IJC
International Journal of Cancer

Potentially functional genetic variants in the complement-related immunity gene-set are associated with non-small cell lung cancer survival

Danwen Qian[1,2*], Hongliang Liu[2*], Xiaomeng Wang[2,3], Jie Ge[2,3], Sheng Luo[4], Edward F. Patz Jr[2,5], Patricia G. Moorman[2,6], Li Su[7], Sipeng Shen[7], David C. Christiani[7,8] and Qingyi Wei[1,2,3]

[1]Cancer Institute, Fudan University Shanghai Cancer Center; Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, 200032, China
[2]Duke Cancer Institute, Duke University Medical Center, Durham, North Carolina 27710, USA
[3]Department of Population Health Sciences, Duke University School of Medicine, Durham, North Carolina 27710, USA
[4]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina 27710 USA
[5]Department of Radiology, Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, North Carolina 27710 USA
[6]Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina, USA
[7]Departments of Environmental Health and Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA
[8]Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

The complement system plays an important role in the innate and adaptive immunity, complement components mediate tumor cytolysis of antibody-based immunotherapy, and complement activation in the tumor microenvironment may promote tumor progression or inhibition, depending on the mechanism of action. In the present study, we conducted a two-phase analysis of two independently published genome-wide association studies (GWASs) for associations between genetic variants in a complement-related immunity gene-set and overall survival of non-small cell lung cancer (NSCLC). The GWAS dataset from Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial was used as the discovery, and multivariate Cox proportional hazards regression with false-positive report probability for multiple test corrections were performed to evaluate associations between 14,699 single-nucleotide polymorphisms (SNPs) in 111 genes and survival of 1,185 NSCLC patients. The identified significant SNPs in a single-locus analysis were further validated with 984 NSCLC patients in the GWAS dataset from the Harvard Lung Cancer Susceptibility (HLCS) Study. The results showed that two independent, potentially functional SNPs in two genes (*VWF* rs73049469 and *ITGB2* rs3788142) were significantly associated with NSCLC survival, with a combined hazards ratio (HR) of 1.22 [95% confidence interval (CI) = 1.07–1.40, $P = 0.002$] and 1.16 (1.07–1.27, 6.45×10^{-4}),

Key words: non-small cell lung cancer, complement pathway, genome-wide association study, single-nucleotide polymorphism, overall survival

Abbreviations: AUC: area under the receiver operating characteristic curve; CFH: complement factor H; CFI: complement factor I; CI: confidence interval; EAF: effect allele frequency; ECM: extracellular-matrix; eQTL: expression quantitative trait loci; FDR: false discovery rate; FPRP: false-positive report probability; GWAS: Genome-Wide Association Study; HLCS: Harvard Lung Cancer Susceptibility; HR: hazards ratio; ITGB2: integrin subunit beta 2; LD: linkage disequilibrium; LFA-1: lymphocyte function-associated antigen-1; Mac-1: macrophage-1 antigen; NSCLC: nonsmall cell lung cancer; OS: overall survival; PLAUR: plasminogen activator, urokinase receptor; PLCO: the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; ROC: receiver operating characteristic; SNPs: single-nucleotide polymorphisms; TCGA: The Cancer Genome Atlas; VWF: von Willebrand factor

Additional Supporting Information may be found in the online version of this article.

Conflicts of interest: The authors declare no conflict of interest.

*D.Q. and H. L. contributed equally to this work

Grant sponsor: NIH; **Grant numbers:** CA090578, CA074386, CA092824, 5U01CA209414; **Grant sponsor:** State Scholarship Fund, China Scholarship Council, the Ministry of Education of the P. R. China through Fudan University Shanghai Cancer Center; **Grant sponsor:** Duke Cancer Institute as part of the P30 Cancer Center Support Grant; **Grant numbers:** NIH/NCI CA014236; **Grant sponsor:** V Foundation for Cancer Research; **Grant numbers:** D2017-19

DOI: 10.1002/ijc.31896

History: Received 28 Jun 2018; Accepted 10 Sep 2018; Online 27 Sep 2018

Correspondence to: Qingyi Wei, Duke Cancer Institute, Duke University Medical Center and Department of Medicine, Duke School of Medicine, 905 S LaSalle Street, Durham, North Carolina 27710, USA, E-mail: qingyi.wei@duke.edu; Tel.: (919) 660-0562

respectively. Finally, we performed expression quantitative trait loci (eQTL) analysis and found that survival-associated genotypes of *VWF* rs73049469 were also significantly associated with mRNA expression levels of the gene. These results indicated that genetic variants of the complement-related immunity genes might be predictors of NSCLC survival, particularly for the short-term survival, possibly by modulating the expression of genes involved in the host immunity.

What's new?

For non-small cell lung cancer (NSCLC), five-year survival remains poor. Immunotherapy approaches are gaining ground, but still do not work in most patients. These authors looked through datasets from two genome-wide association studies for genes involved in the complement system also associated with NSCLC survival. They found that two genetic variants of two genes were associated with survival of non-small cell lung cancer. The survival-associated variant A genotypes of rs73049469 were also associated with mRNA expression levels of the *VWF* gene. These variants could be useful predictors of NSCLC survival, and further functional studies could uncover the roles of these genes in the development of lung cancer.

Introduction

Lung cancer is one of the most common malignancies worldwide, with 222,500 new cases of lung cancer in 2017 in the United States.[1] Despite much-devoted research effort in the treatment of lung cancer in recent decades, lung cancer remains the leading cause of cancer deaths. Non-small cell lung cancer (NSCLC), mainly squamous cell carcinoma and adenocarcinomas, accounts for 80–85% of all lung cancer cases, and most of NSCLC patients have an advanced disease at the time of diagnosis.[2] With the advances in the treatment of NSCLC patients, including surgery, radiotherapy, chemotherapy, molecular targeted therapy and immunotherapy, the prognosis of lung cancer has been improved, but the five-year survival rate remains only 18.1%.[1] Clinically, the known clinicopathological variables, such as age, sex, performance status and most importantly tumor stage, are commonly used for predicting prognosis; however, the response of individuals is heterogeneous, suggesting that genetic factors also account for the variability in treatment response, likely due to genetic variation in drug disposition, pharmacokinetic effects and host immunity. Some studies found that single-nucleotide polymorphisms (SNPs) could influence short-term response and long-term prognosis of cancer patients.[3–5] Therefore, identifying the role of these genetic factors in the prognosis may lead to a better understanding of lung cancer prognosis.

Although genome-wide association studies (GWASs) have identified a number of lung cancer-associated SNPs, biological functions of these SNPs in lung cancer development and progress are still unclear, which limits clinical applications of the GWAS results. Some recommended research strategies in the post-GWAS era include "discovery, expansion and replication", "biological studies" and "epidemiologic studies" (https://epi.grants.cancer.gov/gameon/#funded). By pooling together multiple GWAS datasets, one can identify novel loci with minor but detectable effects, and then examine functional consequence of the loci by additional functional studies to unravel possible mechanisms underlying the observed associations. Another way is to re-analyze the published GWAS datasets by a hypothesize-based gene-set approach in which unnecessary multiple tests may be avoided to increase the study power.

The immune system plays an important role in the development, progression and in some cases inhibition of cancer, for example, complement factor H autoantibodies could kill tumor cells.[6] Recent progress in immunotherapy has become an important therapeutic option for the treatment of cancer, especially check-point inhibitors and chimeric antigen receptor T-cell therapy, but the majority of patients do not respond or become resistant to the treatment.[7–9] Thus, we hypothesize that some host genetic factors account for a heterogeneous anti-tumor immune response, which may have an effect on survival. In the present study, to test the hypothesis, we investigated associations between genetic variants of genes in a complement-related immunity gene-set and NSCLC survival.

The complement system is a complex multistep cascade at the interface of innate and adaptive immunity, which is activated through three pathways: classical, lectin and alternative pathways, to enhance phagocytosis, rigger antibody generation, potentiate inflammation and avoid autologous damages.[10] In tumor immunology, traditionally complement has been considered to mediate tumor cytolysis of antibody-based immunotherapy. For example, clinically available antibodies target tumor cells by complement-dependent cytotoxicity, antibody-dependent cell mediated cytotoxicity and complement-dependent phagocytosis, which directly kill and eliminate tumor cells.[11] However, activation of the complement system in the tumor microenvironment may promote tumor progression. It has been suggested that this is due to induction of tumor-infiltrating immune cell to release pro-inflammatory cytokines, resulting in suppressing the activation of effector T cells and creating an environment favorable for tumor growth.[12] Furthermore, the complement-activated products also promote angiogenesis and facilitate tumor cell migration and invasion.[13] Several studies have shown that some complement components affect lung cancer cells.[14–17] For example, C3 has been proved to be correlated with tissue deposition and NSCLC patients' prognosis,[14] and C5a, which is generated by NSCLC cells, promotes tumor growth and

immunomodulatory effect.[15] However, the roles of genetic variants, such as SNPs, of many other candidate genes in the complement-related gene-set and their functionality involved in tumor growth and progression are still unknown. In the present study, using the publically available GWAS datasets, we performed a complement-related immunity gene-set analysis to evaluate associations between genetic variants of 111 genes and survival of NSCLC patients.

Materials and Methods

Discovery dataset

As shown in the study flowchart (Fig. 1), we used the GWAS dataset from Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial as the discovery with the access approval from the dbGAP from National Cancer Institute (the approval number: PLCO-95). The PLCO dataset is a multicenter randomized study conducted by ten centers in the United States between 1993 and 2011. Nearly 78,000 women and 76,000 men aged 55–74 enrolled in our study, who were randomized to either an intervention arm (received trial screening) or control arm (received standard care).[18] Blood samples as well as individuals' information, such as smoking status, histologic diagnosis, tumor stage and treatment method, were collected in this trial.[19] Since two individuals had missing follow-up information, they were excluded from the eligible subset that contained 1,185 NSCLC patients for survival analysis. Genomic DNA extracted from the blood samples was genotyped for the GWAS with Illumina HumanHap240Sv1.0, HumanHap300v1.1 and HumanHap550v3.0 (dbGaP accession: phs000093.v2.p2 and phs000336.v1.p1).[20,21] All the clinicopathological variables and genotype data of these 1,185 patients were available. The PLCO trial was approved by the institutional review boards of each participating institution, and all subjects signed a written informed consent permitting the research represented here.

Validation dataset

The top hits with potentially functional SNPs obtained from the PLCO discovery analysis were further validated by the GWAS dataset from the Harvard Lung Cancer Susceptibility (HLCS) study. After applying quality control, 984 histology-confirmed Caucasian patients remained in the HLCS study, who were older than 18 years old, with newly diagnosed, histologically confirmed primary NSCLC. Patients' blood samples were used to extract DNA by the Auto Pure Large Sample Nucleic Acid Purification System (QIAGEN Company, Venlo, Limburg, Netherlands) that was genotyped using Illumina Humanhap610-Quad arrays, and the genotyping data were imputed using MaCH1.0 based on the 1000 Genomes Project. All the details of participant recruitment and characteristics have been previously described.[22]

Gene and SNP selection

Genes involved in the complement-related immunity gene-set were selected by the Molecular Signatures Database (http://software.broadinstitute.org/gsea/msigdb/index.jsp), the Human Biological Pathway Unification Database (https://pathcards.genecards.org), and the HUGO Gene Nomenclature Committee (https://www.genenames.org) by the keyword "complement". After removing the duplicated genes and deleting genes in the X chromosome, 111 genes remained as the candidate genes for further analysis (Supporting Information Table S1). All SNPs

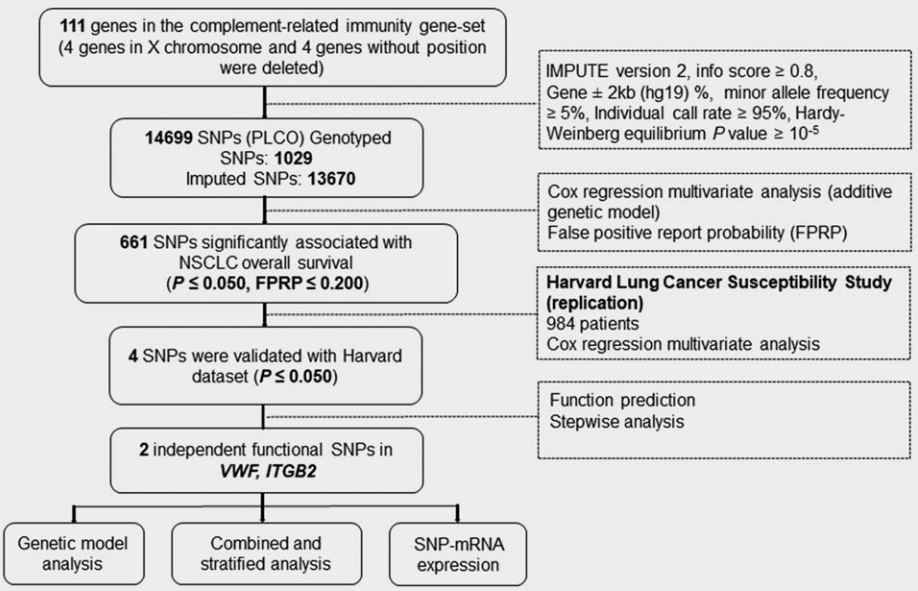

Figure 1. The flowchart of the study. Abbreviations: SNP, single-nucleotide polymorphism; PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial; NSCLC, non-small cell lung cancer; VWF, von Willebrand factor; ITGB2, integrin subunit beta 2.

were first aligned to the plus strand of the reference genome and then used in imputation if the minor allele frequency ≥ 0.05, genotyping rate ≥95% and Hardy–Weinberg equilibrium P value ≥1×10^{-5}. Then we performed imputation for the 111 genes plus two 500 kb flanking buffer regions using IMPUTE2 and the 1000 Genomes Project data (phase 3). After imputation, we extracted all the SNPs in these genes and within their ±2 kb flanking regions according to the following criteria: minor allele frequency ≥0.05, genotyping rate ≥95% and Hardy–Weinberg equilibrium P value ≥1×10^{-5}. As a result, 1,029 genotyped SNPs were chosen from the PLCO GWAS dataset, and additional 13,670 SNPs were imputed.

Statistical analysis

The follow-up time was from the diagnosis of lung cancer to the last follow-up or time of death, and we used overall survival (OS) as the primary end point. In the single-locus analysis, multivariate Cox hazards regression analysis was used to evaluate associations between each of the SNPs and OS (in an additive model) with adjustment for age, sex, smoking status, histology, tumor stage, chemotherapy, radiotherapy, surgery and the first four principal components of the population structures of the PLCO dataset using the GenABEL package of R software.[23] The false discovery rate (FDR) with a cut-off value of 0.200 was used to assess the probability of false positives.[24] Since the majority of SNPs were imputed with a high level of linkage disequilibrium (LD), we also used false-positive report probability (FPRP) with a cutoff value of 0.200 for multiple test corrections. As one Bayesian approach, the FPRP method depends not only on P value but also on prior probability and the statistical power of the test, we assigned a prior probability of 0.100 to detect an HR of 2.0 for an association between genetic variants and survival.[25] Then, we chose SNPs for the validation using the HLCS GWAS dataset, which satisfied any of the following conditions: potentially functional SNPs predicted by SNPinfo and RegulomeDB, and tagging SNPs based on their LD. To identify independent SNPs, we included the validated SNPs in a multivariate stepwise Cox model with adjustment for clinical variables and the first four principal components. Combined-analysis was performed to combine the results of discovery and validation datasets. If the Cochran's Q-test P >0.100 and the heterogeneity statistic (I^2) < 50%, a fixed-effects model was applied. Otherwise, a random-effects model was employed. Kaplan–Meier curve was used to estimate the survival associated with the genotypes, the combination of risk genotypes was conducted to estimate the cumulative effects of identified SNPs.

Expression quantitative trait loci (eQTL) analysis was further performed to assess correlations between SNPs and mRNA expression levels using linear regression analysis with the R software. mRNA expression data of genes were obtained from lymphoblastoid cell lines derived from the 373 European descendants included in the 1000 Genomes Project,[26] and from the whole blood cells and normal lung tissues in the genotype-tissue expression (GTEx, V7) project.[27] Using the data from The Cancer Genome Atlas (TCGA) database (dbGaP Study Accession: phs000178.v9.p8), we examined the differences in mRNA expression levels between paired tumor tissues and adjacent normal tissues by the paired t test.[28] Finally, the receiver operating characteristic (ROC) curve and time-dependent ROC analysis were performed to illustrate the prediction accuracy of models integrating both clinical and genetic variables on NSCLC survival with the timeROC package of R (version 3.5.0) software.[29] Unless specified, all statistical analyses were performed using SAS software (version 9.4; SAS Institute, Cary, NC).

Results

Associations between SNPs in the complement-related immunity gene-set and NSCLC OS in both PLCO and HLCS datasets

The study flowchart is shown in Figure 1, and basic characteristics of 1,185 NSCLC patients have been described previously.[30] In the PLCO discovery with an additive genetic model, the multivariate Cox model with adjustment for age, sex, smoking status, histology, tumor stage, chemotherapy, radiotherapy, surgery and first four principal components (Supporting Information Table S2) identified 661 SNPs that were significantly associated with NSCLC OS after multiple test correction by FPRP ≤0.200 (but no SNP remained significant for FDR ≤ 0.200, because of a high level of LD among these SNPs as a result of imputation). The results are summarized in the Manhattan plot (Supporting Information Fig. S1). The top and potentially functional SNPs were further validated by the HLCS dataset. As a result, four SNPs in four different genes (i.e., rs73049469 in *VWF*, rs3788142 in *ITGB2*, rs4251906 in *PLAUR* and rs116750895 in *CFI*) were validated, of which only rs3788142 was actually genotyped. Further combined-analysis of these SNPs of the two datasets showed a poorer OS associated with the rs73049469 A, rs3788142 A, rs116750895 G or rs4251906 A alleles ($P_{adjusted}$ = 0.002, 6.45×10^{-4}, 3.23×10^{-4} or 3.29×10^{-4}, respectively), and no heterogeneity between the two studies was observed (Table 1).

Identification of independent SNPs associated with OS of NSCLC in the PLCO dataset

Because the HLCS study provided only the summary data, we had to use the PLCO dataset to identify independent SNPs. To identify potentially functional SNPs associated with NSCLC OS with three bioinformatics tools (SNPinfo, RegulomeDB and HaploReg), we found that two of the four validated SNPs were located in the intron regions with some considerable function. In the RegulomeDB, rs73049469 and rs3788142 had a score of 2b and 3a, respectively, while rs4251906 and rs116750895 had no data. Besides, all the four validated SNPs had no function based on the SNPinfo, and in the HaploReg, rs73049469 and rs3788142 may have an effect on histone marks, DNAse and motifs (Supporting Information Table S3).

Table 1. Combined-analysis of four validated SNPs using two previously published NSCLC GWAS datasets

SNP	Allele[1]	Gene	PLCO (n = 1,185)					HLCS (n = 984)			Combined-analysis			
			EAF	HR (95% CI)[2]	*P*[2]	FDR[3]	FPRP[3]	EAF	HR (95% CI)[4]	*P*[4]	P_{het}[5]	I^2	HR (95% CI)[6]	*P*[6]
rs73049469	A/C	*VWF*	0.11	1.20 (1.02–1.42)	0.025	0.819	0.183	0.11	1.27 (1.02–1.59)	0.035	0.676	0	1.22 (1.07–1.40)	0.002
rs3788142	A/G	*ITGB2*	0.24	1.17 (1.04–1.32)	0.008	0.794	0.065	0.24	1.15 (1.01–1.32)	0.030	0.885	0	1.16 (1.07–1.27)	6.45×10^{-4}
rs116750895	G/T	*CFI*	0.06	1.31 (1.07–1.60)	0.008	0.794	0.066	0.06	1.38 (1.04–1.82)	0.024	0.835	0	1.35 (1.14–1.58)	3.23×10^{-4}
rs4251906	G/A	*PLAUR*	0.06	0.73 (0.59–0.89)	0.002	0.794	0.021	0.07	0.76 (0.58–0.99)	0.043	0.839	0	0.74 (0.63–0.87)	3.29×10^{-4}

Abbreviations: SNP, single-nucleotide polymorphism; NSCLC, non-small cell lung cancer; GWAS, genome-wide association study; PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial; HLCS, Harvard Lung Cancer Susceptibility; EAF, effect allele frequency; HR, hazards ratio; CI, confidence interval; FDR, false discovery rate; FPRP, false-positive report probability.

[1]Effect/reference allele.

[2]Obtained from an additive genetic model with adjustment for age, sex, stage, histology, smoking status, chemotherapy, radiotherapy, surgery, PC1, PC2, PC3 and PC4.

[3]FDR and FPRP were available in the PLCO dataset because the HLCS study provided only the summary data.

[4]Obtained from an additive genetic model with adjustment for age, sex, stage, histology, smoking status, chemotherapy, radiotherapy, surgery, PC1, PC2 and PC3.

[5]P_{het}: *P* value for heterogeneity by Cochrane's Q test.

[6]Meta-analysis in the fixed-effects model.

Because the purpose of the present study was to identify potentially functional SNPs, a stringent criterion with RegulomeDB scores ≤3 was used to select SNPs. Therefore, the two SNPs (rs73049469 and rs3788142) were included into a multivariate stepwise Cox model with adjustment for clinical variables and the first four principal components available in the PLCO dataset. As a result, both the two SNPs remained to be independently associated with NSCLC OS (Table 2), and the regional association plot of each SNP is shown in Supporting Information Figure S2.

In the PLCO dataset, patients with rs73049469 A or rs3788142 A allele had an increased risk of death (P_{trend} = 0.025 or 0.008, respectively, Table 3 and Supporting Information Figure S3). Compared to the reference genotype in a dominant genetic model, *VWF* rs73049469 CA + AA and *ITGB2* rs3788142 GA + AA were associated with a significantly increased risk of death (HR = 1.22, 95% CI = 1.03–1.45, *P* = 0.024 for rs73049469 CA + AA and HR = 1.15, 95% CI = 0.99–1.33, *P* = 0.051 for rs3788142 GA + AA).

Combined and stratified analysis of the two independent and functional SNPs in the PLCO dataset

To provide a better estimation of the hazards of survival, we combined the risk genotypes (i.e., rs73049469 CA + AA and rs3788142 GA + AA) into a genetic score, which divided all NSCLC patients into three groups: zero, one and two risk genotype. As shown in Table 3, in the multivariate analysis, an increased genetic score derived from all the risk genotypes was associated with a poorer survival (trend test: *P* = 0.004). To dichotomize the death risk, we re-grouped all the patients into a low-risk group (0 risk genotype) and a high-risk group (1–2 risk genotype). Compared to the low-risk group, patients in the high-risk group had a significantly poorer survival (HR = 1.23, 95% CI = 1.07–1.42, *P* = 0.005). Kaplan–Meier survival curves were present to depict the associations between risk genotypes and NSCLC OS (Fig. 2*a* and 2*b*).

To assess the ability of risk genotypes to predict the survival of NSCLC, we compared the model with the area under the receiver operating characteristic curve (AUC) for clinical variables to that of AUC for both clinical variables and risk genotypes. The addition of risk genotypes to the prediction model of one-year survival increased the AUC from 85.79% to 86.23% (*P* = 0.006, Fig. 2*c*), but this was not found for the five-year survival between the prediction model with or without risk genotypes (Fig. 2*d*). Finally, time-dependent AUC curve was provided to quantitate the ability of risk genotypes to predict NSCLC OS through the entire follow-up period (Fig. 2*e*).

We further performed stratified analysis to evaluate whether the effect of combined risk genotypes on NSCLC OS was modified by age, sex, smoking status, histology, tumor stage, chemotherapy, radiotherapy and surgery (Supporting Information Fig. S3). The results showed that no significant

Table 2. Predictors of OS obtained from stepwise multivariate Cox regression analysis of selected functional variables in the PLCO Trial

Variables[1]	Category[2]	Frequency[3]	HR (95% CI)	*P*
Age	Continuous	1171	1.03 (1.02–1.04)	‹0.001
Sex	Male	694	1.00	
	Female	477	0.78 (0.67–0.91)	0.001
Smoking status	Never	114	1.00	
	Current	417	1.63 (1.22–2.18)	0.001
	Former	640	1.61 (1.23–2.12)	0.001
Histology	AD	574	1.00	
	SC	284	1.17 (0.97–1.41)	0.104
	others	313	1.31 (1.10–1.55)	0.002
Stage	I-IIIA	651	1.00	
	IIIB-IV	520	2.76 (2.28–3.35)	‹0.001
Chemotherapy	No	636	1.00	
	Yes	535	0.58 (0.48–0.69)	‹0.001
Radiotherapy	No	758	1.00	
	Yes	413	0.93 (0.79–1.09)	0.375
Surgery	No	634	1.00	
	Yes	537	0.20 (0.16–0.26)	‹0.001
VWF rs73049469[4]	CC/CA/AA	928/233/10	1.19 (1.01–1.40)	0.037
ITGB2 rs3788142[4]	GG/GA/AA	680/419/72	1.16 (1.04–1.31)	0.011

Abbreviations: OS, overall survival; PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial; HR, hazards ratio; CI, confidence interval; AD, Adenocarcinoma; SC, Squamous cell carcinoma; PC, principal components.
[1]Stepwise analysis included age, sex, smoking status, tumor stage, histology, chemotherapy, radiotherapy, surgery, PC1, PC2, PC3, PC4 and 2 SNPs.
[2]The leftmost was used as the reference.
[3]14 missing date were excluded.
[4]rs73049469 and rs3788142 in an additive genetic model.

interactions were found ($P > 0.05$, Supporting Information Table S4).

In silico functional validation

According to experimental data from the ENCODE project (Supporting Information Fig. S4), we found that SNP rs7304969 to be located in a DNase I hypersensitive site, where the DNase hypersensitivity and histone modification H3K27 acetylation indicated some strong signals for active enhancer and promoter functions. The evidence from the DNase cluster and transcription factor CHIP-seq data predict that rs73049469 is located in the PLAG1 motif and that rs388142 is in the SP4 motif as shown by the position weight matrix (Supporting Information Fig. S4), and the minor allele might affect the binding activity to have an impact on the transcription factors.

To further explore potential functions of these SNPs, we performed the eQTL analysis for the correlation between SNP and mRNA expression using data of the 373 European descendants available in the 1000 Genomes Project. Only the *VWF* rs73049469 A allele showed a significant correlation with decreased mRNA expression levels of the gene ($P = 0.049$, Fig. 3*a*), while this was not the case for the *ITGB2* rs3788142 A allele (Fig. 3*b*), but in the whole blood data of the GTEx project, the rs3788142 A allele was associated with higher expression levels of *ITGB2* (Supporting Information Table S5). Taken together, these findings suggest that *VWF* rs73049469 and *ITGB2* rs3788142 may influence their gene expression at the transcription level.

Finally, to find molecular mechanisms of these genes in the progression of NSCLC, we compared the mRNA expression of these genes in 109 paired NSCLC tumor and adjacent normal tissue samples obtained from the TCGA database. Expression levels of *VWF* and *ITGB2* were both lower in the tumor tissues ($P < 0.001$ for both), compared to the adjacent normal tissues (Supporting Information Fig. S5*a* and S5*b*), and their high expression levels were associated with a better NSCLC OS[31] (Supporting Information Fig. S6*a* and S6*b*).

Discussion

In the present study, we performed the analysis for associations between SNPs in complement-related immunity gene-set and NSCLC OS using two previously published GWAS datasets. We identified and validated four SNPs (i.e., *VWF* rs73049469, *ITGB2* rs3788142, *CFI* rs116750895 and *PLAUR* rs4251906) that were significantly associated with NSCLC OS in European populations. Since *CFI* rs116750895 and *PLAUR* rs4251906 had no functional annotation data in both SNPinfo and RegulomeDB, we excluded these two SNPs in the stepwise analysis, in which both *VWF* rs73049469 and *ITGB2*

Table 3. Associations between two independent and functional SNPs and overall survival of NSCLC in the PLCO Trial

Genotype	No.		Multivariate analysis[1]	
	ALL	Death (%)	HR (95% CI)	*P*
VWF rs73049469 C>A[2]				
CC	928	620 (66.81)	1.00	
CA	233	160 (68.67)	1.22 (1.02–1.45)	0.028
AA	10	7 (70.00)	1.29 (0.61–2.73)	0.510
Trend				0.025
CA + AA	243	167 (68.72)	1.22 (1.03–1.45)	0.024
ITGB2 rs3788142 G>A[3]				
GG	682	445 (65.25)	1.00	
GA	421	291 (69.12)	1.10 (0.97–1.28)	0.244
AA	72	53 (73.61)	1.56 (1.17–2.08)	0.003
Trend				0.008
GA + AA	493	344 (69.78)	1.15 (0.99–1.33)	0.051
Number of risk genotypes[2,4]				
0	541	351 (64.88)	1.00	
1	526	362 (68.82)	1.21 (1.04–1.41)	0.012
2	104	74 (71.15)	1.32 (1.03–1.70)	0.031
Trend				0.004
0	541	351 (64.88)	1.00	
1–2	630	441 (69.21)	1.23 (1.07–1.42)	0.005

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.
[1]Adjusted for age, sex, smoking status, histology, tumor stage, chemotherapy, surgery, and principal components.
[2]14 missing date were excluded.
[3]10 missing date were excluded.
[4]Risk genotypes were *VWF* rs73049469 CA + AA and *ITGB2* rs3788142 GA + AA.

rs3788142 were independently associated with NSCLC OS. In subsequent functional prediction analysis using data from public databases, we found that the *VWF* rs73049469 A allele was associated with a decreased mRNA expression in the established blood cell lines, while the *ITGB2* rs3788142 A allele was associated with an increased mRNA expression in the whole blood in the GTEx project. These results were consistent with those of gene expression analysis between paired tumor and adjacent normal tissue samples and survival analysis in the TCGA database. Specifically, we found that higher expression levels of *VWF* and *ITGB2* were associated with a better survival in the TCGA database. Besides, the prediction model with combined risk genotypes suggested that an improved survival as indicated by ROC curve was only observed for the cut-off time of one year, but not of five years or longer. Therefore, this combined prediction model suggests that these risk genotypes could predict a better short-term survival of NSCLC. Additional data suggest that the expression levels of the genes targeted by these SNPs were associated with NSCLC survival as well.

VWF, located on chromosome 22, encodes a large multimeric plasma protein VWF that is involved in primary and secondary hemostasis.[32] Recent evidence indicates that VWF interacts with complement components, initiates hemostasis by promoting platelet adhesion. It is clear that complement and hemostasis are intrinsically linked at many levels; for instance, the complement factor H (CFH) bound to VWF with a high affinity in Weibel–Palade bodies of vascular endothelial cells, enhancing the cofactor function of CFH in factor I-mediated degradation of complement activation.[33] Another study concluded that normal plasma VWF contributed to factor I-mediated C3b cleavage, whereas large VWF multimers did not have this effect and permitted default complement activation.[34] Most of our knowledge about the coagulation and complement systems comes from studies with the atypical hemolytic uremic syndrome and thrombotic thrombocytopenic purpura, in which VWF interacts with C3b and activates the complement by the alternative pathway, changing the phenotype of microvascular endothelial cells.[35] The molecular mechanism about how VWF regulates the complement components in cancer remains unknown. The immune response, such as a chronic inflammation, may affect the initiation and progression of cancer-based on several experimental and clinical studies, and VWF is thought as an antitumor factor to negatively modulate angiogenesis and apoptosis.[36] However, in lung cancer, VWF has been reported to contribute to the tumorigenesis of lung adenocarcinoma by regulating inflammation.[37] In the present study, it is the first time for us to show

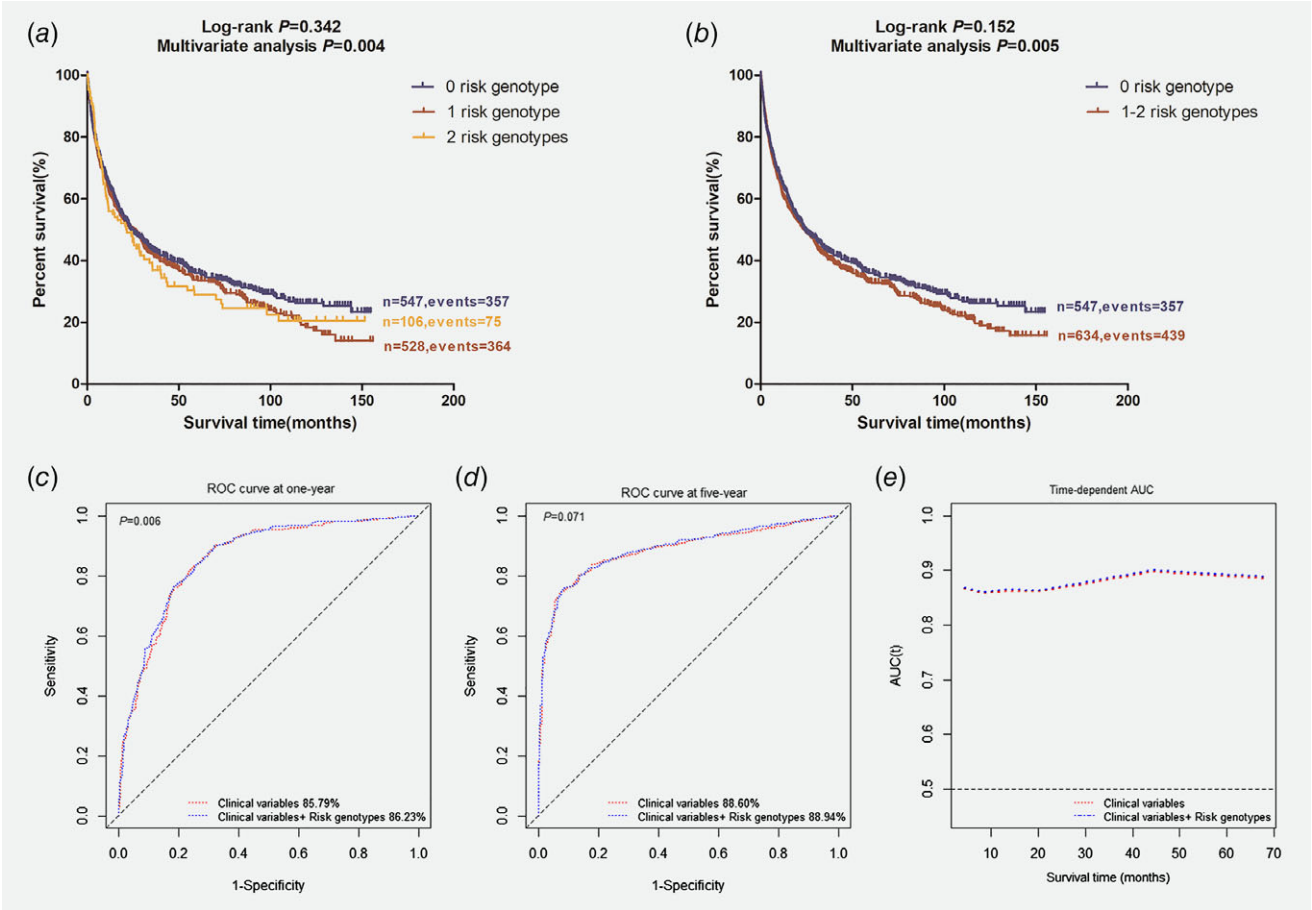

Figure 2. Kaplan–Meier analysis for patients with NSCLC by the combined risk genotypes and receiver operating characteristic (ROC) curve and time-dependent area under the curve (AUC) estimation for prediction of NSCLC survival in the PLCO dataset. (*a*) By 0, 1 and 2 risk genotypes (log-rank test for trend: *P*), (*b*) by 0 and 1–2 risk genotypes (log-rank test and multivariate analysis: *P*) in the PLCO trial, (*c*) one-year NSCLC survival prediction by ROC curve, (d) five-year NSCLC survival prediction by ROC curve, and (*e*) time-dependent AUC estimation: based on age, sex, smoking status, histology, tumor stage, chemotherapy, surgery, principal component and the risk genotypes of the four genes. Abbreviations: NSCLC, non-small cell lung cancer; ROC, receiver operating characteristic; AUC, area under the curve; PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial.

that the *VWF* rs73049469 A allele was associated with a poorer survival of NSCLC, likely due to a decreased mRNA expression. In the 1000 Genomes Project, the *VWF* rs73049469 A allele showed a significant correlation with decreased mRNA expression levels of the gene in 373 normal but transformed lymphoblastoid cells; however, there was a trend for the rs73049469 A allele to be correlated with higher mRNA expression levels ($P > 0.05$) in the GTEx in the whole blood, in addition to a difference in the sample size (GTEx was smaller and thus *P* value was not significant). Moreover, *VWF* mRNA expression was found to be lower in tumor tissues than in normal tissues in the TCGA dataset, and the low expression was associated with a worse survival in NSCLC. According to the ENCODE project, rs73049469 is located in a DNase I hypersensitive site with considerable levels of histone modification H3K27 acetylation, which may lead to an enhanced transcriptional activity. Taken together, it is likely that *VWF* may act as a tumor suppressor gene in NSCLC.

ITGB2 encodes an integrin beta chain that constitutes different integrins by combining with multiple different alpha chains. Lymphocyte function-associated antigen-1 (LFA-1, also known as CD11a/CD18) and macrophage-1 antigen (Mac-1, also known as CD11b/CD18) are the two main members of the β2-integrin family. By interacting with its major ligand ICAM-1 (intercellular adhesion molecules), LFA-1 plays an important role in tumor growth and metastasis. LFA-1 is essential for the cytotoxic immune response against tumors and mediates adhesion of cytotoxic T or NK cell to the target cell.[38] Mac-1 is a receptor for the complement 3(CR3), through an interplay with the complement fragment iC3b, having a great influence on complement-dependent phagocytosis.[39] One study found that in the early adhesive steps of liver metastasis, Mac-1 mediates the adhesion of neutrophils to cancer cells,[40] and another study found that *ITGB2* variant rs2230531 C > T detected by the whole exome sequencing contributed to the susceptibility to chronic lymphocytic leukemia.[41] However, the role of *ITGB2* in

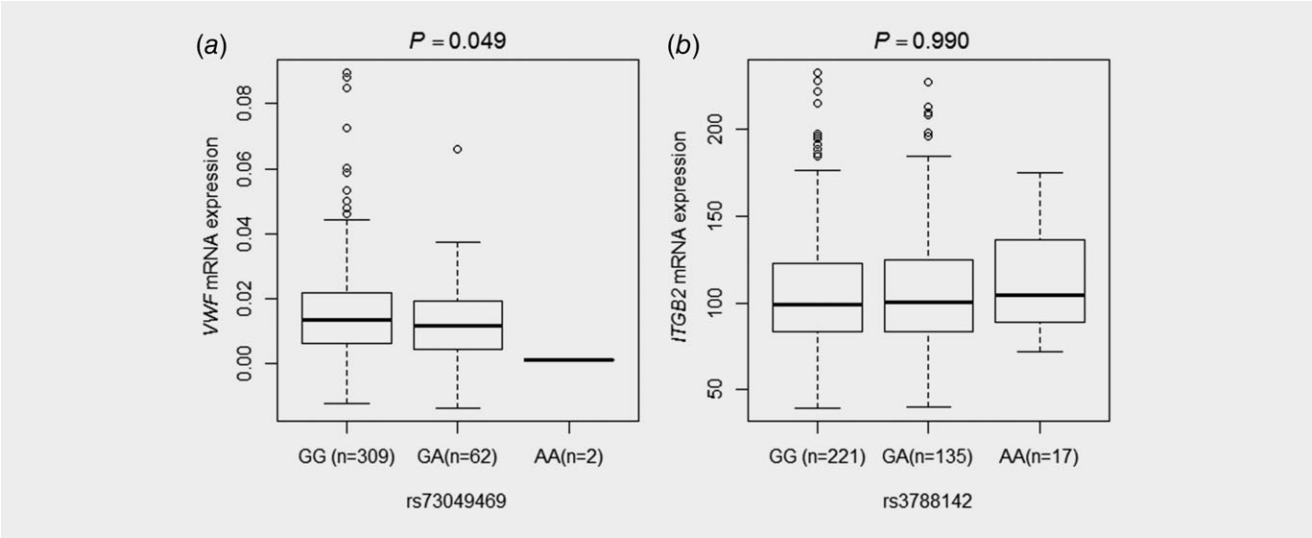

Figure 3. Correlation of the SNPs with mRNA expression in blood cells in the 1,000 Genomes Project. Correlation of (*a*) rs73049469 and (*b*) rs3788142 with mRNA expression levels of *VWF* and *ITGB2*, respectively, in established lymphoblastoid cell-lines derived from each of 373 Europeans individuals.

lung cancer remains unknown. In the present study, we observed that the rs3788142 A allele was associated with an increased mRNA expression of *ITGB2*, while mRNA expression levels were higher in normal lung tissues than in tumor tissues. It is possible that the expression of *ITGB2* in lung cancer may be affected by other factors, such as an imbalanced activation of the complement in the tumor microenvironment, which may cause abnormal expression of *ITGB2*.

In the present study, although we observed associations between genetic variants in several complement-related immunity genes and NSCLS OS backed up with some functional evidence, the exact molecular mechanisms of these SNPs are still unclear. Further biochemical studies and functional experiments are required to validate our results. Because both the discovery and validation datasets used in our study were from Caucasian populations, the results may not be generalizable to other ethnic populations. Although the sample size of PLCO was large enough, the number of patients in subgroups was still small, which would reduce the statistical power to detect a small effect in one particular stratum. Only a few clinical factors were available for additional analysis, and other information, such as performance status and details of treatments, was not available for further adjustment. Finally, detailed genotype and phenotype data of the HLCS study was not accessible for us to do the combined modeling and additional stratified analysis.

In conclusions, two independent functional SNPs (*VWF* rs73049469 C > A and *ITGB2* rs3788142 G > A) were found to be significantly associated with NSCLC OS in both the PLCO trial and HLCS GWAS datasets, possibly by a mechanism of affecting their genes expression. Our findings provided new clues for further functional studies to verify the roles of these SNPs and the genes in the development and progression of lung cancer.

Acknowledgements

The authors thank all the participants of the PLCO Cancer Screening Trial. The authors also thank the National Cancer Institute for providing the 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 dbGaP repository for providing cancer genotyping datasets. The accession numbers for the datasets of 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. Qingyi Wei was supported by the V Foundation for Cancer Research (D2017-19) and also partly supported by the Duke Cancer Institute as part of the P30 Cancer Center Support Grant (Grant ID: NIH/NCI CA014236). Danwen Qian was supported by the State Scholarship Fund, China Scholarship Council, the Ministry of Education of the P. R. China through Fudan University Shanghai Cancer Center. The Harvard Lung Cancer Susceptibility Study was supported by NIH grants 5U01CA209414, CA092824, CA074386 and CA090578 to David C. Christiani.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA Cancer J Clin* 2017;67:7–30.
2. Basumallik N, Agarwal M. *Cancer, lung, small cell (oat cell)*, Treasure Island (FL): StatPearlsed, 2018.
3. Jia M, Zhu M, Zhou F, et al. Genetic variants of JNK and p38alpha pathways and risk of non-small cell lung cancer in an eastern Chinese population. *Int J Cancer* 2017;140:807–17.
4. Zienolddiny S, Skaug V. Single nucleotide polymorphisms as susceptibility, prognostic, and therapeutic markers of nonsmall cell lung cancer. *Lung Cancer (Auckl)* 2012;3:1–14.

5. Stenzel-Bembenek A, Sagan D, Guz M, et al. Single nucleotide polymorphisms in lung cancer patients and cisplatin treatment. *Postepy Hig Med Dosw (Online)* 2014;68:1361–73.
6. Campa MJ, Gottlin EB, Bushey RT, et al. Antibodies from lung cancer patients induce complement-dependent Lysis of tumor cells. Suggesting a Novel Immunotherapeutic Strategy. *Cancer Immunol Res* 2015;3:1325–32.
7. Sharma P, Allison JP. The future of immune checkpoint therapy. *Science* 2015;348:56–61.
8. Rosenberg SA, Restifo NP. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* 2015;348:62–8.
9. June CH, O'Connor RS, Kawalekar OU, et al. CAR T cell immunotherapy for human cancer. *Science* 2018;359:1361–5.
10. Lopez-Lera A, Corvillo F, Nozal P, et al. Complement as a diagnostic tool in immunopathology. *Semin Cell Dev Biol* 2018; [Epub ahead of print].
11. Reis ES, Mastellos DC, Ricklin D, et al. Complement in cancer: untangling an intricate relationship. *Nat Rev Immunol* 2018;18:5–18.
12. Hajishengallis G, Reis ES, Mastellos DC, et al. Novel mechanisms and functions of complement. *Nat Immunol* 2017;18:1288–98.
13. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;144:646–74.
14. Lin K, He S, He L, et al. Complement component 3 is a prognostic factor of nonsmall cell lung cancer. *Mol Med Rep* 2014;10:811–7.
15. Corrales L, Ajona D, Rafail S, et al. Anaphylatoxin C5a creates a favorable microenvironment for lung cancer progression. *J Immunol* 2012;189:4674–83.
16. Ajona D, Ortiz-Espinosa S, Moreno H, et al. A combined PD-1/C5a blockade synergistically protects against lung cancer growth and metastasis. *Cancer Discov* 2017;7:694–703.
17. Zhang Y, Zhang Z, Cao L, et al. A common CD55 rs2564978 variant is associated with the susceptibility of non-small cell lung cancer. *Oncotarget* 2017;8:6216–21.
18. Hocking WG, Hu P, Oken MM, et al. Lung cancer screening in the randomized prostate, lung, colorectal, and ovarian (PLCO) cancer screening trial. *J Natl Cancer Inst* 2010;102: 722–31.
19. Oken MM, Marcus PM, Hu P, et al. Baseline chest radiograph for lung cancer detection in the randomized prostate, lung, colorectal and ovarian cancer screening trial. *J Natl Cancer Inst* 2005;97: 1832–9.
20. Tryka KA, Hao L, Sturcke A, et al. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res* 2014;42:D975–9.
21. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007;39:1181–6.
22. Zhai R, Yu X, Wei Y, et al. Smoking and smoking cessation in relation to the development of coexisting non-small cell lung cancer with chronic obstructive pulmonary disease. *Int J Cancer* 2014; 134:961–70.
23. Aulchenko YS, Ripke S, Isaacs A, et al. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007;23:1294–6.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* 1995;57: 289–300.
25. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004;96: 434–42.
26. Lappalainen T, Sammeth M, Friedlander MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 2013;501:506–11.
27. GTEx Consortium. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015;348:648–60.
28. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;511:543–50.
29. Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med* 2006;25:3474–86.
30. Wang Y, Liu H, Ready NE, et al. Genetic variants in ABCG1 are associated with survival of nonsmall-cell lung cancer patients. *Int J Cancer* 2016;138:2592–601.
31. Gyorffy B, Surowiak P, Budczies J, et al. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One* 2013;8: e82241.
32. Franchini M, Mannucci PM. Von Willebrand factor: another janus-faced hemostasis protein. *Semin Thromb Hemost* 2008;34:663–9.
33. Rayes J, Roumenina LT, Dimitrov JD, et al. The interaction between factor H and VWF increases factor H cofactor activity and regulates VWF prothrombotic status. *Blood* 2014;123:121–5.
34. Feng S, Liang X, Kroll MH, et al. von Willebrand factor is a cofactor in complement regulation. *Blood* 2015;125:1034–7.
35. Bettoni S, Galbusera M, Gastoldi S, et al. Interaction between Multimeric von Willebrand factor and complement: a fresh look to the pathophysiology of microvascular thrombosis. *J Immunol* 2017;199:1021–40.
36. Franchini M, Frattini F, Crestani S, et al. von Willebrand factor and cancer: a renewed interest. *Thromb Res* 2013;131:290–2.
37. Liu C, Zhang YH, Huang T, et al. Identification of transcription factors that may reprogram lung adenocarcinoma. *Artif Intell Med* 2017;83:52–7.
38. Reina M, Espel E. Role of LFA-1 and ICAM-1 in cancer. *Cancers (Basel)* 2017;9:153–66.
39. Mitroulis I, Kang YY, Gahmberg CG, et al. Developmental endothelial locus-1 attenuates complement-dependent phagocytosis through inhibition of mac-1-integrin. *Thromb Haemost* 2014;111:1004–6.
40. Spicer JD, McDonald B, Cools-Lartigue JJ, et al. Neutrophils promote liver metastasis via Mac-1-mediated interactions with circulating tumor cells. *Cancer Res* 2012;72:3919–27.
41. Goldin LR, McMaster ML, Rotunno M, et al. Whole exome sequencing in families with CLL detects a variant in integrin beta 2 associated with disease susceptibility. *Blood* 2016;128:2261–3.