Am J Cancer Res 2020;10(8):2603-2616
www.ajcr.us /ISSN:2156-6976/ajcr0116679

Original Article

Novel genetic variants of *SYK* and *ITGA1* related lymphangiogenesis signaling pathway predict non-small cell lung cancer survival

Lihua Liu[1,2,3], Hongliang Liu[2,3], Sheng Luo[4], Edward F Jr Patz[2,5], Carolyn Glass[2,6], Li Su[7], Lijuan Lin[7], David C Christiani[7,8], Qingyi Wei[2,3,9]

[1]Department of Pulmonary and Critical Care Medicine, The First Affiliated Hospital of Guangxi Medical University, Nanning, Guangxi 530021, China; [2]Duke Cancer Institute, Duke University Medical Center, Durham, NC 27710, USA; Departments of [3]Population Health Sciences, [4]Biostatistics and Bioinformatics, [5]Radiology, Pharmacology and Cancer Biology, [6]Pathology, Duke University School of Medicine, Durham, NC 27710, USA; [7]Departments of Environmental Health and Epidemiology, Harvard School of Public Health, Boston, MA, 02115 USA; [8]Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; [9]Department of Medicine, Duke University School of Medicine, Durham, NC 27710, USA

Received June 21, 2020; Accepted June 28, 2020; Epub August 1, 2020; Published August 15, 2020

Abstract: Although lymphangiogenesis is a vital step in lung cancer metastasis, the association between lymphangiogenesis and non-small cell lung cancer (NSCLC) survival remains unclear. Since single-nucleotide polymorphisms (SNPs) have been reported to predict NSCLC survival, we investigated associations between SNPs in lymphangiogenesis-related pathway genes and NSCLC survival in a discovery genotyping dataset of 1,185 patients from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial and validated the findings in another genotyping dataset of 984 patients from the Harvard Lung Cancer Susceptibility Study. We evaluated associations between 34,509 genetic variants (3252 genotyped and 31,257 imputed) in 247 genes involved in lymphangiogenesis-related pathway and NSCLC survival. After validation, we finally identified two independent SNPs (*SYK* rs11787670 A>G and *ITGA1* rs67715745 T>C) to be significantly associated with NSCLC overall survival (OS), with adjusted hazards ratios of 0.77 and 0.83 (95% confidence interval =0.66-0.90, P=7.20×10^{-4}) and 0.84 (95% confidence interval =0.75-0.92, P=3.50×10^{-4}), respectively. Moreover, an increasing number of combined protective alleles of these two SNPs was significantly associated with an improved NSCLC OS and disease-specific survival (DSS) in the PLCO dataset (P_{trend}=0.011 and 0.006, respectively). Furthermore, the addition of these protective alleles to the prediction model for the 5-year survival increased the time-dependent area under the curve both from 87% to 87.67% for OS (P=0.029) and from 88.54% to 89.06% for DSS (P=0.022). Subsequent expression quantitative trait loci (eQTL) functional analysis revealed that the rs11787670 G allele was significantly associated with an elevated *SYK* mRNA expression in normal tissues. Additional analyses suggested a suppressor role for both *SYK* and *ITGA1* in NSCLC survival. Collectively, these findings indicated that *SYK* rs11787670 A>G and *ITGA1* rs67715745 T>C may be independent prognostic factors for NSCLC survival once further validated.

Keywords: Non-small cell lung cancer, single-nucleotide polymorphism, lymphangiogenesis, survival

Introduction

Although smoking rates across the world have been decreasing, the incidence of lung cancer has plateaued over the past years [1], but the 5-year overall survival (OS) rate for non-small cell lung cancer (NSCLC) remains poor, from 68% in patients with a stage IB disease to 0% to 10% in patients with stage IVA-IVB diseases [2]. In the United States, 228,820 new cases and 135,720 deaths have been estimated to occur in 2020 [3]. Till now, the median OS of patients with advanced NSCLC had increased by only 1.5 months over the past decade despite improved understanding of the biology and the development of new biomarker-targeted therapies [4]. Importantly, interindividual differences in lung cancer survival are commonly

observed among NSCLC patients, even in those with the same clinical tumor stage treated with the same therapeutic regimen [5, 6]. Therefore, it is important to identify additional prognostic factors for NSCLC survival, which could provide a more precise prediction of survival and facilitate treatment decisions for NSCLC patients.

Lymphangiogenesis (formation of new lymphatic vessels), unlike angiogenesis, has been a lesser-focused field in cancer research during the past decades. However, it has been shown that tumour lymphangiogenesis has similarities to that of tumour angiogenesis in cancer progression [7, 8]. Although lymphangiogenesis has an important role in tumor progression and metastasis [9, 10], the detection of lymphangiogenesis is still difficult and impractical, due largely to the lack of specific markers for the lymphatic in human cancers. Lymphatic vessel density (LVD) has been regarded as the most important evaluator for quantifying tumor lymphangiogenesis, and the overexpression of vascular endothelial growth factors-C/D (VEGF-C/D) is significantly correlated with the extent of metastasis in lung cancer [11, 12]. However, there were still contradictory findings of the prognostic effect of lymphangiogenesis on lung cancer. For example, one study indicated that lymphangiogenesis was an independent poor prognostic factor for NSCLC patients [13], whereas another study suggested the opposite [14]. These contradictory results may be related to the lack of standardization of lymphangiogenesis quantification. Generally, it is necessary to have feasible standardization and quantification criteria to establish a useful prognostic factor. Unfortunately, to date, there still has no standardized quantitative biomarker for lymphangiogenesis. Therefore, further exploration of an accurate and feasible predictor for the association between lymphangiogenesis and prognosis in lung cancer is warranted.

Genetic variants are reported to be associated with cancer prognosis [15, 16], and the genome-wide association study (GWAS) is a powerful approach to investigate survival-associated single-nucleotide polymorphisms (SNPs); however, few functional SNPs have been reported to be associated with NSCLC survival at the GWAS level. This is likely because most published GWASs focused strictly on SNPs with the most-significant *P*-values [17] without considering the weighted biological significance of SNPs. Recently, a new biological pathway-based approach, a hypothesis-driven post-GWAS analysis, has been used to identify the causal SNPs in genes involved in some known biological pathways [18, 19]. To date, however, there is no reported study to examine the associations between SNPs in lymphangiogenesis signaling pathway genes and NSCLC survival. Therefore, we hypothesize that genetic variants of the lymphangiogenesis-related pathway genes are associated with NSCLC survival, and we tested this hypothesis by using two existing independent NSCLC GWAS datasets and interrogated the functional relevance of the identified SNPs by looking into other publicly available genomic datasets.

Materials and methods

Study populations

The discovery genotyping dataset used in the present study were obtained from GWAS of the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial from the National Cancer Institute (the approval number: PLCO-95). The PLCO trial, which included 1,185 eligible patients, is a multicenter randomized study performed by ten medical centers in the United States between 1993 and 2011 [20]. Genomic DNA extracted from the whole blood samples of these participants were genotyped with Illumina Human Hap240Sv1.0, Human-Hap300v1.1, and Human Hap550v3.0 (dbGaP accession: phs000093.v2.p2 and phs000336.v1.p1) [21, 22]. The PLCO trial was approved by the institutional review boards of each participating institution, and all subjects provided a written informed consent permitting the use of the datasets.

The identified significant SNPs in the initial analysis among those extracted from the PLCO dataset were further validated by another GWAS dataset from 984 histology-confirmed Caucasian NSCLC patients of the Harvard Lung Cancer Susceptibility (HLCS) study. In the latter study, whole blood samples of all patients were used to extract DNA by the Auto Pure Large Sample Nucleic Acid Purification System (QIAGEN Company, Venlo, Limburg, Netherlands) that were genotyped using Illumina Human hap610-Quad arrays. The genotyping data we-

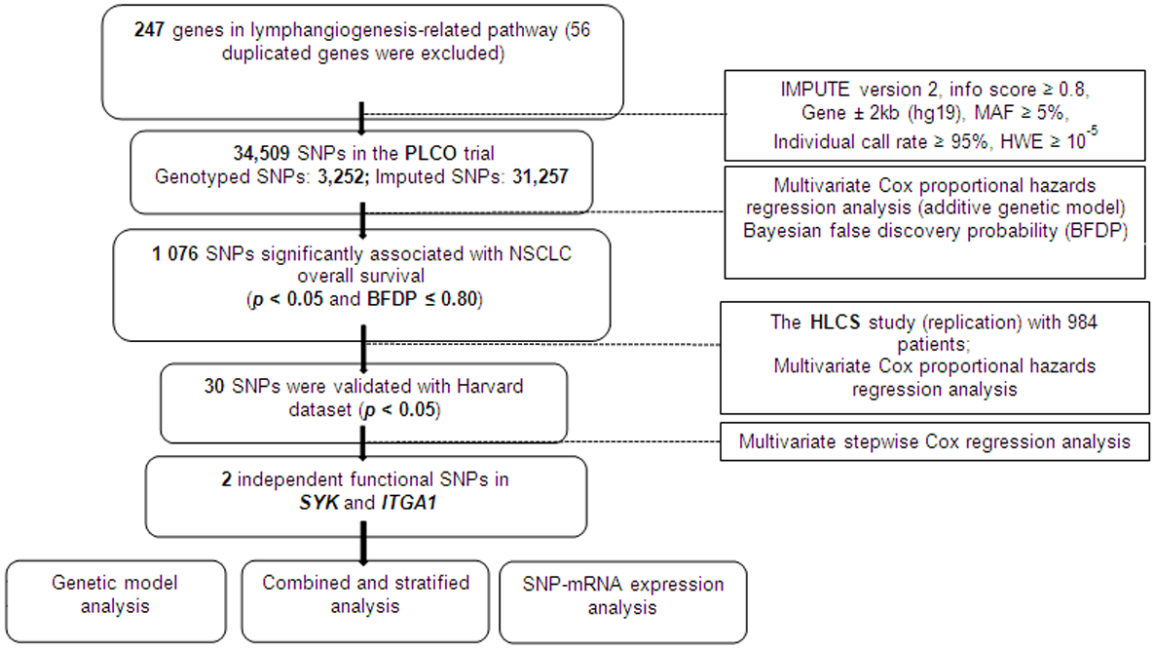

Figure 1. Study flowchart. The overall procedures of the present study. Abbreviations: SNPs, single-nucleotide polymorphisms; MAF: minor allelic frequency; HWE: Hardy-Weinberg Equilibrium; PLCO, The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; HLCS, the Harvard Lung Cancer Susceptibility Study; NSCLC, non-small cell lung cancer.

re imputed using MaCH software based on the 1000 Genomes Project. All the details of participant selection and data collection for both two datasets have been described elsewhere [22, 23].

The use of these two genotyping datasets in this present study was approved by both the Internal Review Board of Duke University School of Medicine (Project #Pro00054575) and the National Center for Biological Information (NCBI) for the access to the NCBI dbGaP database of genotypes and phenotypes (Project #6404). The comparison of the characteristics between the PLCO trial (n=1,185) and the HLCS study (n=984) are presented in Table S1.

Gene selection and SNP imputation

The genes involved in lymphangiogenesis-related pathway were identified by the Molecular Signatures Database (http://software.broadinstitute.org/gsea/msigdb/index.jsp) with the keyword "lymph AND vessel". After excluding 56 duplicated genes, 247 genes were available for further analysis as candidate genes (Table S2). Imputation with IMPUTE2 and the 1,000 Genomes Project data (phase 3) was performed for these candidate genes. All the SNPs in these genes and their ±2 kb flanking regions were extracted in line with the quality criteria listed in **Figure 1**. Finally, a total of 34,509 SNPs (3,252 genotyped and 31,257 imputed) from the PLCO trial were used for further analyses (as shown in Figure S1, imputation info score ≥0.8).

Statistical analyses

The follow-up time in the present study was defined as from the diagnosis of NSCLC to the last follow-up or date of death, and we chose OS as the primary endpoint. In the PLCO dataset, multivariate Cox proportional hazards regression analysis was used to assess the association between all the extracted SNPs in 247 lymphangiogenesis-related pathway genes and OS in an additive model with available covariates (including sex, age, histology, smoking status, tumor stage, chemotherapy, radiotherapy, and surgery) using the GenABEL package of R software [24]. We assigned a prior probability of 0.10 to detect a hazard ratio (HR) of 3.0 for an association variant genotypes or minor alleles of the SNPs with $P<0.05$. Since the strong linkage disequilibrium (LD) within the

imputed SNPs, false discovery rate (FDR) with a cutoff value of 0.20 and Bayesian false discovery probability (BFDP) with a cutoff value of 0.80 for multiple testing correction were performed to lower the probability of potentially false positive results [25]. The associations of the principal components and OS of NSCLC in the PLCO trial are shown in Table S3. Then those identified SNPs in the PLCO discovery dataset were used for further validation by the HLCS dataset. Next, to identify independent SNPs, a multivariable stepwise Cox regression model was used with adjustment for clinical variables and previously published survival-predictive SNPs from the same PLCO trial. Finally, a combined analysis was performed in the PLCO and HLCS combined datasets using PLINK 1.90 with the Cochran's Q statistics and I^2. Since no significant heterogeneity between the two datasets (Q test $P>0.1$, $I^2<25.0\%$) was found, a fixed-effects model of the meta-analysis was applied. The identified SNPs were also visualized by Manhattan plot and regional association plots.

Then, to estimate the survival probability associated with the combined alleles, Kaplan-Meier (KM) survival curve was used. For the stratified analysis in subgroups, we also evaluated interstudy heterogeneity and possible interaction with a x^2-based Q-test. Subsequently, the receiver operating characteristic (ROC) curve and time-dependent area under the curve (AUC) were used to elucidate the prediction accuracy of the models integrating the effects of both clinical and genetic variables on NSCLC survival using the timeROC package of R software (version 3.6.2).

Next, the expression quantitative trait loci (eQTL) analyses were further evaluated the genotype-phenotype correlation between identified SNPs and corresponding mRNA expression with a linear regression model using the data from the 373 European descendants included in the 1,000 Genomes Project [26], the Genotype-Tissue Expression (GTEx) Project (http://www.gtexportal.org/home) [27]. Prediction of bioinformatics function for the tagging SNPs was performed with HaploReg [28] (http://archive.broadinstitute.org/mammals/haploreg). Finally, we explored the correlation between the mRNA expression of SNP related genes and NSCLC survival probability using the KM analysis from an online database (http://kmplot.com/analysis/). The mutation data of those identified genes in lung tumor tissues were also assessed in publically available the database of the cBioPortal for Cancer Genomics (http://www.cbioportal.org). All statistical analyses were performed with SAS software (version 9.4; SAS Institute, Cary, NC) unless specified otherwise.

Results

Associations between SNPs in the lymphangiogenesis-related pathway genes and NSCLC OS in both PLCO trial and HLCS datasets

The workflow of this present study is shown in **Figure 1**. The basic characteristics of 1,185 patients from the PLCO trial and 984 patents in the HLCS study have been described in Table S1 and elsewhere [29]. In the discovery PLCO dataset, we identified 34,509 SNPs (including 3,252 genotyped and 31,257 imputed SNPs) in 247 lymphangiogenesis-related pathway genes, of which 1,076 SNPs were statistically significantly associated with NSCLC OS ($P<0.05$) in univariate Cox proportional hazards regression analyses with a single-locus additive model with multiple test correction (BFDP <0.8). After further validation in the HLCS validation dataset, 30 SNPs remained statistically significant.

Independent SNPs associated with NSCLC OS in the PLCO dataset

To identify which SNPs are independently associated with NSCLC survival, we first used stepwise multivariable Cox regression analysis to evaluate the effects of 30 validated SNPs in the PLCO dataset, because the HLCS dataset did not provide individual genotyping and detailed clinical covariates. In stepwise Cox regression analysis, five SNPs were found to be significantly associated with NSCLC OS. Then, after further adjustment for 28 additional previously published survival-predictive SNPs from the same PLCO dataset, two SNPs (*SYK* rs11787670 A>G and *ITGA1* rs67715745 T>C) remained independently associated with NSCLC OS ($P=0.040$ and $P=0.034$, respectively) (**Table 1**). The results of the meta-analysis for these two independent SNPs in each dataset are present-

Table 1. Two independent SNPs in a multivariate Cox proportional hazards regression analysis with adjustment for other covariates and 28 previously published SNPs for NSCLC in the PLCO Trial

Variables	Category	Frequency	HR (95% CI)[a]	p^a	HR (95% CI)[b]	p^b
Age	Continuous	1,185	1.03 (1.02-1.05)	<0.0001	1.04 (1.03-1.06)	<0.0001
Sex	Male	698	1.00		1.00	
	Female	487	0.77 (0.66-0.90)	0.0008	0.71 (0.61-0.84)	<0.0001
Smoking status	Never	115	1.00		1.00	
	Current	423	1.82 (1.36-2.45)	<0.0001	2.10 (1.55-2.85)	<0.0001
	Former	647	1.74 (1.32-2.29)	<0.0001	1.98 (1.49-2.64)	<0.0001
Histology	Adenocarcinoma	577	1.00		1.00	
	Squamous cell	285	1.15 (0.95-1.38)	0.015	1.22 (1.00-1.48)	0.046
	others	323	1.37 (1.15-1.64)	0.0004	1.40 (1.16-1.67)	0.0003
Tumor stage	I-IIIA	655	1.00		1.00	
	IIIB-IV	528	2.99 (2.46-3.63)	<0.0001	3.46 (2.83-4.24)	<0.0001
Chemotherapy	No	639	1.00		1.00	
	Yes	538	0.57 (0.48-0.68)	<0.0001	0.55 (0.45-0.66)	<0.0001
Radiotherapy	No	762	1.00		1.00	
	Yes	415	1.00 (0.85-1.18)	0.998	1.00 (0.84-1.19)	0.986
Surgery	No	637	1.00		1.00	
	Yes	540	0.22 (0.17-0.28)	<0.0001	0.19 (0.15-0.25)	<0.0001
SYK rs11787670 A>G	AA/AG/GG	1011/154/10	0.78 (0.64-0.95)	0.012	0.81 (0.67-0.99)	0.040
ITGA1 rs67715745 T>C	TT/TC/CC	368/579/248	0.84 (0.73-0.96)	0.010	0.86 (0.74-0.99)	0.034

Abbreviations: SNPs: single-nucleotide polymorphisms; NSCLC, non-small cell lung cancer; PLCO, the Prostate, Lung, Colorectal and Ovarian cancer screening trial; HLCS, Harvard Lung Cancer Susceptibility Study; HR, hazards ratio; CI, confidence interval; [a]Adjusted for age, sex, tumor stage, histology, smoking status, chemotherapy, radiotherapy, surgery, PC1, PC2, PC3 and PC4. [b]Other 28 published SNPs were included for further adjustment: rs779901, rs3806116, rs199731120, rs10794069, rs1732793, rs225390, rs3788142, rs73049469, rs35970494, rs225388, rs7553295, rs1279590, rs73534533, rs677844, rs4978754, rs1555195, rs11660748, rs73440898, rs13040574, rs469783, rs36071574, rs7242481, rs1049493, rs1801701, rs35859010, rs1833970, rs254315 and rs425904.

ed in **Table 2**, showing the absence of heterogeneity across these two datasets.

Specifically, as shown in **Table 3**, both *SYK* rs11787670 G and *ITGA1* rs67715745 C alleles were protective for survival of NSCLC in an allele-dose response manner (*SYK* rs11787670 G: P_{trend}=0.011 for OS and P_{trend}=0.006 for disease-specific survival (DSS) and *ITGA1* rs67715745 C: P_{trend}=0.012 for OS and P_{trend}=0.027 for DSS). We also depicted all the identified SNPs in a Manhattan plot (Figure S2) and each of these two independent SNPs in regional association plots (Figure S3).

Combined and stratified analyses of the two independent SNPs associated with NSCLC survival in the PLCO dataset

To assess the collective effect of two independent SNPs on NSCLC survival, we combined their protective alleles (i.e., *SYK* rs11787670 G and *ITGA1* rs67715745 C alleles) into one variable as a genetic score that was used to categorize patients in the PLCO dataset into four groups (i.e., 0, 1, 2, and 3-4) according to the number of protective alleles (NPA). As shown in **Table 3**, a better NSCLC survival was associated with an increase of NPA (P_{trend}=0.0004 for OS and P_{trend}=0.0006 for DSS). When we dichotomized all the patients into two groups: 0-1 and 2-4 NPA, we found that, compared with the 0-1 group, the 2-4 group had a significantly favorable NSCLC OS (HR=0.65, 95% CI=0.49-0.85 and *P*=0.002) and DSS (HR=0.61, 95% CI=0.45-0.83 and *P*=0.0013). We then used the Kaplan Meire (KM) survival curve to display the significant associations of NPA with NSCLC DSS (**Figure 2A**) and OS (**Figure 2B**) (Log-rank *P*=0.025 and *P*=0.006, respectively). Considering the 2-4 NPA group with relatively limited sample size, we also dichotomized all the patients into 0 and 1-4 NPA groups to evaluate the survival (Table S4). As shown in the KM survival curves that the 0 and 1-4 NPA groups did not show any significant difference in associations of NPA with NSCLC OS (Figure S4A) and DSS (Figure S4B).

Furthermore, to identify whether the effects of NPA on NSCLC OS and DSS were modified by clinical covariates, we performed the stratified

Table 2. Associations of two significant SNPs with of NSCLC overall survival in both discovery and validation datasets from two published GWASs

SNPs	Allele[a]	Gene	PLCO (n=1,185)					HLCS (n=984)			Combined-analysis			
			FDR	BFDP	EAF	HR (95% CI)[b]	p^b	EAF	HR (95% CI)[c]	p^c	$p_{het}{}^d$	I^2	HR (95% CI)[e]	p^e
rs11787670[f]	A>G	*SYK*	0.45	0.66	0.07	0.78 (0.64-0.94)	0.011	0.07	0.75 (0.58-0.96)	0.024	0.800	0.0	0.77 (0.66-0.90)	7.20×10^{-4}
rs67715745[f]	T>C	*ITGA1*	0.46	0.75	0.18	0.84 (0.73-0.96)	0.011	0.18	0.82 (0.70-0.95)	0.010	0.801	0.0	0.83 (0.75-0.92)	3.50×10^{-4}

Abbreviations: SNPs, single-nucleotide polymorphisms; NSCLC, non-small cell lung cancer; GWAS, genome-wide association study; PLCO, the Prostate, Lung, Colorectal and Ovarian cancer screening trial; HLCS, Harvard Lung Cancer Susceptibility Study; FDR, false discovery rate; BFDP, Bayesian false discovery probability; EAF, effect allele frequency; HR, hazards ratio; CI, confidence interval. [a]Reference > effect allele. [b]Adjusted for age, sex, stage, histology, smoking status, chemotherapy, radiotherapy, surgery, PC1, PC2, PC3 and PC4. [c]Adjusted for age, sex, stage, histology, smoking status, chemotherapy, radiotherapy, surgery, PC1, PC2 and PC3. [d]p_{het}: p value for heterogeneity by Cochrane's Q test. [e]Meta-analysis in the fix-effects model. [f]Imputed SNP in the PLCO trial.

Table 3. Associations between the number of protective alleles of two independent SNPs with NSCLC OS and DSS in the PLCO Trial

Alleles	Frequency[a]	OS[b]			DSS[b]		
		Death (%)	HR (95% CI)	*p*	Death (%)	HR (95% CI)	*p*
SYK rs11787670 A>G							
AA	1,011	688 (68.05)	1.00		619 (61.23)	1.00	
AG	154	96 (662.34)	0.81 (0.65-1.00)	0.052	87 (56.49)	0.80 (0.64-1.01)	0.058
GG	10	5 (50.0)	0.45 (0.19-1.10)	0.079	3 (30.00)	0.29 (0.09-0.90)	0.032
Trend test				0.011			0.006
ITGA1 rs67715745 T>C							
TT	794	528 (66.50)	1.00		474 (59.70)	1.00	
TC	350	246 (70.29)	0.89 (0.76-1.04)	0.141	221 (63.14)	0.90 (0.76-1.06)	0.211
CC	31	15 (48.39)	0.52 (0.31-0.88)	0.014	14 (45.16)	0.54 (0.32-0.93)	0.026
Trend test				0.012			0.027
NPA[c]							
0	685	464 (67.74)	1.00		415 (60.58)	1.00	
1	400	270 (67.50)	0.87 (0.75-1.02)	0.081	247 (61.75)	0.89 (0.76-1.05)	0.170
2	84	53 (63.10)	0.63 (0.47-0.84)	0.002	46 (54.76)	0.61 (0.45-0.83)	0.002
3-4	6	2 (33.33)	0.38 (0.10-1.54)	0.176	1 (16.67)	0.21 (0.03-1.50)	0.120
Trend test				0.0004			0.0006
Dichotomized NPA							
0-1	1,085	734 (67.65)	1.00		662 (61.01)	1.00	
2-4	90	55 (61.11)	0.65 (0.49-0.85)	0.002	47 (52.22)	0.61 (0.45-0.83)	0.0013

Abbreviations: SNPs, single nucleotide polymorphisms; NSCLC, non-small cell lung cancer; OS, overall survival; DSS, disease-specific survival. PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial; HR, hazards ratio; CI, confidence interval; NPA, number of protective alleles. [a]10 with missing data were excluded. [b]Adjusted for age, sex, smoking status, histology, tumor stage, chemotherapy, surgery, radiotherapy and principal components. [c]Protective alleles were *SYK* rs11787670 G and *ITGA1* rs67715745 C.

analysis in the PLCO dataset by available clinical covariates, including sex, age, histology, smoking status, tumor stage, chemotherapy, radiotherapy, and surgery. For both 0-1 and 2-4 NPA groups, there were no significant interactions between protective alleles and each of these covariates on NSCLC survival except for patients treated with radiotherapy (*P*=0.032 for OS) (Table S5), while for both 0 and 1-4 NPA groups, no significant interactions were found except for histology (*P*=0.007 for OS and *P*=0.021 for DSS) (Table S6).

Time-dependent AUC and ROC curve of the two independent SNPs for NSCLC survival prediction

We further calculated time-dependent AUC and ROC curves accounted for available clinical covariates to assess the predictive effects of the two independent SNPs on NSCLC survival. The addition of protective alleles to the prediction model of 5-year survival significantly increased AUC from 87.00% to 87.67% for OS (*P*=0.029) (**Figure 2C**, **2D**) and from 88.54% to 89.06% for DSS (*P*=0.022) (**Figure 2E**, **2F**).

Bioinformatics functional prediction of the two independent SNPs

To explore biological functions of the two independent SNPs, we assessed SNP-related genomics data using an online bioinformatics tool (HaploReg, https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php). We found that an A>G change in SYK rs11787670 might alter protein motifs and protein binding activity. Similarly, the *ITGA1* rs67715745 T>C change might also potentially affect the protein-coding function, changing protein motifs (Table S7). Meanwhile, according to experimental data from the Encyclopedia of DNA Elements (ENCODE) project, we found that rs11787670 and rs67715745 were both probably located on the H3K4Me1 regions, DNase cluster and transcription factor CHIP-seq (Figure S5). These findings indicate a strong possibility that these two independent SNPs might

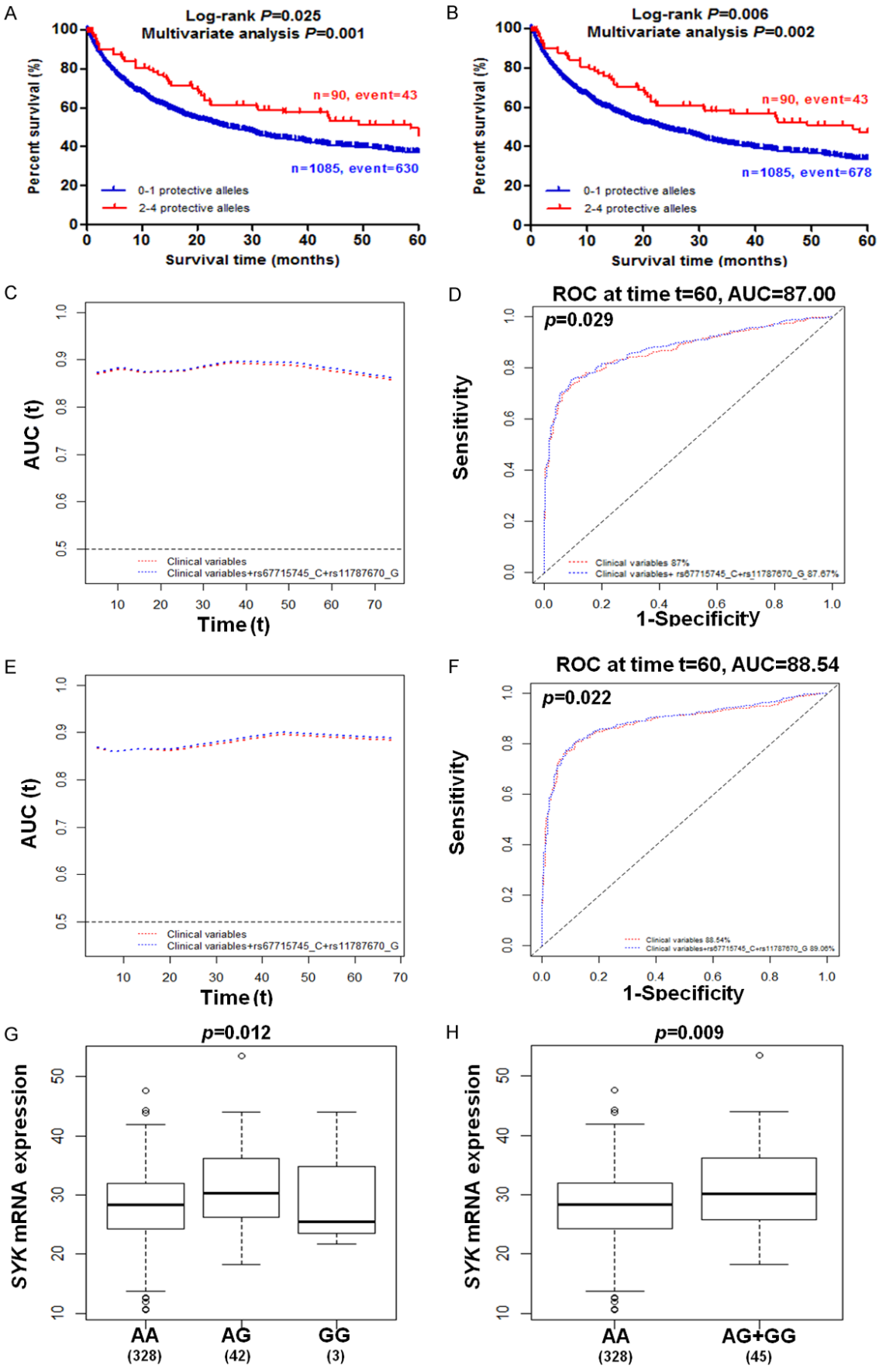
A
Log-rank P=0.025
Multivariate analysis P=0.001
n=90, event=43
n=1085, event=630
0-1 protective alleles
2-4 protective alleles
Percent survival (%)
Survival time (months)
B
Log-rank P=0.006
Multivariate analysis P=0.002
n=90, event=43
n=1085, event=678
0-1 protective alleles
2-4 protective alleles
Percent survival (%)
Survival time (months)
C
AUC (t)
Time (t)
Clinical variables
Clinical variables+rs67715745_C+rs11787670_G
D
ROC at time t=60, AUC=87.00
p=0.029
Sensitivity
1-Specificity
Clinical variables 87%
Clinical variables+ rs67715745_C+rs11787670_G 87.67%
E
AUC (t)
Time (t)
Clinical variables
Clinical variables+rs67715745_C+rs11787670_G
F
ROC at time t=60, AUC=88.54
p=0.022
Sensitivity
1-Specificity
Clinical variables 88.54%
Clinical variables+rs67715745_C+rs11787670_G 89.06%
G
p=0.012
SYK mRNA expression
AA (328)
AG (42)
GG (3)
H
p=0.009
SYK mRNA expression
AA (328)
AG+GG (45)

Figure 2. Two independent SNPs in lymphangiogenesis-related pathway genes predict NSCLC survival and eQTL analysis. Kaplan-Meier survival curves of combined risk alleles of *SYK* rs11787670 A>G and *ITGA1* rs67715745 T>C in the PLCO trial: dichotomized 0-1 protective alleles group and 2-4 protective alleles group in DSS (A) and OS (B). The 60-month NSCLS OS prediction by time-dependent AUC (C) and ROC curve (D) based on clinical variables plus protective alleles. The 60-month NSCLS DSS prediction by time-dependent AUC (E) and ROC curve based on clinical variables plus protective alleles (F). The correlation of rs11787670 genotypes and *SYK* mRNA expression in additive (G) and dominant model (H) from the 1 000 Genomes Project. Abbreviations: PLCO, The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; DSS, disease-specific survival; OS, overall survival; NSCLC, non-small cell lung cancer; AUC, area under receiver curve; ROC, receiver operating characteristic; GTEx, Genotype-Tissue Expression Project.

alter their gene expression through possible transcriptional regulation mechanisms.

eQTL analysis of the two independent SNPs

To further investigate the potential functional relevance of these two SNPs, we performed the eQTL analysis to explore the associations between genotypes of the SNPs and their corresponding mRNA expression levels using data of the 373 European descendants in the 1000 Genomes Project. The *SYK* rs11787670 G genotypes were significantly correlated with *SYK* mRNA expression in both additive (**Figure 2G**) and dominant (**Figure 2H**) models (P=0.012 and P=0.009, respectively), but not in a recessive model (P=0.584) (Figure S6A). Additionally, we also performed the eQTL analysis using data from the GTEx Project, and these genotypes were not significantly correlated with *SYK* mRNA expression levels in either normal lung tissues (n=515) or whole blood samples (n=515) (Figure S6E, S6F). Finally, to investigate the association of SNP-related gene expression and NSCLC survival, we compared mRNA expression levels of SYK in both NSCLC tumor and normal tissues from The Cancer Genome Atlas (TCGA) database (http://ualcan.path.uab.edu/index.html). As shown in Figure S7A, S7B, mRNA expression levels of *SYK* were significantly lower in both lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) tissues than in normal tissues (P=0.003 and P=0.006, respectively). Additionally, the KM survival curves also showed that mRNA expression levels of *SYK* were significantly associated with lung cancer survival (Figure S7E, S7F). These data indicated *SYK* might be a potential prognostic factor for the survival of patients with lung cancer.

Comparatively, the eQTL analysis for *ITGA1* rs67715745 using data of the 373 European descendants obtained in the 1000 Genomes Project indicated that the rs67715745 C genotypes were not correlated with *ITGA1* mRNA expression levels in all additive, dominant and recessive models (Figure S6B-D). Although no *ITGA1* mRNA expression data were available in the GTEx project, we found that *ITGA1* mRNA expression levels were significantly lower in both LUAD (P<0.0001) and LUSC (P<0.0001) tissues than in normal tissues from the TCGA database (Figure S7C, S7D), which could not be depicted in the KM survival curves, because of the absence of available data in this online KM plotter dataset (https://kmplot.com).

Mutation analysis

Given that the effects of gene mutations in tumor tissues may also be involved in tumor metastasis, we investigated the mutation status of *SYK* and *ITGA1* in NSCLC tissues using online cBioPortal for Cancer Genomics (http://www.cbioportal.org). As shown in Figure S8, *SYK* had an extremely low somatic mutation rate in different NSCLC datasets (0.96% in the TCGA 2016; 1.53% in the NSCLC 2017, 2.67% in the MSKCC 2018, and 0.83% in the MSK D1, respectively). Similarly, *ITGA1* also displayed a low somatic mutation rate in different NSCLC datasets (2.27% in the TCGA 2016, 1.53% in the NSCLC 2017, and 1.33% in the MSKCC 2018, respectively). Therefore, these low mutation frequencies in both *SYK* and *ITGA1* unlikely had a significant effect on the expression levels of these two genes in NSCLC.

Discussion

Although the presence of lymphangiogenesis in lung cancer has been linked to metastasis and survival, its routine evaluation in lung tumor tissues and correlation with outcomes remain difficult. In the present study, we assessed the associations between SNPs in lymphangiogenesis-related pathway genes and NSCLC survival using available genotyping data from two previ-

ously published GWASs. Notably, we identified two novel SNPs (i.e., *SYK* rs11787670 A>G and *ITGA1* rs67715745 T>C) that were significantly associated with survival of NSCLC patients of European descendants. Additionally, an increased number of protective alleles of these two independent SNPs were significantly correlated with better NSCLC OS and DSS. Furthermore, the prediction model with combined protective alleles of these two SNPs also showed significantly improved 5-year survival, suggesting that these two independent SNPs may be useful biomarkers of outcomes in NSCLC patients, if validated by future studies. Subsequent analysis for functional relevance of these SNPs indicated that the variant rs11787670 G allele was significantly associated with elevated *SYK* mRNA expression levels, but this association was not found for the variant *ITGA1* rs67715745 C allele. However, mRNA expression levels of *SYK* and *ITGA1* were significantly lower in lung tumor tissues than in adjacent normal lung tissues, while the deceased mRNA expression levels were significantly associated with a poor survival of NSCLC for *SYK* but not for *ITGA1*. These findings provided further support for biological plausibility of the observed associations, particularly for the *SYK* rs11787670 A>G SNP.

SYK, located on chromosome 9q22.2, has 16 exons and encodes a protein known as spleen-associated tyrosine kinase that plays an essential role in the lymphocyte development and activation as well as differentiation of immune cells [30]. Recently, *SYK* was found to be associated with invasion and metastasis in lung cancer cell lines and tissues. As shown in one study, an upregulation of *SYK* expression dramatically promoted the invasion of A549 cells [31]. Other studies detected an extremely low *SYK* mRNA expression in primary tumor tissues, compared with the non-tumor tissues in NSCLC patients [32, 33]. Although few studies investigated the association between *SYK* and NSCLC survival, two studies demonstrated that a lower *SYK* expression was correlated with a poor OS in NSCLC patients [33, 34]. Consistent with these studies, our results suggested that *SYK* is a possible suppressor gene in NSCLC. Moreover, we also found that the survival-associated variant rs11787670 G allele had an allele-dose effect on *SYK* mRNA expression levels in normal lymphoblastoid cells. In bioinformatics analysis using data in the ENCODE Project, rs11787670 seems to be located in a substantial region of the H3K4Me1 layer. A previous study demonstrated that methylation of the *SYK* promoter as well as other epigenetic modifications such as histone methylation/deacetylation have been involved in transcriptional regulation [35]. Therefore, it is also likely that rs11787670 A>G could modify *SYK* mRNA expression through methylation of histone H3. Taken together, the variant rs11787670 G allele may modulate *SYK* mRNA expression and thus survival of NSCLC patients. To our knowledge, this is the first report about the *SYK* rs11787670 G allele as a prognostic factor for NSCLC.

ITGA1, located on chromosome 5q11.2, has 29 exons, encodes one of the important members of integrins, and is involved in cell adhesion, proliferation, tumorigenicity and survival [36]. Prior reports suggest that integrins play a crucial role in tumor development and progression by activating various signaling pathways [37-39]. Although some studies reported that upregulated ITGA1 expression could promote tumorigenicity and tumor progress in colorectal and pancreatic cancer [40, 41]. So far, few integrin members have been explored as prognostic factors for lung cancer. One recent study investigated 30 members of the integrin family to identify prognostic factors for NSCLC, but only ITGA5 and ITGB1 were found to be independent predictors of outcomes in multivariate models [42]. Another study found that integrin α1 knockout mice model showed an inhibitory effect on the primary lung tumors and an increased survival, compared with the wild-type controls [43]. Here, our data showed that rs67715745 T>C was associated with survival of NSCLC, and *ITGA1* mRNA expression was significantly lower in lung cancer tissues than in normal lung tissues from TCGA data. These observations suggest that *ITGA1* may serve as a suppressor gene in NSCLC; however, we did not have data to support the biological plausibility of the association of the variant *ITGA1* rs67715745 C with survival of NSCLC patients.

There are several limitations in this present study. First, the two GWAS datasets used only included Caucasian populations; thus, our results may not be generalizable to other ethnic populations. Second, the exact molecular mechanisms underlying the associations between the two SNPs and NSCLC survival remain

unclear. Thus, further molecular studies and functional experiments are required to verify our findings. Third, although 1 185 participants were recruited in the PLCO trial, the number of each subgroup was relatively small, which might have reduced the statistical power. Fourth, both the PLCO and HLCS datasets contained limited clinical variables for further adjustment and stratification in additional analyses, which could affect the assessment on survival. Finally, the detailed individual genotype and phenotype data of the HLCS study were not available, which made it impossible for us to do the thorough combined and stratified analyses.

Acknowledgements

The authors thank all the participants of the PLCO Cancer Screening Trial; and the National Cancer Institute for providing access to the data collected by the PLCO trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by the National Cancer Institute. The authors would also like to acknowledge the dbGaP repository for providing cancer genotyping datasets. The accession numbers for the datasets for lung cancer are phs000336.v1.p1 and phs000093.v2.p2. A list of contributing investigators and funding agencies for those studies can be found in the Supplemental Data. This work was supported by the National Institute of Health (CA090578, CA074386, CA092824, U01CA209414, R01-NS091307, R56AG062302); The Duke Cancer Institute as part of the P30 Cancer Center Support Grant (NIH/NCI CA014236); and the V Foundation for Cancer Research (D2017-19); The National Natural Science Foundation of China (81760419); 2018 Guangxi One Thousand Young and Middle-Aged College and University Backbone Teachers Cultivation Program to Lihua Liu, P.R. China; "Medical Excellence Award" Funded by the Creative Research Development Grant from the First Affiliated Hospital of Guangxi Medical University to Lihua Liu.

The study was conducted in compliance with the principles of the declaration of Helsinki and local ethical and legal requirements. The protocol and informed consent were approved by the independent ethics committees or institutional review board of all participating sites.

Disclosure of conflict of interest

None.

Abbreviations

SNPs, single nucleotide polymorphisms; NSCLC, Non-small cell lung cancer; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; GWAS, Genome-Wide Association Study; PLCO, the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; HLCS, Harvard Lung Cancer Susceptibility; OS, overall survival; DSS, disease-special survival; LD, linkage disequilibrium; FDR, false discovery rate; BFDP, Bayesian false discovery probability; eQTL, expression quantitative trait loci; TCGA, The Cancer Genome Atlas; ROC, receiver operating characteristic; EAF, effect allele frequency; HR, hazards ratio; CI, confidence interval; AUC, area under the receiver operating characteristic curve; GTEx, genotype-tissue expression project; NPA, number of protective alleles; SYK, spleen tyrosine kinase; ITGA1, integrin subunit alpha 1.

Address correspondence to: Qingyi Wei, Duke Cancer Institute, Duke University Medical Center and Department of Population Health Sciences, Duke University School of Medicine, 905 S LaSalle Street, Durham, NC 27710, USA. Tel: 919-660-0562; E-mail: qingyi.wei@duke.edu

References

[1] Global Burden of Disease Cancer Collaboration, Fitzmaurice C, Akinyemiju TF, Al Lami FH, Alam T, Alizadeh-Navaei R, Allen C, Alsharif U, Alvis-Guzman N, Amini E, Anderson BO, Aremu O, Artaman A, Asgedom SW, Assadi R, Atey TM, Avila-Burgos L, Awasthi A, Ba Saleem HO, Barac A, Bennett JR, Bensenor IM, Bhakta N, Brenner H, Cahuana-Hurtado L, Castañeda-Orjuela CA, Catalá-López F, Choi JJ, Christopher DJ, Chung SC, Curado MP, Dandona L, Dandona R, das Neves J, Dey S, Dharmaratne SD, Doku DT, Driscoll TR, Dubey M, Ebrahimi H, Edessa D, El-Khatib Z, Endries AY, Fischer F, Force LM, Foreman KJ, Gebrehiwot SW, Gopalani SV, Grosso G, Gupta R, Gyawali B, Hamadeh RR, Hamidi S, Harvey J, Hassen HY, Hay RJ, Hay SI, Heibati B, Hiluf MK, Horita N, Hosgood HD, Ilesanmi OS, Innos K, Islami F, Jakovljevic MB, Johnson SC, Jonas JB, Kasaeian A, Kassa TD, Khader YS, Khan EA, Khan G, Khang YH, Khosravi MH, Khubchandani J, Kopec JA, Kumar GA, Kutz M, Lad DP, Lafranconi A, Lan

Q, Legesse Y, Leigh J, Linn S, Lunevicius R, Majeed A, Malekzadeh R, Malta DC, Mantovani LG, McMahon BJ, Meier T, Melaku YA, Melku M, Memiah P, Mendoza W, Meretoja TJ, Mezgebe HB, Miller TR, Mohammed S, Mokdad AH, Moosazadeh M, Moraga P, Mousavi SM, Nangia V, Nguyen CT, Nong VM, Ogbo FA, Olagunju AT, Pa M, Park EK, Patel T, Pereira DM, Pishgar F, Postma MJ, Pourmalek F, Qorbani M, Rafay A, Rawaf S, Rawaf DL, Roshandel G, Safiri S, Salimzadeh H, Sanabria JR, Santric Milicevic MM, Sartorius B, Satpathy M, Sepanlou SG, Shackelford KA, Shaikh MA, Sharif-Alhoseini M, She J, Shin MJ, Shiue I, Shrime MG, Sinke AH, Sisay M, Sligar A, Sufiyan MB, Sykes BL, Tabarés-Seisdedos R, Tessema GA, Topor-Madry R, Tran TT, Tran BX, Ukwaja KN, Vlassov VV, Vollset SE, Weiderpass E, Williams HC, Yimer NB, Yonemoto N, Younis MZ, Murray CJL and Naghavi M. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. JAMA Oncol 2018; 4: 1553-1568.

[2] Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, Nicholson AG, Groome P, Mitchell A and Bolejack V; International Association for the Study of Lung Cancer Staging and Prognostic Factors Committee, Advisory Boards, and Participating Institutions; International Association for the Study of Lung Cancer Staging and Prognostic Factors Committee Advisory Boards and Participating Institutions. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. J Thorac Oncol 2016; 11: 39-51.

[3] Siegel RL, Miller KD and Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020; 70: 7-30.

[4] Bagcchi S. Lung cancer survival only increases by a small amount despite recent treatment advances. Lancet Respir Med 2017; 5: 169.

[5] Perez-Ramirez C, Canadas-Garre M, Molina MA, Robles AI, Faus-Dader MJ and Calleja-Hernandez MA. Contribution of genetic factors to platinum-based chemotherapy sensitivity and prognosis of non-small cell lung cancer. Mutat Res 2017; 771: 32-58.

[6] Massuti B, Sanchez JM, Hernando-Trancho F, Karachaliou N and Rosell R. Are we ready to use biomarkers for staging, prognosis and treatment selection in early-stage non-small-cell lung cancer? Transl Lung Cancer Res 2013; 2: 208-221.

[7] Saharinen P, Tammela T, Karkkainen MJ and Alitalo K. Lymphatic vasculature: development, molecular regulation and role in tumor metastasis and inflammation. Trends Immunol 2004; 25: 387-395.

[8] Paduch R. The role of lymphangiogenesis and angiogenesis in tumor metastasis. Cell Oncol (Dordr) 2016; 39: 397-410.

[9] Stacker SA, Williams SP, Karnezis T, Shayan R, Fox SB and Achen MG. Lymphangiogenesis and lymphatic vessel remodelling in cancer. Nat Rev Cancer 2014; 14: 159-172.

[10] Werynska B, Dziegiel P and Jankowska R. Role of lymphangiogenesis in lung cancer. Folia Histochem Cytobiol 2009; 47: 333-342.

[11] Takanami I. Lymphatic microvessel density using D2-40 is associated with nodal metastasis in non-small cell lung cancer. Oncol Rep 2006; 15: 437-442.

[12] Adachi Y, Nakamura H, Kitamura Y, Taniguchi Y, Araki K, Shomori K, Horie Y, Kurozawa Y, Ito H and Hayashi K. Lymphatic vessel density in pulmonary adenocarcinoma immunohistochemically evaluated with anti-podoplanin or anti-D2-40 antibody is correlated with lymphatic invasion or lymph node metastases. Pathol Int 2007; 57: 171-177.

[13] Wang J, Li K, Wang B and Bi J. Lymphatic microvessel density as a prognostic factor in non-small cell lung carcinoma: a meta-analysis of the literature. Mol Biol Rep 2012; 39: 5331-5338.

[14] Trivella M, Pezzella F, Pastorino U, Harris AL and Altman DG; Prognosis In Lung Cancer (PILC) Collaborative Study Group. Microvessel density as a prognostic factor in non-small-cell lung carcinoma: a meta-analysis of individual patient data. Lancet Oncol 2007; 8: 488-499.

[15] Egan KM, Nabors LB, Olson JJ, Monteiro AN, Browning JE, Madden MH and Thompson RC. Rare TP53 genetic variant associated with glioma risk and outcome. J Med Genet 2012; 49: 420-421.

[16] Lin WY, Camp NJ, Cannon-Albright LA, Allen-Brady K, Balasubramanian S, Reed MW, Hopper JL, Apicella C, Giles GG, Southey MC, Milne RL, Arias-Perez JI, Menendez-Rodriguez P, Benitez J, Grundmann M, Dubrowinskaja N, Park-Simon TW, Dork T, Garcia-Closas M, Figueroa J, Sherman M, Lissowska J, Easton DF, Dunning AM, Rajaraman P, Sigurdson AJ, Doody MM, Linet MS, Pharoah PD, Schmidt MK and Cox A. A role for XRCC2 gene polymorphisms in breast cancer risk and survival. J Med Genet 2011; 48: 477-484.

[17] Liang B, Ding H, Huang L, Luo H and Zhu X. GWAS in cancer: progress and challenges. Mol Genet Genomics 2020; 295: 537-561.

[18] Tang D, Zhao YC, Liu H, Luo S, Clarke JM, Glass C, Su L, Shen S, Christiani DC, Gao W and Wei Q. Potentially functional genetic variants in

PLIN2, SULT2A1 and UGT1A9 genes of the Ketone pathway and survival of non-small cell lung cancer. Int J Cancer 2020; [Epub ahead of print].

[19] Xu Y, Liu H, Liu S, Wang Y, Xie J, Stinchcombe TE, Su L, Zhang R, Christiani DC, Li W and Wei Q. Genetic variant of IRAK2 in the toll-like receptor signaling pathway and survival of non-small cell lung cancer. Int J Cancer 2018; 143: 2400-2408.

[20] Hocking WG, Hu P, Oken MM, Winslow SD, Kvale PA, Prorok PC, Ragard LR, Commins J, Lynch DA, Andriole GL, Buys SS, Fouad MN, Fuhrman CR, Isaacs C, Yokochi LA, Riley TL, Pinsky PF, Gohagan JK, Berg CD and Team PP. Lung cancer screening in the randomized prostate, lung, colorectal, and ovarian (PLCO) cancer screening trial. J Natl Cancer Inst 2010; 102: 722-731.

[21] Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M and Feolo M. NCBI's database of genotypes and phenotypes: dbGaP. Nucleic Acids Res 2014; 42: D975-979.

[22] Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J and Sherry ST. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007; 39: 1181-1186.

[23] Qian D, Liu H, Wang X, Ge J, Luo S, Patz EF Jr, Moorman PG, Su L, Shen S, Christiani DC and Wei Q. Potentially functional genetic variants in the complement-related immunity gene-set are associated with non-small cell lung cancer survival. Int J Cancer 2019; 144: 1867-1876.

[24] Aulchenko YS, Ripke S, Isaacs A and van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics 2007; 23: 1294-1296.

[25] Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 2007; 81: 208-227.

[26] Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlof J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HP, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Geuvadis C, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Hasler R, Syvanen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigo R, Gut IG, Estivill X and Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013; 501: 506-511.

[27] Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015; 348: 648-660.

[28] Ward LD and Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res 2016; 44: D877-881.

[29] Wang Y, Liu H, Ready NE, Su L, Wei Y, Christiani DC and Wei Q. Genetic variants in ABCG1 are associated with survival of nonsmall-cell lung cancer patients. Int J Cancer 2016; 138: 2592-2601.

[30] Sada K, Takano T, Yanagi S and Yamamura H. Structure and function of Syk protein-tyrosine kinase. J Biochem 2001; 130: 177-186.

[31] Sun Q, Peng C, Cong B, Hao Y, Guo J, Zhao Y and Zhao X. Involvement of syk and VEGF-C in invasion of lung adenocarcinoma A549 cells. J Cancer Res Ther 2016; 12: 640-644.

[32] Chuanliang P, Yunpeng Z, Yingtao H, Qifeng S, Xiaogang Z and Bo C. Syk expression in non-small-cell lung cancer and its relation with angiogenesis. J Cancer Res Ther 2016; 12: 663-666.

[33] Kowalczuk O, Laudanski J, Laudanski W, Niklinska WE, Kozlowski M and Niklinski J. Lymphatics-associated genes are downregulated at transcription level in non-small cell lung cancer. Oncol Lett 2018; 15: 6752-6762.

[34] Gao D, Wang L, Zhang H, Yan X, Yang J, Zhou R, Chang X, Sun Y, Tian S, Yao Z, Zhang K, Liu Z and Ma Z. Spleen tyrosine kinase SYK(L) interacts with YY1 and coordinately suppresses SNAI2 transcription in lung cancer cells. FEBS J 2018; 285: 4229-4245.

[35] Bonavida B and Baritaki S. The novel role of yin yang 1 in the regulation of epithelial to mesenchymal transition in cancer via the dysregulated NF-kappaB/Snail/YY1/RKIP/PTEN circuitry. Crit Rev Oncog 2011; 16: 211-226.

[36] Shattil SJ, Kim C and Ginsberg MH. The final steps of integrin activation: the end game. Nat Rev Mol Cell Biol 2010; 11: 288-300.

[37] Chen JC, Chen YJ, Lin CY, Fong YC, Hsu CJ, Tsai CH, Su JL and Tang CH. Correction: amphiregulin enhances alpha6beta1 integrin expression and cell motility in human chondrosarcoma cells through Ras/Raf/MEK/ERK/AP-1 pathway. Oncotarget 2017; 8: 25830.

[38] Boudjadi S, Carrier JC, Groulx JF and Beaulieu JF. Integrin alpha1beta1 expression is con-

trolled by c-MYC in colorectal cancer cells. Oncogene 2016; 35: 1671-1678.
[39] Yan Q, Jiang L, Liu M, Yu D, Zhang Y, Li Y, Fang S, Li Y, Zhu YH, Yuan YF and Guan XY. ANGPTL1 Interacts with Integrin alpha1beta1 to Suppress HCC angiogenesis and metastasis by Inhibiting JAK2/STAT3 Signaling. Cancer Res 2017; 77: 5831-5845.
[40] Li H, Wang Y, Rong SK, Li L, Chen T, Fan YY, Wang YF, Yang CR, Yang C, Cho WC and Yang J. Integrin alpha1 promotes tumorigenicity and progressive capacity of colorectal cancer. Int J Biol Sci 2020; 16: 815-826.
[41] Gharibi A, La Kim S, Molnar J, Brambilla D, Adamian Y, Hoover M, Hong J, Lin J, Wolfenden L and Kelber JA. ITGA1 is a pre-malignant biomarker that promotes therapy resistance and metastatic potential in pancreatic cancer. Sci Rep 2017; 7: 10060.
[42] Zheng W, Jiang C and Li R. Integrin and gene network analysis reveals that ITGA5 and ITGB1 are prognostic in non-small-cell lung cancer. Onco Targets Ther 2016; 9: 2317-2327.
[43] Macias-Perez I, Borza C, Chen X, Yan X, Ibanez R, Mernaugh G, Matrisian LM, Zent R and Pozzi A. Loss of integrin alpha1beta1 ameliorates Kras-induced lung cancer. Cancer Res 2008; 68: 6127-6135.

Table S1. Comparison of characteristics between the PLCO trial and the HLCS study

Characteristics	PLCO		HLCS		p^*
	Frequency	Deaths (%)	Frequency	Deaths (%)	
Total	1,185	798 (67.3)	984	665 (67.5)	
Median overall survival (months)	23.8		39.9		
Age					
≤71	636	400 (62.9)	654	428 (65.4)	<0.0001
>71	549	398 (72.5)	330	237 (71.8)	
Sex					
Male	698	507 (72.6)	507	379 (74.7)	0.0006
Female	487	291 (59.8)	477	286 (59.9)	
Smoking status					
Never	115	63 (54.8)	92	52 (56.5)	0.166
Current	423	272 (64.3)	390	266 (68.2)	
Former	647	463 (71.6)	502	347 (69.1)	
Histology					
Adenocarcinoma	577	348 (60.3)	597	378 (63.3)	<0.0001
Squamous cell carcinoma	285	192 (67.4)	216	156 (72.2)	
Others	323	258 (79.9)	171	131 (76.6)	
Stage					
I-IIIA	655	315 (48.1)	606	352 (58.0)	0.003
IIIB-IV	528	482 (91.3)	377	313 (83.0)	
Missing	2		-		

Abbreviations: PLCO, the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; HLCS, Harvard Lung Cancer Susceptibility Study. *Chi-square test for the comparison of characteristics between the PLCO trial and the HLCS study for each clinical variable.

Table S2. List of 247 selected genes in lymphangiogenesis-related gene-set used in discovery analysis

Dataset	Name of pathway	Selected genes[a]	Number of genes
GO	GO_ENDOTHELIAL_CELL_PROLIFERATION	*ACVRL1, ADAM17, AGGF1, AGTR1, AIMP1, AKT1, AKT3, ALDH1A2, ANG, APELA, APLN, APLNR, APOA1, APOE, APOH, ARG1, ARNT, ATOH8, ATP5F1A, ATP5IF1, BMP2, BMP4, BMP6, BMPER, BMPR2, CAV1, CAV2, CCL11, CCL2, CCL24, CCL26, CCR3, CD34, CDH13, CNMD, CXCL12, CYBA, DLG1, DLL4, DYSF, ECM1, EGFL7, EGR3, EMC10, EPHA2, ERN1, F3, FGF2, FGFBP1, FGFR1, FLT1, FLT4, GATA2, GDF2, GHRL, GHSR, HIF1A, HMGB1, HMGB2, HMOX1, HTR2B, IL10, ITGA4, ITGB1BP1, ITGB3, JCAD, JUN, KDR, KRIT1, LEP, LOXL2, LRG1, MEF2C, MIR101-1, MIR101-2, MIR10A, MIR10B, MIR126, MIR129-1, MIR129-2, MIR130A, MIR132, MIR133B, MIR135B, MIR146A, MIR152, MIR155, MIR15A, MIR15B, MIR16-1, MIR16-2, MIR193A, MIR20B, MIR21, MIR22, MIR222, MIR2355, MIR23A, MIR23B, MIR24-1, MIR24-2, MIR26A1, MIR26A2, MIR27A, MIR27B, MIR29A, MIR29C, MIR30B, MIR30E, MIR329-1, MIR329-2, MIR342, MIR34A, MIR361, MIR410, MIR424, MIR483, MIR487B, MIR492, MIR494, MIR495, MIR497, MIR499A, MIR503, MIR98, MIRLET7B, MMP14, MTOR, MYDGF, NF1, NGFR, NOX5, NR2F2, NR4A1, NRARP, NRAS, NRP1, NRP2, PDCD10, PDCD6, PDCL3, PDGFB, PDPK1, PGF, PIK3CB, PLCG1, PLXNB3, PPARG, PPP1R16B, PRKCA, PRKD1, PRKD2, PRKX, PRL, PROX1, PTPRM, RGCC, RICTOR, RPTOR, SCARB1, SCG2, SEMA5A, SIRT1, SP1, SPARC, STAT1, STAT3, STAT5A, SULF1, SYNJ2BP, TEK, TGFBR1, THAP1, THBS1, THBS4, TNF, TNFSF12, TNMD, VASH1, VASH2, VEGFA, VEGFB, VEGFC, VEGFD, VIP, WNT2, WNT5A, XBP1, XDH, ZNF580*	190
GO	GO_LYMPH_VESSEL_DEVELOPMENT	*ACVR2B, ACVRL1, BMPR2, CCBE1, EFNB2, EPHA2, FLT4, FOXC1, FOXC2, HEG1, LGALS8, NR2F2, PDPN, PKD1, PPP3CB, PROX1, PROX2, PTPN14, SOX18, SYK, TBX1, TMEM204, VASH1, VEGFA, VEGFC*	25
GO	GO_LYMPH_VESSEL_MORPHOGENESIS	*ACVR2B, ACVRL1, BMPR2, CCBE1, EPHA2, FLT4, FOXC1, FOXC2, LGALS8, PDPN, PKD1, PPP3CB, PROX1, PROX2, PTPN14, SOX18, VASH1, VEGFA, VEGFC*	19
GO	GO_LYMPH_NODE_DEVELOPMENT	*CD248, CXCR5, FADD, IL15, IL7R, LTA, LTB, NKX2-3, PDPN, POLB, RC3H1, RC3H2, RIPK3, TGFB1, TNFRSF11A, TOX*	16
GO	GO_LYMPHANGIOGENESIS	*ACVR2B, ACVRL1, BMPR2, CCBE1, EPHA2, FLT4, FOXC1, FOXC2, PDPN, PPP3CB, PROX1, PROX2, PTPN14, SOX18, VASH1, VEGFC*	16
GO	GO_LUNG_VASCULATURE_DEVELOPMENT	*ERRFI1, FOXF1, ID1, LIF, STRA6, TCF21*	6
GO	GO_LYMPHATIC_ENDOTHELIAL_CELL_DIFFERENTIATION	*ACVR2B, ACVRL1, BMPR2, NR2F2, PDPN, PROX1, PROX2, SOX18*	8
PID	PID_LYMPH_ANGIOGENESIS_PATHWAY	*AKT1, COL1A1, COL1A2, CREB1, CRKFLT4, FN1, GRB2, ITGA1, ITGA2, ITGA4, ITGA5, ITGB1, MAP2K4, MAPK1, MAPK11, MAPK14, MAPK3, PIK3CA, PIK3R1, RPS6KA1, SHC1, SOS1, VEGFC, VEGFD*	25
KEGG	-	-	0
BIOCARTA	-	-	0
REACTOME	-	-	0
Total		*PDCD6, CRK, TMEM204, FOXC1, MIR22, MIR132, PKD1, MIR483, PDPK1, LRG1, MYDGF, MIR101-2, BMP2, MIR497, TNFSF12, PTPRM, BMP6, ERRFI1, SEMA5A, MIR34A, ITGB1BP1, ADAM17, GHRL, MTOR, MAP2K4, PPARG, PDPN, MIR24-2, MIR27A, MIR23A, FGFBP1, EPHA2, TBX1, ANG, MAPK1, PRL, EGR3, LOXL2, MMP14, RIPK3, RPS6KA1, MIR155, TEK, ATP5IF1, FLT1, XBP1, NF1, MIR193A, PRKD1, MAPK3, ID1, JCAD, LIF, HMGB1, LTA, TNF, LTB, XDH, CCL2, CCL11, ITGB1, NRP1, MIR499A, BMPER, HMOX1, IL7R, MAPK14, PPP1, R16B, MIR26A1, FGFR1, ACVR2B, RICTOR, SOS1, PDGFB, PLCG1, THBS1, STAT5A, STAT3, MIR30E, DLL4, TGFB1, RGCC, POLB, THAP1, MIR129-2, ATP5F1A, VEGFA, CXCL12, ITGB3, APOE, MIR152, CCR3, MIRLET7B, MIR10A, PRKD2, NGFR, COL1A1, GDF2, MIR16-1, MIR15A, MAPK11, EMC10, MIR133B, ITGA1, ITGA2, ACVRL1, NR4A1, CNMD, SP1, BMP4, ITGA5, WNT5A, KDR, ZNF580, APLNR, CCBE1, MIR130A, MIR21, MIR26A2, ALDH1A2, JUN, TOX, TNFRSF11A, ERN1, HIF1A, SOX18, VEGFB, APOH, PRKCA, MIR101-1, CD248, PIK3R1, NOX5, SIRT1, FADD, SULF1, SYNJ2BP, DYSF, GRB2, STRA6, PPP3CB, PROX2, CCL26, PGF, CCL24, AGGF1, VASH1, RPTOR, THBS4, CDH13, ATOH8, FOXF1, FOXC2, MEF2C, CYBA, KRIT1, SYK, COL1A2, F3, MIR492, NR2F2, MIR23B, MIR27B, MIR24-1, MIR342, PDCL3, NKX2-3, MIR329-1, MIR329-2, MIR494, MIR495, MIR487B, MIR410, TGFBR1, AKT1, EFNB2, AIMP1, NRAS, CAV2, CAV1, APOA1, WNT2, CXCR5, FGF2, HEG1, SCARB1, RC3H2, MIR129-1, LEP, GATA2, MIR29A, ARG1, TCF21, MIR30B, PIK3CB, EGFL7, MIR126, NRARP, IL15, AGTR1, ECM1, ARNT, SPARC, VIP, SHC1, MIR146A, MIR15B, MIR16-2, APELA, PDCD10, GHSR, RC3H1, HMGB2, MIR10B, VEGFC, PIK3CA, FLT4, ITGA4, STAT1, DLG1, BMPR2, MIR135B, NRP2, IL10, MIR2355, MIR29C, CD34, CREB1, VASH2, PROX1, PTPN14, FN1, SCG2, HTR2B, LGALS8, TNMD, AKT3APLN, MIR20B, MIR222, MIR361, MIR424, MIR503, MIR98, PLXNB3, PRKX, VEGFD*	247[b]

[a]Genes were selected based on online datasets (http://software.broadinstitute.org/gsea/msigdb/search.jsp) and literatures; [b]56 duplicated genes had been excluded; Keyword: lymph; lymph AND vessel; Organism: Homo sapiens.

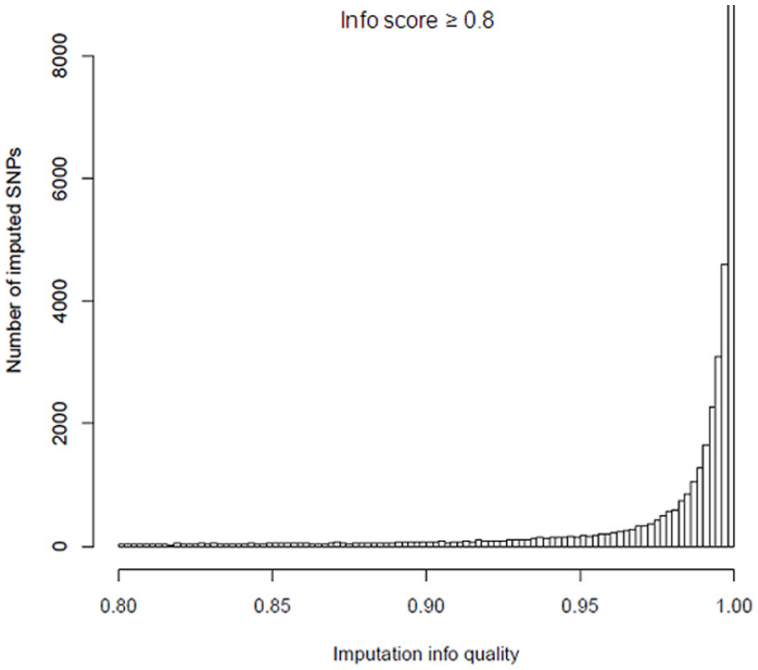

Figure S1. The distribution of the imputation info score in the present study.

Table S3. Associations of the first 10 principal components and OS of NSCLC in the PLCO trial

PC*	Parameter Estimate	Standard Error	Chi-Square	*p*
PC1	4.821	1.353	12.697	<0.001
PC2	-0.681	1.228	0.308	0.579
PC3	-3.054	0.949	10.351	0.001
PC4	-2.837	1.246	5.184	0.023
PC5	-0.910	1.232	0.546	0.460
PC6	1.355	1.252	1.172	0.279
PC7	-0.236	1.218	0.038	0.846
PC8	-1.684	1.322	1.622	0.203
PC9	-1.886	1.267	2.216	0.137
PC10	0.347	1.240	0.078	0.180

Abbreviations: OS, overall survival; NSCLC, non-small cell lung cancer; PLCO, the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; PC, principal component. *The first 4 PC were used for adjustment for population stratification in the multivariate analysis.

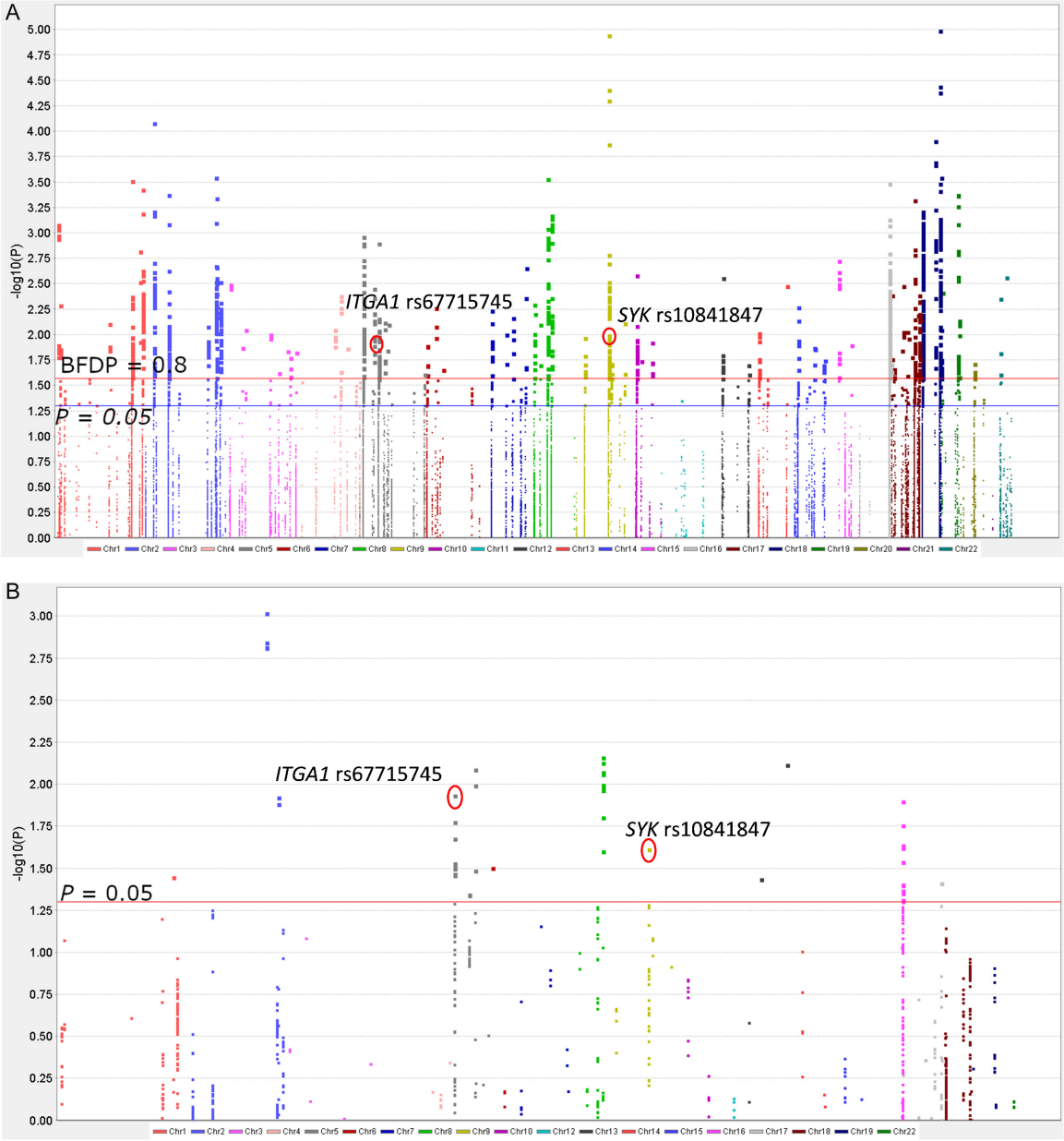

Figure S2. Manhattan plot. Manhattan plot for 34,509 SNPs of lymphangiogenesis-related pathway genes in the PLCO trial (A). Manhattan plot for 1,076 SNPs in the HLCS dataset (B). The blue horizontal line indicates *p*=0.05 and the red line indicates BFDP=0.80. Abbreviations: PLCO, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; HLCS, the Harvard Lung Cancer Susceptibility Study; BFDP, Bayesian false-discovery probability.

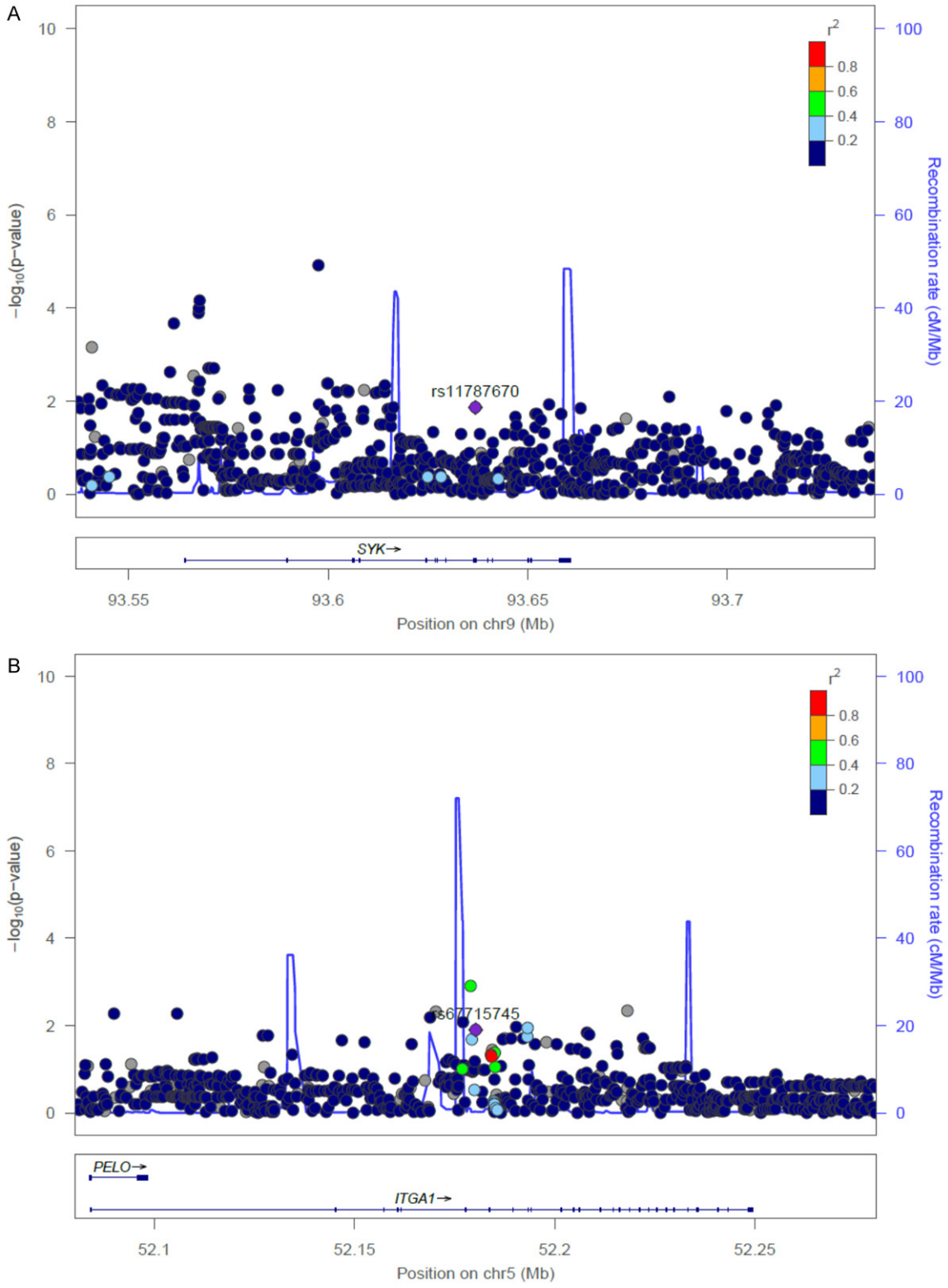

Figure S3. Regional association plots for two independent SNPs in the lymphangiogenesis-related pathway genes. Regional association plots contained 100 kb up or downstream of *SYK* (A) and *ITGA1* (B). Data points are colored according to the level of LD of each pair of SNPs based on the hg19/1000 Genomes European population. The left-hand y-axis shows the association *P*-value of individual SNPs in the discovery dataset, which is plotted as -log10 (*P*) against chromosomal base-pair position. The right-hand y-axis shows the recombination rate estimated from HapMap Data Rel 22/phase II European population. The Regional association plots were generated using Locus Zoom (http://locuszoom.org/). Abbreviations: SNPs, single-nucleotide polymorphisms; LD, linkage disequilibrium.

Table S4. Associations between the number of protective alleles of two independent SNPs with NSCLC OS and DSS in the PLCO Trial

Alleles	Frequency[a]	OS[b]			DSS[b]		
		Death (%)	HR (95% CI)	*p*	Death (%)	HR (95% CI)	*p*
SYK rs11787670 A>G							
AA	1,011	688 (68.05)	1.00		619 (61.23)	1.00	
AG	154	96 (662.34)	0.81 (0.65-1.00)	0.052	87 (56.49)	0.80 (0.64-1.01)	0.058
GG	10	5 (50.0)	0.45 (0.19-1.10)	0.079	3 (30.00)	0.29 (0.09-0.90)	0.032
Trend test				0.011			0.006
ITGA1 rs67715745 T>C							
TT	794	528 (66.50)	1.00		474 (59.70)	1.00	
TC	350	246 (70.29)	0.89 (0.76-1.04)	0.141	221 (63.14)	0.90 (0.76-1.06)	0.211
CC	31	15 (48.39)	0.52 (0.31-0.88)	0.014	14 (45.16)	0.54 (0.32-0.93)	0.026
Trend test				0.012			0.027
NPA[c]							
0	685	464 (67.74)	1.00		415 (60.58)	1.00	
1	400	270 (67.50)	0.87 (0.75-1.02)	0.081	247 (61.75)	0.89 (0.76-1.05)	0.170
2	84	53 (63.10)	0.63 (0.47-0.84)	0.002	46 (54.76)	0.61 (0.45-0.83)	0.002
3-4	6	2 (33.33)	0.38 (0.10-1.54)	0.176	1 (16.67)	0.21 (0.03-1.50)	0.120
Trend test				0.0004			0.0006
Dichotomized NPA							
0	685	464 (67.74)	1.00		415 (60.58)	1.00	
1-4	490	325 (66.33)	0.81 (0.70-0.94)	0.005	294 (60.00)	0.82 (0.71-0.96)	0.012

Abbreviations: SNP, single nucleotide polymorphism; NSCLC, non-small cell lung cancer; PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial; HR, hazards ratio; CI, confidence interval; OS, overall survival; DSS, disease-specific survival. NPA: number of protective alleles. [a]10 with missing data were excluded. [b]Adjusted for age, sex, smoking status, histology, tumor stage, chemotherapy, surgery, radiotherapy and principal components. [c]Protective alleles were *SYK* rs11787670_G and *ITGA1* rs67715745_C.

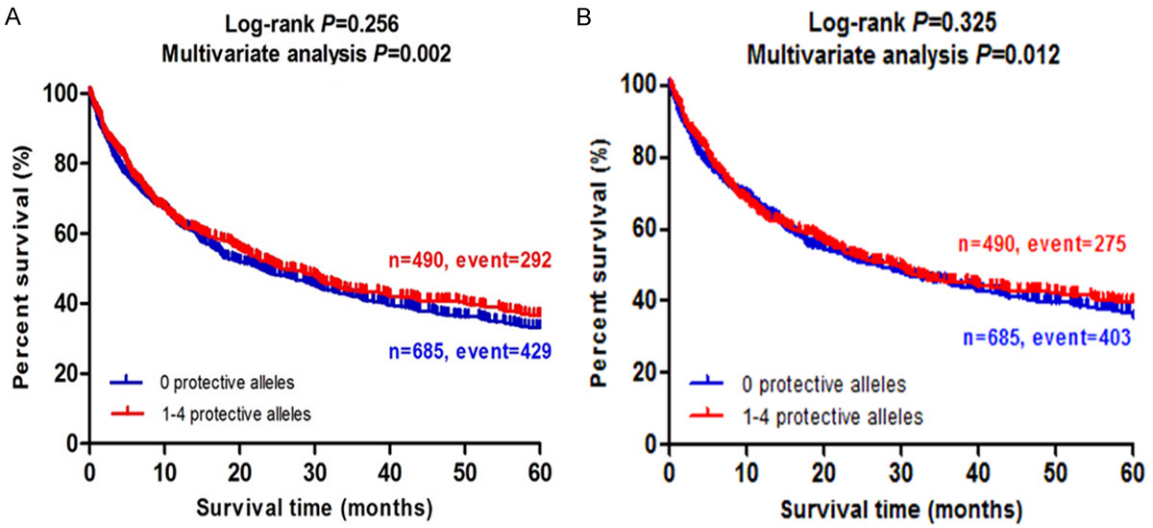

Figure S4. Kaplan-Meier survival curve of combined protective alleles of *SYK* rs11787670 A>G and *ITGA1* rs67715745 T>C in the PLCO trial: dichotomized 0 protective alleles group and 1-4 protective alleles group in OS (A), dichotomized 0 protective alleles group and 1-4 protective alleles group in DSS (B). Abbreviations: PLCO, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial.

Table S5. Stratified analysis for associations between (0-1 vs. 2-4) protective alleles and NSCLC survival in the PLCO trial

Characteristics	0-1 protective alleles	2-4 protective alleles	Multivariate Analysis[b] for OS			Multivariate Analysis[b] for DSS		
	Frequency[a]	Frequency[a]	HR (95% CI)	*p*	p_{inter}[c]	HR (95% CI)	*p*	p_{inter}[c]
Age (years)								
≤71	591	43	0.85 (0.57-1.27)	0.429		0.67 (0.42-1.07)	0.091	
>71	494	47	0.55 (0.37-0.81)	0.002	0.140	0.60 (0.40-0.89)	0.011	0.668
Sex								
Male	635	60	0.72 (0.52-1.00)	0.047		0.71 (0.50-1.00)	0.052	
Female	450	30	0.48 (0.28-0.83)	0.009	0.274	0.44 (0.25-0.80)	0.007	0.188
Smoking status								
Never	106	8	0.54 (0.14-2.08)	0.373		0.59 (0.17-2.02)	0.402	
Current	382	35	0.55 (0.35-0.88)	0.013		0.51 (0.31-0.85)	0.010	
Former	597	47	0.77 (0.53-1.11)	0.165	0.271	0.74 (0.49-1.10)	0.132	0.316
Histology								
Adeno	535	40	0.79 (0.51-1.22)	0.284		0.63 (0.39-1.04)	0.069	
Squamous	258	26	0.54 (0.30-0.95)	0.033		0.49 (0.25-0.92)	0.028	
Others	292	24	0.57 (0.35-0.92)	0.022	0.913	0.68 (0.42-1.11)	0.123	0.380
Tumor stage								
I-IIIA	599	55	0.65 (0.43-0.98)	0.042		0.56 (0.34-0.93)	0.025	
IIIB-IV	486	35	0.65 (0.44-0.94)	0.024	0.678	0.70 (0.48-1.02)	0.062	0.714
Chemotherapy								
No	594	44	0.72 (0.47-1.12)	0.149		0.56 (0.32-0.96)	0.035	
Yes	491	46	0.66 (0.46-0.95)	0.027	0.836	0.70 (0.49-1.01)	0.057	0.193
Radiotherapy								
No	700	61	0.88 (0.63-1.25)	0.484		0.78 (0.53-1.15)	0.206	
Yes	385	29	0.45 (0.28-0.72)	0.0009	0.032	0.50 (0.31-0.80)	0.005	0.203
Surgery								
No	588	47	0.61 (0.44-0.85)	0.003		0.67 (0.48-0.93)	0.018	
Yes	497	43	0.78 (0.47-1.31)	0.354	0.245	0.49 (0.24-1.00)	0.052	0.622

Abbreviations: OS, overall survival; DSS, disease-specific survival; NSCLC, non-small cell lung cancer; PLCO, the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; HR, hazards ratio; CI, confidence interval. [a]10 with missing data were excluded; [b]Adjusted for age, sex, stage, histology, smoking status, chemotherapy, radiotherapy, surgery, PC1, PC2, PC3, and PC4; [c]p_{inter}: *p* value for interaction analysis between characteristics and protective alleles.

Table S6. Stratified analysis for associations between (0 vs. 1-4) protective alleles and NSCLC survival in the PLCO trial

Characteristics	0 protectivealleles	1-4 protectivealleles	Multivariate Analysisb for OS			Multivariate Analysis[b] for DSS		
	Frequency[a]	Frequency[a]	HR (95% CI)	*p*	p_{inter}[c]	HR (95% CI)	*p*	p_{inter}[c]
Age (years)								
≤71	369	265	0.80 (0.65-0.78)	0.028		0.78 (0.63-0.97)	0.027	
>71	316	225	0.85 (0.69-1.04)	0.118	0.810	0.87 (0.70-1.08)	0.207	0.985
Sex								
Male	414	281	0.76 (0.63-0.91)	0.003		0.78 (0.65-0.95)	0.013	
Female	271	209	0.90 (0.70-1.14)	0.377	0.144	0.92 (0.71-1.18)	0.498	0.269
Smoking status								
Never	70	44	0.84 (0.47-1.47)	0.535		0.89 (0.50-1.57)	0.688	
Current	247	170	0.77 (0.60-1.00)	0.046		0.81 (0.62-1.07)	0.135	
Former	368	276	0.86 (0.71-1.04)	0.120	0.732	0.84 (0.69-1.03)	0.092	0.963
Histology								
Adeno	352	223	0.75 (0.60-0.94)	0.012		0.73 (0.58-0.92)	0.008	
Squamous	159	125	0.78 (0.58-1.05)	0.103		0.81 (0.59-1.12)	0.209	
Others	174	142	0.96 (0.74-1.24)	0.767	0.007	0.99 (0.77-1.31)	0.991	0.021
Tumor stage								
I-IIIA	397	257	0.87 (0.69-1.09)	0.230		0.92 (0.71-1.18)	0.498	
IIIB-IV	288	233	0.85 (0.70-1.02)	0.078	0.525	0.85 (0.70-1.02)	0.085	0.472
Chemotherapy								
No	384	254	0.78 (0.62-0.96)	0.022		0.78 (0.62-1.00)	0.045	
Yes	301	236	0.66 (0.46-0.95)	0.027	0.530	0.90 (0.74-1.10)	0.302	0.454
Radiotherapy								
No	448	313	0.86 (0.71-1.03)	0.108		0.86 (0.70-1.06)	0.156	
Yes	237	177	0.81 (0.65-1.02)	0.070	0.505	0.84 (0.67-1.06)	0.134	0.755
Surgery								
No	365	270	0.87 (0.73-1.03)	0.099		0.89 (0.74-1.06)	0.172	
Yes	320	220	0.76 (0.58-1.00)	0.052	0.524	0.72 (0.53-0.99)	0.044	0.781

Abbreviations: OS, overall survival; DSS, disease-specific survival; NSCLC, non-small cell lung cancer; PLCO, the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; HR, hazards ratio; CI, confidence interval. [a]10 with missing data were excluded; [b]Adjusted for age, sex, stage, histology, smoking status, chemotherapy, radiotherapy, surgery, PC1, PC2, PC3 and PC4; [c]p_{inter}: *p* value for interaction analysis between characteristics and protective alleles.

Table S7. Function prediction for *SYK* rs11787670 and *ITGA1* rs67715745

SNP	Gene	Chr	Type	Haploreg v4.1[a]						
				Promoter histone marks	Enhancer histone marks	DNAse	Proteinsbound	Motifs changed	Selected eQTLhits	dbSNP func annot
rs11787670	*SYK*	9	imputed	--	--		CTCF	5 altered motifs	--	--
rs67715745	*ITGA1*	5	imputed	--	--	--	--	HDAC2, Zfp105	--	--

Abbreviations: SNP, single nucleotide polymorphism; Chr, chromosome; DNAse, deoxyribonuclease; eQTL, expression quantitative trait loci; dbSNP func annot, dbSNP function annotation; [a]Haploreg: https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php.

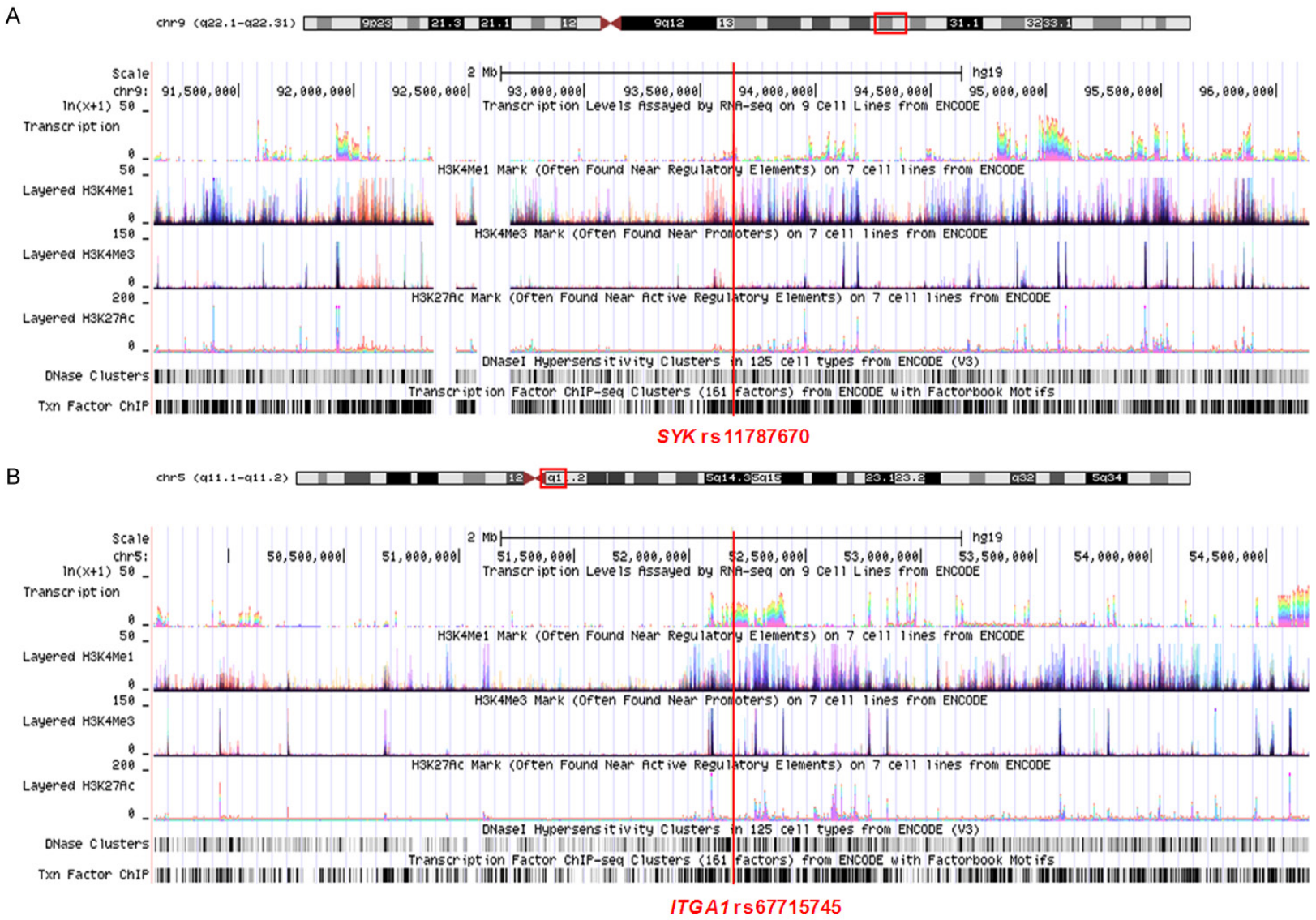

Figure S5. Functional prediction of two independent SNPs in lymphangiogenesis-related pathway genes in the ENCODE data. Location and functional prediction of *SYK* rs11787670 (A). Location and functional prediction of *ITGA1* rs67715745 (B). The H3K4Me3, H3K4Me1, and H3K27Ac tracks showed the genome-wide levels of enrichment of acetylation of lysine 27, the mono-methylation of lysine 4, and tri-methylation of lysine 4 of the H3 histone protein. DNase clusters track showed DNase hypersensitivity areas. Txn factor track showed regions of transcription factor binding of DNA.

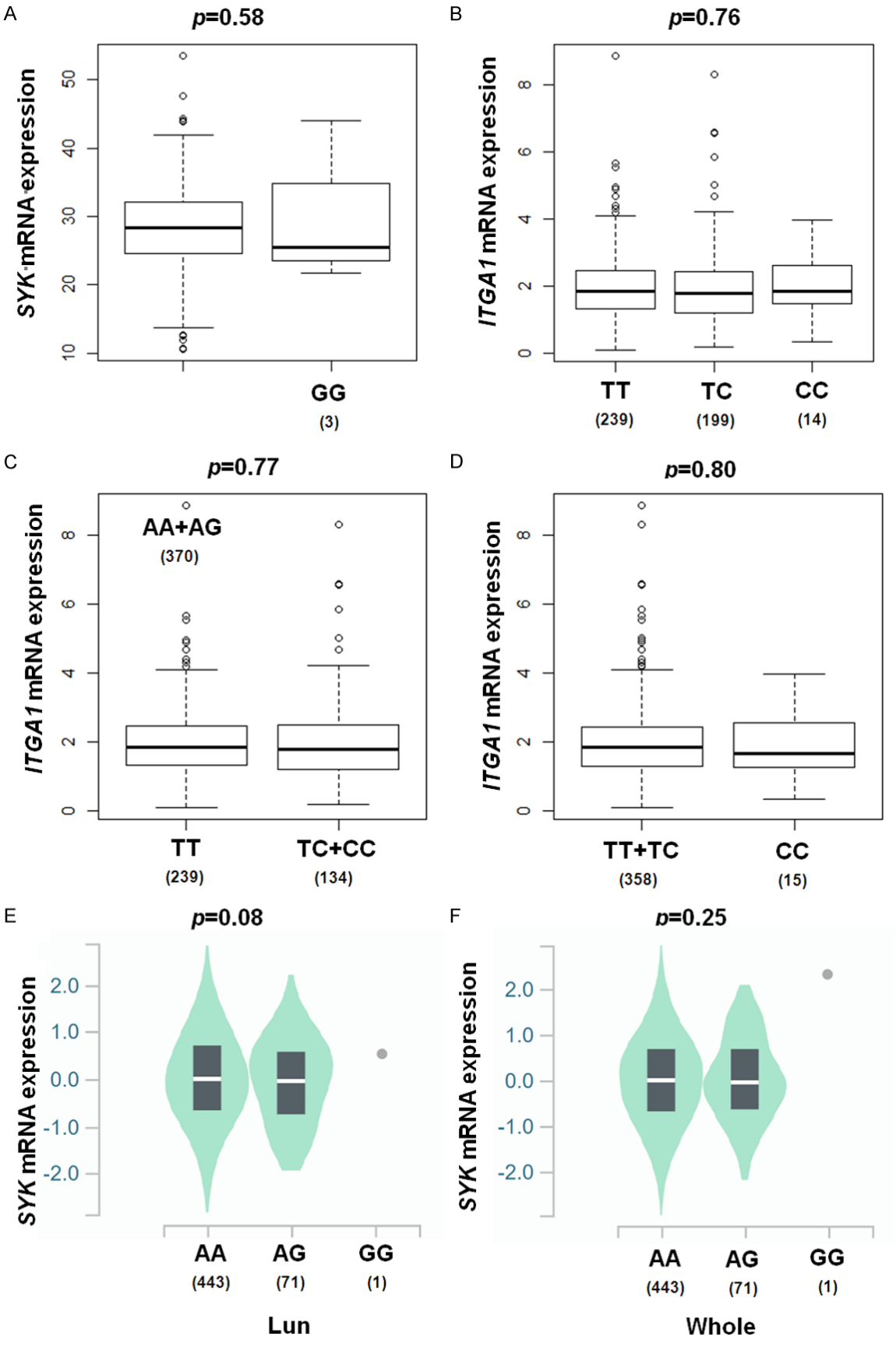
A
p=0.58
SYK mRNA expression
10
20
30
40
50
GG
(3)
B
p=0.76
ITGA1 mRNA expression
0
2
4
6
8
TT
(239)
TC
(199)
CC
(14)
C
p=0.77
ITGA1 mRNA expression
0
2
4
6
8
AA+AG
(370)
TT
(239)
TC+CC
(134)
D
p=0.80
ITGA1 mRNA expression
0
2
4
6
8
TT+TC
(358)
CC
(15)
E
p=0.08
SYK mRNA expression
2.0
1.0
0.0
-1.0
-2.0
AA
(443)
AG
(71)
GG
(1)
Lun
F
p=0.25
SYK mRNA expression
2.0
1.0
0.0
-1.0
-2.0
AA
(443)
AG
(71)
GG
(1)
Whole

Figure S6. The eQTLs analysis for *SYK* rs11787670 and *ITGA1* rs67715745. The correlation of rs11787670 genotypes and *SYK* mRNA expression in the recessive model (A). The correlation of rs67715745 genotypes and *ITGA1* mRNA expression in the additive model (B), the dominant model (C), and the recessive model (D) from the 1,000 Genomes Project. The correlation of rs11787670 genotypes and *SYK* mRNA expression in normal lung tissues (E) and whole blood samples (F) from the GTEx database. Abbreviations: eQTLs, expression quantitative trait loci; GTEx, Genotype-Tissue Expression project.

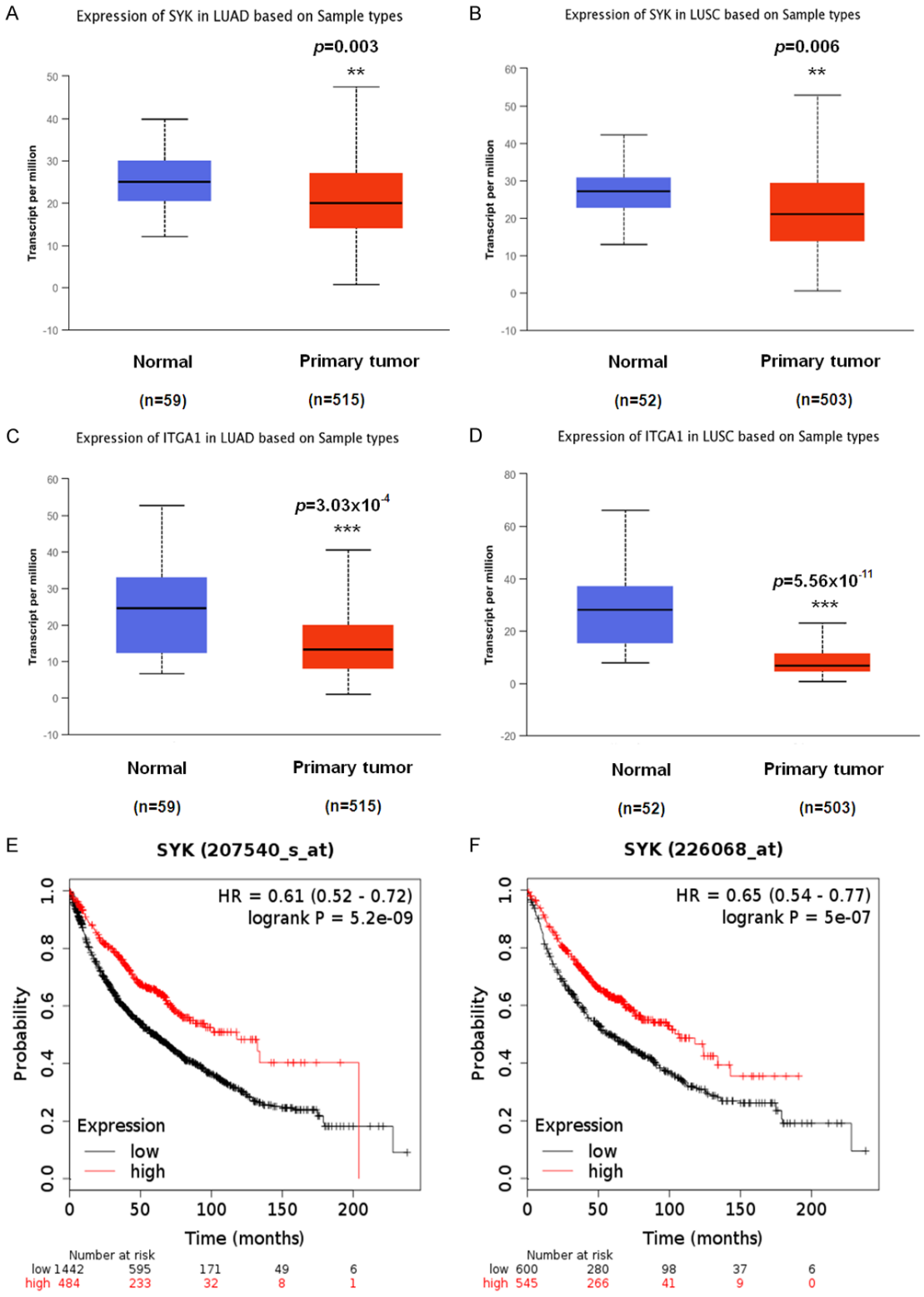

Figure S7. mRNA expression analysis and survival analysis of *SYK* and *ITGA1*. The difference of *SYK* mRNA expression between normal tissues and LUAD tissues in the TCGA database (A); The difference of *SYK* mRNA expression between normal tissues and LUSC tissues in the TCGA database (B); The difference of *ITGA1* mRNA expression between normal tissues and LUAD tissues in the TCGA database (C); The difference of *ITGA1* mRNA expression between normal tissues and LUSC tissues in the TCGA database (D); *SYK* mRNA expression showed significant correlation with lung cancer survival probability (E, F). Abbreviations: LUAD, Lung adenocarcinoma; TCGA, The Cancer Genome Atlas; LUSC, Lung squamous cell carcinoma. The online Kaplan-Meier plotter was obtained from http://kmplot.com/analysis/.

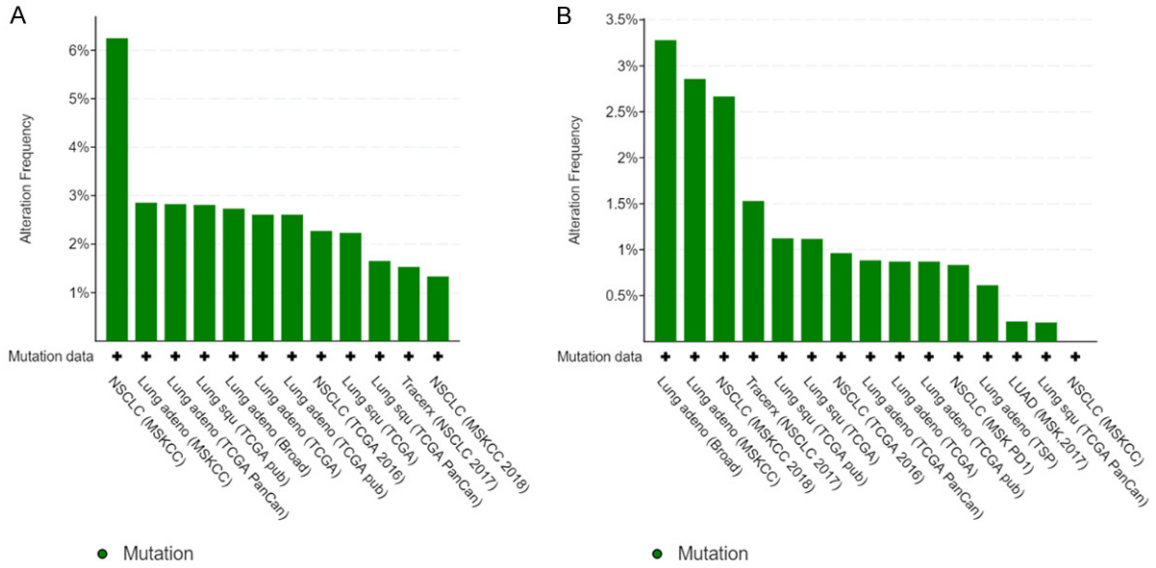

Figure S8. Mutation frequency of *SYK* and *ITGA1* in non-small cell lung tumor tissues. Mutation frequency of *SYK* in NSCLC, LUAD and LUSC using the online database of the cBioPortal for Cancer Genomics (A). Mutation frequency of *ITGA1* in NSCLC, LUAD and LUSC using the online database of the cBioPortal for Cancer Genomics (B). Abbreviations: NSCLC, non-small cell lung cancer; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma. The online cBioPortal for Cancer Genomics database was obtained from http://www.cbioportal.org.