Published in final edited form as:

Cancer Immunol Immunother. 2021 October ; 70(10): 2819–2833. doi:10.1007/s00262-021-02877-9.

Potentially functional variants of *ERAP1*, *PSMF1* and *NCF2* in the MHC-I related pathway predict non-small cell lung cancer survival

Sen Yang^{1,2,3,*}, Dongfang Tang^{2,3,*}, Yu Chen Zhao^{2,3}, Hongliang Liu^{2,3}, Sheng Luo⁴, Thomas E. Stinchcombe^{2,5}, Carolyn Glass^{2,6}, Li Su⁷, Sipeng Shen⁷, David C. Christiani^{7,8}, Qiming Wang^{1,**}, Qingyi Wei^{2,3,5,**}

¹Department of Internal Medicine, Affiliated Cancer Hospital of Zhengzhou University, Henan Cancer Hospital, Zhengzhou, China

²Duke Cancer Institute, Duke University Medical Center, Durham, NC 27710, USA

³Department of Population Health Sciences, Duke University School of Medicine, Durham, NC 27710, USA

⁴Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA

⁵Department of Medicine, Duke University Medical Center, Durham, NC 27710, USA

⁶Department of Pathology, Duke University School of Medicine, Durham, NC 27710, USA

⁷Departments of Environmental Health and Department of Epidemiology, Harvard School of Public Health, Boston, MA, 02115 USA

⁸Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

Abstract

Background: Cellular immunity against tumor cells is highly dependent on antigen presentation by major histocompatibility complex class I (MHC-I) molecules. However, few published studies

**Correspondence to: Qiming Wang, Department of Internal Medicine, Affiliated Cancer Hospital of Zhengzhou University, Henan Cancer Hospital, Zhengzhou, China; qimingwang1006@126.com; or Qingyi Wei, Duke Cancer Institute, Duke University Medical Center and Department of Population Health Sciences, Duke University School of Medicine, 905 S LaSalle Street, Durham, NC 27710, USA, Tel.: (919) 660-0562, qingyi.wei@duke.edu.

Authors' contributions

Sen Yang, Dongfang Tang, Qiming Wang, Qingyi Wei contributed to the study conception and design. Material preparation, data collection and analysis were performed by Sen Yang, Dongfang Tang, Yu Chen Zhao, Hongliang Liu, Sheng Luo, Thomas E. Stinchcombe., Carolyn Glass, Li Su, Sipeng Shen, David C. Christiani, Qiming Wang, Qingyi Wei. The first draft of the manuscript was written by Sen Yang and Dongfang Tang and critically reviewed and revised by Qiming Wang and Qingyi Wei, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

*Sen Yang and Dongfang Tang contributed equally to this work.

Conflicts of interest/Competing interests

The authors declare no potential conflicts of interest.

Availability of data and material

The PLCO dataset used for the analyses described in the present study was obtained from dbGaP (<http://www.ncbi.nlm.nih.gov/gap>) through dbGaP accession number phs000336.v1.p1 and phs000093.v2.p2. The HLCS dataset used for the replication will be made available upon reasonable request.

Code availability

Not applicable.

have investigated associations between functional variants of MHC-I-related genes and clinical outcomes of lung cancer patients.

Methods: We performed a two-phase multivariate Cox proportional hazards regression analysis by using two previously published genome-wide association studies to evaluate associations between variants in the MHC-I-related gene-set and the survival of non-small cell lung cancer (NSCLC) patients, followed by expression quantitative trait loci analysis.

Results: Of the 7,811 single-nucleotide polymorphisms (SNPs) identified in 89 genes of 1,185 NSCLC patients in the discovery, 24 SNPs remained statistically significant after validation in additional 984 NSCLC patients from the Harvard Lung Cancer Susceptibility study. In a multivariate stepwise Cox model, three independent functional SNPs (*ERAPI* rs469783 T>C, *PSMFI* rs 13040574 C>A and *NCF2* rs36071574 G>A) remained significant with an adjusted hazards ratio (HR) of 0.83 [95% confidence interval (CI)=0.77-0.89, $P=8.0\times 10^{-7}$], 0.86 (0.80-0.93, $P=9.4\times 10^{-5}$) and 1.31 (1.11-1.54, $P=0.001$) for overall survival (OS), respectively. Further combined genotypes revealed a poor survival in a dose-response manner in association with the number of unfavorable genotypes ($P_{\text{trend}} < 0.0001$ and 0.0002 for OS and disease-specific survival, respectively). Also, *ERAPI* rs469783C and *PSMFI* rs13040574A alleles were associated with higher mRNA expression levels of their genes.

Conclusion: These potentially functional SNPs of the MHC-I-related genes may be biomarkers for NSCLC survival, possibly through modulating the expression of corresponding genes.

Keywords

Non-small cell lung cancer; MHC-I pathway; survival; genome-wide association study; single-nucleotide polymorphisms

Introduction

Lung cancer is one of the most common malignancies in humans, with the highest cancer-related mortality worldwide [1]. It is estimated that there will be approximately 228,150 new cases and 142,670 deaths from lung cancer in the United States in 2019 [2]. Non-small cell lung cancer (NSCLC) is the most common histological type, accounting for approximately 90% of all lung cancer patients [3]. Current treatment options for advanced NSCLC include chemotherapy, radiotherapy and target therapy, but the first two treatments only modestly improve survival, while the patients inevitably develop resistance to the latter treatment targeting several driver mutations responsible for tumor progression [4–6]. In recent years, the role of the immune system in cancer development and progression has been widely recognized [7–9]. Immunotherapy is now established as the “fourth pillar” of cancer treatment alongside surgery, radiation, and chemotherapy [10]. Immunotherapy alone in patients with a high level of PD-L1 expression or in combination with chemotherapy is now the standard first-line therapy for patients with metastatic NSCLC [11–13]. However, many patients do not benefit from the current immunotherapy, and thus there is an urgent need to identify survival-related predictive biomarkers to maximize the benefits of immunotherapy.

Major histocompatibility complex class I (MHC-I) proteins are central mediators in cellular immunity as they govern cytotoxic T lymphocyte (CTL) function through the process of

antigen presentation and serve as markers for natural killer (NK) cells. For example, an effective CTL response relies on the ability of MHC-I proteins to present a diverse array of peptides [14]. On the other hand, the killing effect of the cellular immunity in tumor cells is highly dependent on MHC-I activation of CTL on the surface of cancer cells and dendritic cells, which means MHC-I molecules expose the intracellular protein content on the cell surface, allowing T cells to detect foreign or mutated peptides [15]. Due to the important role that MHC-I proteins play in cellular immunity, we hypothesize that genetic variants of the genes involved in the MHC-I pathway in the process of tumor antigen presentation are associated with NSCLC survival.

Since the hypothesis-free genome-wide association study (GWAS) only focuses on the top or most-significant SNPs/genes with a stringent P value after correction for multiple tests, most of the identified top SNPs lack of functional annotations. Up to date, few novel functional SNPs have been identified in associations with prognosis of lung cancer patients in GWASs in populations of European descent [16–19]; thus, as a promising hypothesis-driven method in the post GWAS era, the biological pathway-based approach has been applied to reanalyze published GWAS datasets to assess the cumulative effect of SNPs across multiple genes in the same biological pathway. Since much fewer SNPs in candidate genes of a significant biological pathway will be included in the analysis, unnecessary multiple tests for SNPs that may hold no apparent biological significance would be avoided, which improves the overall study power to identify both statistically significant and biologically important associations.

Material and Methods

Study populations

In the two-phase analysis, we used the GWAS dataset of lung cancer patients of European descent from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial as the discovery dataset. The PLCO is a randomized control study conducted by the National Cancer Institute (NCI), which included 77,500 men and 77,500 women aged 55–74 years, enrolled between the years of 1993 and 2011 from 10 medical centers in the United States. All the participants were randomized into either the intervention arm that received a trial screening or the control arm that received standard care instead, who were then followed up for at least 13 years after the enrollment [20]. Blood samples and personal information about smoking status and family history were collected at enrollment, and cancer diagnosis, histopathology, tumor stage and treatment method were collected in the follow-up period [21]. In these 150,000 patients, a total of 1,185 patients with NSCLC were eligible for survival analysis after excluding two individuals who had no follow-up information. Genomic DNA extracted from the whole blood samples of these participants were genotyped with Illumina HumanHap240Sv1.0 and HumanHap550v3.0 (dbGaP accession: phs000093.v2.p2 and phs000336.v1.p1) [22,23]. All the participants had provided a written informed consent permitting for the PLCO trial to use of the datasets, and each institutional review board of participating institutions had approved the use of the datasets.

Another GWAS dataset of 984 histology-confirmed Caucasian NSCLC patients from the Harvard Lung Cancer Susceptibility (HLCS) Study initiated in 1992 was used as the validation [24]. In the HLCS study, the whole blood samples and personal information were collected after the diagnosis, and DNA from the blood samples were extracted with Auto Pure Large Sample nucleic acid purification system (QIAGEN Company, Venlo, Limburg, Netherlands) and genotyped by using the Illumina Humanhap610-Quad array. The genotyping data were used for imputation with the Mach3 software based on the sequencing data from the 1,000 Genomes Project [25].

The use of these two GWAS datasets in the present study was approved by both the Internal Review Board of Duke University School of Medicine (Project #Pro00054575) and the National Center for Biological Information (NCBI) for the access to the NCBI dbGaP database of genotypes and phenotypes (Project #6404). The comparison of the characteristics between the PLCO trial (n=1185) and the HLCS study (n=984) is presented in Supplementary Table 1.

Gene and SNP selection

The genes involved in the MHC-I-related pathway were selected by the Molecular Signatures Database (<http://software.broadinstitute.org/gsea/msigdb/index.jsp>) with the keyword “MHC class I AND peptide”. After removal of 85 duplicated genes, one pseudogene and three genes in the X chromosome, 89 genes remained as candidate genes for further analysis (Supplementary Table 2). Imputation with IMPUTE2 and the 1,000 Genomes Project data (phase 3) was performed for these candidate genes. After that, SNPs within these genes and their ± 2 kb flanking regions were extracted with the following criteria: an imputation info score ≥ 0.8 (Supplementary Figure 1), a genotyping rate $\geq 95\%$, a minor allelic frequency (MAF) $\geq 5\%$, and Hardy–Weinberg equilibrium (HWE) $\geq 1 \times 10^{-5}$. As a result, a total of 7,811 SNPs (527 genotyped and 7,284 imputed) were selected from the PLCO GWAS dataset (dbGaP accession: phs000093.v2.p2 and phs000336.v1.p1) for further analysis. The HLCS study was used as the validation dataset, and the SNP inclusion criteria for the HLCS genotyping data was the same as the PLCO genotyping data.

Statistical analysis

For the survival analysis, the time to event was from the date patient was diagnosed as having NSCLC to the date of patient death. Participants were enrolled between 1993 and 2001 and subsequently followed up locally by that screening center through 2011. Two patients were lost during the follow-up. The covariates to be adjusted in both discovery and validation included age, sex, smoking status and histology, and the age used was age at diagnosis. In the single-locus analysis, we first used multivariate Cox proportional hazards regression analysis to assess the association between each of the 7,811 SNP and NSCLC survival in an additive genetic model, with adjustment for age, sex, smoking status, histology, tumor stage, chemotherapy, radiotherapy, surgery and the first four of the ten principal components (PCs) in the PLCO dataset (Supplementary Table 3) by using the GenABEL package of R software [26]. For controlling multiple comparison in the discovery, we chose the measure from the strictest (Bonferroni correction) to the least strict [Bayesian false discovery probability (BFDP)]. After the failure from Bonferroni

Correction and false discovery rate of 0.2, we chose BFDLP to maximize the number of SNPs to be validated. Due to the implementation of imputation, we used BFDLP with a cut-off value of 0.80 for multiple testing correction to lower the probability of discovering potentially false positive results [27]. We assigned a prior probability of 0.10 to detect a hazards ratio (HR) of 3.0 for an association with variant genotypes or minor alleles of the SNPs with $P < 0.05$. After that, we validated those chosen SNPs by using the HLCS GWAS dataset. Next, we performed an inverse variance weighted meta-analysis to combine the results of both discovery and validation datasets. In the meta-analysis, Cochran's Q-test and the heterogeneity statistic (I^2) were used to assess the inter-study heterogeneity. If no heterogeneity was observed between the two datasets ($P_{\text{het}} > 0.10$ and $I^2 < 50\%$), a fixed-effects model was implemented; otherwise, a random-effects model was applied. Furthermore, a multivariate stepwise Cox model, including available demographic and clinical variables as well as the first four principal components of the PLCO dataset, was performed to identify independent SNPs, and then the model was further adjusted for previously published survival-predictive SNPs from the same PLCO dataset. The results of selected SNPs are summarized in Manhattan plots and the regional association plots.

Next, we used the combined genotypes to evaluate the cumulative effects of the identified SNPs and the Kaplan-Meier curve to estimate the overall survival (OS) and disease-specific survival (DSS) probability associated with the combined genotypes. We also assessed possible interactions with a χ^2 -based Q-test between subgroups in the stratified analysis. We then performed the receiver operating characteristic (ROC) curve and time-dependent area under the curve (AUC) with the timeROC package of R software (version 3.5.0) to illustrate the prediction accuracy of the model integrating the effects of both clinical and genetic variables on NSCLC survival [28]. To evaluate the correlations between SNPs and the corresponding mRNA expression levels, we performed the expression quantitative trait loci (eQTL) analyses with a linear regression model by using the R software. The mRNA expression data were obtained from two sources: 373 European individuals included in the 1,000 Genomes Project as well as 369 whole blood samples and 383 normal lung tissues included in the genotype-tissue expression (GTEx) project [29,30]. Additional bioinformatics functional prediction for the tagging SNPs were performed with SNPinfo [31], RegulomeDB [32] (<http://www.regulomedb.org>) and HaploReg [33] (<http://archive.broadinstitute.org/mammals/haploreg/haploreg.php>). We also evaluated associations between the mRNA expression levels of the three genes and those of *CD8A* and *CD274* in normal lung tissues in the TCGA cohorts.

Finally, the differences in mRNA expression levels were examined in 109 pairs of lung cancer tissues and adjacent normal tissues from the Cancer Genome Atlas (TCGA) database by using a paired *t* test model. We also assessed the differences in mRNA expression levels in a larger, but not paired, dataset from TCGA (<http://ualcan.path.uab.edu>), and Kaplan-Meier survival analysis was performed to assess the association between the mRNA expression levels and survival probability (<http://kmplot.com/analysis/index.php?p=service&cancer=lung>). All statistical analyses were performed with a statistical significance of $P < 0.05$ by using the SAS software (version 9.4; SAS Institute, Cary, NC, USA), unless otherwise indicated. Bonferroni correction statistical significance was defined as $P > 0.05$.

Data availability

The PLCO dataset used for the analyses described in the present study was obtained from dbGaP (<http://www.ncbi.nlm.nih.gov/gap>) through dbGaP accession number phs000336.v1.p1 and phs000093.v2.p2. The HLCS dataset used for the replication will be made available upon reasonable request.

Results

Associations between SNPs in the MHC-I-related pathway genes and NSCLC survival

The workflow chart of the present study is shown in Figure 1. The basic characteristics of 1,185 NSCLC patients from the PLCO trial and 984 NSCLC patients from the HLCS study have been described elsewhere [34] and are also shown in Supplementary Table 1. In the discovery PLCO genotype dataset, a single-locus multivariate Cox regression analysis was performed for the selected 7,811 SNPs. For multiple testing correction, none of the SNPs passed Bonferroni Correction ($P > 0.05$) or false discovery rate ($P > 0.20$). This is likely due to the high LD among the SNPs generated by imputation. Besides, our purpose of using this pre-screening was to identify functional candidate SNPs for further analysis. Therefore, we used the BFDP method as recommended by the authors of the method. With multiple testing correction by BFDP 0.80, we identified 206 SNPs to be significantly associated with NSCLC OS ($P < 0.05$), of which 24 SNPs remained significant after further validated by the HLCS genotype dataset (Figure 1). Subsequently, we performed a combined-analysis of the PLCO and HLCS datasets for these 24 newly identified SNPs and found that a better survival was associated with SNPs in *ERAPI* and *PSMF1*, while a poorer survival was associated with SNPs in *NCF2*, without heterogeneity between the two studies (Table 1).

Independent SNPs associated with NSCLC survival in the PLCO dataset

When the 24 validated SNPs were all included in the multivariate stepwise Cox regression model for the PLCO dataset (because the HLCS study dataset did not have the detailed genotyping data), only three SNPs were left to be significantly associated with survival. Then, we expanded the model by further including other 15 previously reported survival-predictive SNPs from the same PLCO dataset, and these three newly identified SNPs remained significantly associated with survival (Table 2). The results of selected SNPs are summarized in a Manhattan plot (Supplementary Figure 2), and the regional association plot for each of these three SNPs is shown in Supplementary Figure 3.

In the PLCO dataset with complete adjustment for available covariates, patients with the *ERAPI* rs469783C allele or *PSMF1* rs13040574A allele had a decreased risk of death or a better survival [$P_{\text{trend}}=0.0003$ for OS; $P_{\text{trend}}=0.0008$ for DSS and $P_{\text{trend}}=0.006$ for OS; $P_{\text{trend}}=0.014$ for DSS, respectively], while patients with the *NCF2* rs36071574A allele had an increased risk of death or a worse survival ($P_{\text{trend}}=0.018$ for OS and $P_{\text{trend}}=0.012$ for DSS) (Table 3). Specifically, compared with the TT genotype, the *ERAPI* rs469783 C variant genotypes were associated with a better survival (TC: HR=0.94, 95% CI=0.81-1.10, $P=0.462$ for OS and 0.95, 0.81-1.12, 0.539 for DDS; CC: 0.65, 0.52-0.80, <0.0001 for OS and 0.64, 0.51-0.81, 0.0002 for DSS; and TC+CC: 0.86, 0.74-0.99, 0.041 for OS and 0.86, 0.74-1.01, 0.063 for DSS), while compared with the CC genotype, *PSMF1* rs13040574 A

variant genotypes were associated with a better survival (CA: HR=0.79, 95% CI=0.66-0.93 and $P=0.005$ for OS and 0.83, 0.69-0.99, $P=0.038$ for DSS; AA: 0.75, 0.61-0.92, 0.007 for OS and 0.76, 0.61-0.95, 0.015 for DDS; and CA+AA: 0.78, 0.66-0.91, 0.002 for OS and 0.81, 0.68-0.96, 0.015 for DSS) (Table 3). In contrast, compared with the GG genotype, *NCF2* rs36071574 A variant genotypes were associated with a worse survival (GA: HR=1.26, 95% CI=1.00-1.60, $P=0.051$ for OS and 1.29, 1.01-1.65, 0.043 for DDS; AA: 2.35, 0.87-6.38, 0.093 for OS and 2.68, 0.99-7.29, 0.053 for DSS; and GA+AA: HR=1.29, 95% CI=1.03-1.62, $P=0.030$ for OS and 1.32, 1.04-1.68, 0.022 for DSS) (Table 3).

Combined effects of the three independent SNPs in the PLCO dataset

Because the available validation dataset from the HLCS study lacked detailed genotyping data, we used the PLCO dataset to assess the combined effect of the three independent SNPs on NSCLC OS and DSS. For the combined analysis, we reversed some of the alleles based on their actual effect direction to align the allele effect direction of different loci. First, we combined the unfavorable genotypes (i.e., *ERAP1* rs469783 TT, *PSMF1* rs13040574 CC and *NCF2* rs136071574 GA+AA) into a number of unfavorable genotypes (NUGs) score. As shown in Table 3, the increased NUG score was associated with a worse survival as assessed in the multivariate analysis of the PLCO dataset ($P_{\text{trend}} < 0.0001$ and 0.0002 for OS and DSS, respectively) (Table 3).

Then, to facilitate further stratification analysis, we used the dichotomized NUG score to divide all the patients into 0-1 score and 2-3 scores. Compared with the 0-1 score group, the score 2-3 group had a significantly worse survival (HR=1.52, 95% CI=1.24–1.86, $P < 0.0001$ for OS and 1.51, 1.22-1.87, 0.0001 for DSS) (Table 3). We further presented Kaplan-Meier survival curves to depict these associations between unfavorable genotypes and NSCLC OS and DSS (Figure 2a–2d).

We also analyzed the associations between genetic score and survival of NSCLC in the PLCO dataset. The genetic score was calculated by adding number of unfavorable genotypes and weighted by HR. As shown in Supplementary Table 4, an increasing genetic score was associated with a decreasing survival as assessed in the multivariate analysis of the PLCO dataset ($P_{\text{trend}} < 0.0001$ and 0.0002 for OS and DSS, respectively). We also dichotomized this genetic score into two groups of 0-1.29 and 2.44-3.73 scores. Compared with the 0-1.29 group, the 2.44-3.73 group had a significantly worse survival (HR=1.50, 95% CI=1.20-1.87, $P=0.0004$ for OS and 1.47, 1.16-1.86, 0.002 for DSS).

Stratified analysis for associations between NUGs and NSCLC survival

To evaluate whether the combined effect of unfavorable genotypes on NSCLC OS and DSS were modified by other covariates, we performed stratified analysis by age, sex, smoking status, histology, tumor stage, chemotherapy, radiotherapy and surgery in the PLCO dataset. As a result, there was no obvious difference in survival between the strata of these covariates, and no significant interactions were found on both NSCLC OS and DSS between the strata ($P > 0.05$ for all strata, Supplementary Table 5).

The ROC curves and time dependent AUC

We further assessed the predictive value of the three SNPs with time-dependent AUC and ROC curves at the 60th month (or five-year survival) in the PLCO dataset (with the follow-up time between 0.03 and 155.83 months and the median follow-up time of 19.80 months). Compared with the model of covariates including age, sex, smoking status, histology, tumor stage, chemotherapy, radiotherapy, surgery and first four principal components, the time-dependent AUC plot with addition of the independent SNPs did not improve prediction performance of the model at the 60th month (or five-year survival). However, when we performed the time-dependent AUC and ROC curves at the 120th month (or ten-year survival) with tumor grade as an additional covariate in the PLCO dataset, the prediction performance of the model was improved significantly. The AUCs changed from 84.41% to 86.68% ($P=0.022$) for OS and from 84.88% to 87.08% ($P=0.030$) for DSS (Supplementary Figure 4), suggesting that these newly identified SNPs contributed to the overall prediction of survival of NSCLC patients in the PLCO dataset.

The eQTL analysis

Subsequently, we performed the eQTL analysis to explore the correlations between genotypes of the newly identified three survival predictive SNPs and their corresponding mRNA expression levels. In the RNA-Seq data of lymphoblastoid cell lines from 373 European descendants available in the 1,000 Genomes Project, the *ERAP1* rs469783C and *PSMF1* rs13040574A alleles both showed a significant correlation with increased mRNA expression levels of their genes ($P<2.0\times10^{-16}$ and $P=0.004$, respectively; Figure 2e and 2f) [29]; however, the correlation between the *NCF2* rs36071574A allele and the mRNA expression levels was not significant ($P=0.701$) (Supplementary Figure 5a) [29]. Then, we performed eQTL by using the data of 369 whole blood samples and 383 normal lung tissue from the GTEx project and found that the rs469783C allele remained correlated with higher expression levels of *ERAP1* in lung normal tissues ($P=6.18\times10^{-15}$) and whole blood ($P=1.18\times10^{-19}$) (Figure 2g and 2h) [30]; however, there was no significant correlation between the *PSMF1* rs13040574A allele with the mRNA expression in both normal lung tissues ($P=0.933$) and whole blood ($P=0.498$) (Supplementary Figure 5d and 5e), nor for the *NCF2* rs36071574A allele and the mRNA expression levels in normal lung tissues ($P=0.326$) or whole blood ($P=0.671$) (Supplementary Figure 5b and 5c) [30]. Next, we assessed the eQTL for the SNPs that are in high linkage disequilibrium (LD) ($r^2>0.80$) with *NCF2* rs36071574 to determine if SNPs in high LD with rs36071574 could have an effect on the mRNA expression of *NCF2*. We found that the *NCF2* rs12094228 G allele, which is in high LD with the *NCF2* rs36071574 A allele ($r^2=0.93$), was associated with a lower expression level of *NCF2* in lung normal tissues ($P=0.027$) (Supplementary Figure 6b) [30]. SNPs in high LD with *NCF2* rs36071574 are summarized in Supplementary Table 6 [32].

Finally, we performed functional prediction for the four identified SNPs (including rs12094228 in high LD with rs36071574) using the online tools of SNPinfo, RegulomeDB, and Haploreg to predict their bioinformatics function. Although all these four SNPs had no obvious functions as predicted by SNPinfo, they had some predicted bioinformatics functions based on RegulomeDB and Haploreg. For example, *ERAP1* rs469783 T>C had an effect on enhancer histone marks, DNase and motifs; *PSMF1* rs13040574 C>A had

an effect on motifs; *NCF2* rs36071574 G>A had an effect on enhancer histone marks and motifs; and *NCF2* rs12094228 T>G had an effect on enhancer histone marks. Details on their corresponding biological function prediction are summarized in Supplementary Figure 7 and Supplementary Table 7.

Differential mRNA expression analysis

Finally, we assessed the mRNA expression levels of the three genes identified by the SNPs in 109 pairs of tumor and adjacent normal tissue samples in NSCLC obtained from the TCGA database and in non-paired tumor and normal tissue samples in the UALCAN database (<http://ualcan.path.uab.edu/>). We also assessed the association between the mRNA expression levels and survival probability in the Kaplan-Meier Plotter database (www.kmplot.com). The probes were chosen according to the web recommendation (probes for *ERAPI*, *PSMF1* and *NCF2* were 209788_s_at, 236012_at and 209949_at; respectively). As shown in Figure 3a and Supplementary Figure 8a and 8b, compared with adjacent normal tissues, tumor tissues had a lower mRNA expression level of *ERAPI* in all the samples, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) ($P=0.0001$, $P=0.029$ and $P=0.002$, respectively). In the UALCAN (<http://ualcan.path.uab.edu/>) database, the mRNA expression levels of *ERAPI* were also much lower in tumor tissues in LUAD ($P=0.059$) and in LUSC ($P=7.8\times10^{-5}$) (Supplementary Figure 8c and 8d). Moreover, the higher expression levels of *ERAPI* mRNA were associated with a better NSCLC survival (Figure 3d). Similarly, as shown in Figure 3b and Supplementary Figure 9a and 9b, compared with adjacent normal tissues, tumor tissues had a higher mRNA expression level of *PSMF1* in the combined LUAD and LUSC samples ($P=0.006$) and in LUSC ($P<0.0001$) but not in LUAD ($P=0.571$). In the UALCAN (<http://ualcan.path.uab.edu/>) database, the results were also similar; that is, mRNA expression levels of *PSMF1* in tumor tissues were higher in LUSC tissues ($P=3.3\times10^{-7}$), but not in LUAD ($P=0.340$), compared with normal tissues (Supplementary Figure 9c and 9d). However, the higher expression levels of *PSMF1* mRNA were associated with a better NSCLC survival, which may be due to PSMF1 as part of proteasome inhibitor PI31 was upregulated in tumor tissue indirectly (Figure 3e). Lastly, as shown in Figure 3c and Supplementary Figure 10a and 10b, compared with adjacent normal tissues, tumor tissues had a lower mRNA expression level of *NCF2* in the combined LUAD and LUSC samples ($P<0.0001$), in LUAD ($P<0.0001$) and in the LUSC ($P<0.0001$). In the UALCAN (<http://ualcan.path.uab.edu/>) database, the results were also similar; that is, mRNA expression levels of *NCF2* in tumor tissues were lower in LUAD ($P<1.0\times10^{-12}$) and in LUSC ($P=1.6\times10^{-12}$), compared with normal tissues (Supplementary Figure 10c and 10d). Besides, the higher expression levels of *NCF2* mRNA were associated with a better NSCLC survival as well (Figure 3f). Since the three genes have been shown to be associated with *CD8A* and *CD274*, [35] we also evaluated associations between expressions levels of the identified genes and those of *CD8A* and *CD274* in normal lung tissues in the TCGA dataset (Supplementary Figure 11), and we found that the expression levels of *PSMF1* were negatively correlated with the expression levels of *CD274* ($P=0.020$), but such a correlation was not observed for other pairs of the genes.

Discussion

In the present study, we tested the hypothesis that SNPs in the MHC-I-related gene-set are associated with NSCLC survival. By using two previously published and publicly available lung cancer GWAS datasets, we identified and validated three independent SNPs (i.e., *ERAPI* rs460783 T>C, *PSMF1* rs13040574 C>A and *NCF2* rs36071574 G>A) that were significantly associated with NSCLC survival in Caucasian populations. In subsequent functional genotype-mRNA expression correlation analysis, we found that the low death-risk-associated rs469783C and rs13040574A alleles were also associated with higher mRNA expression levels of *ERAPI* and *PSMF1* in lymphoblastoid cell lines, respectively, while the rs12094228G allele, which is in high LD with the *NCF2* rs3743254A risk allele, was associated with lower mRNA expression levels of *NCF2* in normal lung tissues. These SNP may affect the mRNA expression by having effect on enhancer histone marks, DNase or motifs according to the functional prediction for these SNPs with the online tools of SNPinfo, RegulomeDB, and HaploReg (Supplementary Figure 7 and Supplementary Table 7).

We identified *ERAPI* as a potential suppresser gene in NSCLC, considering that *ERAPI* mRNA expression levels were higher in normal tissues than in tumor tissues and that a higher level of *ERAPI* expression was associated with a better survival. These findings were consistent with the results of the low death-risk associated *ERAPI* rs469783C allele being associated with higher mRNA expression levels of *ERAPI* and low risk of death in NSCLC patients from both the PLCO trial and HLCS study.

PSMF1 is also more likely to be a potential suppresser gene in NSCLC, considering that higher levels of *PSMF1* mRNA expression were associated with a better survival in patients with NSCLC, and the expression levels of *PSMF1* were negatively correlated with the expression levels of *CD274* that is also known as *PD-L1*. Higher *PSMF1* mRNA expression in LUSC might be due to deregulation of the ubiquitin proteasome system in tumor tissues, because PSMF1, as a proteasome inhibitor, is expected to be upregulated in cancer cells [36]. These counterintuitive results might be due to the fact that these findings were from different study populations. The actual mechanism underlying this phenomenon warrants further investigations. The findings associated with *PSMF1* were also consistent with the results of the protective effect from the *PSMF1* rs13040574A allele being associated with a higher mRNA expression of *PSMF1* and a lower risk of death in patients with NSCLC from both the PLCO trial and the HLCS study.

NCF2 is also identified as a potential suppresser gene in NSCLC – *NCF2* mRNA expression levels were higher in normal tissues than in tumor tissues, and a higher expression level of *NCF2* was associated with a better survival. This observation was consistent with the results that the death-risk associated *NCF2* rs3743254A allele was associated with lower mRNA expression levels of *NCF2* and high risk of death in NSCLC patients from both the PLCO trial and HLCS study.

Overall, these findings suggest that functional genetic variants in the MHC-I-related pathway genes may have played roles in NSCLC progression, possibly through a mechanism

of modulating the expression of the genes, such as *ERAP1*, *PSMF1* and *NCF2*, which may provide new scientific insights into the management and treatment of NSCLC patients, if validated in additional investigations.

ERAP1, located on chromosome 5q15, belongs to the oxytocinase subfamily of M1 metalloproteases and is a critical gene involved in protein processing and transport [37]. ERAP1 could trim peptides within the endoplasmic reticulum so that they can be loaded onto MHC-I. These peptides are attached to MHC-I in the endoplasmic reticulum and exported to the cell surface, where they are presented to the immune system. After the immune system recognizes the peptides (such as viral or bacterial peptides), it responds by triggering the self-destruction process of the cell [38–40]. Defects in the expression and function of *ERAP1* have been detected in various solid and hematological tumors, including melanoma, leukemia-lymphomas and cancers of the breasts, colon, lung, skin, chorion, cervix, prostate, kidneys and bladder [41]. Low expression levels of *ERAP1* have been associated with a poor clinical outcome of patients affected by triple-negative breast cancer [42] as well as with a worse OS and metastases in cervical carcinoma patients [43]. In esophageal carcinoma lesions, the decreased expression of *ERAP1* was significantly associated with the depth of tumor invasion; additionally, the expression of ERAP1 was either lost or reduced in 20 and 28% of the patients, respectively [44]. In a recent study of the role of ERAP1 in T-cell mediated tumor rejection, ERAP1 was found to control the efficacy of adoptive T-cell transfer in a mouse model of genetically depleted ERAP1 by operating both on the direct antigen presentation in tumor cells and on the antigen cross-presentation in the adoptive T-cell transfer recipient's cells [45], suggesting that ERAP1 is required for the proliferation of CD8⁺ T cell after adoptive T-cell transfer and its lack in transplanted recipients results in a failure of adoptive T-cell therapy [45]. Although the expression of ERAP1 is frequently altered in tumors as compared to their normal counterparts, with low expression associated with poor prognosis, the contribution of these enzymes to tumor growth and the activation of anti-tumor immune responses are still not well understood.

PSMF1, located on chromosome 20p13, encodes the proteasome inhibitor PI31 subunit that inhibits the proteasome activities by either direct binding to the outer rings of the 20S proteasome or by competing with the activating particles for 20S binding [46–48]. The ubiquitin proteasome system (UPS) has emerged as an important regulator for the targeted degradation of proteins involved in diverse cellular processes such as cell cycle control, gene transcription, DNA repair, and apoptosis induction [36]. Deregulation of the UPS has been reported in numerous types of cancer [49,50]. Some of the main players of the UPS and the mechanisms are postulated to drive cancer formation [36]. Based on evidence presented in the past and the results we presented here, we consider *PSMF1* a proteasome inhibitor and a potential tumor suppressor gene.

NCF2, located on chromosome 1q25.3, encodes a protein neutrophil cytosol factor 2, a subunit of a multi-protein complex known as nicotinamide adenine dinucleotide phosphate (NADPH) oxidase. This oxidase produces a burst of superoxide that is delivered to the lumen of the neutrophil phagosome. NADPH oxidase has been shown to regulate antigen processing and MHC-I cross-presentation in dendritic cells; however, whether NADPH

oxidase regulates this process by modulating the phagosome pH or redox microenvironment is currently under debate [51].

These results shown that these identified survival-associated loci are likely to have an effect on prognosis of NSCLC patients, and these loci could be used as potential biomarkers of prognosis for clinically monitoring the outcomes of the treatment. However, clinical utility of these variants in the progression of NSCLC need to be further validated by other investigators. There are several limitations in the present study. Firstly, although several genetic variants backed up with *in silico* functional evidence in the MCH-I-related genes were found to be associated with NSCLS survival, the exact molecular mechanisms of these SNPs underlying the observed associations are still unclear. Secondly, both discovery and validation datasets were from Caucasian populations; therefore, our results may not be generalizable to other ethnic populations. Thirdly, though some clinical factors were available in the analysis of the PLCO dataset, there are still some information such as performance status, nutritional status and specific treatments, such as immunotherapy, that were not available for further adjustments and stratification analysis. Finally, some of the significant SNPs from the PLCO trial were not validated in the HLCS study, which might be due to the minor yet noticeable differences between either the characteristics or limited sample sizes of the two included study populations. Additional validation by studies with larger sample sizes are needed to confirm these findings; however, the findings in the present study could provide new insights for additional functional studies that will unravel the potential of these genetic variants of MHC-I pathway genes as promising predictors of survival in NSCLC patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank all the participants of the PLCO Cancer Screening Trial. We 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 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 these studies can be found in the supplemental data.

Funding

This work was supported by the National Institutes of Health (grant number: CA092824, 5U01CA209414, CA074386, CA090578, R01NS091307 and R56AG062302); the V Foundation for Cancer Research (grant number: D2017-19); the Duke Cancer Institute as part of the P30 Cancer Center Support Grant (Grant ID: NIH/NCI CA014236); the short-term international training program for PhD from Zhengzhou University, P.R. China.

Abbreviations:

MHC-I	major histocompatibility complex class I
NSCLC	Non-small cell lung cancer
SNPs	single nucleotide polymorphisms

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-specific survival
LD	linkage disequilibrium
FDR	false discovery rate
BFDp	Bayesian false discovery probability
eQTL	expression quantitative trait loci
TCGA	the Cancer Genome Atlas
ROC	receiver operating characteristic
ERAP1	endoplasmic reticulum aminopeptidase 1
PSMF1	proteasome inhibitor subunit 1
NCF2	neutrophil cytosolic factor 2
EAF	effect allele frequency
HR	hazards ratio
CI	confidence interval
AUC	area under the receiver operating characteristic curve
ROC	receiver operating characteristic curve
LUAD	lung adenocarcinoma
LUSC	lung squamous cell carcinoma
NADPH	nicotinamide adenine dinucleotide phosphate

References

- [1]. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin*2015; 65: 87–108 [PubMed: 25651787]
- [2]. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin*2019; 69: 7–34 [PubMed: 30620402]
- [3]. Goldstraw P, Ball D, Jett JR, Le Chevalier T, Lim E, Nicholson AG, et al. Non-small-cell lung cancer. *Lancet*2011; 378: 1727–40 [PubMed: 21565398]
- [4]. Kobayashi S, Boggon TJ, Dayaram T, Janne PA, Kocher O, Meyerson M, et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N Engl J Med*2005; 352: 786–92 [PubMed: 15728811]

- [5]. Katayama R, Shaw AT, Khan TM, Mino-Kenudson M, Solomon BJ, Halmos B, et al. Mechanisms of acquired crizotinib resistance in ALK-rearranged lung Cancers. *Sci Transl Med*2012; 4: 120ra17
- [6]. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med*2004; 350: 2129–39 [PubMed: 15118073]
- [7]. Liu Y, Zeng G. Cancer and innate immune system interactions: translational potentials for cancer immunotherapy. *J Immunother*2012; 35: 299–308 [PubMed: 22495387]
- [8]. Ostrand-Rosenberg S. Immune surveillance: a balance between protumor and antitumor immunity. *Curr Opin Genet Dev*2008; 18: 11–8 [PubMed: 18308558]
- [9]. Schreiber RD, Old LJ, Smyth MJ. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science*2011; 331: 1565–70 [PubMed: 21436444]
- [10]. Ryu R, Ward KE. Atezolizumab for the First-Line Treatment of Non-small Cell Lung Cancer (NSCLC): Current Status and Future Prospects. *Front Oncol*2018; 8: 277 [PubMed: 30087855]
- [11]. Paz-Ares L, Luft A, Vicente D, Tafreshi A, Gumus M, Mazieres J, et al. Pembrolizumab plus Chemotherapy for Squamous Non-Small-Cell Lung Cancer. *N Engl J Med*2018; 379: 2040–2051 [PubMed: 30280635]
- [12]. Gandhi L, Rodriguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, et al. Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. *N Engl J Med*2018; 378: 2078–2092 [PubMed: 29658856]
- [13]. Reck M, Rodriguez-Abreu D, Robinson AG, Hui R, Czoszi T, Fulop A, et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *N Engl J Med*2016; 375: 1823–1833 [PubMed: 27718847]
- [14]. Hughes AL, Hughes MK. Natural selection on the peptide-binding regions of major histocompatibility complex molecules. *Immunogenetics*1995; 42: 233–43 [PubMed: 7672817]
- [15]. Kobayashi KS, van den Elsen PJ. NLRC5: a key regulator of MHC class I-dependent immune responses. *Nat Rev Immunol*2012; 12: 813–20 [PubMed: 23175229]
- [16]. Huang YT, Heist RS, Chirieac LR, Lin X, Skaug V, Zienoldiny S, et al. Genome-wide analysis of survival in early-stage non-small-cell lung cancer. *J Clin Oncol*2009; 27: 2660–7 [PubMed: 19414679]
- [17]. Wu X, Ye Y, Rosell R, Amos CI, Stewart DJ, Hildebrandt MA, et al. Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy. *J Natl Cancer Inst*2011; 103: 817–25 [PubMed: 21483023]
- [18]. Xun WW, Brennan P, Tjonneland A, Vogel U, Overvad K, Kaaks R, et al. Single-nucleotide polymorphisms (5p15.33, 15q25.1, 6p22.1, 6q27 and 7p15.3) and lung cancer survival in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Mutagenesis*2011; 26: 657–66 [PubMed: 21750227]
- [19]. Wu X, Wang L, Ye Y, Aakre JA, Pu X, Chang GC, et al. Genome-wide association study of genetic predictors of overall survival for non-small cell lung cancer in never smokers. *Cancer Res*2013; 73: 4028–38 [PubMed: 23704207]
- [20]. Hocking WG, Hu P, Oken MM, Winslow SD, Kvale PA, Prorok PC, 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 [PubMed: 20442215]
- [21]. Oken MM, Marcus PM, Hu P, Beck TM, Hocking W, Kvale PA, 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 [PubMed: 16368945]
- [22]. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res*2014; 42: D975–9 [PubMed: 24297256]
- [23]. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*2007; 39: 1181–6 [PubMed: 17898773]
- [24]. Asomaning K, Miller DP, Liu G, Wain JC, Lynch TJ, Su L, et al. Second hand smoke, age of exposure and lung cancer risk. *Lung Cancer*2008; 61: 13–20 [PubMed: 18191495]

- [25]. Zhai R, Yu X, Wei Y, Su L, 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: 961–70 [PubMed: 23921845]
- [26]. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*2007; 23: 1294–6 [PubMed: 17384015]
- [27]. Wakefield JA Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet*2007; 81: 208–27 [PubMed: 17668372]
- [28]. Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med*2006; 25: 3474–86 [PubMed: 16220486]
- [29]. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*2013; 501: 506–11 [PubMed: 24037378]
- [30]. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*2015; 348: 648–60 [PubMed: 25954001]
- [31]. Xu Z, Taylor JA. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res*2009; 37: W600–5 [PubMed: 19417063]
- [32]. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*2012; 22: 1790–7 [PubMed: 22955989]
- [33]. Ward LD, 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: D877–81 [PubMed: 26657631]
- [34]. Wang Y, Liu H, Ready NE, Su L, Wei Y, Christiani DC, et al. Genetic variants in ABCG1 are associated with survival of nonsmall-cell lung cancer patients. *Int J Cancer*2016; 138: 2592–601 [PubMed: 26757251]
- [35]. Kim S, Jang JY, Koh J, Kwon D, Kim YA, Paeng JC, et al. Programmed cell death ligand-1-mediated enhancement of hexokinase 2 expression is inversely related to T-cell effector gene expression in non-small-cell lung cancer. *J Exp Clin Cancer Res*2019; 38: 462 [PubMed: 31718692]
- [36]. Mofers A, Pellegrini P, Linder S, D'Arcy P. Proteasome-associated deubiquitinases and cancer. *Cancer Metastasis Rev*2017; 36: 635–653 [PubMed: 29134486]
- [37]. Hattori A, Tsujimoto M. Endoplasmic reticulum aminopeptidases: biochemistry, physiology and pathology. *J Biochem*2013; 154: 219–28 [PubMed: 23946506]
- [38]. Serwold T, Gonzalez F, Kim J, Jacob R, Shastri N. ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature*2002; 419: 480–3 [PubMed: 12368856]
- [39]. York IA, Chang SC, Saric T, Keys JA, Favreau JM, Goldberg AL, et al. The ER aminopeptidase ERAP1 enhances or limits antigen presentation by trimming epitopes to 8-9 residues. *Nat Immunol*2002; 3: 1177–84 [PubMed: 12436110]
- [40]. Saric T, Chang SC, Hattori A, York IA, Markant S, Rock KL, et al. An IFN-gamma-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides. *Nat Immunol*2002; 3: 1169–76 [PubMed: 12436109]
- [41]. Stratikos E, Stamogiannos A, Zervoudi E, Fruci D. A role for naturally occurring alleles of endoplasmic reticulum aminopeptidases in tumor immunity and cancer pre-disposition. *Front Oncol*2014; 4: 363 [PubMed: 25566501]
- [42]. Pedersen MH, Hood BL, Beck HC, Conrads TP, Ditzel HJ, Leth-Larsen R. Downregulation of antigen presentation-associated pathway proteins is linked to poor outcome in triple-negative breast cancer patient tumors. *Oncoimmunology*2017; 6: e1305531 [PubMed: 28638726]
- [43]. Mehta AM, Jordanova ES, Kenter GG, Ferrone S, Fleuren GJ. Association of antigen processing machinery and HLA class I defects with clinicopathological outcome in cervical carcinoma. *Cancer Immunol Immunother*2008; 57: 197–206 [PubMed: 17622526]
- [44]. Ayshamgul H, Ma H, Ilyar S, Zhang LW, Abulizi A. Association of defective HLA-I expression with antigen processing machinery and their association with clinicopathological characteristics

in Kazak patients with esophageal cancer. *Chin Med J (Engl)*2011; 124: 341–6 [PubMed: 21362330]

- [45]. Schmidt K, Keller C, Kuhl AA, Textor A, Seifert U, Blankenstein T, et al. ERAP1-Dependent Antigen Cross-Presentation Determines Efficacy of Adoptive T-cell Therapy in Mice. *Cancer Res*2018; 78: 3243–3254 [PubMed: 29559473]
- [46]. Cho-Park PF, Steller H. Proteasome regulation by ADP-ribosylation. *Cell*2013; 153: 614–27 [PubMed: 23622245]
- [47]. Zaiss DM, Standera S, Holzhutter H, Kloetzel P, Sijts AJ. The proteasome inhibitor PI31 competes with PA28 for binding to 20S proteasomes. *FEBS Lett*1999; 457: 333–8 [PubMed: 10471803]
- [48]. McCutchen-Maloney SL, Matsuda K, Shimbara N, Binns DD, Tanaka K, Slaughter CA, et al. cDNA cloning, expression, and functional characterization of PI31, a proline-rich inhibitor of the proteasome. *J Biol Chem*2000; 275: 18557–65 [PubMed: 10764772]
- [49]. Micel LN, Tentler JJ, Smith PG, Eckhardt GS. Role of ubiquitin ligases and the proteasome in oncogenesis: novel targets for anticancer therapies. *J Clin Oncol*2013; 31: 1231–8 [PubMed: 23358974]
- [50]. D’Arcy P, Linder S. Molecular pathways: translational potential of deubiquitinases as drug targets. *Clin Cancer Res*2014; 20: 3908–14 [PubMed: 25085788]
- [51]. Gardiner GJ, Deffit SN, McLetchie S, Perez L, Walline CC, Blum JS. A role for NADPH oxidase in antigen presentation. *Front Immunol*2013; 4: 295 [PubMed: 24069023]

Figure 1: The flowchart of the present study.

Abbreviations: SNP, single-nucleotide polymorphism; PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial; NSCLC, non-small cell lung cancer; HLCS, Harvard lung cancer susceptibility study; *ERAP1*, endoplasmic reticulum aminopeptidase 1; *PSMF1*, proteasome inhibitor subunit 1; *NCF2*, neutrophil cytosolic factor 2.

Figure 2: Prediction of survival with combined unfavorable genotypes and eQTL analysis for SNPs in *ERAP1* and *PSMF1*.

Kaplan–Meier survival curves for OS in the PLCO dataset for (a) the combined unfavorable genotypes and (b) dichotomized groups of the NUGs; Kaplan–Meier survival curves for DSS in the PLCO dataset for (c) the combined unfavorable genotypes and (d) dichotomized groups of the NUGs. *ERAP1* rs469783C allele was associated with higher mRNA expression of *ERAP1* (e) in 373 Europeans from the 1000 Genomes Project, (g) in normal lung tissue and (h) whole blood from GTEx project; *PSMF1* rs13040574A allele was associated with higher mRNA expression of *PSMF1* (f) in 373 Europeans from the 1000 Genomes Project. Abbreviations: NUG, number of unfavorable genotypes; PLCO, The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; *ERAP1*, endoplasmic reticulum aminopeptidase 1; *PSMF1*, proteasome inhibitor subunit 1.

Figure 3: Differential mRNA expression analysis and overall survival analysis of the three genes in the TCGA database.

(a) Higher *ERAP1* mRNA expression levels were found in the adjacent normal tissues of 109 paired NSCLC tissue; (b) higher expression levels of *PSMF1* were found in the tumor tissue while (c) higher *NCF2* mRNA expression levels were found in the adjacent normal tissues; (d) higher *ERAP1* mRNA expression levels, (e) higher *PSMF1* mRNA expression levels and higher *NCF2* mRNA expression levels were all correlated with better survival. Abbreviations: TCGA, The Cancer Genome Atlas; NSCLC, non-small cell lung cancer; *ERAP1*, endoplasmic reticulum aminopeptidase 1; *PSMF1*, proteasome inhibitor subunit 1; *NCF2*, neutrophil cytosolic factor 2.

Table 1.

Associations of 24 validated significant SNPs with overall survival in both discovery and validation datasets from two previously published NSCLC GWASs

SNP	Allele	Gene	PLCO (n=1185)			HLCS (n=984)			Combined-analysis					
			FDR	BFDP	EAF	HR (95% CI) ^a	P ^a	EAF	HR (95% CI) ^b	P ^b	P _{het} ^c	I ²	HR (95% CI) ^d	P ^d
rs469783	T>C	ERAP1	0.29	0.16	0.42	0.83 (0.75-0.92)	0.0003	0.45	0.83 (0.74-0.92)	0.0007	0.989	0	0.83 (0.77-0.89)	8.0E-7
rs246456	T>C	ERAP1	0.29	0.16	0.33	0.83 (0.74-0.92)	0.0006	0.39	0.86 (0.77-0.97)	0.011	0.636	0	0.85 (0.78-0.91)	2.9E-5
rs30379 ^e	G>T	ERAP1	0.29	0.16	0.33	0.83 (0.74-0.92)	0.0006	0.35	0.87 (0.78-0.97)	0.013	0.573	0	0.85 (0.78-0.92)	3.6E-5
rs151964	T>G	ERAP1	0.29	0.16	0.33	0.83 (0.74-0.92)	0.0006	0.35	0.87 (0.78-0.97)	0.013	0.572	0	0.85 (0.78-0.92)	3.7E-5
rs30187 ^f	C>T	ERAP1	0.29	0.16	0.33	0.83 (0.74-0.92)	0.0006	0.35	0.87 (0.78-0.97)	0.013	0.571	0	0.85 (0.78-0.92)	3.7E-5
rs246455	T>C	ERAP1	0.29	0.16	0.33	0.83 (0.74-0.92)	0.0006	0.35	0.87 (0.78-0.97)	0.015	0.555	0	0.85 (0.79-0.92)	4.0E-5
rs168674	C>T	ERAP1	0.29	0.16	0.33	0.83 (0.74-0.92)	0.0006	0.41	0.87 (0.77-0.98)	0.017	0.568	0	0.85 (0.78-0.92)	4.6E-5
rs27524	G>A	ERAP1	0.34	0.61	0.35	0.85 (0.77-0.95)	0.004	0.37	0.87 (0.78-0.98)	0.019	0.719	0	0.86 (0.80-0.93)	0.0001
rs34755	T>G	ERAP1	0.43	0.61	0.49	0.88 (0.79-0.97)	0.009	0.52	0.88 (0.79-0.98)	0.020	0.998	0	0.88 (0.82-0.95)	0.0007
rs10911362	A>G	NCF2	0.61	0.80	0.06	1.27 (1.03-1.56)	0.023	0.05	1.36 (1.06-1.73)	0.014	0.678	0	1.31 (1.12-1.53)	0.0009
rs36071574	G>A	NCF2	0.59	0.77	0.05	1.30 (1.05-1.62)	0.018	0.05	1.32 (1.03-1.70)	0.029	0.916	0	1.31 (1.11-1.54)	0.001
rs4142354	C>T	PSMF1	0.34	0.72	0.47	0.86 (0.78-0.96)	0.006	0.47	0.85 (0.76-0.95)	0.006	0.931	0	0.86 (0.79-0.92)	7.1E-5
rs6134012	C>T	PSMF1	0.38	0.69	0.47	0.87 (0.78-0.96)	0.008	0.47	0.86 (0.77-0.96)	0.006	0.839	0	0.86 (0.80-0.93)	0.0002
rs6077915	C>T	PSMF1	0.34	0.72	0.47	0.86 (0.78-0.96)	0.006	0.47	0.86 (0.77-0.96)	0.007	0.986	0	0.86 (0.80-0.93)	8.7E-5
rs13040574	C>A	PSMF1	0.34	0.72	0.47	0.86 (0.78-0.96)	0.006	0.47	0.86 (0.77-0.96)	0.008	0.991	0	0.86 (0.80-0.93)	9.4E-5
rs2284371	A>T	PSMF1	0.34	0.72	0.47	0.86 (0.78-0.96)	0.006	0.47	0.88 (0.79-0.99)	0.034	0.720	0	0.87 (0.81-0.94)	0.0004

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 Study; FDR, false discovery rate; BFDP, Bayesian false discovery probability; EAF, effect allele frequency; HR, hazards ratio; CI, confidence interval;

^aObtained from an additive genetic model with adjustment for age, sex, stage, histology, smoking status, chemotherapy, radiotherapy, surgery, PC1, PC2, PC3, and PC4;

^bObtained from an additive genetic model with adjustment for age, sex, stage, histology, smoking status, chemotherapy, radiotherapy, surgery, PC1, PC2, and PC3;

^cP_{het}: P-value for heterogeneity by Cochran's Q test;

^dMeta-analysis in the fixed-effects model.

^eSNPs rs30380, rs469758, rs30378, rs246453, rs246454 are high LD with rs30379 and have same results.

SNPs rs26510, rs27710, rs27529 are high LD with rs30187 and have same results.
SNPs rs469783, rs30187, rs27524, rs6077915 are genotyped. The other SNPs are imputed.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Three independent SNPs in multivariate Cox proportional hazards regression analysis with adjustment for other covariates and previous published SNPs in the PLCO dataset

Table 2.

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 (1.02-1.05)	<0.0001
Sex	Male	698	1.00		1.00	
	Female	487	0.77 (0.66-0.89)	0.0005	0.85 (0.64-0.88)	0.0003
Smoking status	Never	115	1.00		1.00	
	Current	423	1.67 (1.25-2.24)	0.0006	1.89 (1.40-2.55)	<0.0001
	Former	647	1.64 (1.25-2.16)	0.0004	1.87 (1.41-2.49)	<0.0001
Histology	AD	577	1.00		1.00	
	SC	285	1.15 (0.95-1.39)	0.144	1.19 (0.98-1.44)	0.083
	Others	323	1.30 (1.10-1.54)	0.003	1.33 (1.11-1.58)	0.002
Stage	I-III A	655	1.00		1.00	
	IIIB-IV	528	2.92 (2.40-3.55)	<0.0001	3.14 (2.57-3.82)	<0.0001
Chemotherapy	No	639	1.00		1.00	
	Yes	538	0.57 (0.47-0.67)	<0.0001	0.57 (0.47-0.68)	<0.0001
Radiotherapy	No	762	1.00		1.00	
	Yes	415	0.94 (0.80-1.11)	0.453	0.94 (0.80-1.12)	0.490
Surgery	No	637	1.00		1.00	
	Yes	540	0.21 (0.16-0.27)	<0.0001	0.19 (0.15-0.25)	<0.0001
<i>ERAP1</i> rs469783 T>C	TT/TC/CC	401/580/204	0.83 (0.75-0.92)	0.0003	0.81 (0.73-0.90)	<0.0001
<i>PSMF1</i> rs13040574 C>A	CC/CA/AA	318/616/251	0.86 (0.77-0.95)	0.004	0.84 (0.75-0.93)	0.001
<i>NCF2</i> rs36071574 G>A	GG/GA/AA	1063/116/6	1.33 (1.07-1.65)	0.011	1.33 (1.07-1.67)	0.011

Abbreviations: SNP: single-nucleotide polymorphism; PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial; GWAS, genome-wide association study; HR: hazards ratio; CI: confidence interval.

^a Stepwise analysis included age, sex, smoking status, tumor stage, histology, chemotherapy, radiotherapy, surgery, PC1, PC2, PC3, PC4 and SNPs.

^b Fifteen published SNPs were used for post-stepwise adjustment. Five SNPs were reported in previous publication (PMID: 27557513); One SNP was reported in the previous publication (PMID: 29978465); Two SNPs were reported in the previous publication (PMID: 30259978); Two SNPs were reported in the previous publication (PMID: 26757251); Three SNPs were reported in the previous publication (PMID: 30650190); Two SNPs were reported in the previous publication (PMID: 30989732);

Table 3.

Associations between three independent SNPs and survival of NSCLC in the PLCO Trial

Genotype	Frequency	OS ^d			DSS ^d		
		Death (%)	HR (95% CI)	P	Death (%)	HR (95% CI)	P
ERAP1/rs469783 T>C ^b							
TT	397	273 (68.77)	1.00		246 (61.96)	1.00	
TC	578	399 (69.03)	0.94 (0.81-1.10)	0.462	361 (62.46)	0.95 (0.81-1.12)	0.539
CC	200	117 (58.50)	0.65 (0.52-0.80)	<0.0001	102 (51.00)	0.64 (0.51-0.81)	0.0002
Trend test				0.0003			0.0008
Dominant							
TT	397	273 (68.77)	1.00		246 (61.96)	1.00	
TC+CC	778	516 (66.32)	0.86 (0.74-0.99)	0.041	463 (59.51)	0.86 (0.74-1.01)	0.063
Or reverse							
TC+CC	778	516 (66.32)	1.00		463 (59.51)	1.00	
TT	397	273 (68.77)	1.16 (1.01-1.35)	0.041	246 (61.96)	1.16 (0.99-1.35)	0.063
PSMF1/rs13040574 C>A ^c							
CC	315	210 (66.67)	1.00		183 (58.10)	1.00	
CA	610	405 (66.39)	0.79 (0.66-0.93)	0.005	373 (61.15)	0.83 (0.69-0.99)	0.038
AA	250	174 (69.60)	0.75 (0.61-0.92)	0.007	153 (61.20)	0.76 (0.61-0.95)	0.015
Trend test				0.006			0.014
Dominant							
CC	315	210 (66.67)	1.00		183 (58.10)	1.00	
CA+AA	860	579 (67.33)	0.78 (0.66-0.91)	0.002	526 (61.16)	0.81 (0.68-0.96)	0.015
Or reverse							
CA+AA	860	579 (67.33)	1.00		526 (61.16)	1.00	
CC	315	210 (66.67)	1.28 (1.10-1.52)	0.002	183 (58.10)	1.23 (1.04-1.47)	0.015
NCF2/rs36071574 G>A ^d							
GG	1057	706 (66.79)	1.00		633 (59.89)	1.00	
GA	113	79 (69.91)	1.26 (1.00-1.60)	0.051	72 (63.72)	1.29 (1.01-1.65)	0.043
AA	5	4 (80.00)	2.35 (0.87-6.38)	0.093	4 (80.00)	2.68 (0.99-7.29)	0.053

Genotype	Frequency	OS ^d			DSS ^d		
		Death (%)	HR (95% CI)	P	Death (%)	HR (95% CI)	P
Trend test				0.018			0.012
Dominant							
GG	1057	706 (66.79)	1.00		633 (59.89)	1.00	
GA+AA	118	83 (70.34)	1.29 (1.03-1.62)	0.030	76 (64.41)	1.32 (1.04-1.68)	0.022
NUG ^{e,f}							
0	510	343 (67.25)	1.00		312 (61.18)	1.00	
1	511	334 (65.36)	1.15 (0.98-1.34)	0.082	295 (57.73)	1.12 (0.95-1.32)	0.176
2	143	104 (72.73)	1.58 (1.26-1.98)	<0.0001	96 (67.13)	1.57 (1.25-1.99)	0.0001
3	11	8 (72.73)	2.61 (1.28-5.32)	0.008	6 (54.55)	2.12 (0.94-4.82)	0.072
Trend test				<0.0001			0.0002
0-1	1021	677 (66.31)	1.00		607 (59.45)	1.00	
2-3	154	112 (72.73)	1.52 (1.24-1.86)	<0.0001	102 (66.23)	1.51 (1.22-1.87)	0.0001

Abbreviations: SNP, single nucleotide polymorphism; NSCLC, non-small cell lung cancer; PLCO, Prostate, Lung, Colorectal and Ovarian cancer screening trial; OS, overall survival; DSS, disease-specific survival; HR, hazards ratio; CI, confidence interval; NUG: number of unfavorable genotypes.

^a Adjusted for age, sex, smoking status, histology, tumor stage, chemotherapy, surgery, radiotherapy and principal components.

^b 10 missing date were excluded;

^c 10 missing date were excluded;

^d 10 missing date were excluded

^e 10 missing date were excluded.

^f Unfavorable genotypes were *ERAP1* rs469783 TT, *PSMF1* rs13040574 CC, *NCF2* rs36071574 GA+AA.