Potentially functional genetic variants in the c-Myb signaling pathway genes are associated with survival of non-small cell lung cancer patients in the PLCO Cancer Screening Trial

Yi Guo

Shanghai Jiao Tong University

Yun Feng

Shanghai Jiao Tong University School of medicine

Hongliang Liu

Duke University School of Medicine

Sheng Luo

Duke university

Patricia G. Moorman

Duke university

Edward F. Patz

Duke university

Li Su

Harvard University

Sipeng Shen

Harvard University

David C. Christiani

Harvard University

Qingyi Wei (✉ qingyi.wei@duke.edu)

Department of medicine, Duke university

Research article

Keywords: Non-small cell lung cancer, Single nucleotide polymorphism, Survival, C-Myb; ETS1, ZFH3

Posted Date: August 4th, 2019

DOI: <https://doi.org/10.21203/rs.2.11515/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background The c-Myb signaling pathway plays a critical role in regulating expression of genes involved in differentiation, proliferation of cells and cell death, which are related to cancer progression and metastasis. **Methods** To test the hypothesis that genetic variants in the c-Myb pathway genes are associated with survival of non-small cell lung cancer (NSCLC) patients, we used genotyping data from 1,185 NSCLC patients in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial as the discovery dataset and 984 NSCLC patients in the Harvard Lung Cancer Susceptibility (HLCS) Study as the validation dataset. With genotyping data for 8587 SNPs in 83 genes of the c-Myb pathway from the PLCO dataset, we performed Cox proportional hazards regression analysis to evaluate association between each SNP and survival, and resultant significant SNPs were validated in the HLCS dataset and further evaluated for their functional relevance. **Results** Specifically, we found that ETS1 rs11606640 GA+AA and ZFHX3 rs62053220 CG+GG genotypes were associated with a favorable overall survival in NSCLC patients ($P_{\text{dom}}=0.004$ and 0.001 , respectively) in a multivariable Cox model. These two validated SNPs were predicted by the RegulomeDB score to be potentially functional. In addition, ETS1 and ZFHX3 mRNA expression levels in lung cancer tissues were significantly lower than that in adjacent normal lung tissues ($P<0.05$ for both SNPs). Furthermore, the TCGA-based expression quantitative trait loci (eQTL) analysis showed that ETS1 rs11606640 genotypes were significantly associated with their corresponding mRNA expression levels in lung cancer tissues in both additive and recessive models ($P_{\text{add}}=0.045$ and $P_{\text{rec}}=0.042$), but such a phenotype-genotype correlation for ZFHX3 rs62053220 genotypes was found in only a dominant model ($P_{\text{dom}}=0.048$). **Conclusion** We identified two potentially functional SNPs (ETS1 rs11606640 G>A and ZFHX3 rs62053220 C>G) that are associated with survival in NSCLC patients. Additional larger studies are needed to validate our findings.

Background

Lung cancer is the second most common cancer and the leading cause of cancer deaths in the United States. Non-small cell lung cancer (NSCLC) accounts for more than 80% of all the lung cancer cases. Despite recent progress in the diagnosis and treatment of NSCLC, the prognosis remains poor with a five-year survival of only 18%.¹ Therefore, it is important to understand genetic predictors that may underlie the response to treatment and variation in the outcomes of NSCLC patients.

Published genome-wide association studies (GWASs) have identified a number of single nucleotide polymorphisms (SNPs) associated with lung cancer survival using the “20-50 most-significant SNPs” strategy.² However, the nature of these hypothesis-free studies has limitations in revealing functional SNPs from the relevant interactive genes. Recently, a pathway-based approach has been proposed, which not only jointly considered multiple variants in interactive genes but also considered their potential biological functions and mechanisms that may underlie the associations with survival in NSCLC.³ Therefore, by using a hypothesis-driven pathway-based analysis, we aimed to identify additional functional SNPs associated with survival in NSCLC.

C-Myb, as a DNA-binding transcription factor, functions in the regulation of haematopoietic cell self-renewal and multi-lineage differentiation. C-Myb binds to the specific sequence t/cAAcT/gG, known as a Myb-binding site (MBS), within regulatory regions of the target genes. Therefore, the c-Myb signaling pathway plays a critical role in regulating expression of the target genes involved in differentiation, proliferation of cells, and cell death. Because some of the target genes, such as *discoïdin domain receptor 2*, activated by the c-Myb pathway are related to cancer progression and metastasis, overexpression of these genes has been associated with development and progression of cancers of the lung, colon, breasts and stomach as well as nasopharyngeal carcinoma. However, an association between genetic variants in the c-Myb pathway genes and survival of NSCLC patients remains unclear. Therefore, we performed the present study to identify genetic variants in the c-Myb signaling genes and determined their effects on survival of NSCLC patients.

Methods

Discovery dataset

In the discovery phase, we used a GWAS dataset from the Prostate, Lung, Colorectal and Ovarian, (PLCO) Cancer Screening Trial, a large population-based randomized trial designed to evaluate the effectiveness of cancer screening and to investigate etiologic factors and early markers of cancers. The PLCO trial included 1,185 Caucasian NSCLC patients who were recruited at 10 centers in the United States between 1993 and 2001. In this trial, DNA was extracted from the blood samples and genotyped using the Illumina HumanHap240Sv1.0, HumanHap300v1.1 and HumanHap550v3.0 (dbGaP accession: phs000093.v2.p2 and phs000336.v1.p1). The follow-up information available for analysis included overall survival (OS), disease-specific survival (DSS) and other clinical variables. The study protocol was reviewed and approved by the institutional review board of the National Cancer Institute (NCI), and a written informed consent was obtained from each participant.

Validation dataset

After identifying representative significant SNPs in the PLCO GWAS database through single-locus analysis, we conducted a validation analysis using the GWAS dataset from the Harvard Lung Cancer Susceptibility (HLCS) Study of 984 histology-confirmed Caucasian NSCLC patients. The genotyping data from blood samples were obtained using Illumina Humanhap610-Quad arrays, and imputation was performed using MaCH v1.0 based on the 1000 Genomes Project. Multivariate Cox regression analysis was also applied to estimate HRs. Only those significant SNPs discovered in the PLCO dataset and validated by the HLCS dataset were subjected to further functional validation.

Selection of candidate genes in the c-Myb pathway

Based on the Molecular Signatures Database (MSigDB), (<http://www.broadinstitute.org/gsea/msigdb/search.jsp>), a collection of annotated gene sets for enrichment analysis, we selected candidate genes in the c-Myb pathway using

'c-Myb' as a keyword for the search. Duplicate genes, pseudogenes and genes located on chromosome X were removed before the analyses. The 2-kb upstream and downstream sequences of the identified genes were also included to cover the promoter regions for all the candidate genes.

Genotyping data extraction and imputation

We extracted genotyping data for all the common SNPs located on those selected candidate genes using PLINK 1.07 with the quality control criteria: (1) minor allele frequency (MAF) ≥ 0.05 ; (2) genotyping call rate $\geq 95\%$; (3) P -value for Hardy-Weinberg Equilibrium test $\geq 1 \times 10^{-5}$. Then, we performed imputation with IMPUTE2 according to the 1000 Genomes Project (Northern Europeans from Utah) data (phase III release October 2014), and the imputed SNPs with an info score ≥ 0.8 were used in further analyses. The Manhattan plot of all of the candidate SNPs was generated by using Haploview software (v4.2). The regional association plot was made for the final significant genes to show all genotyped and imputed SNPs in that gene.

Statistical analysis

In the single locus analysis, we used multivariate Cox proportional hazards regression analysis 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 to evaluate the associations between extracted SNPs and survival in an additive genetic model using the GenABEL package in R software. We estimated hazards ratio (HR) and 95% confidence interval (CI) for each SNP. We first used the false discovery rate (FDR) method for multiple testing correction; nevertheless, considering the vast majority of the SNPs in the present study were imputed and thus not independent as required by FDR or other correction methods such as the Bonferroni correction, the false positive report probability (FPRP) method was then chosen to perform the multiple testing correction with a strict cut-off of 0.10. We assigned a prior probability of 0.10 to detect an HR of 2.0 for an association with variant genotypes or minor alleles of the SNPs with $P < 0.05$. The SNPs that were identified with a P -value < 0.05 by Cox regression analysis and passed the FPRP threshold were subjected to linkage disequilibrium (LD) analysis using Haploview software. Pairwise $r^2 > 0.80$ was considered in high LD, and the representative SNPs ($r^2 > 0.80$ in LD with other SNPs in the same gene) were used for further population and functional validation. Then, we assessed the associations of those SNPs, identified in the discovery PLCO dataset and validated the HLCS dataset, with OS/DSS of NSCLC using additive, dominant and recessive genetic models. To assess their collective effect, we also combined the protective genotypes of those validated SNPs as a genetic score for the number of protective genotypes in association with OS/DSS.

To assess any possible interactions, we first performed stratified analysis of the combined effect of protective genotypes as defined by the genetic score on NSCLC OS/DSS by clinical characteristics in the PLCO dataset, including age, sex, smoking status, histology, tumor stage, chemotherapy, radiotherapy and surgery, because the HLCS did not have such detailed clinical information. To combine the two GWAS datasets by a meta-analysis approach, a fixed-effects model was used when no heterogeneity was found between two studies (Q-test P -value > 0.10 and I^2 < 50.0%); otherwise, a random-effects model was applied. Kaplan-Meier survival curves, plotted using GraphPad Prism 6.0 (GraphPad Software Inc, San Diego, CA, USA), and the log-ranks tests were used to estimate the effects of protective genotypes on the cumulative probability of OS/DSS. In all the findings, a P -value < 0.05 was considered statistically significant. All statistical analyses were performed with SAS software (v9.4; SAS Institute, Cary, NC, USA), if not specified otherwise.

Functional annotation

We used an *in silico* approach, with use of SNPinfo(<http://snpinfo.niehs.nih.gov>), RegulomeDB (<http://regulomedb.org>), and HaploReg v4.1

(<http://archive.broadinstitute.org/mammals/haploreg/haploreg.php>), three online bioinformatics tools, to predict potential functions of survival-associated SNPs. Survival analysis relevant to the mRNA expression of the validated genes in lung cancer tissues were also performed using available online data (<http://kmplot.com/analysis>).

The expression data of lung cancer tissues from the Cancer Genome Atlas (TCGA) database (dbGaP Study Accession: phs000178.v1.p1), obtained from the Broad TCGA GDAC site (<http://gdac.broadinstitute.org>), were used for the expression quantitative trait loci (eQTL) analysis of the validated SNPs. We also compared differential expression levels of validated genes in lung cancer tissues with that in the adjacent normal lung tissues.

Additional eQTL analysis was performed, in which the correlations between the genotypes of those validated SNPs and their mRNA expression levels were explored using RNA sequencing data from lymphoblastoid cells of 373 European descendants available in the 1000 Genomes Project.

Results

Study population characteristics

Demographic and clinical characteristics of NSCLC patients from the PLCO trial have been previously described as also shown in Supplementary Table 1. There were 798 (67.3%) deaths during the follow-up with a median survival time of 23.77 months. There were 11 subjects with missing clinical information, including one for pack years, two for tumor stage, and eight for chemotherapy/radiotherapy/surgery).

Overall, subjects with age 71 years, being female and never smoker, having a lower stage and chemotherapy or surgery tended to have a better survival.

The identified SNPs obtained from the PLCO discovery analysis were further validated by the GWAS dataset from the 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, as previously described¹⁵.

Gene and SNP extraction from the PLCO trial

The overall workflow is shown in Figure 1. Eighty-four candidate genes in the c-Myb signaling pathway were identified from MSigDB, with one gene located on chromosome X removed from further analyses (Supplementary Table 2). As a result, a total of 8,587 SNPs (980 genotyped and 7,607 imputed) were extracted from the PLCO dataset. The Manhattan plot shows the locations of all of the candidate SNPs (Supplementary Figure 1.) in the final 83 genes.

Survival association analyses of the PLCO dataset

In the single locus analysis using the PLCO dataset of 8,587 SNPs, 692 SNPs were found to be significantly associated with NSCLC OS ($P < 0.05$) after adjustment for age, sex, smoking status, histology, tumor stage, chemotherapy, radiotherapy, surgery and the first four principal components, of which 195 SNPs remained significant after the multiple testing correction ($FPRP \leq 0.10$). Finally, 42 representative SNPs were selected for validation based on additional LD analysis ($r^2 > 0.80$) for each gene (Figure 1).

Validation by the HLCS GWAS dataset and the combined analysis

The resultant 42 SNPs identified in the discovery underwent validation using the HLCS GWAS dataset. Two validated SNPs, *ETS1* rs11606640 and *ZFHX3* rs62053220, remained significantly associated with survival of NSCLC patients in both of the datasets ($P < 0.05$ for both PLCO and HLCS). Specifically, the results showed that, for *ETS1* rs11606640, HR was 0.82 (95% CI 0.71-0.95, $P = 0.006$) in the PLCO discovery and 0.73 (95% CI 0.62-0.85, $P = 7.2 \times 10^{-5}$) in the HLCS validation, whereas, for *ZFHX3* rs62053220, HR was 0.71 (95% CI 0.58-0.86, $P = 4.4 \times 10^{-4}$) in the PLCO discovery and 0.73 (95% CI 0.60-0.89, 0.002) in the HLCS validation (Supplementary Table 3). In the combined analysis of two datasets for these two validated SNPs showed an improved OS of NSCLC associated with the rs11606640 A and rs62053220 G alleles [HR=0.78 (0.70-0.86) with $P = 2.55 \times 10^{-6}$ and HR=0.73 (0.63-0.83) with $P = 2.54 \times 10^{-6}$, respectively] (Table 1). All genotyped and imputed SNPs are shown in the regional association plots with an expansion of 500 KB in the flanks of the gene region, in which the two validated SNPs are each labeled in purple (Supplementary Figure 2A).

Independent SNPs and further survival analyses with genetic models

We further conducted a stepwise Cox regression analysis of selected clinical variables from the PLCO dataset plus the two validated SNPs to confirm whether these SNPs were independent predictors of survival. We did this using the PLCO dataset, because only the PLCO dataset, but not the HLCS dataset, had detailed clinical variables, including age, sex, smoking status, tumor stage, tumor histology, chemotherapy, radiotherapy, and surgery. The Cox regression analysis also included top four principal components calculated based on all the SNPs in the 83 genes. The final stepwise analysis included 1,163 patients, because there were 10 patients with missing data for clinical variables (two for tumor stage and eight for chemotherapy/radiotherapy/surgery), eight patients with missing data for *ETS1* rs11606640, and four patients with missing data for *ZFHX3* rs62053220. Both of the two SNPs remained significant in the final multivariate model (Table 2) .

Survival analysis with different genetic models for the two independent SNPs was also performed in the PLCO dataset accordingly. Multivariate Cox regression analyses found, under an additive genetic model, that *ETS1* rs11606640 A and *ZFHX3* rs62053220 G variant alleles were associated with a decreased death risk of NSCLC, with a variant-allele attributed HR of 0.83 (95% CI=0.72-0.95, $P=0.008$) and 0.71 (95% CI=0.58-0.86, $P<0.001$), respectively; *ETS1* rs11606640 GA+AA and *ZFHX3* rs62053220 CG+GG genotypes were found to predict a favorable survival in NSCLC patients in a dominant model ($P<0.005$ for both). The univariate and multivariate Cox regression analyses with different genetic models of the two SNPs are presented in Table 3. Alternatively, as shown in Supplementary Table 4, *ETS1* rs11606640 GA+AA and *ZFHX3* rs62053220 CG+GG genotypes were significantly associated with a favorable DSS survival in the multivariate analysis ($P<0.01$ for both).

Combined analyses of the two SNPs

Because both *ETS1* rs11606640 GA+AA and *ZFHX3* rs62053220 CG+GG genotypes were associated with a favorable prognosis in NSCLC patients, we combined these genotypes into a genetic score to assess their collective effect. All the NSCLC patients were divided into three groups with zero, one and two of this genetic score. As shown in Table 3, the increase per score unit was correlated with an improved OS after adjustment for age, sex, smoking, tumor stage, histology, chemotherapy, radiotherapy, surgery and the top four principal components ($P_{\text{trend}}<0.001$). In the multivariate OS analysis with the dichotomized genetic score of 0 and 1-2, the latter was associated with a significant decrease in risk of death (HR=0.70, 0.60-0.81, $P<0.001$). Similar results of univariate and multivariate DSS analysis of the combined protective genotypes are shown in Supplementary Table 4.

Assessment of interaction and stratification analysis

Stratification analysis between the two groups of protective genotypes (score 0 versus 1-2) was performed with the PLCO dataset by other covariates, such as male vs. female, patients ≤ 71 vs. >71 years, patients with early vs. late tumor stage, surgery vs. no surgical therapy, with vs. without radiotherapy/chemotherapy, and the results are depicted by Kaplan-Meier survival curves and log-rank

tests, as shown in Figure 2. The association between scores 1-2 and a favorable survival was more evident for male patients, patients aged >71years, patients with late tumor stage and without surgery (all $P<0.001$), but a multiplicative interaction was found only between the genetic scores with age and radiotherapy ($P=0.002$ and 0.040 , respectively) (Supplementary Table 5). Hence, as shown in Supplementary Tables 6 and 7, we further reassessed the associations of *ETS1* rs11606640 and *ZFHX3* rs62053220 with OS of NSCLC using different genetic models by subgroups of age ≤ 71 or > 71 years and with or without radiotherapy.

The eQTL analyses

Different mRNA expression levels of *ETS1* and *ZFHX3* between lung cancer tissues and adjacent normal lung tissues were compared using the TCGA dataset. The results showed that *ETS1* and *ZFHX3* mRNA expression levels were significantly lower in lung cancer tissues than in adjacent normal lung tissues, adenocarcinoma tissues as well as squamous tissues (all $P<0.05$) (Figure 3).

In addition, the *ETS1* rs11606640 genotypes were significantly associated with the corresponding mRNA expression levels in lung squamous cancer tissues in an additive and recessive model ($P_{\text{additive}}=0.045$, $P_{\text{recessive}}=0.042$, respectively), which was not found in lung adenocarcinoma tissues; *ZFHX3* rs62053220 genotypes were significantly associated with the corresponding mRNA expression levels in adenocarcinoma tissues only in a dominant model ($P_{\text{dominant}}=0.048$) (Figure 4). Furthermore, based on survival analysis of available online data, low mRNA expression levels of *ETS1* and *ZFHX3* in lung cancer tissue were associated with a poor outcome of the lung cancer patients (Supplementary Figure 2B).

However, additional eQTL analysis using datasets in the 1000 Genomes Project showed that *ETS1* rs11606640 and *ZFHX3* rs62053220 genotypes were not significantly associated with their levels of the corresponding mRNA expression in the lymphoblastoid cells in an additive model ($P_{\text{additive}} = 0.459$ and 0.979 , respectively) (Supplementary Figure 2C).

Functional validation analyses *in silico*

In a functional prediction analysis of the two validated SNPs, we found a RegulomeDB score of 4 for both of *ETS1* rs11606640 and *ZFHX3* rs62053220. The results of predicted functional relevance by SNPinfo, RegulomeDB and HaploReg v4.1 are shown in Supplementary Table 8.

Discussion

In recent years, a pathway-based hypothesis-driven strategy for the post-GWAS analysis has generated some interesting results. Multiple functionally related genes in the same biological pathway may work together to confer susceptibility to disease development and progression, but independently they may not reach a significant threshold as defined in any single GWAS. Mechanistic studies in addition to the

pathway-based analysis can further help elucidate potential underlying mechanisms the SNPs may provide for the observed associations.

The c-Myb signaling pathway regulates different target genes in distinct cell types, suggesting a tissue-specific transcriptional regulation. The c-Myb protein has a molecular weight of approximately 75kDa, consisting of three major domains: an N-terminal DNA binding domain (DBD), a central transactivation domain (TAD), and a C-terminal-negative regulatory domain (NRD) for its function. Initial research of the c-Myb protein focused on its oncogenic potential in leukemia, but more recently c-Myb overexpression has been shown to be relevant to the development and metastasis of certain epithelial cancers, including cancers of the lung, breasts and colon.^{11,12} Considering that c-Myb signaling may play a possible key role in lung carcinogenesis and progression, we explored the associations between thousands of SNPs in this pathway genes with survival of NSCLC patients.

In the present study, we found that variant genotypes of two SNPs (i.e., *ETS1* rs11606640 GA+AA and *ZFX3* rs62053220 CG+GG), as well as their combined effect, were independently and significantly associated with a favorable survival outcome of NSCLC patients. Additional eQTL analysis of the target tissues and tissue-specific survival analysis suggests that the *ETS1* genes may be tumor suppressors, because *ETS1* mRNA expression was lower in lung cancer tissue than that in adjacent normal tissue, and the variant *ETS1* rs11606640 A genotypes were associated with an increased mRNA expression in tumor tissues and because the variant-associated high expression improved survival of the NSCLC patients. However, the fact that we failed to find a correlation between these two SNPs and their mRNA levels in lymphoblastoid cells may further reflect that such gene expression levels are indeed tissue-specific. Therefore, the molecular mechanisms underlying the observed associations needs additional investigations.

ETS1 rs11606640 is located on chromosome 11q24.3, encoding the ETS1 protein, a founding member of the ETS family of transcription factors, which likely binds to DNA through the action of an 85 amino acid DNA-binding domain. *ETS1* is known to target a wide variety of genes involved in proliferation, angiogenesis, apoptosis differentiation, migration and invasion of lung cancer. A previous GWAS reported that another two *ETS1* SNPs (rs4937362 and rs4938573) were associated with risk of haematopoietic malignancies; but there is no report for an association of *ETS1* SNPs with lung cancer survival. In the present study, we reported an association between *ETS1* rs11606640 and survival in NSCLC patients.

However, *ETS1* is considered as a marker of poor prognosis in several other cancer. For example, an elevated *ETS1* expression has been observed in several invasive and metastatic solid tumors, such as lymph node metastasis. In addition, ETS1 functions as a tissue-specific transcription factor, targeting genes including *cyclin D1*, *cyclin D3*, *MMP-1*, *MMP-3* and *MMP-9*, which are key elements in the process of carcinogenesis and metastasis. Therefore, the exact mechanisms underlying lung cancer progression remain to be investigated.

ZFHX3 rs62053220 is located on chromosome 16q22.2-q22.3, encoding a zinc finger homeobox, also named *ATBF1* (the AT motif binding factor 1), which is a large transcription factor with 23 zinc fingers and four homeodomains. *ZFHX3* functions in multiple biological processes, including embryonic development, neuronal differentiation, and neuronal death in response to DNA damage or oxidative stress, and *ZFHX3* abnormalities play a role in multiple human diseases, including cancer. Several published studies on *ZFHX3* SNPs have focused on atrial fibrillation and obesity; although the present study identified *ZFHX3* rs62053220 to be associated with survival in NSCLC patients.

The existence of four homeoboxes in *ZFHX3* suggests its dynamic function in biological processes, because few homeobox genes have more than one homeobox that interacts with other transcription factors to regulate the transcription of many genes. *ZFHX3* is thought to be a tumor suppressor through multiple mechanisms, because its expression has been frequently reported to be reduced in multiple types of cancers; which is consistent with our results using the TCGA database and also supported by animal studies. For example, the deletion of *ZFHX3* in mice led to disturbance of epithelial homeostasis and neoplastic morphology. Because *ZFHX3* is normally localized in the nucleus, mis-localization to the cytoplasm in cancer cells has been reported to be associated with disease progression and decreased survival. In a study that assessed *ZFHX3* mRNA expression in NSCLC tissue samples from 140 patients, a poor five-year overall survival rate among patients was associated with low levels of *ZFHX3* expression. Therefore, additional investigations are needed to unravel the role of *ZFHX3* in tumor progression.

The present study has some limitations. First, because the GWAS datasets were only available for Caucasian populations and because genetic variant analysis can be influenced by different ethnic backgrounds, our findings may not be generalizable to other ethnic populations. Second, some significant top SNPs in the PLCO dataset were not validated in the HLCS dataset possibly as a result of bias in patient selection as shown by different distributions of baseline characteristics in the two datasets. Third, although the eQTL analysis in the TCGA database showed *ETS1* rs11606640 and *ZFHX3* rs62053220 genotypes were significantly associated with their corresponding mRNA expression levels, these results were only found in tumor tissues but not in lymphoblastoid cells, reflecting a genetically determined tissue-specific gene expression that needs additional investigations. Finally, because the discovery and validation datasets had limited clinical information for a solid analysis for the independent contribution of the identified SNPs to the survival prediction, larger studies with more detailed clinical information are needed to further validate our findings.

Abbreviations

NSCLC non-small cell lung cancer

SNP single nucleotide polymorphism

PLCO Prostate, Lung, Colorectal and Ovarian cancer trial

GWASs genome-wide association studies

MSigDB Molecular signatures database

OS overall survival

DSS disease specific survival

HR hazard ratio

95%CI 95%Confidence interval

FPRP false positive report probability

LD linkage disequilibrium

eQTL expression quantitative trait loci

TCGA the Cancer Genome Atlas

Declarations

Ethics approval and consent to participate

The study protocol was reviewed and approved by the institutional review board of the National Cancer Institute (NCI), and a written informed consent was obtained from each participant.

Consent for publication

Not applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article [nd its supplementary information files.

Competing interests

The authors declare that they have no competing interests.

Funding

Qingyi Wei was supported by Lung Cancer Initiative and 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). The Harvard Lung Cancer Susceptibility Study was supported by NIH grants CA092824, CA074386, and CA090578 to David C. Christiani.

Authors' contributions

YG: Code validation, data analysis, manuscript preparation. YF: Data analysis, manuscript preparation. HLL: Data acquisition, code validation, data interpretation. SL: Data analysis. PGM: Data interpretation. EFPG: Manuscript preparation. LS: Data analysis. SS: Data interpretation. DCC: Data interpretation. QYW: Data acquisition, code validation, data interpretation, manuscript preparation.

All authors read and approved the final manuscript.

Acknowledgments

The authors thank the National Cancer Institute for access to data collected by the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI. The author would also like to acknowledge dbGaP repository for providing the cancer genotyping datasets. The accession numbers for the datasets of lung cancer are phs000336.v1.p1 and phs000093.v2.p2. Funding support for the GWAS of Lung Cancer and Smoking was provided through the NIH Genes, Environment and Health Initiative [GEI] (Z01 CP 010200). The human subjects participating in the GWAS derive from The Environment and Genetics in Lung Cancer Etiology (EAGLE) case-control study and the Prostate, Lung Colon and Ovary Screening Trial and these studies are supported by intramural resources of the National Cancer Institute. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies, GENEVA Coordinating Center (U01HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Johns Hopkins University Center for Inherited Disease Research, was provided by the NIH GEI (U01HG004438). PLCO was also supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. The authors thank PLCO screening center investigators and staff, and the staff of Information Management Services Inc. and Westat Inc. Most importantly, we acknowledge trial participants for their contributions that made this study possible. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute The Cancer Genome Atlas (TCGA) Research Network can be found at "<http://cancergenome.nih.gov>". The TCGA SNP data analyzed here are requested through dbGAP (accession#: phs000178.v1.p1).

References

1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2017. *CA Cancer J Clin.* 2017;67:7-30.
2. Torre LA, Siegel RL, Jemal A. Lung Cancer Statistics. *Adv Exp Med Biol.* 2016;893:1-19.
3. Wang JZ, Xiang JJ, Wu LG, Bai YS, Chen ZW, Yin XQ, et al. A genetic variant in long non-coding RNA MALAT1 associated with survival outcome among patients with advanced lung adenocarcinoma: a survival cohort analysis. *BMC Cancer.* 2017;17(1):167.
4. Boland JM, Maleszewski JJ, Wampfler JA, Voss JS, Kipp BR, Yang P, et al. Pulmonary invasive mucinous adenocarcinoma and mixed invasive mucinous/nonmucinous adenocarcinoma- a clinicopathological and molecular genetic study with survival analysis. *Hum Pathol.* 2018;71:8-19.
5. Fan Y, Zhu X, Xu Y, Lu X, Xu Y, Wang M, et al. Cell-cycle and DNA-damage response pathway is involved in leptomeningeal metastasis of non-small cell lung cancer. *Clin Cancer Res.* 2018;1;24(1):209-16.
6. Sausville LN, Jones CC, Aldrich MC, Blot WJ, Pozzi A, Williams SM. Genetic variation in the eicosanoid pathway is associated with non-small-cell lung cancer (NSCLC) survival. *PLoS One.* 2017;12(7):e0180471.
7. Farrell KA, Withers SB, Holt CM. C-Myb function in the vessel wall. *Front Biosci (Elite Ed).* 2011;3:968-77.
8. Ramsay RG, Gonda TJ. MYB function in normal and cancer cells. *Nat Rev Cancer.* 2008;8(7):523–34.
9. Nakano K, Uchimar K, Utsunomiya A, Yamaguchi K, Watanabe T. Dysregulation of c-Myb pathway by aberrant expression of proto-oncogene MYB provides the basis for malignancy in adult T-cell Leukemia/lymphoma Cells. *Clin Cancer Res.* 2016;1;22(23):5915-28.
10. Kim D, You E, Jeong J, Ko P, Kim JW, Rhee S. DDR2 controls the epithelial mesenchymal-transition-related gene expression via c-Myb acetylation upon matrix stiffening. *Sci Rep.* 2017;28:7(1):6847.
11. Wang W, Wu S, Shi Y, Miao Y, Luo X, Ji M, et al. c-Myb regulates cell growth and DNA damage repair through modulating MiR-143. *FEBS let.* 2015;589(5):555-64.
12. Hayes RB, Sigurdson A, Moore L, Peters U, Huang WY, Pinsky P, et al. Methods for etiologic and early marker investigations in the PLCO trial. *Mutat Res.* 2005;592:147-54.
13. 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(10):722-31.
14. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic acids research.* 2014;42:D975-9.
15. 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.
16. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet.* 2008;17:R156-65.
17. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet.* 2013;93(5):779-97.
18. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 2005;21(2):263-5.
19. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 2010;26(18):2336-7.
20. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics.* 2007;23:1294-6.
21. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res.* 2001;125(1-2):279-84.
22. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt.* 2014;34(5):502-8.
23. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst.* 2004;96(6):434-42.
24. Xu ZL, Taylor JA. SNPinfo: integrating GWAS and candidate

gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* 2009;37:W600-5. 25. 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. 26. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 2012;40:D930-4. 27. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* 2014;511:543-50. 28. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.* 2013;501(7468):506-11. 29. 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(11):2592-601. 30. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genome wide association studies. *Am. J. Hum. Genet.* 2007;81,1278–83. 31. Bengtson M, Klepper K, Gundersen S, Cuervo I, Drabløs F, Hovig E, et al. c-Myb binding sites in haematopoietic chromatin landscapes. *PLoS One.* 2015;10(7):e0133280. 32. Pattabiraman DR, Gonda TJ. Role and potential for therapeutic targeting of MYB in leukemia. *Leukemia.* 2013;27:269–77. 33. Malaterre J, Carpinelli M, Ernst M, Alexander W, Cooke M, Sutton S, et al. c-Myb is required for progenitor cell homeostasis in colonic crypts. *Proc Natl Acad Sci USA.* 2007;104,3829–34. 34. Li Y, Jin K, Van Pelt GW, van Dam H, Yu X, Mesker WE, et al. c-Myb Enhances Breast Cancer Invasion and Metastasis through the Wnt/beta-Catenin/Axin2 Pathway. *Cancer research.* 2016;76(11):3364-75. 35. Werner MH, Clore M, Fisher CL, Fisher RJ, Trinh L, Shiloach J, et al. The solution structure of the human ETS1-DNA complex reveals a novel mode of binding and true side chain intercalation. *Cell.* 1995;83:761-71. 36. Higuchi T, Bartel FO, Masuya M, Deguchi T, Henderson KW, Li R, et al. Thymomegaly, microsplenial, and defective homeostatic proliferation of peripheral lymphocytes in p51-Ets1 isoform-specific null mice. *Mol Cell Biol.* 2007;27: 3353–66. 37. Skibola CF, Berndt SI, Vijai J, Conde L, Wang Z, Yeager M, et al. Genome-wide association study identifies five susceptibility loci for follicular lymphoma outside the HLA region. *American journal of human genetics.* 2014; 95(4):462–71. 38. Katayama S, Nakayama T, Ito M, Naito S, Sekine I. Expression of the ets-1 proto-oncogene in human breast carcinoma: differential expression with histological grading and growth pattern. *Histol Histopathol.* 2005;20:119–26. 39. Seth A, Watson DK. ETS transcription factors and their emerging roles in human cancer. *Eur J Cancer.* 2005;41:2462–78. 40. Sementchenko VI, Watson DK. Ets target genes: past, present and future. *Oncogene.* 2000;19: 6533–49. 41. Miura Y, Tam T, Ido A, Morinaga T, Miki T, Hashimoto T, et al. Cloning and characterization of an ATBF1 isoform that expresses in a neuronal differentiation-dependent manner. *J. Biol. Chem.* 1995;270,26840–8. 42. Li M, Zhao D, Ma G, Zhang B, Fu X, Zhu Z, et al. Upregulation of ATBF1 by progesterone-PR signaling and its functional implication in mammary epithelial cells. *Biochem Biophys Res Commun.* 2012;430:358–63. 43. Burgner D, Davila S, Breunis WB, Ng SB, Li Y, Bonnard C, et al. A genome-wide association study identifies novel and functionally related susceptibility loci for Kawasaki disease. *PLoS Genet.* 2009;5(1):e1000319. 44. Zaw KTT, Sato N, Ikeda S, Thu KS, Mieno MN, Arai T, et al. Association of ZFHX3 gene variation with atrial fibrillation, cerebral infarction, and lung thromboembolism: An autopsy study. *J Cardiol.* 2017;70(2):180-4. 45. SA Yang. Association study between ZFHX3 gene polymorphisms and obesity in Korean population. *Journal of Exercise Rehabilitation.* 2017;13(4):491-4. 46. Yasuda H, Mizuno A, Tamaoki T,

Morinaga T. ATBF1, a multiple-homeodomain zinc finger protein, selectively down-regulates AT-rich elements of the human α -fetoprotein gene. *Mol. Cell Biol.* 1994;14, 1395–1401. 47. Miura Y, Kataoka H, Joh T, Tada T, Asai K, Nakanishi M, et al. Susceptibility to killer T cells of gastric cancer cells enhanced by Mitomycin-C involves induction of ATBF1 and activation of p21 (Waf1/Cip1) promoter. *Microbiol Immunol.* 2004;48(2):137–45. 48. Mabuchi M, Kataoka H, Miura Y, Kim TS, Kawaguchi M, Ebi M, et al. Tumor suppressor, AT motif binding factor 1 (ATBF1), translocates to the nucleus with runt domain transcription factor 3 (RUNX3) in response to TGF- β signal transduction. *Biochemical and Biophysical Research Communications.* 2010;398:321–5. 49. Kim CJ, Song JH, Cho YG, Cao Z, Lee YS, Nam SW, et al. Down-regulation of ATBF1 is a major inactivating mechanism in hepatocellular carcinoma. *Histopathology.* 2008;52:552–9. 50. Sun X, Xing C, Fu X, Li J, Zhang B, Frierson HF Jr, et al. Additive effect of *Zfhx3/Atbf1* and *Pten* deletion on mouse prostatic tumorigenesis. *J. Genet. Genomics.* 2015;42:373–82. 51. Minamiya Y, Saito H, Ito M, Imai K, Konno H, Takahashi N, et al. Suppression of Zinc Finger Homeobox 3 expression in tumor cells decreases the survival rate among non-small cell lung cancer patients. *Cancer Biomark.* 2012;11(4):139–46.

Table And Additional File Details

Due to technical limitations, Tables 1-3 are only available as a download in the supplemental file section.

Supplementary Figure 1: Manhattan plot of 8587 SNPs of the c-Myb pathway genes in the PLCO trial. The statistical values across the autosomes for associations between 8587 SNPs and overall survival are plotted as $-\log_{10}$ P values. The blue horizontal line indicates $P=0.05$ and the red line indicates FPRP = 0.10.

Supplementary Figure 2: A. Regional association plots. B. Kaplan-Meier analysis for patients with NSCLC by the two genes. C. Correlations between the two validated SNPs and mRNA expression levels of the associated genes by using the data of 373 European descents in the 1000 Genomes project. A. The left-hand Y-axis shows the $-\log_{10}$ transformation of P-value of individual SNPs, which is plotted against the chromosomal base-pair position with an expansion of 500 KB in the flanks of the gene region. The right-hand Y-axis shows the recombination rate estimated for European populations from HapMap Data Rel 22/phase II. a. *ETS1* rs11606640; b. *ZFHX3* rs62053220. B. Based on online survival analysis software (www.kmplot.com/analysis), a. low *ETS1* expression were associated with poorer survival of NSCLC; b. low *ZFHX3* expression were associated with poorer survival of NSCLC. C. a. *ETS1* rs11606640 in an additive model, ($P=0.459$); b. *ZFHX3* rs62053220 in an additive model, ($P=0.979$).

Figures

Figure 1

Study workflow chart. (SNP = single nucleotide polymorphism; PLCO = Prostate, Lung, Colorectal and Ovarian cancer trial; FPRP = false-positive report probability; eQTL = expression quantitative trait loci)

Figure 2

Kaplan-Meier (KM) survival curves for NSCLC patients of two validated SNPs and combined protective genotypes in the PLCO trial. a. Dichotomized groups of combined protective genotypes for overall survival in male patients ($P=0.034$); b. Dichotomized groups of combined protective genotypes for overall survival in patients >71 years ($P=0.007$); c. Dichotomized groups of combined protective genotypes for overall survival in patients with tumor stage IIIB-V ($P=0.002$); d. Dichotomized groups of combined protective genotypes for overall survival in patients with no surgery ($P=0.002$).

Figure 3

Comparison of mRNA expression levels of ETS1 and ZFH3 between lung cancer tissue and normal lung tissues in TCGA. a. The ETS1 mRNA expression levels in lung cancer tissues were significantly lower than that in normal lung tissues ($P < 0.001$); b. The ETS1 mRNA expression levels in lung adenocarcinoma tissues were significantly lower than that in normal lung tissues ($P < 0.001$); c. The ETS1 mRNA expression levels in lung squamous tissues were significantly lower than that in normal lung tissues ($P < 0.001$); d. The ZFH3 mRNA expression levels in lung cancer tissues were significantly lower than that in normal

lung tissues ($P=0.018$); e. The ZFHX3 mRNA expression levels in lung adenocarcinoma tissues were significantly lower than that in normal lung tissues ($P=0.037$); f. The ZFHX3 mRNA expression levels in lung squamous tissues were significantly lower than that in normal lung tissues ($P=0.002$); ETS1_t= lung cancer tissues; ETS1_n= adjacent normal lung tissues; ZFHX3_t= lung cancer tissues; ZFHX3_n= adjacent normal lung tissues.

Figure 4

eQTL analyses of ETS1 rs11606640, ZFHX3 rs62053220 genotypes and corresponding gene mRNA expression. All the data were from TCGA dataset. a. rs11606640 additive model in lung squamous cancer tissue ($P=0.045$); b. rs11606640 recessive model in lung squamous cancer tissue ($P=0.042$); c. rs62053220 dominant model in lung adenocarcinoma tissue ($P=0.048$).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigures12.pdf](#)
- [SupplementaryTables18.pdf](#)
- [Tables13.pdf](#)