–1.40)	0.611
GLC[6]	481	478	1.00	1.01 (0.83–1.24)	0.881	0.80	1.06 (0.85–1.33)	0.584
AD	186	478	1.00	0.88 (0.67–1.16)	0.368	0.80	0.92 (0.68–1.25)	0.609
SQ	97	478	1.00	1.20 (0.85–1.70)	0.299	0.80	1.17 (0.80–1.70)	0.426
Ever smoking	433	258	1.00	1.01 (0.78–1.32)	0.920	0.80	1.06 (0.79–1.41)	0.701
Never smoking	35	220	1.00	0.99 (0.54–1.82)	0.978	0.80	0.90 (0.47–1.70)	0.736
Discovery combined	12160	16838		0.94 (0.91–0.98)	0.002		0.94 (0.90–0.98)	0.003
Harvard[7]	984	970	1.00	1.01 (0.88–1.17)	0.857	0.83	1.02 (0.89–1.17)	0.791
AD	597	970	1.00	1.00 (0.86–1.18)	0.952	0.83	1.04 (0.88–1.21)	0.673
SQ	216	970	1.00	1.03 (0.82–1.31)	0.781	0.83	1.00 (0.79–1.27)	0.967
Ever smoking	892	809	1.00	1.00 (0.86–1.16)	0.962	0.83	0.97 (0.84–1.13)	0.687
Never smoking	92	161	1.00	1.15 (0.76–1.76)	0.502	0.83	1.51 (0.99–2.29)	0.053
deCOD[8]	1319	26380	1.00	0.98 (0.91–1.07)	0.689	0.89	0.99 (0.91–1.09)	0.911
AD	547	26380	1.00	0.96 (0.85–1.09)	0.524	0.89	1.02 (0.88–1.17)	0.834
SQ	259	26380	1.00	0.92 (0.77–1.10)	0.373	0.89	0.84 (0.69–1.03)	0.095
Replication combined	2303	27350		0.99 (0.92–1.06)	0.803		1.00 (0.93–1.08)	0.960
Overall	14463	44188		0.95 (0.92–0.98)	0.004		0.95 (0.92–0.99)	0.009
Overall AD combined	4862	43221		0.95 (0.91–1.00)	0.053		0.96 (0.91–1.01)	0.114
Overall SQ combined	3897	43365		0.92 (0.87–0.97)	0.002		0.93 (0.88–0.99)	0.016
Overall ever smoking combined	10420	9317		0.96 (0.91–1.00)	0.043		0.96 (0.91–1.00)	0.064
Overall never smoking combined	731	3230		0.95 (0.83–1.09)	0.467		1.00 (0.86–1.16)	0.959

Table 2. Summary of the association results of two SNPs in the eight lung cancer GWAS studies. AD: adenocarcinoma, SQ: squamous cell carcinoma. The combined OR and *P* value were estimated using a fixed-effects model. [1]ICR: the Institute of Cancer Research Genome-wide Association Study, UK. [2]MDACC: the MD Anderson Cancer Center Genome-wide Association Study, US. [3]IARC: the International Agency for Research on Cancer Genome-wide Association Study, France. [4]NCI: the National Cancer Institute Genome-wide Association Study, US. [5]Toronto: the Samuel Lunenfeld Research Institute Genome-wide Association Study, Toronto, Canada. [6]GLC: German Lung Cancer Study, Germany. [7]Harvard: Harvard Lung Cancer Study, USA. [8]deCODE: Icelandic Lung Cancer Study, Iceland.

PTPN2 plays a dual role in development and progression of cancer. Proliferation and cell cycle assays demonstrated that overexpression of PTPN2 would decrease serum requirement, increase formation of larger colonies in soft agar, alter morphology, and rapidly progress through G1 and S phases and the rate of cell division[25,26]. Another study showed that the proliferation rate would reduce in TCPTP (−/−), compared to TCPTP (+/+),

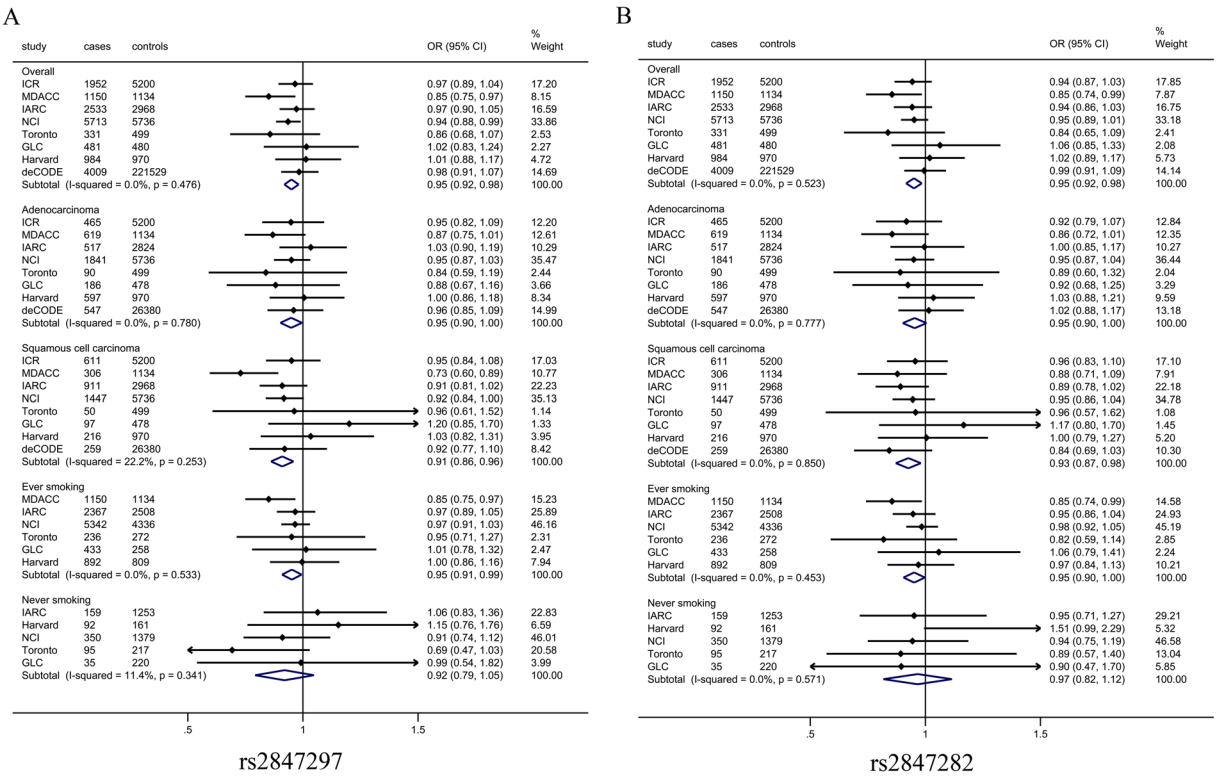

Figure 3. Forest plots of effect size and direction for tagSNPs from TRICL-ILCCO consortium. *PTPN2* rs2847297 $P_{combined} = 0.004$ in all individuals; $P_{combined} = 0.052$ in overall adenocarcinoma individuals; $P_{combined} = 0.002$ in overall squamous cell carcinoma individuals; $P_{combined} = 0.042$ in overall ever smoking individuals; $P_{combined} = 0.465$ in overall never smoking individuals (**A**); *PTPN2* rs2847282$P_{combined} = 0.009$ in all individuals; $P_{combined} = 0.114$ in overall adenocarcinoma individuals; $P_{combined} = 0.016$ in overall squamous cell carcinoma individuals; $P_{combined} = 0.066$ in overall ever smoking individuals; $P_{combined} = 0.960$ in overall never smoking individuals (**B**); Each box and horizontal line represent the OR point estimate and 95% CI derived from the additive model. The area of each box is proportional to the statistical weight of the study. Diamonds represent the ORs obtained from the combined analysis with 95% confidence intervals indicated by their widths. The meta-analysis includes eight GWAS studies [the Institute of Cancer Research (ICR) GWAS, the MD Anderson Cancer Center (MDACC) GWAS, the International Agency for Research on Cancer (IARC) GWAS, the National Cancer Institute (NCI) GWAS, the Lunenfeld-Tanenbaum Research Institute (Toronto) GWAS, German Lung Cancer Study (GLC) GWAS, Harvard Lung Cancer Study (Harvard) GWAS, Icelandic Lung Cancer Study (deCODE) GWAS]. NCI GWAS includes four sub-studies: the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC), the Cancer Prevention Study II Nutrition Cohort (CPS-II), the Environment and Genetics in Lung Cancer Etiology (EAGLE), and the Prostate, Lung, Colon, Ovary Screening Trial (PLCO).

lymphocytes[27]. We found that *PTPN2* mRNA expression levels in matched lung cancer tissues were increased compared to adjacent normal tissues from the TCGA database, some other studies also demonstrated that PTPN2 expression levels were higher in lung AD[28,29] and SQ[30,31] than in normal lung tissues. These findings provided oncogenic evidence of *PTPN2* and were consistent with our results that the two susceptibility loci of *PTPN2* were associated with a decreased lung cancer risk as a result of a decreased mRNA expression level of the gene. In addition, we found that the eQTL analysis result of rs2847297 in lung tissue was also significant in the GTEx analysis ($P = 4.0E10–7$) (http://www.gtexportal.org/home/eqtls/bySnp?snpId=rs2847297&tissueName=All). This result is also consistent with the eQTL analysis from the lymphoblastoid cell lines in the present study. However, it has been reported that overexpression of PTPN2 induces apoptosis in the p53 + A549 and MCF-7 cells but not in p53- HeLa cells, also consistent with features of a tumor suppressor[32]. Another study demonstrated that PTPN2 was absent in a large proportion of "triple-negative" primary human breast cancers and PTPN2 overexpression would suppress tumor growth[33].

In subgroup analysis we found that the two SNPs were more likely to be associated with SQ risk, and the risk associated with rs2847297 G allele was more likely to be among ever smoking. Cigarette smoke is the major risk factor for lung cancer, especially for SQ. Study showed that smoking led to an increased expression of Nkx2[34], which is the transcription factor (TF) of *PTPN2*. Therefore, it is likely that the locus has the possibility of influencing lung cancer risk of ever smokers through changing the expression of PTPN2.

Our study has some limitations. First, genes in the TCPTP pathway were identified mainly from the Molecular Signatures Database and Genecards. Although we did search some relative articles to complete the list of genes in the pathway, some newly discovered genes in the pathway might have been missed. Second, although we

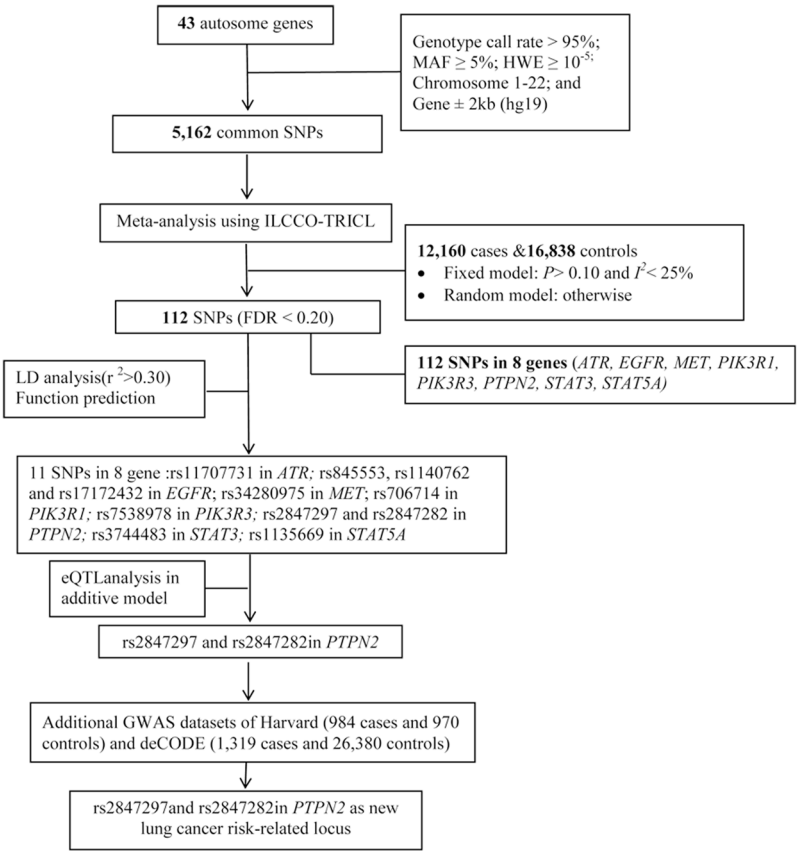

Figure 4. Flowchart of SNP selection among the TCPTP pathway genes.

demonstrated the association of thetwo novel potentially functional loci in *PTPN2* with lung cancer risk with functional evidence from eQTL analyses, the exact biochemical and molecular mechanisms are still unclear. Third, our eQTL analyses were limited to publicly available data from lymphoblastoid cell lines but target tissues, which could provide more direct correlation results between the two SNPs and PTPN2 expression.

Taken together, the present study revealed two novel, potentially functional susceptibility loci in *PTPN2* associated with lung cancer risk in European populations, particularly among ever smokers and squamous carcinoma. Further validation and functional evaluation of these genetic variants are warranted to verify our findings.

Materials and Methods

Study populations. The present study first used genotyping data from the TRICL-ILCCO consortium, which included 12,160 lung cancer cases and 16,838 controls (all Europeans) of six previously published GWAS studies: The University of Texas MD Anderson Cancer Center (MDACC), Institute of Cancer Research (ICR), National Cancer Institute (NCI), International Agency for Research on Cancer (IARC), Toronto study from Samuel Lunenfeld Research Institute study (Toronto), and German Lung Cancer Study (GLC). The expanded analysis included additional two GWAS studies of European ancestry from the Harvard Lung Cancer Study (984 cases and 970 controls)[35] and the Icelandic Lung Cancer Study (deCODE) (1,319 cases and 26,380 controls)[36] of the ILCCO. Details of the study populations are presented in the supplementary file. A written informed consent was obtained by all participating GWAS studies. All methods were performed in accordance with the relevant guidelines and regulations for each of the participating institutions, and the present study followed the study protocols approved by Duke University Health System Institutional Review Board.

Selection of Genes and SNPs from TCPTP pathway. Genotyping in these GWAS studies was performed by one of Illumina HumanHap 317, 317 + 240 S, 370Duo, 550, 610 or 1 M arrays. IMPUTE2 v2.1.1 or MaCH v1.0 software was used for imputation. Genes in the TCPTP pathway were identified from the Molecular Signatures Database (http://www.broadinstitute.org/gsea/index.jsp)[37] and Genecards (http://www.genecards.org/). Overall, 43 genes located on autosomal chromosomes were selected (detailed in Supplementary Table S1). The final meta-analysis contained 5,162 SNPs with the following inclusion criteria: genotyping rate >95%, minor allele frequency (MAF) ≥ 5%, and Hardy-Weinberg Equilibrium (HWE) exact *P* value ≥ 10^{-5}. The detailed workflow is shown in Fig. 4.

***In silico* functional prediction and validation.** We use three *in silico* tools, SNPinfo (http://snpinfo.niehs.nih.gov/snpinfo/snpfunc.htm)[38], RegulomeDB (http://regulomedb.org/)[39], and HaploReg (http://www.

broadinstitute.org/mammals/haploreg/haploreg.php)[40] to predict potential functions. The expression quantitative trait loci (eQTL) analysis was performed in the 1000 Genomes Project[41]. The mRNA expression of lung cancer tissue samples was performed in TCGA[42].

Statistical analysis. Odds ratios (ORs) and their 95% confidence intervals (CIs) were calculated using Stata (v10, State College, Texas, USA) and PLINK (v1.06) software. A meta-analysis with the inverse variance method was employed on the 5,162 SNPs. We used Cochran's Q statistic to test for heterogeneity and I^2 statistic for the proportion of the total variation[43]. The fixed-effects model was used when there was no heterogeneity among GWAS studies (Q-test $P > 0.100$ and $I2 < 25\%$); otherwise, the random-effects model was used. The false discovery rate (FDR) was performed to control for multiple testing with a threshold <0.20[44]. The genes mRNA expression levels in lung cancer and adjacent tissues from TCGA database were performed by paired t-test. Regional association plots were performed by LocusZoom[45]. Haploview v4.2 was used to generate the Manhattan plot and LD plots[46]. All other analyses were conducted with SAS (Version 9.3; SAS Institute, Cary, NC, USA).

References

1. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA: a cancer journal for clinicians.* **65**, 87–108, doi:10.3322/caac.21262 (2015).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA: a cancer journal for clinicians.* **66**, 7–30, doi:10.3322/caac.21332 (2016).
3. Smith, C. Genomics: SNPs and human disease. *Nature.* **435**, 993–993, doi:10.1038/435993a (2005).
4. Amos, C. I. *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature genetics.* **40**, 616–622, doi:10.1038/ng.109 (2008).
5. Dong, J. *et al.* Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nature genetics.* **44**, 895–899, doi:10.1038/ng.2351 (2012).
6. Hu, Z. *et al.* A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nature genetics.* **43**, 792–796, doi:10.1038/ng.875 (2011).
7. Hung, R. J. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature.* **452**, 633–637, doi:10.1038/nature06885 (2008).
8. Lan, Q. *et al.* Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nature genetics.* **44**, 1330–1335, doi:10.1038/ng.2456 (2012).
9. McKay, J. D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nature genetics.* **40**, 1404–1406, doi:10.1038/ng.254 (2008).
10. Miki, D. *et al.* Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. *Nature genetics.* **42**, 893–896, doi:10.1038/ng.667 (2010).
11. Shiraishi, K. *et al.* A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nature genetics.* **44**, 900–903, doi:10.1038/ng.2353 (2012).
12. Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nature genetics.* **40**, 1407–1409, doi:10.1038/ng.273 (2008).
13. Dong, J. *et al.* Genome-wide association study identifies a novel susceptibility locus at 12q23.1 for lung squamous cell carcinoma in han chinese. *PLoS genetics.* **9**, e1003190, doi:10.1371/journal.pgen.1003190 (2013).
14. Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nature genetics.* **46**, 736–741, doi:10.1038/ng.3002 (2014).
15. Zhang, R. *et al.* A genome-wide gene-environment interaction analysis for tobacco smoke and lung cancer susceptibility. *Carcinogenesis.* **35**, 1528–1535, doi:10.1093/carcin/bgu076 (2014).
16. Tiganis, T., Bennett, A. M., Ravichandran, K. S. & Tonks, N. K. Epidermal growth factor receptor and the adaptor protein p52Shc are specific substrates of T-cell protein tyrosine phosphatase. *Molecular and cellular biology.* **18**, 1622–1634, doi:10.1128/MCB.18.3.1622 (1998).
17. Mattila, E., Auvinen, K., Salmi, M. & Ivaska, J. The protein tyrosine phosphatase TCPTP controls VEGFR2 signalling. *Journal of cell science.* **121**, 3570–3580, doi:10.1242/jcs.031898 (2008).
18. Persson, C. *et al.* Site-selective regulation of platelet-derived growth factor beta receptor tyrosine phosphorylation by T-cell protein tyrosine phosphatase. *Molecular and cellular biology.* **24**, 2190–2201, doi:10.1128/MCB.24.5.2190-2201.2004 (2004).
19. Ten Hoeve, J. *et al.* Identification of a nuclear Stat1 protein tyrosine phosphatase. *Molecular and cellular biology.* **22**, 5662–5668, doi:10.1128/MCB.22.16.5662-5668.2002 (2002).
20. Yamamoto, T. *et al.* The nuclear isoform of protein-tyrosine phosphatase TC-PTP regulates interleukin-6-mediated signaling pathway through STAT3 dephosphorylation. *Biochemical and biophysical research communications.* **297**, 811–817, doi:10.1016/S0006-291X(02)02291-X (2002).
21. Lu, X. *et al.* T-cell protein tyrosine phosphatase, distinctively expressed in activated-B-cell-like diffuse large B-cell lymphomas, is the nuclear phosphatase of STAT6. *Molecular and cellular biology.* **27**, 2166–2179, doi:10.1128/MCB.01234-06 (2007).
22. Galic, S. *et al.* Regulation of insulin receptor signaling by the protein tyrosine phosphatase TCPTP. *Molecular and cellular biology.* **23**, 2096–2108, doi:10.1128/MCB.25.2.819-829.2005 (2003).
23. Domanska, D. *et al.* STAT3 rs3816769 polymorphism correlates with gene expression level and may predispose to nonsmall cell lung cancer: a preliminary study. *Polskie Archiwum Medycyny Wewnetrznej.* **123**, 672–679 (2013).
24. Uzunoglu, F. G. *et al.* Vascular endothelial growth factor receptor 2 gene polymorphisms as predictors for tumor recurrence and overall survival in non-small-cell lung cancer. *Annals of surgical oncology.* **19**, 2159–2168, doi:10.1245/s10434-012-2227-4 (2012).
25. Ganapati, U. *et al.* A nuclear protein tyrosine phosphatase induces shortening of G1 phase and increase in c-Myc protein level. *Experimental cell research.* **265**, 1–10, doi:10.1006/excr.2001.5158 (2001).
26. Radha, V., Nambirajan, S. & Swarup, G. Overexpression of a nuclear protein tyrosine phosphatase increases cell proliferation. *FEBS letters.* **409**, 33–36, doi:10.1016/S0014-5793(97)00471-7 (1997).
27. Dupuis, M., De Jesus Ibarra-Sanchez, M., Tremblay, M. L. & Duplay, P. Gr-1 + myeloid cells lacking T cell protein tyrosine phosphatase inhibit lymphocyte proliferation by an IFN-gamma- and nitric oxide-dependent mechanism. *Journal of immunology.* **171**, 726–732, doi:10.4049/jimmunol.171.2.726 (2003).
28. Landi, M. T. *et al.* Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS one.* **3**, e1651, doi:10.1371/journal.pone.0001651 (2008).
29. Okayama, H. *et al.* Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer research.* **72**, 100–111, doi:10.1158/0008-5472.CAN-11-1403 (2012).
30. Hou, J. *et al.* Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PloS one.* **5**, e10312, doi:10.1371/journal.pone.0010312 (2010).
31. Weiss, J. *et al.* Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. *Science translational medicine* **2**, 62ra93–62ra93, doi:10.1126/scitranslmed.3001451 (2010).

32. Radha, V., Sudhakar, C. & Swarup, G. Induction of p53 dependent apoptosis upon overexpression of a nuclear protein tyrosine phosphatase. *FEBS letters.* **453**, 308–312, doi:10.1016/S0014-5793(99)00734-6 (1999).
33. Shields, B. J. *et al*. TCPTP regulates SFK and STAT3 signaling and is lost in triple-negative breast cancers. *Molecular and cellular biology.* **33**, 557–570, doi:10.1128/MCB.01016-12 (2013).
34. Tong, M. *et al*. Differential Contributions of Alcohol and Nicotine-Derived Nitrosamine Ketone (NNK) to White Matter Pathology in the Adolescent Rat Brain. *Alcohol and alcoholism.* **50**, 680–689, doi:10.1093/alcalc/agv102 (2015).
35. Su, L. *et al*. Genotypes and haplotypes of matrix metalloproteinase 1, 3 and 12 genes and the risk of lung cancer. *Carcinogenesis.* **27**, 1024–1029, doi:10.1093/carcin/bgi283 (2006).
36. Thorgeirsson, T. E. *et al*. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature.* **452**, 638–642, doi:10.1038/nature06846 (2008).
37. Liberzon, A. *et al*. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems.* **1**, 417–425, doi:10.1016/j.cels.2015.12.004 (2015).
38. Xu, Z. L. & Taylor, J. A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res* **37**, W600–W605, doi:10.1093/nar/gkp290 (2009).
39. Boyle, A. P. *et al*. Annotation of functional variation in personal genomes using RegulomeDB. *Genome research.* **22**, 1790–1797, doi:10.1101/gr.137323.112 (2012).
40. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930–934, doi:10.1093/nar/gkr917 (2012).
41. Lappalainen, T. *et al*. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.* **501**, 506–511, doi:10.1038/nature12531 (2013).
42. Rodgers, K. & Network, C. G. A. R. Comprehensive molecular profiling of lung adenocarcinoma (vol. 511, pg 543, 2014). *Nature.* **514** (2014).
43. Higgins, J. P., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ.* **327**, 557–560, doi:10.1136/bmj.327.7414.557 (2003).
44. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289–300 (1995).
45. Pruim, R. J. *et al*. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* **26**, 2336–2337, doi:10.1093/bioinformatics/btq419 (2010).
46. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* **21**, 263–265, doi:10.1093/bioinformatics/bth457 (2005).

Acknowledgements

As Duke Cancer Institute members, QW, KO, and NR acknowledge support from the Duke Cancer Institute as part of the P30 Cancer Center Support Grant (Grant ID: NIH CA014236). QW was also supported by the start-up funds from Duke Cancer Institute, Duke University Medical Center. This work was supported by the Transdisciplinary Research in Cancer of the Lung (TRICL) Study and, U19-CA148127 on behalf of the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network. The Toronto study was supported by Canadian Cancer Society Research Institute (020214), Ontario Institute of Cancer and Cancer Care Ontario Chair Award to RH The ICR study was supported by Cancer Research UK (C1298/A8780 andC1298/A8362-Bobby Moore Fund for Cancer Research UK) and NCRN, HEAL and Sanofi-Aventis. Additional funding was obtained from NIH grants (5R01CA055769, 5R01CA127219, 5R01CA133996, and 5R01CA121197). The Liverpool Lung Project (LLP) was supported by The Roy Castle Lung Cancer Foundation, UK. The ICR and LLP studies made use of genotyping data from the Wellcome Trust Case Control Consortium 2 (WTCCC2); a full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Sample collection for the Heidelberg lung cancer study was in part supported by a grant (70–2919) from the Deutsche Krebshilfe. The work was additionally supported by a Helmholtz-DAAD fellowship (A/07/97379 to MNT) and by the NIH (U19CA148127). The KORA Surveys were financed by the GSF, which is funded by the German Federal Ministry of Education, Science, Research and Technology and the State of Bavaria. The Lung Cancer in the Young study (LUCY) was funded in part by the National Genome Research Network (NGFN), the DFG (BI576/2-1; BI 576/2-2), the Helmholtzgemeinschaft (HGF) and the Federal office for Radiation Protection (BfS: STSch4454). Genotyping was performed in the Genome Analysis Center (GAC) of the Helmholtz ZentrumMuenchen. Support for the Central Europe, HUNT2/Tromsø and CARET genome-wide studies was provided by Institut National du Cancer, France. Support for the HUNT2/Tromsø genome-wide study was also provided by the European Community (Integrated Project DNA repair, LSHG-CT- 2005–512113), the Norwegian Cancer Association and the Functional Genomics Programme of Research Council of Norway. Support for the Central Europe study, Czech Republic, was also provided by the European Regional Development Fund and the State Budget of the Czech Republic (RECAMO, CZ.1.05/2.1.00/03.0101). Support for the CARET genome-wide study was also provided by grants from the US National Cancer Institute, NIH (R01 CA111703 and UO1 CA63673), and by funds from the Fred Hutchinson Cancer Research Center. Additional funding for study coordination, genotyping of replication studies and statistical analysis was provided by the US National Cancer Institute (R01 CA092039). The lung cancer GWAS from Estonia was partly supported by a FP7 grant (REGPOT245536), by the Estonian Government (SF0180142s08), by EU RDF in the frame of Centre of Excellence in Genomics and Estoinian Research Infrastructure's Roadmap and by University of Tartu (SP1GVARENG). The work reported in this paper was partly undertaken during the tenure of a Postdoctoral Fellowship from the IARC (for MNT). The Environment and Genetics in Lung Cancer Etiology (EAGLE), the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC), and the Prostate, Lung, Colon, Ovary Screening Trial (PLCO)studies and the genotyping of ATBC, the Cancer Prevention Study II Nutrition Cohort (CPS-II) and part of PLCO were supported by the Intramural Research Program of NIH, NCI, Division of Cancer Epidemiology and Genetics. ATBC was also supported by US Public Health Service contracts (N01-CN-45165, N01-RC-45035 and N01-RC-37004) from the NCI. PLCO was also supported by individual contracts from the NCI to the University of Colorado Denver (NO1-CN-25514), Georgetown University(NO1-CN-25522), Pacific Health Research Institute (NO1-CN-25515), Henry Ford Health System (NO1-CN-25512), University of Minnesota(NO1-CN-25513), Washington University(NO1-CN-25516), University of Pittsburgh (NO1-CN-25511), University of Utah

(NO1-CN-25524), Marshfield Clinic Research Foundation (NO1-CN-25518), University of Alabama at Birmingham (NO1-CN-75022, Westat, Inc. NO1-CN-25476), University of California, Los Angeles (NO1-CN-25404). The Cancer Prevention Study II Nutrition Cohort was supported by the American Cancer Society. The NIH Genes, Environment and Health Initiative (GEI) partly funded DNA extraction and statistical analyses (HG-06-033-NCI-01 andRO1HL091172-01), genotyping at the Johns Hopkins University Center for Inherited Disease Research (U01HG004438 and NIH HHSN268200782096C) and study coordination at the GENEVA Coordination Center (U01 HG004446) for EAGLE and part of PLCO studies. Funding for the MD Anderson Cancer Study was provided by NIH grants (P50 CA70907, R01CA121197, R01CA127219, U19 CA148127, R01 CA55769, and K07CA160753) and CPRIT grant (RP100443). Genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is funded through a federal contract from the NIH to The Johns Hopkins University (HHSN268200782096C). The Harvard Lung Cancer Study was supported by the NIH (National Cancer Institute) grants CA092824, CA090578, and CA074386. deCODE: The project was funded in part by GENADDICT: LSHMCT-2004-005166), the National Institutes of Health (R01-DA017932).

Author Contributions

Y.F. and Q.W. designed and conceived the experiments. H.L., Z.L. C.M. and Y.R. helped to analyze the data. Y.H., R.J.H., Y.B., J.M., P.B., H.B., A.R., R.S.H., N.C., M.T.L., I.B., Y.Y., X.W., D.C.C. and C.I. Amos collected the samples and provided data. All authors reviewed the paper.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-00850-0

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.