Am J Cancer Res 2021;11(5):2264-2277
www.ajcr.us /ISSN:2156-6976/ajcr0129427

Original Article

Genetic variants of *DOCK2*, *EPHB1* and *VAV2* in the natural killer cell-related pathway are associated with non-small cell lung cancer survival

Hailei Du[1,2,3], Lihua Liu[2,3], Hongliang Liu[2,3], Sheng Luo[4], Edward F Patz Jr[2,5], Carolyn Glass[2,6], Li Su[7], Mulong Du[7], David C Christiani[7,8], Qingyi Wei[2,3,9,10]

[1]Department of Thoracic Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, P. R. China; [2]Duke Cancer Institute, Duke University Medical Center, Durham, NC 27710, USA; [3]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC 27710, USA; [4]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA; [5]Departments of Radiology, Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC 27710, USA; [6]Department of Pathology, Duke University School of Medicine, Durham, NC 27710, USA; [7]Departments of Environmental Health and Epidemiology, Harvard TH Chan 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; [10]Duke Global Health Institute, Duke University, Durham, Durham, NC 27710, USA

Received January 1, 2021; Accepted March 14, 2021; Epub May 15, 2021; Published May 30, 2021

Abstract: Although natural killer (NK) cells are a known major player in anti-tumor immunity, the effect of genetic variation in NK-associated genes on survival in patients with non-small cell lung cancer (NSCLC) remains unknown. Here, in 1,185 with NSCLC cases of a discovery dataset, we evaluated associations of 28,219 single nucleotide polymorphisms (SNPs) in 276 NK-associated genes with their survival. These patients were from the reported genome-wide association study (GWAS) from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. We further validated the findings in an additional 984 cases from the Harvard Lung Cancer Susceptibility (HLCS) Study. We identified three SNPs (i.e., *DOCK2* rs261083 G>C, *VAV2* rs2519996 C>T and *EPHB1* rs36215 *A*>G) to be independently associated with overall survival (OS) in NSCLC cases with adjusted hazards ratios (HRs) of 1.16 (95% confidence interval [CI] = 1.07-1.26, $P = 3.34\times10^{-4}$), 1.28 (1.12-1.47, $P = 4.57\times10^{-4}$) and 0.75 (0.67-0.83, $P = 1.50\times10^{-7}$), respectively. Additional joint assessment of the unfavorable genotypes of the three SNPs showed significant associations with OS and disease-specific survival of NSCLC cases in the PLCO dataset (P_{trend}<0.0001 and <0.0001, respectively). Moreover, the survival-associated *DOCK2* rs261083 C allele had a significant correlation with reduced *DOCK2* transcript levels in lung adenocarcinoma (LUAD), while the rs36215 G allele was significantly correlated with reduced *EPHB1* transcript levels in lymphoblastoid cell lines in the 1000 Genomes Project. These results revealed that *DOCK2* and *EPHB1* genetic variants may be prognostic biomarkers of NSCLC survival, likely via transcription regulation of respective genes.

Keywords: Non-small cell lung cancer, single-nucleotide polymorphism, genetic variant, natural killer cell, survival analysis

Introduction

As the deadliest malignancy worldwide, lung cancer has contributed to a large portion of cancer-related deaths globally and caused about 228,820 new cases and 135,720 deaths in the United States in 2020 [1, 2]. To date, lung cancer remains a heavy burden on public health of the general population. Accounting for about 85% of all lung cancer cases, the heterogeneous non-small cell lung cancer (NSCLC) comprises two broad subtypes: lung squamous cell carcinomas (LUSC) and lung adenocarcinomas (LUAD) [3]. Despite advances made in the treatment approaches, the prognosis of lung cancer patients remains poor, with <20% of individuals surviving up to 5 years after the diagnosis [4]. Moreover, clinical differences in

the patients' response to the same treatment have been observed [5], suggesting that host factors, including genetic factors such as single nucleotide polymorphisms (SNPs), could also play an essential role in outcomes of the patients; therefore, it is paramount to determine important genetic variants as biomarkers for evaluating therapeutic response and outcomes of NSCLC cases.

Because SNPs can affect their gene expression and thus functions, they may have a significant impact on cancer prognosis [6, 7]. However, genome-wide association studies (GWASs) have detected only few novel and functional SNPs that predict survival of NSCLC patients, because GWASs on survival are exploratory and have focused on SNPs or genes with a stringent *P*-values of 5×10^{-8} as a result of multiple testing comparison; meanwhile, the vast majority of the reported SNPs lack of functional annotations [8]. As a promising hypothesis-driven approach in the post-GWAS era, a biological pathway-based method has been used to reanalyze previously reported GWAS datasets and to evaluate the joint impact of SNPs on many genes of the same molecular pathway, which drastically improves the study power and facilitate subsequent functional analysis [9].

Cancer immunotherapy is a promising treatment strategy against cancer, through activating or boosting the immune system by mobilizing anti-cancer immunity, which potentially inhibits some types of cancer. For example, immune checkpoint inhibitors, e.g., programmed death 1 (PD-1) and ligand programmed death-ligand 1 (PD-L1), are applied as the first-line therapy for some advanced or metastatic NSCLC not expressing targetable genetic mutations [10]. In addition, natural killer (NK) cells, as the first line of defense, bridge and orchestrate immune responses, playing a vital anti-cancer role in the host response [11]. Since not all patients benefit from such immunotherapies, a major challenge is to identify individuals with inherited differences in the highly complex interaction involving the tumor and immune cells, including NK cells, for personalized treatment. Hence, we hypothesize that gene variants in the NK cell-related pathway involved in the anti-cancer immune response are associated with survival of NSCLC patients. By using two available published GWAS datasets of NSCLC cases, we tested our hypothesis.

Materials and methods

Study populations

The discovery genotype dataset was from the GWAS of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial with 1,185 eligible Caucasian cases. The PLCO trial was a large randomized cancer screening study performed in ten American medical centers from 1993 to 2011, which was funded by the National Cancer Institute (NCI). A total of 155,000 participants (77,500 men and 77,500 women of 55-74 years old) were included in the screening trial. All the participants were randomized to the intervention arm (with screening) and control (standard care) arms [12]. Genomic DNA extraction from whole blood specimens was performed, and the genotyping utilized Illumina Human Hap240Sv1.0 and Human Hap550v3.0 (dbGaP accession: phs000093.v2.p2 and phs000336.v1.p1) [13, 14]. The PLCO trial was approved by the institutional review boards of all institutions involved, and informed consent was obtained from all participants.

The validation genotype dataset was from another GWAS performed in the Harvard Lung Cancer Susceptibility (HLCS) Study, which included 984 Caucasian NSCLC cases with available whole blood samples for DNA extraction. Genotyping was carried out with Illumina Human hap610-Quad arrays, and MaCH3.0 was utilized for analysis by referring to the sequencing data for Caucasians in the 1000 Genomes Project [15].

Utilization of both GWAS datasets had the approval from the Internal Review Board of Duke University School of Medicine (Project #Pro00054575) and the National Center for Biological Information (NCBI) for access to the de-identified database of genotypes and phenotypes (dbGaP; Project #6404). The features of the PLCO trial (n = 1185) and the HLCS study (n = 984) are shown in Table S1.

Gene selection and SNP imputation

Genes of the NK cell-associated pathway were identified in the Molecular Signatures Database (http://software.broadinstitute.org/gsea/msigdb/index.jsp) using "natural" AND "killer" AND "cell" as keywords. After excluding 369 duplicated and eight unavailable genes in NCBI, 267

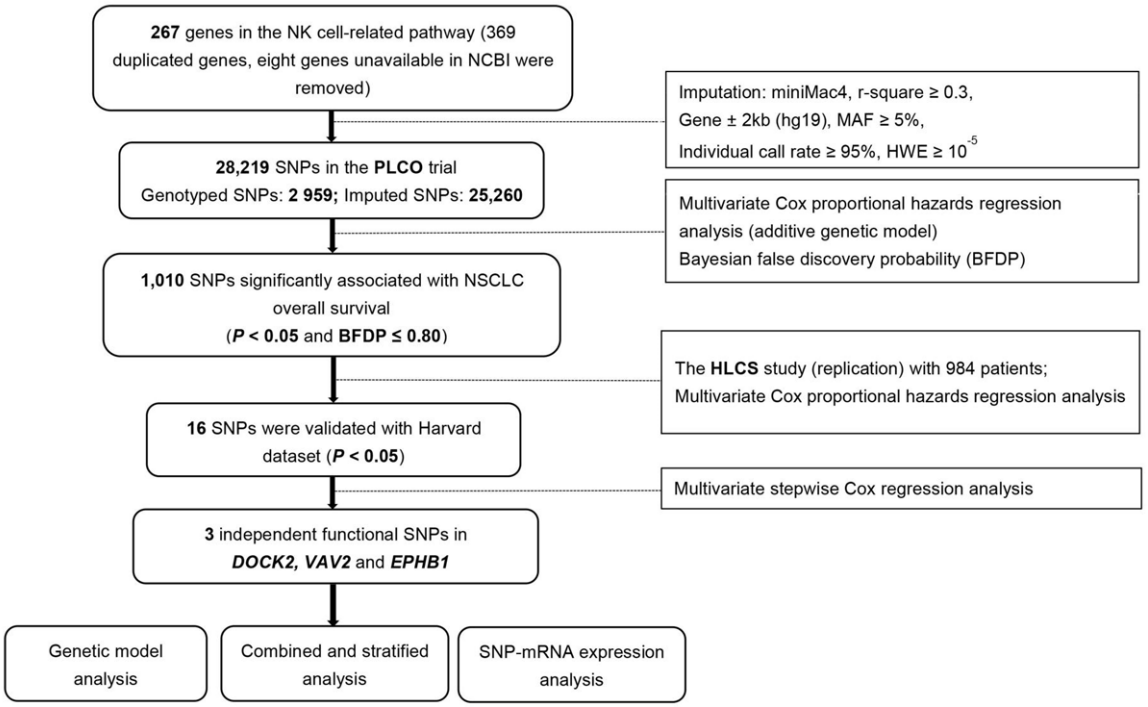


Figure 1. Study flowchart. The overall procedures of the current study. Abbreviations: NK, Natural killer; SNPs, single-nucleotide polymorphisms; MAF: minor allelic frequency; HWE: Hardy-Weinberg Equilibrium; PLCO, The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; HLCS, the Harvard Lung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

genes were considered candidate genes for further analysis (Table S2). Imputation was carried out for the selected candidate genes with Minimac4 and the reference panel of the 1000 Genomes Project data (phase 3). Then, all the SNPs in the candidate genes and alongside the respective ±2 kb flanking regions were extracted and depicted in **Figure 1**. Finally, a total of 28,219 SNPs (2,959 genotyped and 25,260 imputed) from the PLCO trial were selected for further analyses (r-square ≥0.3, individual call rate ≥95%, minor allelic frequency ≥5%, and Hardy-Weinberg equilibrium ≥10^{-5}).

Statistical analyses

The follow-up duration for the participants in both GWAS datasets spanned from NSCLC diagnosis to final follow-up or death. OS of NSCLC cases constituted the primary endpoint. In single-locus analysis, multivariable Cox proportional hazards regression was utilized for evaluating associations of all SNPs in the 267 NK cell-related pathway genes with OS in the additive genetic model after adjusting for age, sex, smoking status, tumor stage, histology, radiotherapy, chemotherapy, surgery, and the major four principal components (Table S3) of the PLCO genotyping dataset using the GenABEL package of R [16]. Because most of the investigational SNPs were in high linkage disequilibrium (LD) as a result of imputation, we then utilized Bayesian false discovery probability (BFDP) with a cut-off of 0.80 for multiple testing correction for reducing false-positives according to previous reports [17, 18]. A prior probability of 0.10 was used for detecting a hazards ratio (HR) of 3.0 for associations with variant genotypes/minor alleles of investigational SNPs at P<0.05. Next, the chosen SNPs in the PLCO discovery dataset were subjected to subsequent validation with the HLCS dataset. To identify independent SNPs, a multivariate stepwise Cox regression model was used, adjusting for patient features and 41 SNPs predicting survival that were previously reported in the PLCO trial dataset. Finally, inverse variance weighted meta-analysis was performed for combining data from the discovery and validation sets using PLINK 1.90 with Cochran's Q statistics and I^2. As both datasets showed no heterogeneity (P_{het}>0.10 and I^2<50%), the fixed-

effects model was used for data assessment. The identified SNPs were also visualized by Manhattan plots and regional association plots.

Then, the combined unfavorable genotypes of the identified SNPs were used to estimate their collective effects on survival probability. The Kaplan-Meier technique was utilized for survival function estimate. The Cochran's Q-test was also carried out for comparing effect differences between subgroups in the stratified analysis. Subsequently, we carried out expression quantitative trait loci (eQTL) analysis for evaluating correlations of SNPs with the corresponding mRNA transcript levels utilizing a linear regression model in the R v3.6.2. The transcriptional expression data originated from three sources, including 373 European patients of the 1,000 Genomes Project [19], whole blood specimens from 369 subjects and non-diseased lung samples from 383 cases in the genotype-tissue expression (GTEx) project [20], and lung cancer tissues specimens from the The Cancer Genome Atlas (TCGA) database (dbGaP Study Accession phs000178.v10.p8) [21]. Next, bioinformatics functional prediction for the validated SNPs was performed with SNPinfo [22]. (https://snpinfo.niehs.nih.gov), RegulomeDB [23] (http://www.regulomedb.org), and HaploReg [24] (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php) were utilized for predicting the functions of these validated SNPs. Differences in mRNA transcript levels between paired tumor specimens and adjacent noncancerous tissue samples from the TCGA database were evaluated by paired Student *t*-test. Moreover, associations of transcript levels with survival propability were assessed by the Kaplan-Meier method (http://kmplot.com/analysis/index.php?p=service&cancer=lung). SAS 9.4 (SAS Institute, USA) was utilized for data analysis, unless specified otherwise.

Results

Associations of SNPs in NK cell-asscociated pathway genes with NSCLC OS in the PLCO and HLCS datasets

The overall flowchart is presented in **Figure 1**. Baseline features for respective 1,185 and 984 NSCLC cases from the PLCO and HLCS study were reported previously [25]. The discovery PLCO dataset encompassed 28,219 SNPs (including 2,959 genotyped; 25,260 imputed SNPs) in the 267 NK cell-related pathway genes available for the initial analysis, of which 1,010 SNPs were statistically significant associated with OS in NSCLC ($P<0.05$) following multiple testing correction by BFDP≤0.8. After subsequent validation in the HLCS validation dataset, 16 SNPs still showed statistical significance.

Associations between independent SNPs and with NSCLC OS in the PLCO dataset

To determine whether the above-mentioned 16 SNPs were independently associated with NSCLC survival, adjustment was made for other covariates and previously reported SNPs in the same dataset, using the same Cox regression model. However, the individual genotyping and clinical data were only available for patients from the PLCO trial but not for those from the HLCS study. Therefore, we first performed stepwise multivariate Cox regression analysis for evaluating the impact of the 16 validated SNPs on OS only in the PLCO dataset. In stepwise Cox regression analysis, nine SNPs were identified with significant and independent associations with NSCLC OS. Subsequently, we further adjusted for other 41 additional SNPs that were previously reported to be associated with OS in the same PLCO dataset. As a result, three SNPs (*DOCK2* rs261083 G>C, *VAV2* rs2519996 C>T and *EPHB1* rs36215 A>G) were still statistically significant in associations with NSCLC OS (P = 0.004, P = 0.019 and $P<0.0001$, respectively, in the additive model) (**Table 1**). Subsequent meta-analysis of the above-mentioned three independent SNPs yielded consistent results ($P = 3.34\times10^{-4}$, 4.57×10^{-4} and 5.00×10^{-4}, respectively) without heterogeneity across these two datasets (**Table 2**).

Specifically, individuals with the *EPHB1* rs36215 G allele had a favorable OS and DSS (P_{trend} = 0.017 and 0.018, respectively, in the trend test), while patients with both *DOCK2* rs261083 C and *VAV2* rs2519996 T alleles showed an elevated risk of death ($P_{trend}<0.007$ and P_{trend} = 0.016 for OS, respectively, and P_{trend} = 0.015 and P_{trend} = 0.048 for DSS, respectively, in the trend test) (**Table 3**). We also visualized the gene locations in the genome for all

Table 1. Three independent SNPs in a multivariate Cox proportional hazards regression analysis with adjustment for other covariates and 41 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	1185	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	0.76 (0.66-0.89)	0.005	0.70 (0.59-0.82)	<0.0001
Smoking status	Never	115	1.00		1.00	
	Current	423	1.69 (1.26-2.27)	0.0004	2.10 (1.55-2.85)	<0.0001
	Former	647	1.67 (1.27-2.20)	0.0003	2.03 (1.52-2.70)	<0.0001
Histology	Adenocarcinoma	577	1.00		1.00	
	Squamous cell	285	1.20 (0.99-1.44)	0.059	1.20 (0.99-1.46)	0.071
	others	323	1.32 (1.12-1.57)	0.001	1.41 (1.18-1.69)	0.0002
Tumor stage	I-IIIA	655	1.00		1.00	
	IIIB-IV	528	2.84 (2.34-3.45)	<0.0001	3.40 (2.78-4.16)	<0.0001
Chemotherapy	No	639	1.00		1.00	
	Yes	538	0.57 (0.48-0.68)	<0.0001	0.54 (0.45-0.65)	<0.0001
Radiotherapy	No	762	1.00		1.00	
	Yes	415	0.93 (0.79-1.10)	0.415	0.98 (0.83-1.16)	0.821
Surgery	No	637	1.00		1.00	
	Yes	540	0.20 (0.15-0.26)	<0.0001	0.18 (0.14-0.23)	<0.0001
DOCK2 rs261083 G>C	GG/GC/CC	694/416/75	1.14 (1.01-1.27)	0.029	1.20 (1.06-1.36)	0.004
VAV2 rs2519996 C>T	CC/CT/TT	1056/124/5	1.27 (1.03-1.55)	0.024	1.31 (1.05-1.63)	0.019
EPHB1 rs36215 A>G	AA/AG/GG	1030/150/5	0.79 (0.64-0.97)	0.024	0.63 (0.50-0.78)	<0.0001

Abbreviations: SNP: single-nucleotide polymorphisms; NSCLC, non-small cell lung cancer; PLCO, the Prostate, Lung, Colorectal and Ovarian cancer screening trial; HR: hazards ratio; CI: confidence interval. [a]Adjusted for age, sex, tumor stage, histology, smoking status, chemotherapy, radiotherapy, surgery, two other identified SNPs, PC1, PC2, PC3 and PC4. [b]Other 41 previously published SNPs from the same GWAS dataset were also 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, rs425904, rs35385129, rs4487030, rs60571065, rs13213007, rs115613985, rs9673682, rs2011404, rs7867814, rs2547235, rs4733124, rs11225211, rs11787670 and rs67715745.

the significant SNPs by the Manhattan (Figure S1) and regional association plots (Figure S2).

Combined and stratified analyses of the three SNPs independently associated with NSCLC OS in the PLCO dataset

To evaluate the collective impact of the above-mentioned three SNPs on NSCLC survival, the respective unfavorable genotypes (i.e., *DOCK2* rs261083 GC+CC, *VAV2* rs2519996 CT+TT, and *EPHB1* rs36215 AA) were included to build a genetic score to categorize the totality of NSCLC patients into four groups based on the number of unfavorable genotypes (NUG). As illustrated in **Table 3**, multivariable Cox analysis showed that an elevated genetic score independently predicted an elevated risk of death (trend test: *P*<0.0001 for both OS and DSS). Next, the totality of patients were dichotomized into the low-risk (0-1 NUG) and high-risk groups (2-3 NUGs). In comparison with low-risk cases, high-risk cases showed a significantly elevated risk of death (OS: HR = 1.35, 95% CI = 1.17-1.56 and *P*<0.0001 and DSS: HR = 1.38, 95% CI = 1.19-1.60 and *P*<0.0001). We also generated KM survival curves to depict the associations of unfavorable genotypes with survival in NSCLC (**Figure 2A-D**).

Then, stratification analysis was carried out to assess whether the impact of unfavorable genotypes assessed in combination on NSCLC survival were altered by sex, age, smoking status, tumor stage, histology, radiotherapy, chemotherapy and surgery in the PLCO dataset. Obvious differences in survival or interactions were not observed among the strata of these covariates in OS or DSS in NSCLC (*P*>0.05, Table S4).

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

SNPs	Allele[a]	Gene	Chr	Position	PLCO (n = 1185)			HLCS (n = 984)			Combined-analysis			
					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
rs261083	G>C	*DOCK2*	5	169278751	0.24	1.17 (1.05-1.31)	0.006	0.21	1.16 (1.02-1.31)	0.023	0.879	0.0	1.16 (1.07-1.26)	3.34×10^{-4}
rs2519996	C>T	*VAV2*	9	136769348	0.06	1.29 (1.05-1.58)	0.016	0.07	1.27 (1.06-1.54)	0.012	0.933	0.0	1.28 (1.12-1.47)	4.57×10^{-4}
rs36215	A>G	*EPHB1*	3	134627850	0.07	0.76 (0.61-0.94)	0.010	0.05	0.75 (0.59-0.94)	0.015	0.879	0.0	0.75 (0.67-0.83)	5.00×10^{-4}

Abbreviations: SNPs, single-nucleotide polymorphisms; Chr: chromosome; 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; EAF, effect allele frequency; HR, hazards ratio; CI, confidence interval; FDR, false discovery rate; BFDP, Bayesian false discovery probability; LD, linkage disequilibrium. [a]Effect/reference 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.

Table 3. Associations between the number of unfavorable genotypes of three independent SNPs with OS and DSS of NSCLC in the PLCO Trial

Alleles	Frequency[a]	OS[b]			DSS[b]		
		Death (%)	HR (95% CI)	*P*	Death (%)	HR (95% CI)	*P*
DOCK2 rs261083 G>C							
GG	691	463 (67.00)	1.00		415 (60.06)	1.00	
GC	412	277 (67.23)	1.28 (1.10-1.49)	0.001	254 (61.65)	1.32 (1.13-1.55)	0.0006
CC	72	49 (68.06)	1.17 (0.87-1.58)	0.298	40 (55.56)	1.07 (0.77-1.48)	0.699
Trend test				0.007			0.015
Dominant							
GG	691	463 (67.00)	1.00		415 (60.06)	1.00	
GC+CC	484	326 (67.36)	1.27 (1.09-1.46)	0.002	294 (60.74)	1.28 (1.10-1.49)	0.0016
VAV2 rs2519996 C>T							
CC	1049	699 (66.63)	1.00		628 (59.87)	1.00	
CT	121	86 (71.07)	1.32 (1.05-1.65)	0.016	78 (64.46)	1.30 (1.03-1.65)	0.031
TT	5	4 (80.00)	1.33 (0.49-3.59)	0.575	3 (60.00)	1.02 (0.33-3.21)	0.971
Trend test				0.016			0.048
Dominant							
CC	1049	699 (66.63)	1.00		628 (59.87)	1.00	
CT+TT	126	90 (71.43)	1.32 (1.06-1.65)	0.014	81 (64.29)	1.29 (1.02-1.63)	0.034
EPHB1 rs36215 A>G							
AA	1020	700 (68.63)	1.00		629 (61.67)	1.00	
AG	150	84 (56.00)	0.74 (0.59-0.93)	0.010	76 (50.67)	0.74 (0.58-0.94)	0.013
GG	5	5 (100.00)	0.97 (0.39-2.38)	0.943	4 (80.00)	0.87 (0.32-2.35)	0.776
Trend test				0.017			0.018
Dominant							
AA	1020	700 (68.63)	1.00		629 (61.67)	1.00	
AG+GG	155	89 (57.42)	0.75 (0.60-0.937)	0.011	80 (51.61)	0.74 (0.59-0.94)	0.013
Change reference genotypes							
AG+GG	155	89 (57.42)	1.00		80 (51.61)	1.00	
AA	1020	700 (68.63)	1.33 (1.07-1.67)	0.011	629 (61.67)	1.35 (1.06-1.70)	0.013
NUG[c]							
0	85	48 (56.47)	1.00		44 (51.76)	1.00	
1	591	395 (66.84)	1.39 (1.03-1.88)	0.032	350 (59.22)	1.36 (0.99-1.87)	0.055
2	458	317 (69.21)	1.78 (1.31-2.41)	0.0002	291 (63.54)	1.80 (1.31-2.48)	0.0003
3	41	29 (70.73)	2.10 (1.32-3.35)	0.0018	24 (58.54)	1.87 (1.13-3.08)	0.015
Trend test				<0.0001			<0.0001
Dichotomized NUG							
0-1	676	443 (65.53)	1.00		394 (58.28)	1.00	
2-3	499	346 (69.34)	1.35 (1.17-1.56)	<0.0001	315 (63.13)	1.38 (1.19-1.60)	<0.0001

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. NUG: number of unfavorable genotypes. [a]10 missing date were excluded. [b]Adjusted for age, sex, smoking status, histology, tumor stage, chemotherapy, surgery, radiotherapy and principal components. [c]Unfavorable genotypes were *DOCK2* rs261083 GC+CC, *VAV2* rs2519996 CT+TT and *EPHB1* rs36215 AA.

Expression quantitative trait loci (eQTL) analysis

To investigate possible functions of these three SNPs, eQTL was performed to explore correlations of SNP genotypes with the respective mRNA transcript levels. In the TCGA dataset, the *DOCK2* rs261083 C allele was significantly correlated with a reduced mRNA transcript levels in LUAD tissue specimens in both additive ($P_{additive}$ = 0.025, **Figure 3A**) and dominant ($P_{dominant}$ = 0.038, **Figure 3B**) models, but not observed in the 1000 Genomes Project database and LUSC from the TCGA dataset (Figure S3A-F). The *EPHB1* rs36215 G allele was significantly correlated with decreased mRNA

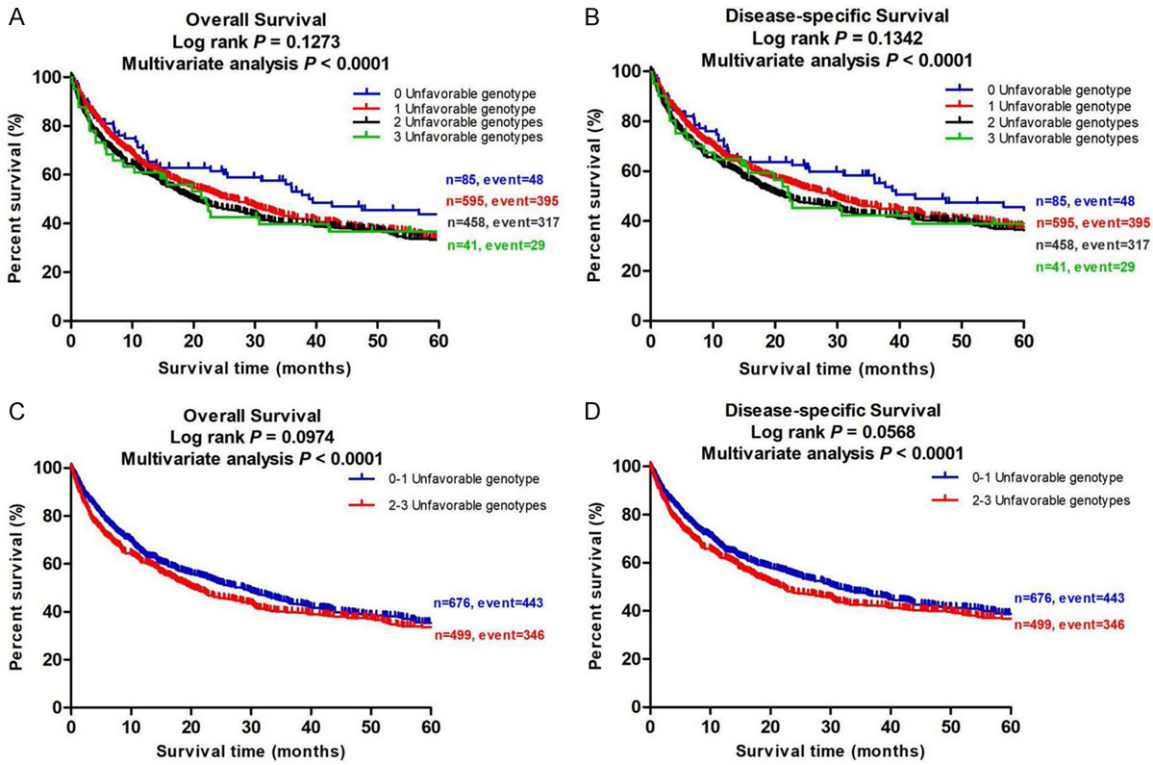


Figure 2. Kaplan-Meier (KM) survival curves for NSCLC patients based on the combined unfavorable genotypes of the three replicated SNPs in the PLCO trial. A. Based on 0, 1, 2, and 3 unfavorable genotypes in OS. B. Based on 0, 1, 2, and 3 unfavorable genotypes in DSS. C. Dichotomized groups of the unfavorable genotypes divided into 0-1 and 2-3 in OS from the PLCO trial. D. Dichotomized groups of the unfavorable genotypes divided into 0-1 and 2-3 in DSS from the PLCO trial. Abbreviations: OS, overall survival; DSS, disease-specific survival; PLCO, The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial.

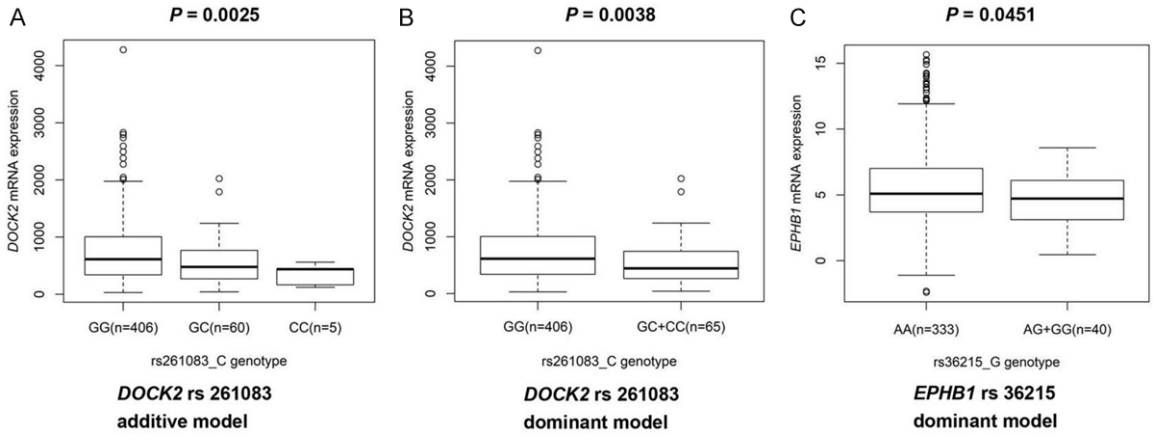


Figure 3. Associations of significant SNP genotypes with their respective mRNA amounts. A. rs261083 additive model in LUAD from the TCGA dataset. B. rs261083 dominant model in LUAD from the TCGA dataset. C. rs36215 dominant model from the 1,000 Genomes Project dataset. Abbreviations: NSCLC, non-small cell lung cancer; PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial.

transcript levels of the gene in the dominant model (P = 0.045, **Figure 3C**) in the 1000 Gemomes Project database but not in another two models (Figure S3G, S3H) and no correlation between the *EPHB1* rs36215 G allele and mRNA amounts was observed in genetic models in LUAD (Figure S3I) and LUSC (Figure S3J-L) from the TCGA dataset. And there was no

significant correlation between the *VAV2* rs2519996 T allele and mRNA transcript levels in all three genetic models in the 1000 Genomes Project (Figure S3M-O) and the TCGA dataset (Figure S3P-S). Additionally, eQTL analysis of the GTEx Project dataset was carried out, and these genotypes had no significant correlations with the respective mRNA transcript levels in either normal lung tissue specimens (n = 515) or whole blood samples (n = 515) (Figure S4).

Differential mRNA expression analysis in target tissues

In comparison with paired adjacent noncancerous tissues, *DOCK2* expression in tumor tissues samples was markedly lower in LUAD, LUSC and combined LUAD+LUSC samples (all $P<0.001$; **Figure 4A**), and elevated *DOCK2* transcript levels were also associated with reduced risk of death as depicted by a KM survival curve of lung cancer found in the online data (Figure S5A) (http://kmplot.com/analysis/index.php?p=service&cancer=lung). In contrast, *EPHB1* transcript levels were higher in LUAD, LUSC and combined LUAD+LUSC samples ($P = 0.001$, <0.001 and <0.001, respectively) in comparison with paired adjacent normal lung tissue specimens (**Figure 4B**). Meanwhile, these elevated transcript levels had no association with elevated risk of death (Figure S5B). Likewise, *VAV2* gene expression was remarkably elevated in LUAD and combined LUAD+LUSC samples ($P = 0.0013$ and 0.001, respectively) but not in LUSC ($P = 0.1411$; **Figure 4C**); meanwhile, elevated *VAV2* mRNA transcript levels had no association with a high risk of death (Figure S5C).

Bioinformatics analyses for functional prediction

In silico prediction of the functions of the abovementioned three significant SNPs was carried out with online bioinformatics tools (i.e., SNPinfo, RegulomeDB, and HaploReg). We found that a *DOCK2* rs261083 G>A change might alter protein motifs, while *VAV2* rs2519996 C>T and *EPHB1* rs36215 A>G changes might potentially affect enhancer histone marks (Table S5). Additionally, according to experimental data in the Encyclopedia of DNA Elements (ENCODE) project, no potential functions for these independent SNPs were predicted (Figure S6).

Discussion

In the current study, we demonstrated that *DOCK2* rs261083 G>C, *VAV2* rs2519996 C>T, and *EPHB1* rs36215 A>G in the NK cell-related gene-set significantly predicted OS and DSS in Caucasian NSCLC patients. Additionally, the *DOCK2* rs261083 C allele affected *DOCK2* mRNA transcript levels in LUAD specimens in the TCGA dataset, while the *EPHB1* rs36215 G allele was significantly correlated with mRNA transcript levels in lymphoblastoid cells from 373 European individuals in the 1000 Genomes Project.

Considering gene expression data in the TCGA database, *DOCK2* might represent a suppressor gene, because reduced *DOCK2* mRNA transcript levels were correlated with an elevated risk of death in patients with both LUAD and LUSC. However, the rs261083 C allele appeared to affect *DOCK2* mRNA expression levels in LUAD but not LUSC. This might result from the small number of tumor specimens available for analysis; alternatively, there might be distinct molecular mechanisms and genetic heterogeneity between these subtypes of NSCLC [26, 27]. On the other hand, both *EPHB1* and *VAV2* seemed to have some oncogenic features, because higher mRNA transcript levels were associated with an elevated risks of death in NSCLC cases. Moreover, the rs36215 G allele predicted a reduced risk of death and low *EPHB1* mRNA transcript levels in normal lymphoblastic cell lines, although the rs2519996 T allele did not predict *VAV2* gene expression. Jointly, these findings suggest that the observed associations of genetic variants in the NK cell-associated genes with NSCLC survival are biologically plausible.

DOCK2, also called dedicator of cytokinesis2 and located on chromosome 5, is critical for lymphocyte migration and regulates T cell responsiveness [28]. Functionally, DOCK2 regulates not only the differentiation of NKs, plasmacytoid dendritic cells and T helper 2 cells [29] but also cell motility, polarity, adhesion, proliferation and apoptosis by activating Rac [30]. *DOCK2* is also likely involved in the carcinogenesis. For instance, a report showed *DOCK2* overexpression is associated with good

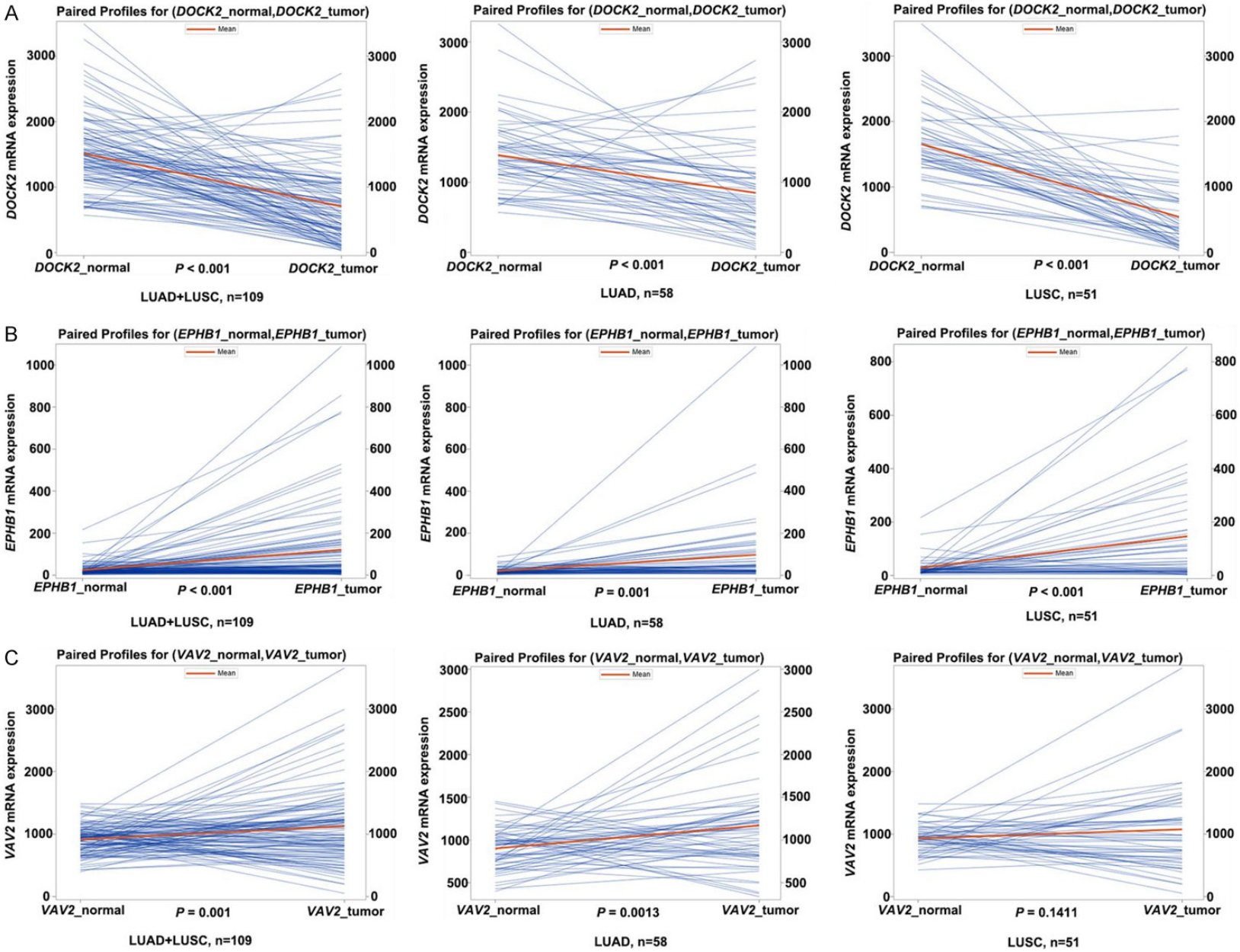

A
Paired Profiles for (DOCK2_normal,DOCK2_tumor)
Mean
DOCK2 mRNA expression
DOCK2_normal
P < 0.001
DOCK2_tumor
LUAD+LUSC, n=109
Paired Profiles for (DOCK2_normal,DOCK2_tumor)
Mean
DOCK2 mRNA expression
DOCK2_normal
P < 0.001
DOCK2_tumor
LUAD, n=58
Paired Profiles for (DOCK2_normal,DOCK2_tumor)
Mean
DOCK2 mRNA expression
DOCK2_normal
P < 0.001
DOCK2_tumor
LUSC, n=51
B
Paired Profiles for (EPHB1_normal,EPHB1_tumor)
Mean
EPHB1 mRNA expression
EPHB1_normal
P < 0.001
EPHB1_tumor
LUAD+LUSC, n=109
Paired Profiles for (EPHB1_normal,EPHB1_tumor)
Mean
EPHB1 mRNA expression
EPHB1_normal
P = 0.001
EPHB1_tumor
LUAD, n=58
Paired Profiles for (EPHB1_normal,EPHB1_tumor)
Mean
EPHB1 mRNA expression
EPHB1_normal
P < 0.001
EPHB1_tumor
LUSC, n=51
C
Paired Profiles for (VAV2_normal,VAV2_tumor)
Mean
VAV2 mRNA expression
VAV2_normal
P = 0.001
VAV2_tumor
LUAD+LUSC, n=109
Paired Profiles for (VAV2_normal,VAV2_tumor)
Mean
VAV2 mRNA expression
VAV2_normal
P = 0.0013
VAV2_tumor
LUAD, n=58
Paired Profiles for (VAV2_normal,VAV2_tumor)
Mean
VAV2 mRNA expression
VAV2_normal
P = 0.1411
VAV2_tumor
LUSC, n=51

Figure 4. *DOCK2, EPHB1* and *VAV2* mRNA expression in lung cancer tissue and adjacent noncancerous lung tissues specimens in the TCGA dataset. A. Elevated DOCK2 expression was detected in noncancerous tissue samples versus both LUAD tissues and LUSC tissue specimens. B. Elevated EPHB1 expression was found in both LUAD and LUSC tumor tissue specimens in comparison with the noncancerous tissue samples. C. Elevated expression of VAV2 was found in the LUAD tissue samples in comparison with the noncancerous tissue samples. Abbreviations: NSCLC, non-small cell lung cancer; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

acute myeloid leukemia (AML) prognosis [31]. Additionally, *DOCK2* overexpression was reported to be correlated with a favorable prognosis in colorectal cancer with overtly enhanced CD8+ lymphocytes infiltration [32]. The above findings suggest that *DOCK2* might be a potential tumor suppressor, because *DOCK2* was suppressed in tumor tissues in comparison with benign tissues and that elevated *DOCK2* gene expression predicted good prognosis in LUAD cases.

EPHB1 (Eph receptor B1), an Eph family member, modulates not only adult angiogenesis but also embryonic vascular system development [33]. A previous study revealed that high *EPHB1* transcript levels in NSCLC tissue specimens in comparison with adjacent noncancerous normal lung tissue specimens; Furthermore, *EPHB1* overexpression induces migration and invasion and is also associated with reduced survival of lung cancer patient [34]. These findings indicate that *EPHB1* may have an oncogenic effect in line with the above finding that the *EPHB1* transcript levels were elevated in lung cancer tissue specimens in comparison with paired noncancerous lung tissue specimens from the same patients. However, no study has reported associations of genetic variants of *EPHB1* with NSCLC survival. As shown above, the *EPHB1* rs36215 variant G allele predicted a reduced risk of death in NSCLC cases, compared with individuals with the wildtype allele, possibly because the *EPHB1* rs36215 G allele decreased *EPHB1* gene expression. Since OS was comparable between the higher and lower expression of *EPHB1* groups of NSCLC cases, additional molecular mechanisms might control *EPHB1* expression in cancer, which needs further investigation.

VAV2, (vav guanine nucleotide exchange factor 2), located on chromosome 9, is one of the essential regulators of immune function [35]. The newly described *VAV2* has abnormal expression in several malignancies [36-38], and the VAV2-Rac1 pathway is involved in cancer development [39]. Reports assessing *VAV2* in NSCLC are scarce. A report suggested *VAV2* modulates vimentin-associated FAK activation and controls lung cancer cell adhesion *in vitro* [40]; Another study showed that miR-331-3p suppresses epithelial-to-mesenchymal transition (EMT), migration and metastatic potential by interacting with ErbB2 and VAV2 via the Rac1/PAK1/β-catenin signaling in NSCLC [41]. Furthermore, host Vav2 deficiency decreases microvascular density and tumor growth and/or survival, in Lewis lung carcinoma models [42]. These findings suggest that *VAV2* might have an oncogenic function in the tumor microenvironment in the NSCLC biology, corroborating the above-mentioned finding of *VAV2* overexpression in lung cancer specimens in comparison with paired noncancerous tissue samples form the same patients. However, no reports have assessed the roles of *VAV2* genetic variants in NSCLC prognosis. As shown above, the *VAV2* rs2519996 variant T allele was not found to affect *VAV2* mRNA transcript levels, although it predicted an elevated risk of death in NSCLC, in comparison with the wildtype. Furthermore, *VAV2* mRNA amounts were markedly increased in lung cancer tissue specimens compared with noncancerous counterparts, although VAV2 upregulation was not correlated with reduced survival. Thus, the molecular mechanisms of *VAV2* in predicting survival of NSCLC patients need to be further investigated.

The limitations of the current study should be mentioned. First, only Caucasian patients were included in the two available GWAS datasets, which may have reduced the generalizability of the findings to other ethnic groups. Secondly, the exact molecular mechanisms underpinning the above-mentioned associations of these independent SNPs with NSCLC survival remain unknown and should be further explored. Thirdly, no detailed information about the therapies administered to the patients was available in the PLCO trial, which makes it impossible to be adjusted, such as immunotherapies, in the analysis. Finally, despite the relatively large sample size of the PLCO trial, the numbers of patients in various subgroups remain

relatively small, likely reducing the statistical power in the stratified or subgroup analyses.

Overall, we demonstrated that three novel independent functional SNPs (*DOCK2* rs261083 G>C, *VAV2* rs2519996 C>T, and *EPHB1* rs36215 A>G) were associated with survival in NSCLC cases in both the PLCO trial and the HLCS study. We also showed that *DOCK2* rs261083 C likely has an effect on survival in NSCLC cases, possibly by altering the targeted gene expression. Our findings indicate that these three SNPs might represent new prognostic biomarkers of NSCLC survival, once further validated in additional lung cancer patient populations and upon further mechanistic assessment.

Acknowledgements

The current study was funded by the National Institute of Health (CA090578, CA074386, CA092824 and U01CA209414); 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 authors are grateful to all the participants of the PLCO Cancer Screening Trial; and the National Cancer Institute (NCI) for providing access to PLCO trial data. The statements included in this report are solely those of the authors and do not represent or imply concurrence or endorsement by the NCI. The authors also acknowledge the dbGaP repository for providing cancer genotyping datasets. The accession numbers for lung cancer datasets are phs000336.v1.p1 and phs000093.v2.p2. A list of contributing investigators and funding agencies for those trials is included in the Supplemental Data.

Disclosure of conflict of interest

None.

Abbreviations

SNP, single nucleotide polymorphism; NK, natural killer; 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; eQTL, expression quantitative trait loci; TCGA, The Cancer Genome Atlas; HR, hazards ratio; CI, confidence interval; GTEx, genotype-tissue expression project; NUG, number of unfavorable genotypes; DOCK2, Dedicator of cytokinesis 2; EPHB1, Erythropoietin-producing hepatocellular carcinoma (Eph) Receptor B1; VAV2, vav guanine nucleotide exchange factor 2.

Address correspondence to: Qingyi Wei, Duke Cancer Institute, Duke University Medical Center, Durham, NC 27710, USA; 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] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68: 394-424.

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

[3] Molina JR, Yang P, Cassivi SD, Schild SE and Adjei AA. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc 2008; 83: 584-594.

[4] Mao Y, Yang D, He J and Krasna MJ. Epidemiology of lung cancer. Surg Oncol Clin N Am 2016; 25: 439445.

[5] Zhao J, Lin G, Zhuo M, Fan Z, Miao L, Chen L, Zeng A, Yin R, Ou Y, Shi Z, Yin J, Gao W, Chen J, Zhou X, Zeng Y, Liu X, Xu H, Chen R, Xia X and Carbone DP. Next-generation sequencing based mutation profiling reveals heterogeneity of clinical response and resistance to osimertinib. Lung Cancer 2020; 141: 114-118.

[6] Mullany LE, Herrick JS, Wolff RK and Slattery ML. Single nucleotide polymorphisms within MicroRNAs, MicroRNA targets, and MicroRNA biogenesis genes and their impact on colorectal cancer survival. Genes Chromosomes Cancer 2017; 56: 285-295.

[7] Zhao D, Wu YH, Zhao TC, Jia ZF, Cao DH, Yang N, Wang YQ, Cao XY and Jiang J. Single-nucleotide polymorphisms in Toll-like receptor genes are associated with the prognosis of gastric cancer and are not associated with Helicobacter pylori infection. Infect Genet Evol 2019; 73: 384-389.

[8] Fridley BL and Biernacka JM. Gene set analysis of SNP data: benefits, challenges, and future directions. Eur J Hum Genet 2011; 19: 837-43.

[9] Gallagher MD and Chen-Plotkin AS. The post-GWAS era: from association to function. Am J Hum Genet 2018; 102: 717-30.
[10] Peters S, Reck M, Smit EF, Mok T and Hellmann MD. How to make the best use of immunotherapy as first-line treatment of advanced/metastatic non-small-cell lung cancer. Ann Oncol 2019; 30: 884-896.
[11] Hodgins JJ, Khan ST, Park MM, Auer RC and Ardolino M. Killers 2.0: NK cell therapies at the forefront of cancer control. J Clin Inves 2019; 129: 3499-3510.
[12] Weissfeld JL, Schoen RE, Pinsky PF, Bresalier RS, Doria-Rose VP, Laiyemo AO, Church T, Yokochi LA, Yurgalevitch S, Rathmell J, Andriole GL, Buys S, Crawford ED, Fouad M, Isaacs C, Lamerato L, Reding D, Prorok PC and Berg CD; PLCO Project Team. Flexible sigmoidoscopy in the randomized prostate, lung, colorectal, and ovarian (PLCO) cancer screening trial: added yield from a second screening examination. J Natl Cancer Inst 2012; 104: 280-9.
[13] 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.
[14] 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.
[15] Zhai R, Yu X, Wei Y, Su L and Christiani DC. Smoking and smoking cessation in relation to the development of co-existing non-small cell lung cancer with chronic obstructive pulmonary disease. Int J Cancer 2014; 134: 961-970.
[16] Aulchenko YS, Ripke S, Isaacs A and van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics 2007; 23: 1294-1296.
[17] Park JH, Geum DI, Eisenhut M, van der Vliet HJ and Shin JI. Bayesian statistical methods in genetic association studies: empirical examination of statistically non-significant Genome Wide Association Study (GWAS) meta-analyses in cancers: a systematic review. Gene 2019; 685: 170-178.
[18] Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 2007; 81: 208-227.
[19] Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf 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 Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X and Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013; 501: 506-151.
[20] Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015; 348: 648-60.
[21] Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014; 511: 543-550.
[22] Xu Z and Taylor JA. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Res 2009; 37: W600-605.
[23] Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM and Snyder M. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 2012; 22: 1790-1797.
[24] 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.
[25] 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.
[26] Relli V, Trerotola M, Guerra E and Alberti S. Abandoning the notion of non-small cell lung cancer. Trends Mol Med 2019; 25: 585-594.
[27] McGranahan N and Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell 2017; 168: 613-628.
[28] Kunisaki Y, Tanaka Y, Sanui T, Inayoshi A, Noda M, Nakayama T, Harada M, Taniguchi M, Sasazuki T and Fukui Y. DOCK2 is required in T cell precursors for development of Valpha14 NK T cells. J Immunol 2006; 176: 4640-4645.
[29] Dobbs K, Domínguez Conde C, Zhang SY, Parolini S, Audry M, Chou J, Haapaniemi E, Keles S, Bilic I, Okada S, Massaad MJ, Rounioja S, Alwahadneh AM, Serwas NK, Capuder K, Çiftçi E, Felgentreff K, Ohsumi TK, Pedergnana V, Boisson B, Haskoloğlu Ş, Ensari A, Schuster M,

Moretta A, Itan Y, Patrizi O, Rozenberg F, Lebon P, Saarela J, Knip M, Petrovski S, Goldstein DB, Parrott RE, Savas B, Schambach A, Tabellini G, Bock C, Chatila TA, Comeau AM, Geha RS, Abel L, Buckley RH, İkincioğulları A, Al-Herz W, Helminen M, Doğu F, Casanova JL, Boztuğ K and Notarangelo LD. Inherited DOCK2 deficiency in patients with early-onset invasive infections. N Engl J Med 2015; 372: 2409-2422.

[30] Guo X and Chen SY. Dedicator of cytokinesis 2 in cell signaling regulation and disease development. J Cell Physiol 2017; 232: 1931-1940.

[31] Hu N, Pang Y, Zhao H, Si C, Ding H, Chen L, Wang C, Qin T, Li Q, Han Y, Dai Y, Zhang Y, Shi J, Wu D, Zhang X, Cheng Z and Fu L. High expression of DOCK2 indicates good prognosis in acute myeloid leukemia. J Cancer 2019; 10: 6088-6094.

[32] Miao S, Zhang RY, Wang W, Wang HB, Meng LL, Zu LD and Fu GH. Overexpression of dedicator of cytokinesis 2 correlates with good prognosis in colorectal cancer associated with more prominent $CD8^+$ lymphocytes infiltration: a colorectal cancer analysis. J Cell Biochem 2018; 119: 8962-8970.

[33] Wei W, Wang H and Ji S. Paradoxes of the EphB1 receptor in malignant brain tumors. Cancer Cell Int 2017; 17: 21.

[34] Wang L, Peng Q, Sai B, Zheng L, Xu J, Yin N, Feng X and Xiang J. Ligand-independent EphB1 signaling mediates TGF-β-activated CDH2 and promotes lung cancer cell invasion and migration. J Cancer 2020; 11: 4123-4131.

[35] Rodriguez-Fdez S and Bustelo XR. The Vav GEF family: an evolutionary and functional perspective. Cells 2019; 8: 465.

[36] Citterio C, Menacho-Márquez M, García-Escudero R, Larive RM, Barreiro O, Sánchez-Madrid F, Paramio JM and Bustelo XR. The rho exchange factors Vav2 and Vav3 control a lung metastasis-specific transcriptional program in breast cancer cells. Sci Signal 2012; 5: ra71.

[37] Jiang Y, Prabakaran I, Wan F, Mitra N, Furstenau DK, Hung RK, Cao S, Zhang PJ, Fraker DL and Guvakova MA. Vav2 protein overexpression marks and may predict the aggressive subtype of ductal carcinoma in situ. Biomark Res 2014; 2: 22.

[38] Wang R, Zhao N, Li S, Fang JH, Chen MX, Yang J, Jia WH, Yuan Y and Zhuang SM. MicroRNA-195 suppresses angiogenesis and metastasis of hepatocellular carcinoma by inhibiting the expression of VEGF, VAV2, and CDC42. Hepatology 2013; 58: 642-653.

[39] Wang P, Liu GZ, Wang JF and Du YY. SNHG3 silencing suppresses the malignant development of triple-negative breast cancer cells by regulating miRNA-326/integrin α5 axis and inactivating Vav2/Rac1 signaling pathway. Eur Rev Med Pharmacol Sci 2020; 24: 5481-5492.

[40] Havel LS, Kline ER, Salgueiro AM and Marcus AI. Vimentin regulates lung cancer cell adhesion through a VAV2-Rac1 pathway to control focal adhesion kinase activity. Oncogene 2015; 34: 1979-1990.

[41] Li X, Zhu J, Liu Y, Duan C, Chang R and Zhang C. MicroRNA-331-3p inhibits epithelial-mesenchymal transition by targeting ErbB2 and VAV2 through the Rac1/PAK1/β-catenin axis in non-small-cell lung cancer. Cancer Sci 2019; 110: 1883-1896.

[42] Brantley-Sieders DM, Zhuang G, Vaught D, Freeman T, Hwang Y, Hicks D and Chen J. Host deficiency in Vav2/3 guanine nucleotide exchange factors impairs tumor growth, survival, and angiogenesis in vivo. Mol Cancer Res 2009; 7: 615-623.

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

Characteristics	PLCO		HLCS		*P**
	Frequency	Deaths (%)	Frequency	Deaths (%)	
Total	1185	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 Suceptibility Study. *Chi-square test for the comparison of the characteristics between the PLCO trial and Harvard study for each clinical variable.

Table S2. List of 267 selected genes in the Natural killer cell related gene-set used in the discovery analysis

Dataset	Name of pathway	Selected genes[a]	Number of genes
BIOCARTA	BIOCARTA_INFLAM_PATHWAY	*CD4, CSF1, CSF2, CSF3, CXCL8, HLA-DRA, HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, IFNA1, IFNB1, IFNG, IL10, IL11, IL13, IL15, IL1A, IL2,IL3, IL4, IL5, IL6, IL7, PDGFA, TGFB1, TGFB2, TGFB3, TNF*	29
BIOCARTA	BIOCARTA_NO2IL12_PATHWAY	*CCR5, CD2, CD247, CD3D, CD3E, CD3G, CD4, CXCR3, IFNG, IL12RB1, IL12RB2, JAK2, LILRB1, NOS2, STAT4, TYK2*	16
GO	GO_IMMUNOLOGICAL_SYNAPSE_FORMATION	*CCL19, CCL21, CCR7, CD6, CD81, DLG1, DOCK2, DOCK8, EPHB1, HAVCR2, LGALS3, MSN, NCK2, PRF1*	14
GO	GO_NATURAL_KILLER_CELL_ACTIVATION	*AP1G1, AXL, BAG6, BLOC1S3, CASP8, CD2, CD244, CLNK, CORO1A, ELF4, FGR, FLT3LG, GAS6, HAVCR2, HLA-E, HLA-F, HNF1A, ID2, IFNA1, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA2, IFNA21, IFNA4, IFNA5, IFNA6, IFNA7, IFNA8, IFNB1, IFNE, IFNK, IFNW1, IL12A, IL12B, IL15, IL18, IL18R1, IL2, IL21R, IL23A, IL23R, ITGB2, KIR3DS1, KLRC4-KLRK1, KLRF2, KLRK1, LAMP1, LEP, MERTK, MICA, NCR1, NCR3, PGLYRP1, PGLYRP2, PGLYRP3, PGLYRP4, PIBF1, PIK3CD, PRDM1, PRDX1, PTPN22, PTPRC, RAB27A, RASGRP1, RHBDD3, SLAMF7, SNX27, SP3, STAT5B, TICAM1, TOX, TUSC2, TYRO3, TYROBP, ULBP1, ULBP2, ULBP3, UNC13D, VAMP7, ZBTB1, ZNF683*	84
GO	GO_NATURAL_KILLER_CELL_ACTIVATION_INVOLVED_IN_IMMUNE_RESPONSE	*AP1G1, CD244, CORO1AM, HLA-F, IFNA1, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA2, IFNA21, IFNA4, IFNA5, IFNA6, IFNA7, IFNA8, IFNB1, IFNE, IFNK, IFNW1, IL12B, KLRF2, LAMP1, PGLYRP1, PGLYRP2, PGLYRP3, PGLYRP4, RAB27A, UNC13D, VAMP7, ZNF683*	32
GO	GO_NATURAL_KILLER_CELL_CHEMOTAXIS	*CCL2, CCL3, CCL4, CCL5, CCL7, CXCL14, KLRC4-KLRK1, KLRK1, PIK3CD, PIK3CG, XCL1*	11
GO	GO_NATURAL_KILLER_CELL_DIFFERENTIATION	*AXL, FLT3LG, GAS6, HNF1A, ID2, IL15, MERTK, PGLYRP1, PGLYRP2, PGLYRP3, PGLYRP4, PIK3CD, PRDM1, PTPRC, RASGRP1, SP3, STAT5B, TOX, TUSC2, TYRO3, ZBTB1, ZNF683*	22
GO	GO_NATURAL_KILLER_CELL_LECTIN_LIKE_RECEPTOR_BINDING	*HLA-E, MICA, MICB, RAET1E, RAET1G, ULBP1, ULBP2, ULBP3*	8
GO	GO_NATURAL_KILLER_CELL_MEDIATED_IMMUNE_RESPONSE_TO_TUMOR_CELL	*CD160, CD226, CEACAM1, CRTAM, HAVCR2, IL12A, IL12B, NECTIN2, PVR*	9
GO	GO_NATURAL_KILLER_CELL_PROLIFERATION	*ELF4, FLT3LG, HLA-E, IL12B, IL15, IL18, IL23A, IL23R, LEP, PTPN22, STAT5B*	11
GO	GO_NEGATIVE_REGULATION_OF_NATURAL_KILLER_CELL_ACTIVATION	*CLNK, FGR, HAVCR2, HLA-F, MICA, PGLYRP1, PGLYRP2, PGLYRP3, PGLYRP4, PIBF1, RHBDD3*	11
GO	GO_NEGATIVE_REGULATION_OF_NATURAL_KILLER_CELL_MEDIATED_IMMUNITY	*ARRB2, CD96, CEACAM1, CLEC12B, CRK, HAVCR2, HLA-E, HLA-F, HLA-G, KIR2DL4, LGALS9, LILRB1, MICA, SERPINB4, SERPINB9*	15
GO	GO_NK_T_CELL_ACTIVATION	*CD300A, ELF4, HSPH1, IL12A, IL12B, IL15, IL18, IL23A, IL23R, RASAL3, ZBTB7B*	11
GO	GO_POSITIVE_REGULATION_OF_NATURAL_KILLER_CELL_ACTIVATION	*AP1G1, AXL, BLOC1S3, FLT3LG, GAS6, HLA-E, HLA-F, IL12A, IL12B, IL15, IL18, IL23A, IL23R, LAMP1, RASGRP1, STAT5B, TICAM1, TOX, TYROBP, ZBTB1*	20
GO	GO_POSITIVE_REGULATION_OF_NATURAL_KILLER_CELL_CHEMOTAXIS	*CCL3, CCL4, CCL5, CCL7, CXCL14, XCL1*	6
GO	GO_POSITIVE_REGULATION_OF_NATURAL_KILLER_CELL_DIFFERENTIATION	*AXL, FLT3LG, GAS6, IL15, RASGRP1, STAT5B, TOX, ZBTB1*	8
GO	GO_POSITIVE_REGULATION_OF_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	*AP1G1, CADM1, CD160, CD226, CRTAM, HLA-E, HLA-F, IL12A, IL12B, IL18RAP, IL21, KLRC4-KLRK1, KLRK1, LAG3, LAMP1, NCR3, NECTIN2, PVR, RAET1E, RASGRP1, SH2D1A, SLAMF6, STAT5B, VAV1*	24
GO	GO_POSITIVE_REGULATION_OF_NATURAL_KILLER_CELL_MEDIATED_IMMUNE_RESPONSE_TO_TUMOR_CELL	*CD160, CD226, CRTAM, IL12A, IL12B, NECTIN2, PVR*	7
GO	GO_POSITIVE_REGULATION_OF_NATURAL_KILLER_CELL_MEDIATED_IMMUNITY	*AP1G1, CADM1, CD160, CD226, CLNK, CRTAM, HLA-E, HLA-F, HLA-G, IL12A, IL12B, IL18RAP, IL21, KIR2DL4, KLRC4-KLRK1, KLRK1, LAG3, LAMP1, NCR3, NECTIN2, PVR, RAET1E, RAET1G, RASGRP1, SH2D1A, SH2D1B, SLAMF6, STAT5B, VAV1*	29
GO	GO_POSITIVE_REGULATION_OF_NATURAL_KILLER_CELL_PROLIFERATION	*FLT3LG, HLA-E, IL12B, IL15, IL18, IL23A, IL23R, STAT5B*	8
GO	GO_POSITIVE_REGULATION_OF_NK_T_CELL_ACTIVATION	*HSPH1, IL12A, IL12B, IL18, IL23A, IL23R, RASAL3*	7

GO	GO_REGULATION_OF_NATURAL_KILLER_CELL_ACTIVATION	*AP1G1, AXL, BLOC1S3, CLNK, FGR, FLT3LG, GAS6, HAVCR2, HLA-E, HLA-F, IL12A, IL12B, IL15, IL18, IL23A, IL23R, LAMP1, LEP, MICA, PGLYRP1, PGLYRP2, PGLYRP3, PGLYRP4, PIBF1, PRDM1, PTPN22, RASGRP1, RHBDD3, STAT5B, TICAM1, TOX, TYROBP, ZBTB1, ZNF683*	34
GO	GO_REGULATION_OF_NATURAL_KILLER_CELL_CHEMOTAXI	*CCL2, CCL3, CCL4, CCL5, CCL7, CXCL14, KLRC4-KLRK1, KLRK1, XCL1*	9
GO	GO_REGULATION_OF_NATURAL_KILLER_CELL_DIFFERENTIATION	*AXL, FLT3LG, GAS6, IL15, PGLYRP1, PGLYRP2, PGLYRP3, PGLYRP4, PRDM1, RASGRP1, STAT5B, TOX, ZBTB1, ZNF683*	14
GO	GO_REGULATION_OF_NATURAL_KILLER_CELL_DIFFERENTIATION_INVOLVED_IN_IMMUNE_RESPONSE	*PGLYRP1, PGLYRP2, PGLYRP3, PGLYRP4, ZNF683*	5
GO	GO_REGULATION_OF_NATURAL_KILLER_CELL_MEDIATED_IMMUNITY	*AP1G1, ARRB2, CADM1, CD160, CD226, CD96, CEACAM1, CLEC12B, CLNK, CRK, CRTAM, HAVCR2, HLA-E, HLA-F, HLA-G, IL12A, IL12B, IL18RAP, IL21, KIR2DL4, KLRC4-KLRK1, KLRK1, LAG3, LAMP1, LEP, LGALS9, LILRB1, MICA, NCR1, NCR3, NECTIN2, PIK3R6, PVR, RAET1E, RAET1G, RASGRP1, SERPINB4, SERPINB9, SH2D1A, SH2D1B, SLAMF6, STAT5B, VAV1*	43
GO	GO_REGULATION_OF_NK_T_CELL_ACTIVATION	*CD300A, HSPH1, IL12A, IL12B, IL18, IL23A, IL23R, RASAL3, ZBTB7B*	9
GO	GO_REGULATION_OF_NK_T_CELL_DIFFERENTIATION	*AP3B1, AP3D1, PRDM1, TGFBR2, ZBTB16, ZNF683*	6
GO	GO_REGULATION_OF_NK_T_CELL_PROLIFERATION	*IL12B, IL18, IL23A, RASAL3, ZBTB7B*	5
KEGG	KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	*ARAF, BID, BRAF, CASP3, CD244, CD247, CD48, CHP1, CHP2, CSF2, FAS, FASLG, FCER1G, FCGR3A, FCGR3B, FYN, GRB2, GZMB, HCST, HLA-A, HLA-B, HLA-C, HLA-E, HLA-G, HRAS, ICAM1, ICAM2, IFNA1, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA2, IFNA21, IFNA4, IFNA5, IFNA6, IFNA7, IFNA8, IFNAR1, IFNAR2, IFNB1, IFNG, IFNGR1, IFNGR2, ITGAL, ITGB2, KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL4, KIR2DL5A, KIR2DS1, KIR2DS3, KIR2DS4, KIR2DS5, KIR3DL1, KIR3DL2, KLRC1, KLRC2, KLRC3, KLRD1, KLRK1, KRAS, LAT, LCK, LCP2, MAP2K1, MAP2K2, MAPK1, MAPK3, MICA, MICB, NCR1, NCR2, NCR3, NFAT5, NFATC1, NFATC2, NFATC3, NFATC4, NRAS, PAK1, PIK3CA, PIK3CB, PIK3CD, PIK3CG, PIK3R1, PIK3R2, PIK3R3, PIK3R5, PLCG1, PLCG2, PPP3CA, PPP3CB, PPP3CC, PPP3R1, PPP3R2, PRF1, PRKCA, PRKCB, PRKCG, PTK2B, PTPN11, PTPN6, RAC1, RAC2, RAC3, RAET1E, RAET1G, RAET1L, RAF1, SH2D1A, SH2D1B, SH3BP2, SHC1, SHC2, SHC3, SHC4, SOS1, SOS2, SYK, TNF, TNFRSF10A, TNFRSF10B, TNFRSF10C, TNFRSF10D, TNFSF10, TYROBP, ULBP1, ULBP2, ULBP3, VAV1, VAV2, VAV3, ZAP70*	137
REACTOME	-	-	0
PID	-	-	0
Total			267[b]

[a]Genes were selected based on online datasets (http://software.broadinstitute.org/gsea/msigdb/search.jsp) and literatures; [b]369 duplicated genes and 8 gene unavailble in NCBI had been removed; Keyword: Natural AND killer AND cell; Organism: Homo sapiens.

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 the adjustment for population stratification in the multivariate analysis.

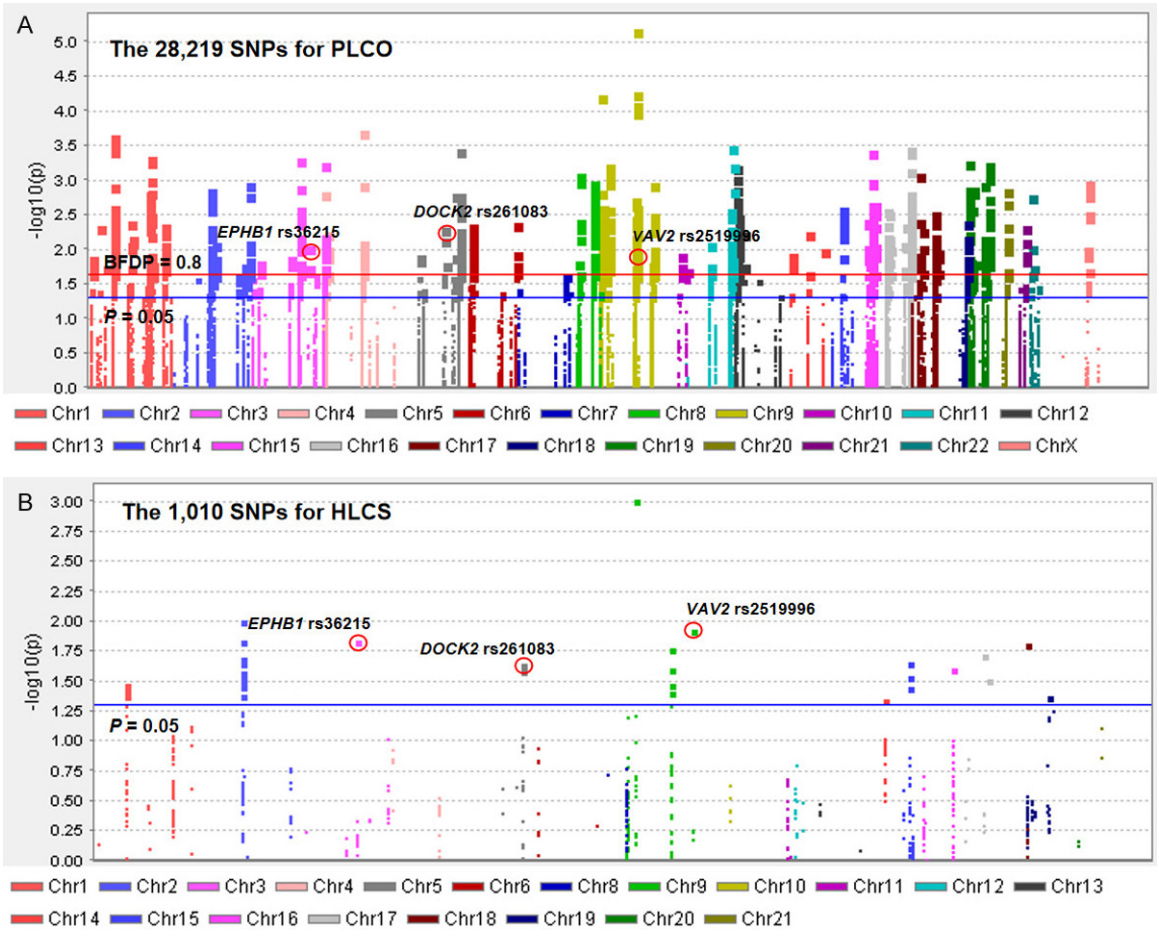


Figure S1. Manhattan plot. Manhattan plots for 28,219 SNPs and 1,010 SNPs of natural killer cell-related pathway genes in the PLCO trial (A) and the HLCS study (B), respectively. The blue horizontal line indicates *P* = 0.05, and the red line indicates BFDP = 0.80. Abbreviations: PLCO, Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; HLCS, the Harvard Lung Cancer Susceptibility Study; BFDP, Bayesian false-discovery probability.

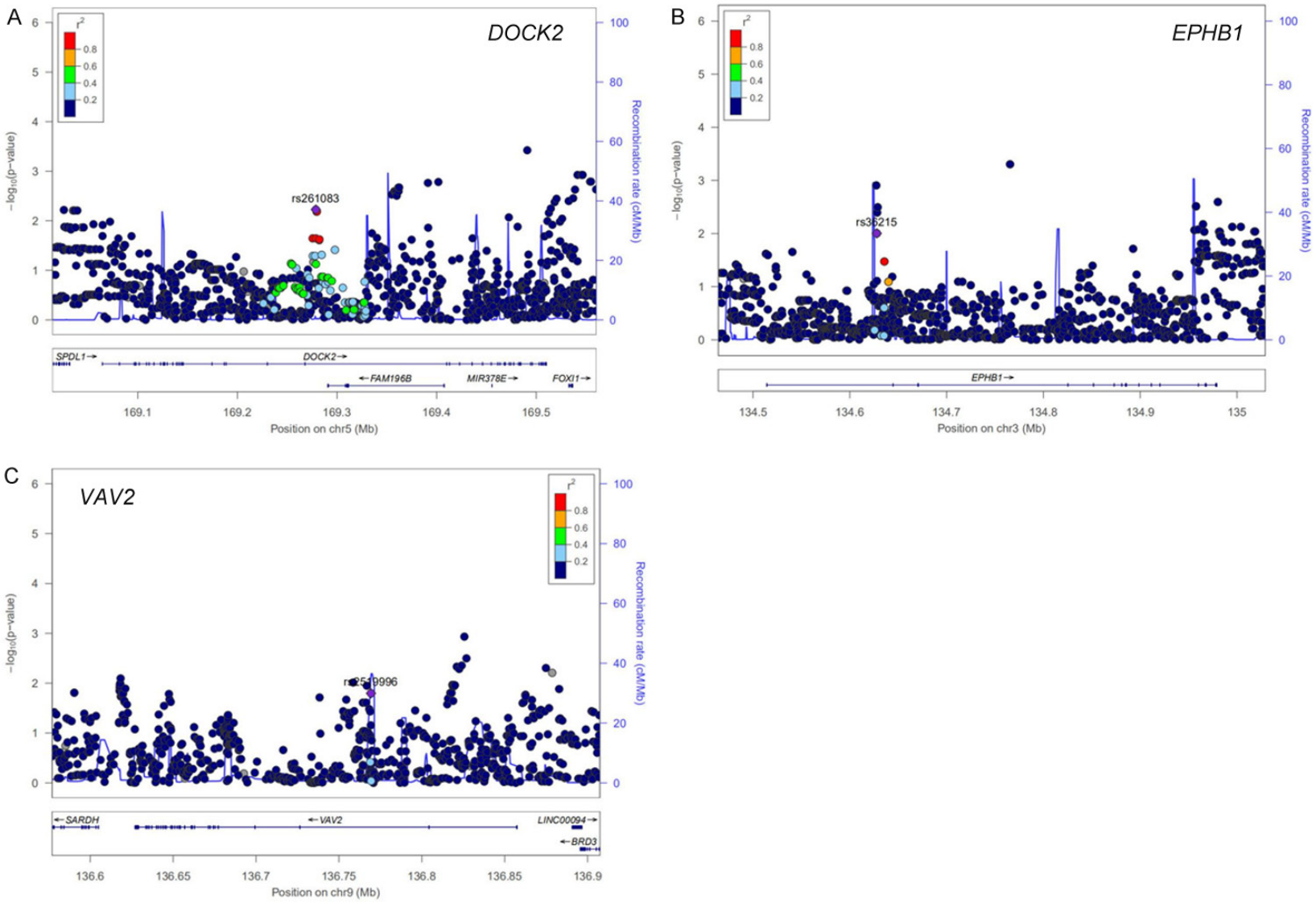
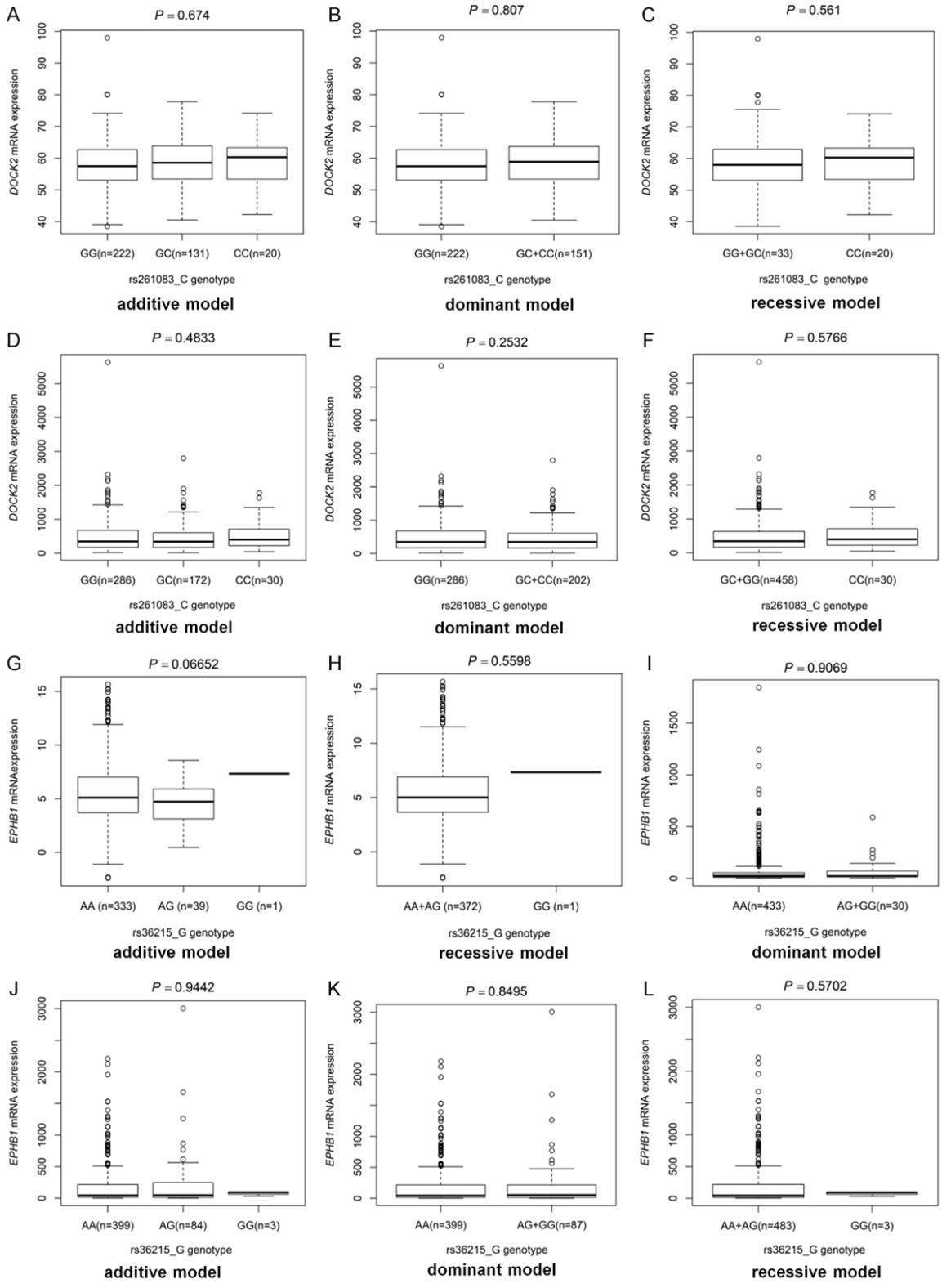


Figure S2. Regional association plots for the three independent SNPs in the natural killer cell pathway genes. Regional association plots contained 50 kb up or downstream of *DOCK2* (A), *EPHB1* (B) and *VAV2* (C). Data points are colored according to the level of linkage disequilibrium of each pair 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.

Table S4. Stratified analysis for associations between the number of unfavorable genotypes and survival of NSCLC in the PLCO trial

Characteristics Frequency[a]	0-1 unfavorable genotype	2-3 unfavorable genotype	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	366	268	1.36 (1.11-1.67)	0.0033		1.34 (1.08-1.66)	0.0085	
>71	310	231	1.34 (1.09-1.64)	0.0060	0.4278	1.43 (1.15-1.77)	0.0013	0.1835
Sex								
Male	391	304	1.37 (1.14-1.64)	0.0007		1.42 (1.17-1.73)	0.0004	
Female	285	185	1.29 (1.02-1.64)	0.0371	0.8579	1.27 (0.99-1.63)	0.0572	0.6836
Smoking status								
Never	66	48	0.74 (0.41-1.34)	0.3259		0.70 (0.38-1.27)	0.2339	
Current	244	173	1.59 (1.23-2.05)	0.0004		1.69 (1.29-2.21)	0.0001	
Former	366	278	1.38 (1.15-1.67)	0.0007	0.0919	1.41 (1.16-1.72)	0.0007	0.1019
Histology								
Adeno	328	247	1.57 (1.26-1.95)	<0.0001		1.53 (1.22-1.91)	0.0002	
Squamous	170	114	1.17 (0.86-1.59)	0.3260		1.38 (0.99-1.93)	0.0569	
Others	178	138	1.22 (0.94-1.57)	0.1301	0.1646	1.20 (0.92-1.57)	0.1866	0.2658
Tumor stage								
I-IIIA	379	275	1.07 (0.85-1.34)	0.5761		1.17 (0.91-1.51)	0.2152	
IIIB-IV	297	224	1.49 (1.24-1.80)	<0.0001	0.1269	1.43 (1.20-1.76)	0.0001	0.9597
Chemotherapy								
No	362	276	1.28 (1.04-1.59)	0.0181		1.36 (1.08-1.71)	0.0085	
Yes	314	223	1.43 (1.17-1.75)	0.0005	0.4078	1.40 (1.14-1.72)	0.0014	0.7379
Radiotherapy								
No	434	327	1.28 (1.06-1.55)	0.0112		1.36 (1.11-1.67)	0.0032	
Yes	242	172	1.45 (1.16-1.81)	0.0011	0.3383	1.43 (1.13-1.79)	0.0025	0.6561
Surgery								
No	368	267	1.41 (1.19-1.68)	<0.0001		1.41 (1.18-1.68)	0.0001	
Yes	308	232	1.13 (0.86-1.48)	0.3793	0.1423	1.22 (0.90-1.65)	0.1962	0.3776

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 missing date 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 characteristic and protective alleles.


A
P = 0.674
DOCK2 mRNA expression
GG(n=222)
GC(n=131)
CC(n=20)
rs261083_C genotype
additive model
B
P = 0.807
DOCK2 mRNA expression
GG(n=222)
GC+CC(n=151)
rs261083_C genotype
dominant model
C
P = 0.561
DOCK2 mRNA expression
GG+GC(n=33)
CC(n=20)
rs261083_C genotype
recessive model
D
P = 0.4833
DOCK2 mRNA expression
GG(n=286)
GC(n=172)
CC(n=30)
rs261083_C genotype
additive model
E
P = 0.2532
DOCK2 mRNA expression
GG(n=286)
GC+CC(n=202)
rs261083_C genotype
dominant model
F
P = 0.5766
DOCK2 mRNA expression
GC+GG(n=458)
CC(n=30)
rs261083_C genotype
recessive model
G
P = 0.06652
EPHB1 mRNAexpression
AA (n=333)
AG (n=39)
GG (n=1)
rs36215_G genotype
additive model
H
P = 0.5598
EPHB1 mRNA expression
AA+AG (n=372)
GG (n=1)
rs36215_G genotype
recessive model
I
P = 0.9069
EPHB1 mRNA expression
AA(n=433)
AG+GG(n=30)
rs36215_G genotype
dominant model
J
P = 0.9442
EPHB1 mRNA expression
AA(n=399)
AG(n=84)
GG(n=3)
rs36215_G genotype
additive model
K
P = 0.8495
EPHB1 mRNA expression
AA(n=399)
AG+GG(n=87)
rs36215_G genotype
dominant model
L
P = 0.5702
EPHB1 mRNA expression
AA+AG(n=483)
GG(n=3)
rs36215_G genotype
recessive model

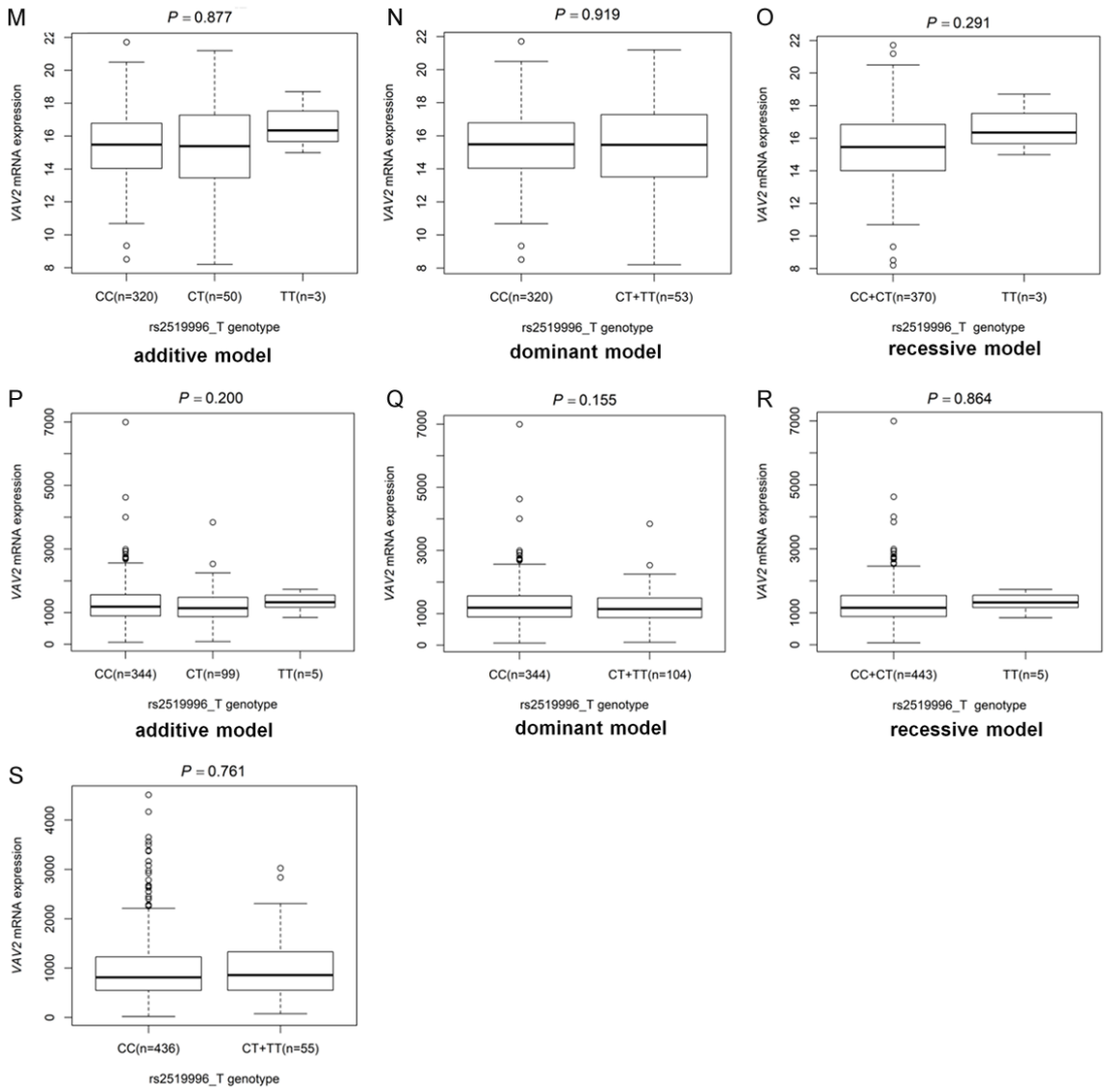


Figure S3. Correlation of genotypes with the mRNA expression levels of the corresponding genes in different genetic models. The correlation between rs261083 genotypes and *DOCK2* mRNA expression levels from the 1000 Genomes Project in the additive (A), dominant (B) and recessive (C) models and in LUSC from the TCGA dataset in additive (D), dominant (E) and recessive (F) models. The correlation between rs36215 genotypes and *EPHB1* mRNA expression levels from the 1000 Genomes Project in additive (G) and recessive (H) models. The correlation between rs36215 genotypes and *EPHB1* mRNA expression levels in LUAD in dominant (I) and LUSC in additive (J), dominant (K) and recessive (L) models from the TCGA dataset. The correlation between rs2519996 genotypes and *VAV2* mRNA expression levels from the 1000 Genomes Project in the additive (M), dominant (N) and recessive (O) models. The correlation between rs2519996 genotypes and *VAV2* mRNA expression levels in LUAD in additive (P), dominant (Q) and recessive (R) and LUSC in dominant (S) models from the TCGA dataset.

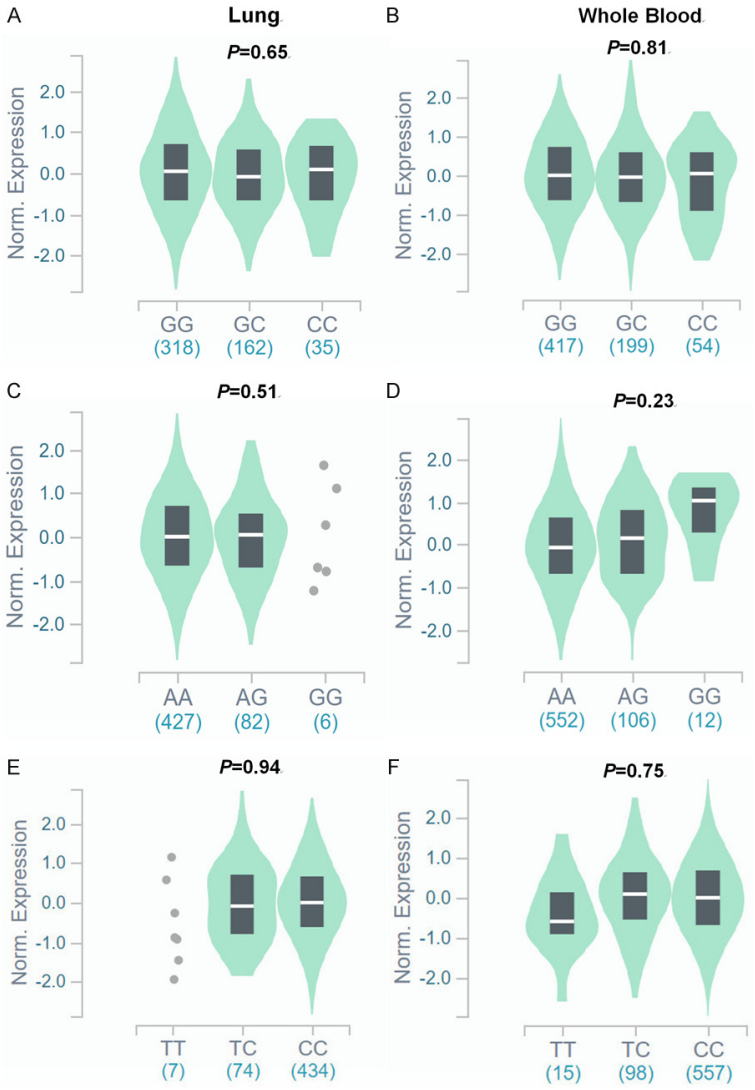


Figure S4. Correlation of genotypes with the mRNA expression levels of the corresponding genes from the GTEx database. The correlation between rs261083 C genotypes and *DOCK2* mRNA expression levels in normal lung tissues (A) and whole blood samples (B); The correlation between rs36215 G genotypes and *EPHB1* mRNA expression levels in normal lung tissues (C) and whole blood samples (D); The correlation between rs2519996 T genotypes and *VAV2* mRNA expression levels in normal lung tissues (E) and whole blood samples (F). Abbreviations: GTEx, Genotype-Tissue Expression project.

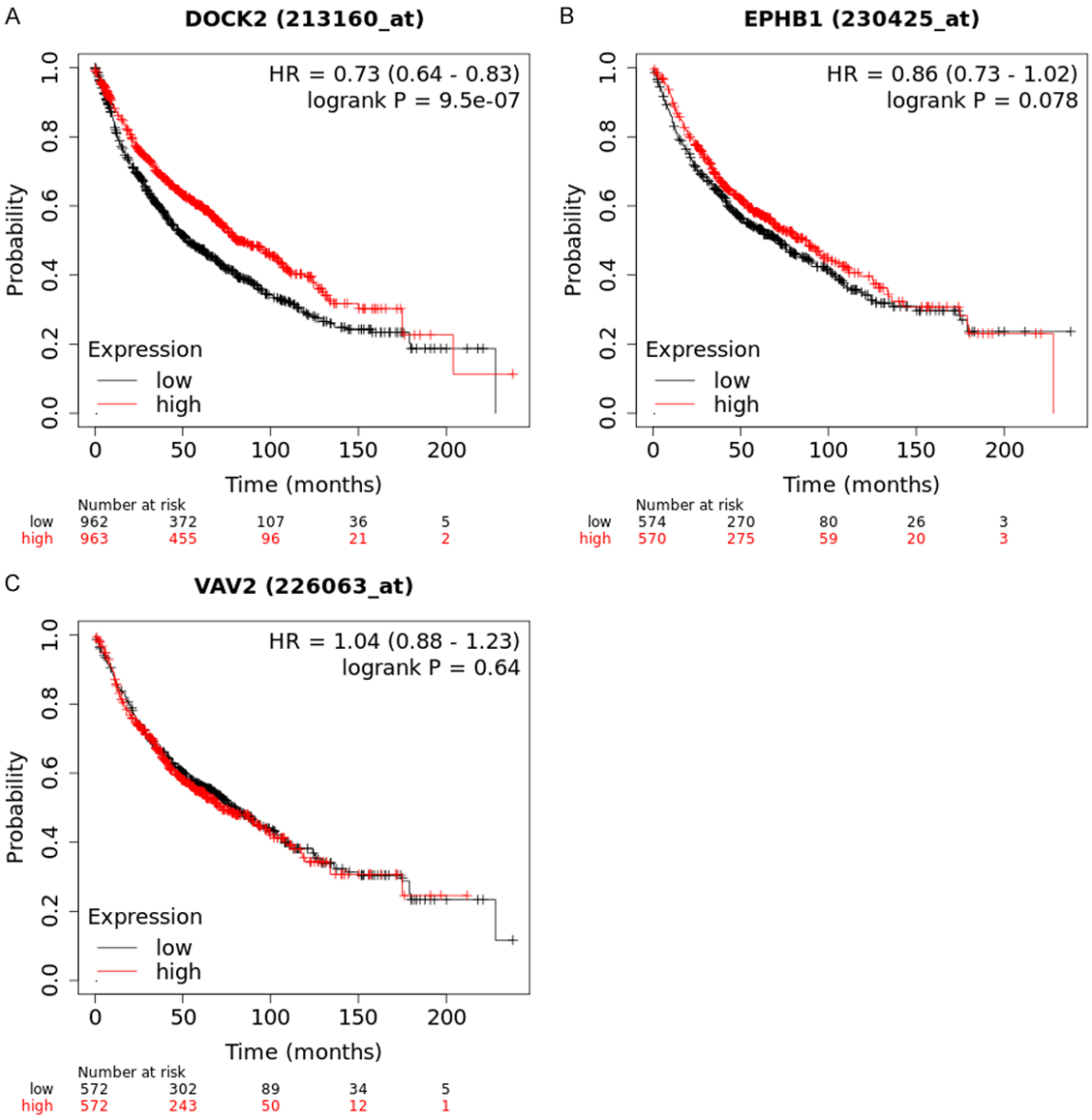


Figure S5. Kaplan-Meier analysis for patients with NSCLC by expression levels of the three genes. Based on online survival analysis software (www.kmplot.com/analysis). A. High DOCK2 expression was associated with a better survival of NSCLC; B. EPHB1 expression were not associated with over survival of NSCLC significantly; C. VAV2 expression were not associated with over survival of NSCLC significantly.

Table S5. Function prediction for *DOCK2* rs261083, *VAV2* rs2519996 and *EPHB1* rs36215

SNP	Gene	Chr	RegDB[1]	Haploreg v4.1[2]					
				Promoter histone marks	Enhancer histone marks	DNAse	Motifs changed	Selected eQTL hits	dbSNP func annot
rs261083	*DOCK2*	5	6	--	--	BLD	EWSR1-FLI1, VDR	2 hits	Intronic
rs2519996	*VAV2*	9	5	--	BRST, BRN, MUS	BRST, BLD	--	2 hits	Intronic
rs36215	*EPHB1*	3	5	--	BRN	--	--	2 hits	Intronic

Abbreviations: SNP, single nucleotide polymorphism; Chr, chromosome; dbSNP func annot, dbSNP function annotation; DNase, deoxyribonuclease; eQTL, expression quantitative trait loci. [1]RegulomeDB: http://regulomedb.org/. [2]Haploreg: https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php.

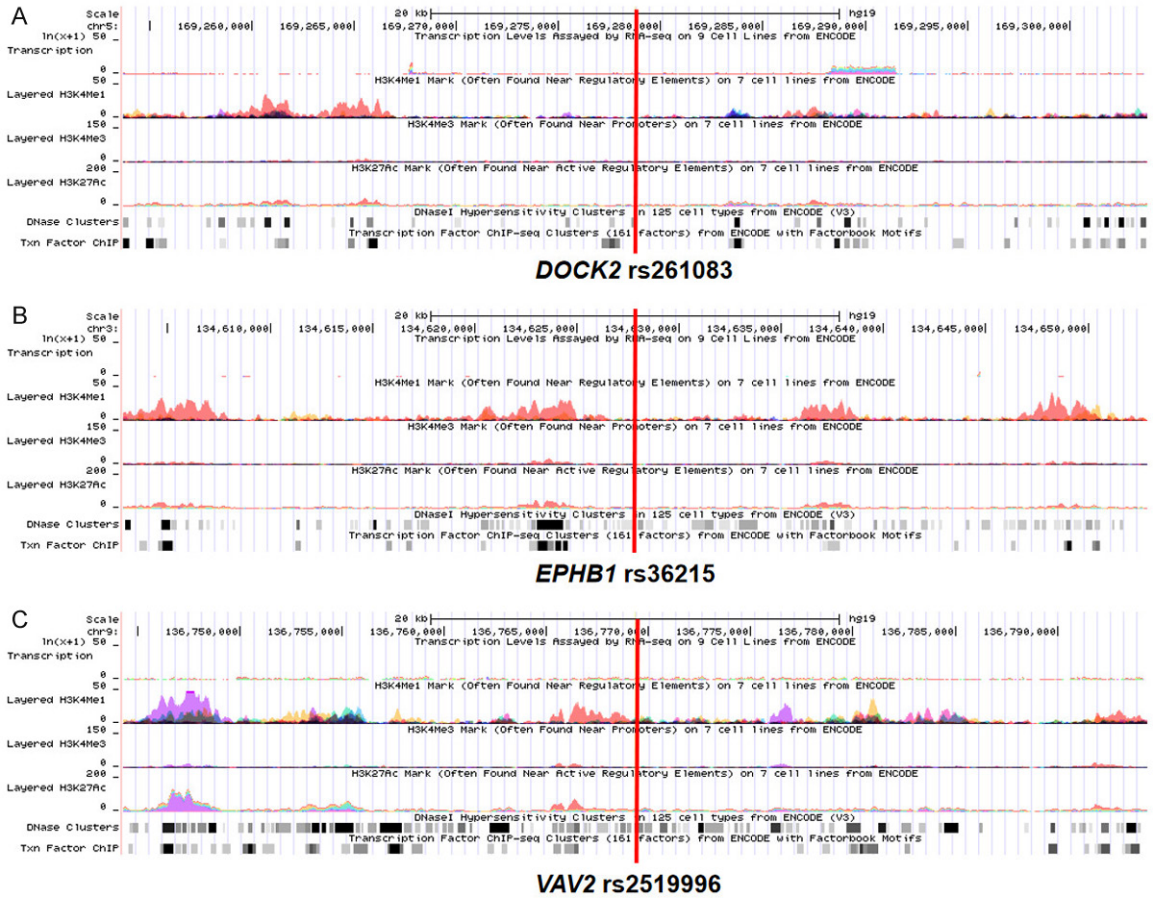


Figure S6. Functional prediction of three independent SNPs in natural killer cell-related pathway genes in the ENCODE data. Location and functional prediction of *DOCK2* rs261083 (A), *EPHB1* rs36215 (B) and*VAV2* rs2519996 (C). The H3K4Me3, H3K4Me1, and H3K27Ac tracks showed the genome-wide levels of 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. Tnx factor track showed regions of transcription factor binding of DNA.

Supplemental Data

We wish to thank all of the investigators and funding agencies that enabled the deposition of data in dbGaP and PLCO that we used in the present study:

The datasets used for the analyses described in the present study were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000336.v1.p1 and phs000093.v2.p2. Principal Investigators are Maria Teresa Landi, Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA and Neil E. Caporaso. Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA.

Funding support for the GWAS of Lung Cancer and Smoking was provided through the NIH Genes, Environment and Health Initiative [GEI] (Z01 CP 010200). The human subjects participating in the GWAS were derived from The Environment and Genetics in Lung Cancer Etiology (EAGLE) case-control study and the Prostate, Lung Colon and Ovary Screening Trial, and these studies are supported by intramural resources of the National Cancer Institute. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies, GENEVA Coordinating Center (U01HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438).

PLCO was also supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. The authors thank PLCO screening center investigators and staff, and the staff of Information Management Services Inc. and Westat Inc. Most importantly, we acknowledge trial participants for their contributions that made this study possible.