Analysis of immune-related signatures of lung adenocarcinoma identified two distinct subtypes: implications for immune checkpoint blockade therapy

Qinghua Wang¹, Meiling Li¹, Meng Yang², Yichen Yang², Fengju Song¹, Wei Zhang³, Xiangchun Li², Kexin Chen¹

¹Department of Epidemiology and Biostatistics, National Clinical Research Center for Cancer, Key Laboratory of Molecular Cancer Epidemiology of Tianjin, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300060, China
²Tianjin Cancer Institute, National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy of Tianjin, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300060, China
³Wake Forest Baptist Comprehensive Cancer Center, Wake Forest Baptist Medical Center, Winston-Salem, NC 27157, USA

Correspondence to: Kexin Chen, Xiangchun Li; email: chenkexin@tmu.edu.cn, lixiangchun@tmu.edu.cn

Keywords: lung adenocarcinoma, immune subtyping, immunotherapy, prognosis

Received: September 18, 2019   Accepted: January 27, 2020   Published: February 24, 2020

Copyright: Wang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ABSTRACT

Immune checkpoint blockade (ICB) therapies have revolutionized the treatment of human cancers including lung adenocarcinoma (LUAD). However, our understanding of the immune subtyping of LUAD and its association with clinical response of immune checkpoint inhibitor remains incomplete. Here we performed molecular subtyping and association analysis of LUAD from the Cancer Genome Atlas (TCGA) and validated findings from TCGA cohort in 9 independent validation cohorts. We conducted consensus molecular subtyping with nonnegative matrix factorization (NMF). Potential response of ICB therapy was estimated with Tumor Immune Dysfunction and Exclusion (TIDE) algorithm. We identified 2 distinct subtypes of LUAD in TCGA cohort that were characterized by significantly different survival outcomes (i.e., high- and low-risk subtypes). The high-risk subtype was featured by lower TIDE score, upregulation of programmed death-ligand 1 (PD-L1) expression, and higher tumor mutation burden (TMB). The high-risk subtype also harbored significantly elevated cell cycle modulators CDK4/CDK6 and TP53 mutation. These observations were validated in 9 independent LUAD cohorts. Our findings suggest that immune checkpoint blockade therapy may be efficacious for high-risk subtype of LUAD patients.

INTRODUCTION

Lung adenocarcinoma (LUAD) is a leading cause of cancer related death worldwide and accounts for about 40% of lung cancer patients [1]. Over the last decade, treatment regimens that target epidermal growth factor receptor and anaplastic lymphoma kinase have only benefitted a small fraction of LUAD patients [2, 3]. Clinically defined molecular subtyping of LUAD is in urgent need for precise treatment.

Programmed cell death protein 1/programmed cell death-ligand 1 (PD-1/PD-L1) axis is a critical immune checkpoint pathway that could downregulate response of immune system in lung cancer [4]. Enhanced PD-L1 expression level on tumor cells or tumor infiltrating lymphocytes leads to T cell exhaustion [5], therefore decreasing tumor-specific immune capacity and promoting tumor proliferation [6, 7]. Immune checkpoint blockade (ICB) therapy aimed at PD-1/PD-L1 could repress the negative regulatory signaling and unleash T
cell from exhausted status [7]. ICB therapy can reverse the immunosuppressive microenvironment by decreasing the potential of tumor immune escape, therefore, resulting in noteworthy improvement of prognosis [8, 9]. High expression level of PD-L1 was approved by The Food and Drug Administration (FDA) as a biomarker to receive pembrolizumab therapy in non-small-cell lung cancer (NSCLC) [8].

Tumor mutation burden (TMB) is an indicator implicated in better response to ICB treatment, which is independent of PD-L1 expression level [10, 11]. An integrated analysis of 27 cancer types reported positive association between TMB and benefit of ICB treatment [12]. In present clinical practice, the proportion of patients benefitted from ICB therapy remains low. Therefore, the development of new biomarkers to select patient for ICB response rate is definitely required [13, 14].

Tumor Immune Dysfunction and Exclusion (TIDE) algorithm is a computational method using gene expression profile to predict the ICB response in NSCLC and melanoma [15]. TIDE uses a set of gene expression markers to estimate 2 distinct mechanisms of tumor immune evasion, including dysfunction of tumor infiltration cytotoxic T lymphocytes (CTL) and exclusion of CTL by immunosuppressive factors. Patients with higher TIDE score have a higher chance of antitumor immune escape, thus exhibiting lower response rate of ICB treatment [15]. The TIDE score was shown to have a higher accuracy than PD-L1 expression level and TMB in predicting survival outcome of cancer patients treated with ICB agents [15–18]. Several recent studies have reported its utility in predicting or evaluating the ICB therapy efficacy [18–22].

In this study, we performed consensus clustering analysis for LUAD based on immune gene expression signatures and identified 2 subtypes. These 2 subtypes were characterized by significantly different survival outcomes (i.e., low- versus high-risk subtypes), TIDE scores, PD-L1 expression, TMB and enrichment of cell cycle signaling. The high-risk subtype is presumably efficacious towards ICB treatment.

RESULTS

Identification of LUAD subtypes with prognostic significance

A list of 2,995 immune-related genes (See Materials and Methods) was compiled, 433 of which were significantly associated with overall survival (all FDR < 0.05; Supplementary Table 1) in the TCGA LUAD cohort (n = 502). Using this 433-gene panel, we identified 2 distinct LUAD subtypes (Figure 1A and Supplementary Figure 1) by performing consensus clustering analysis. These 2 subtypes were significantly different in survival (HR: 1.99, 95% CI: 1.43-2.76, Log rank P < 0.001; Figure 1B) and we referred them as high-risk and low-risk subtypes. After adjusting for age, gender, stage and smoking status, this association remained statistically significant (HR: 1.80, 95% CI: 1.28-2.53, P < 0.001; Figure 1C).

To verify the findings in TCGA LUAD cohort, we applied consensus clustering analysis using the aforementioned 433 immune-related genes and performed survival analysis in 9 additional independent LUAD cohorts. Our analysis showed that these 2 distinct subtypes identified in TCGA LUAD cohort were also identified in these 9 validation datasets, in that high-risk group was associated with worse survival outcome (HR range: 1.69 [95% CI: 1.29-2.21, Log rank P < 0.001] to 4.29 [95% CI: 2.14-8.65, Log rank P < 0.001]; Figure 2A–2I). These associations remained statistically significant after controlling for other confounding factors (Supplementary Figure 2A–2I). Result of another LUAD dataset OSE81089 (n = 108) showed a consistent trend although did not reach statistical significance (HR: 1.59, 95% CI: 0.89-2.85, Log rank P = 0.11; Supplementary Figure 3A). No significant association was observed for lung squamous cell carcinoma from TCGA dataset (HR: 1.03, 95% CI: 0.78-1.35, Log rank P = 0.86; Supplementary Figure 3B).

We obtained an accuracy of 0.947 with 126 genes (Supplementary Figure 4A and Supplementary Table 2) by applying recursive feature elimination to reduce the number of genes. The prognostic significance of high-risk versus low-risk subtypes was maintained using these 126 genes in TCGA and 9 validation cohorts (HR range: 1.84 [95% CI: 1.37-2.48, Log rank P < 0.001] to 5.99 [95% CI: 2.72-12.23, Log rank P < 0.001]; Supplementary Figure 4B–4K).

Predictive ICB response of identified LUAD subtypes

In the TCGA LUAD cohort, the TIDE score was significantly lower in high-risk subtype compared with low-risk subtype (Wilcoxon rank-sum test, P < 0.001; Figure 3A). The difference remained statistically significant after adjusting for age, gender, stage and smoking status (OR: 0.13, 95% CI: 0.08-0.20, P < 0.001; Figure 3B). This association was verified in 9 independent cohorts using univariate analysis (Wilcoxon rank-sum test, P < 0.05; Figure 3C), and 8 of these 9 cohorts showed the same association through multivariate logistic model (P < 0.001; Supplementary Figure 5A–5I). These discoveries suggested that patients of high-risk subtype may be more sensitive to ICB therapy as judged by the TIDE score.
Differences of PD-L1 expression and TMB between 2 LUAD subtypes

The expression of PD-L1 was significantly higher in high-risk group versus low-risk group in TCGA (Wilcoxon rank-sum test, $P = 0.003$; Figure 4A). Consistent association was also observed in 7 of 9 validation cohorts (Wilcoxon rank-sum test, $P < 0.05$; Figure 4B). We could not validate this association in the other 2 cohorts (GSE68465 and GSE11969) due to the lack of PD-L1 probes on the expression chip used.

Patients in high-risk subtype had a significantly higher mutation load in TCGA LUAD samples (Wilcoxon rank-sum test, $P < 0.001$; Figure 4C). Associations between high-risk subtype and TMB remained statistically significant (OR: 3.99, 95% CI: 2.19-7.46, $P < 0.001$; Figure 4D) after taking into account mutational signatures implicated in age-related deamination of 5-methylcytosine (signature 1), overactivity of mRNA-editing enzyme APOBEC (signature 2) and tobacco smoking-related C>A mutations (signature 4) (Supplementary Figure 6), and mutations in genes related to DNA damage repair (DDR) such as POLE, BRCA1/2, and TP53. Additionally, we found that high-risk subtype harbored a significantly higher nonsynonymous mutation load (Wilcoxon rank-sum test, $P < 0.001$; Figure 4E).

Figure 1. Identification of high-risk and low-risk subtypes of LUAD in TCGA cohort by consensus clustering. (A) The relationship between cophenetic, dispersion and silhouette coefficients with respect to number of clusters. (B) Kaplan-Meier survival plot of the high-risk versus the low-risk subtype. (C) Forest plot representation of multivariate Cox model depicted association between overall survival and LUAD subtypes with other clinical factors taken into account.
In TCGA LUAD cohort, association of TMB with survival was not significant. Associations of TIDE score and PD-L1 expression with survival were not significant across TCGA LUAD cohort and the 9 validation cohorts (Supplementary Figure 7).

**Functional characterization of the high-risk subtype**

Cell cycle relevant signaling pathways were frequently activated in high-risk group as compared with low-risk group (FDR < 0.001; Figure 5A). Molecular markers involved in cell cycle checkpoint were significantly upregulated in the high-risk subtype (Wilcoxon rank-sum test, P < 0.001; Figure 5B). The cell cycle checkpoint markers analyzed include *CCND1*, *CCNE1*, *CDK2*, *CDK4*, and *CDK6* in G1/S checkpoint, and *CCNA2* and *CDK1* in G2/M checkpoint. Previous studies reported that CDK4 and CDK6 inhibitors could enhance T cell activity [23] and reverse the T cell exclusion signature [24] to obtain better ICB treatment response.

---

**Figure 2.** Kaplan-Meier plots of high-risk and low-risk subtypes of LUAD in 9 validation cohorts of (A) GSE72094, (B) GSE68465, (C) GSE50081, (D) GSE42127, (E) GSE41271, (F) GSE31210, (G) GSE30219, (H) GSE13213, and (I) GSE11969.
Figure 3. Distribution of TIDE scores in high-risk subtype of LUAD versus low-risk subtype of LUAD. (A) Boxplot representation of TIDE scores in the high-risk group versus low-risk group in TCGA LUAD cohort. (B) Forest plot representation of multivariate model with adjustment for confounding factors in TCGA cohort. (C) Distribution of TIDE scores in 9 independent validation cohorts.
Mutation patterns of SMGs in relation to LUAD subtypes

We identified 23 significantly mutated genes (SMGs) in TCGA LUAD cohort (Supplementary Figure 8). TP53, NAV3, COL11A1, KEAP1 and SMARCA4 were more frequently mutated in high-risk subtype (Fisher exact test, OR > 1, P < 0.05; Figure 6). Association of these SMGs with high-risk subtype remained significant after including age, gender, stage and smoking status (OR > 1, P < 0.01; Supplementary Figure 9). Higher mutation frequency of TP53 in high-risk subtype was observed in 3 validation cohorts that had TP53 mutation data (OR > 1, P < 0.01; Supplementary Figure 10). Mutation data of NAV3, COL11A1, KEAP1 and SMARCA4 of the aforementioned 3 validation cohorts was unavailable. In addition, we found that patients in high-risk subtype were more likely to be male, current smoker, and at advanced clinical stage (chi-square test, P < 0.001; Supplementary Table 3).

Figure 4. Distribution of PD-L1 expression and tumor mutation burden and their associations with the high-/low-risk subtypes of LUAD. (A, B) Difference in the PD-L1 expression in TCGA and validation cohorts stratified by high-/low-risk subtypes of LUAD. (C–E) Distribution and association of mutation burden in the high-risk group versus low-risk group.
DISCUSSION

We uncovered 2 prognostic subtypes of LUAD by analyzing 2,300 LUAD samples from TCGA LUAD cohort and 9 independent validation cohorts. The high-risk subtype has significantly lower TIDE score, higher PD-L1 expression and higher TMB. Pathway analysis suggested that high-risk group was featured by activated cell cycle signaling. TP53 mutation was more frequently mutated in the high-risk subtype.

We observed that TIDE score was significantly lower in high-risk subtype than low-risk subtype via univariate regression analysis in TCGA LUAD cohort and 9 validation cohorts. The difference remained statistically significant in multivariate analysis among these cohorts except GSE30219, this is probably due to smaller sample size of GSE30219 cohort (n = 85) as compared with the other LUAD cohorts. These findings suggested that patients of high-risk subtype may be more sensitive to ICB treatment.

Based on the KEYNOTE-001 clinical trial, PD-L1 high expression was an essential condition for the use of pembrolizumab in NSCLC [8, 25]. Several studies have reported that PD-L1 high expression was correlated to elevated response rate and survival benefit in ICB therapy of NSCLC, chronic lymphocytic leukemia and urothelial cancer [26–28]. However, in the clinical trial CHECKMATE-032 that enrolled patients with urothelial cancer, no significant difference was observed in response rate between PD-L1 positive and negative subgroups [29]. This suggests that our understanding of association between PD-L1 and ICB

![Functional annotation in high-/low-risk subtypes of LUAD](image)

**Figure 5.** Functional annotation in high-/low-risk subtypes of LUAD. (A) Dysregulation of signaling pathways stratified by identified LUAD subtypes. (B) Expression profiles of cell cycle checkpoint markers in the high-risk group versus low-risk group.
response is far from complete and that new markers associated with ICB response are needed. A possible explanation for this inconsistency is attributed to different criterion used for evaluating high and low PD-L1 expression [30]. TMB is emerging as a potential biomarker for predicting response of ICB therapy. Three clinical trials including KEYNOTE-001, CHECKMATE-026 and CHECKMATE-227 showed that patients with high TMB could benefit more from ICB therapy in NSCLC [31–33]. In our study, the high-risk subtype had significantly higher PD-L1 expression and TMB, suggesting a greater potential of ICB therapy response.

Consistent with our findings, Seo et al. identified the immune-defective and immune-competent subtypes in LUAD, and the immune-competent subtype was characterized by activated microenvironment and elevated expression of immune checkpoint genes [34]. In our study, we observed a significant prognostic difference between the high-risk and low-risk subtypes. However, a survival difference was not observed for the 2 subtypes proposed by Seo et al. A recent study from Song et al. also reported identification of the high-risk and low-risk subtypes of LUAD [35]. The high-risk subtype in this study exhibited worse survival outcome and a higher tumor mutation load, which supports our result that the high-risk subtype may be more responsive to immune checkpoint therapy because of the higher tumor mutation load. The prognostic significance of high-risk and low-risk subtypes from our study was validated in GSE31210, a validation dataset used in Song et al. research, whereas subtypes proposed by Song et al. was not significant. Taken together, the gene panels used in our study for molecular subtyping are more generalizable and robust given that we included more validation sets (n=9) as compared with 4 validation cohorts used by Song et al.

Cell cycle relevant pathways and checkpoint markers contributed mostly to the worse prognosis of high-risk subtype were significantly upregulated. These observations suggested the high-risk patients may be suitable to receive cell cycle inhibitors. Previous studies...

Figure 6. Mutational landscape of SMGs in TCGA LUAD cohort stratified by high-/low-risk subtypes. SMGs with significantly different mutation rate were highlighted in bold.
showed that CDK4 and CDK6 inhibitors enhanced the response of ICB in mouse models due to their ability to elevate expression of endogenous retroviruses [36] that was associated with T cell activity [23] and ICB response [37]. Another study reported that the T cell exclusion signature was predictive of poor ICB response, however, CDK4 and CDK6 inhibitors could reverse this signature to get a better response in \textit{in vitro} experiments in melanoma [24]. The combination therapy of cell cycle inhibitors and ICB agents may be more effective for the high-risk patients.

\textit{TP53} was frequently mutated in the high-risk subtype and its mutation was reported to be associated with poorer prognosis [38, 39]. Patients harbored mutations of \textit{TP53} had a higher TMB owing to loss of DNA repair function (Figure 4D). Recent study reported that \textit{TP53} mutations significantly induced the expression of immune checkpoints and activated T-effector and interferon-\(\gamma\) signature in LUAD, suggesting \textit{TP53} mutation patients would be more responsive to checkpoint blockade [40]. Patients of high-risk subtype were more likely to be current smokers. Higher response rate of ICB treatment was observed in smokers of NSCLC [41], typically due to high mutation burden generated by the mutagenic effects of cigarette smoke [42]. These 2 factors may underlie the response to ICB therapy for patients of high-risk group.

There are several limitations in this study. Firstly, gene expression data used in our study was from different platforms, this difference may introduce bias in the analysis procedure. Secondly, results derived from TCGA LUAD mutational landscape were not validated in independent datasets due to the unavailability of mutation data in validation sets. Thirdly, we lacked an in-house validation set.

In summary, we identified 2 prognostically and clinically relevant subtypes of LUAD. Molecular markers suggest that patients from the high-risk subtype may be more responsive to ICB therapy, which needs to be tested in future clinical trials.

\textbf{MATERIALS AND METHODS}

\textbf{Collection of genomic data}

We collected gene expression profiles of 502 LUAD samples from TCGA-LUAD cohort (https://gdc.cancer.gov) and 1,798 samples from 9 cohorts in Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/geo/) (Supplementary Table 4). In total, we obtained 2,300 LUAD samples from 10 independent cohorts (Supplementary Table 5). We also collected clinical data for each cohort. All gene expression data were uniformly normalized. For genes with multiple probes, their expression levels were calculated as the mean expression level of these probes. Only TCGA cohort contained genomic mutation data used in our study. A flow chart to depict the study design was shown in Supplementary Figure 11.

\textbf{Identification of immune-related genes associated with prognosis}

We obtained 2,995 immune-related genes from 160 immune signatures curated in the TCGA pan-cancer immune landscape project [43], which was based on 11 immune relevant studies [44–54] (Supplementary Materials and Methods). We used univariate Cox proportional hazards model to examine the associations between gene expression and overall survival. Genes with false discovery rate (FDR) less than 0.05 were considered to be statistically significant and included in consensus clustering analysis. Further feature selection was conducted by using recursive feature elimination (RFE) with random forest as classifier and 10-fold cross-validation method in \textit{R} package \textit{caret} (version 6.0-82).

\textbf{Consensus molecular subtyping with NMF}

We used nonnegative matrix factorization (NMF) to perform molecular subtyping [55, 56]. Specifically, NMF was applied to gene expression matrix \(A\) that contained prognostically significant immune-related genes aforementioned. Matrix \(A\) was factorized into 2 nonnegative matrices \(W\) and \(H\) (i.e., \(A=WH\)). Repeated factorization of matrix \(A\) was performed and its outputs were aggregated to obtain consensus clustering of LUAD samples. The optimal number of subtypes was selected according to cophenetic, dispersion, and silhouette coefficients [57]. The consensus clustering was conducted with \textit{R} package \textit{NMF} (version 0.21.0) [58]. The NMF method was also used to extract mutational signatures based on the framework proposed by a previously study [59]. The \textit{R} code was available in Supplementary Materials and Methods.

\textbf{Prediction of ICB therapy response}

Potential ICB response was predicted with TIDE algorithm [15].

\textbf{Gene set enrichment analysis}

The \textit{R} package \textit{limma} (version 3.38.3) [60] and \textit{DESeq2} (version 1.22.2) [61] were used to calculate the differential expressed \(t\) statistics for microarray and RNA sequencing data. We used \(t\) statistic as input to \(R\) function \textit{fgsea} that implemented in \textit{fgsea} package.
(version 1.6.0) to perform gene set enrichment analysis (GSEA). Signaling pathways in Molecular Signatures Database (MSigDB) were used in GSEA [62].

**Identification of significantly mutated genes**

Significantly mutated genes (SMGs) were identified using MutSigCV algorithm [63]. The significant enrichment of non-silent somatic mutations of a gene was measured by MutSigCV through addressing mutational context specific background mutation rates. A gene was considered an SMG if it meets these criteria: statistically significant (q < 0.1), expressed in TCGA LUAD data [64] and encyclopedia of cell lines [65], and mutation rate greater than 3%.

**Statistical analyses**

R software 3.5.1 was applied in this study for the statistical analyses. Univariate and multivariate Cox proportional hazards model were used to analyze the association between subtypes and prognosis with R survival package. Survival curve was drawn with Kaplan-Meier method and Log-rank test was used to evaluate difference between survival curves. Association between mutation rate of SMGs and 2 subtypes was evaluated by Fisher’s exact test. Multivariate logistic regression was performed to test the association between SMGs and identified LUAD subtypes by taking into account confounding factors. The continuous and categorical variables between 2 subtypes were compared using two-sided Wilcoxon rank-sum test and chi-square test, respectively. Benjamini-Hochberg method was used to adjust for multiple hypothesis testing [61].

**Abbreviations**


**AUTHOR CONTRIBUTIONS**

KXC and XCL designed this study; KXC, XCL, and QHW developed the methodology and acquired the related data; XCL, QHW, MLL, MY, YCY, and FJS performed data analysis and interpretation; XCL, KXC, WZ and QHW drafted and revised the manuscript; KXC and XCL supervised this study.

**ACKNOWLEDGMENTS**

We would like to thank Prof. Wei Zhang at the Wake Forest Baptist Comprehensive Cancer Center for his valued suggestions and editing of this manuscript.

**CONFLICTS OF INTEREST**

All authors declare no conflicts of interest.

**FUNDING**

The study has been supported by the Program for Changjiang Scholars and Innovative Research Team in University in China (IRT_14R40) and National Natural Science Foundation of China (NSFC) (no. 31801117).

**REFERENCES**


www.aging-us.com 3321 AGING


SUPPLEMENTARY MATERIALS

2995 immune-related genes

We obtained these 2995 immune-related genes from 160 immune signatures curated in a previous study [1], which was based on 11 immune relevant studies. Eighty-three of these 160 immune signatures were cancer immune response-related signatures, whereas the remaining 77 signatures are of general validity for immunity. These 83 signatures consisted of 68 gene sets from a earlier study [2], 9 signatures derived from TCGA gene expression data (immune metagene attractors) [3, 4], 3 signatures representing immune contexture function [5–7], and 3 signatures from a recent study [8]. The remaining 77 signatures comprised of 45 immune-cell specific signatures from 2 sources [9, 10], and 32 signatures from the ImmuneSigDB [11, 12].

Main R codes used in this study

# Nonnegative matrix factorization (NMF)
library(NMF)
estim.r <- nmf(gene_expression_matrix, 2:6, 
nrun = 200, method = 'brunet')
plot(estim.r)
consensusmap(estim.r)
fit <- nmf(gene_expression_matrix, 2, nrun = 200, method = "brunet")
subtype.result <- predict(fit)

# Gene set enrichment analysis (GSEA)
library(fgsea)
library(ggplot2)
fgsea.result <- fgsea(pathways = annotation_pathways,
                         stats = genes_rank_list,
                         minSize = 15,
                         maxSize = 500,
                         nperm = 1000000)
plotEnrichment(path[["pathway_name"]],
genes_rank_list)

# Waterfall plot
library(GenVisR)
waterfall(mutation_data, plotGenes = genes_to_plot,
          mainDropMut = TRUE, coverageSpace = 300000000,
          clinDat = clinical_data)

# Multivariate regression model
library(forestmodel)
library(survival)
forest_model(coxph(Surv(survival_time, survival_end) ~ variables, related_data),
             factor_separate_line = T)
forest_model(glm(categorical variable ~ variables,
               binomial(), related_data), factor_separate_line = T))

REFERENCES

PMID: 29628290

PMID: 24516633

PMID: 23468608

PMID: 23596202

PMID: 26967649

PMID: 23890060

PMID: 28344865


Supplementary Figures

Supplementary Figure 1. Heatmap representation of NMF clustering.
**Supplementary Figure 2.** Prognostic significance of high-risk versus low-risk subtypes using multivariate Cox model in 9 validation cohorts of (A) GSE72094, (B) GSE68465, (C) GSE50081, (D) GSE42127, (E) GSE41271, (F) GSE31210, (G) GSE30219, (H) GSE13213, and (I) GSE11969.
Supplementary Figure 3. (A) Kaplan-Meier plot of identified 2 subtypes in GSE81089 dataset. (B) Kaplan-Meier plot of identified TCGA LUSC subtypes.
Supplementary Figure 4. (A) Relation between classification accuracy and selected genes via recursive feature elimination algorithm. (B–K) Kaplan-Meier plot of identified subtypes using 126 genes in TCGA and 9 validation cohorts.
**Supplementary Figure 5.** Association of identified LUAD subtypes with TIDE score using multivariate logistic analysis in 9 validation cohorts of (A) GSE72094, (B) GSE68465, (C) GSE50081, (D) GSE42127, (E) GSE41271, (F) GSE31210, (G) GSE30219, (H) GSE13213, and (I) GSE11969.
Supplementary Figure 6. (A) Mutational signatures extracted from TCGA LUAD cohort and (B) their cosine similarity with COSMIC mutational signatures.
Supplementary Figure 7. Kaplan-Meier plots with respect to TIDE score, PD-L1 expression, and TMB in TCGA and 9 validation cohorts.
Supplementary Figure 8. Mutational landscape of SMGs in TCGA LUAD cohort.

Supplementary Figure 9. Associations between mutations in **TP53**, **NAV3**, **CLO11A1**, **KEAP1** and **SMARCA4** and identified 2 LUAD subtypes using multivariate logistic analysis.
Supplementary Figure 10. Associations of TP53 mutation with high-risk subtype in (A) GSE72094, (B) GSE13213 and (C) GSE11969.

Supplementary Figure 11. Flow chart of our study. TCGA and 9 public LUAD cohorts containing 2300 samples were included to perform relevant analyses.
Supplementary Tables

Please browse Full Text version to see the data of Supplementary Table 1

Supplementary Table 1. 433 genes with FDR less than 0.05 in univariate survival analysis.

Supplementary Table 2. 126 genes were obtained through RFE features selection.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Entrezid</th>
<th>Gene</th>
<th>Entrezid</th>
<th>Gene</th>
<th>Entrezid</th>
<th>Gene</th>
<th>Entrezid</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSPAN32</td>
<td>10077</td>
<td>NCR3</td>
<td>259197</td>
<td>MCM10</td>
<td>55388</td>
<td>CCNB1</td>
<td>891</td>
</tr>
<tr>
<td>RASGRP2</td>
<td>10235</td>
<td>ASPM</td>
<td>259266</td>
<td>DEPDC1</td>
<td>55635</td>
<td>SELENBP1</td>
<td>8991</td>
</tr>
<tr>
<td>CDKN3</td>
<td>1033</td>
<td>STAP1</td>
<td>26228</td>
<td>PBK</td>
<td>55872</td>
<td>PRC1</td>
<td>9055</td>
</tr>
<tr>
<td>NDC80</td>
<td>10403</td>
<td>AMPD1</td>
<td>270</td>
<td>KIF15</td>
<td>56992</td>
<td>CD1B</td>
<td>910</td>
</tr>
<tr>
<td>CENPA</td>
<td>1058</td>
<td>HPGDS</td>
<td>27306</td>
<td>PTGDS</td>
<td>5730</td>
<td>CD1C</td>
<td>911</td>
</tr>
<tr>
<td>CENPE</td>
<td>1062</td>
<td>GNG7</td>
<td>2788</td>
<td>SPC25</td>
<td>57405</td>
<td>CD1D</td>
<td>912</td>
</tr>
<tr>
<td>RAD51AP1</td>
<td>10635</td>
<td>LINC00926</td>
<td>283663</td>
<td>DNASE2B</td>
<td>58511</td>
<td>CD1E</td>
<td>913</td>
</tr>
<tr>
<td>PLK4</td>
<td>10733</td>
<td>C11orf21</td>
<td>29125</td>
<td>RGS13</td>
<td>6003</td>
<td>CCNB2</td>
<td>9133</td>
</tr>
<tr>
<td>KIF2C</td>
<td>11004</td>
<td>RACGAP1</td>
<td>29127</td>
<td>RRM2</td>
<td>6241</td>
<td>EXO1</td>
<td>9156</td>
</tr>
<tr>
<td>UBE2C</td>
<td>11065</td>
<td>HLA-DOB</td>
<td>3112</td>
<td>BLK</td>
<td>640</td>
<td>CD5</td>
<td>921</td>
</tr>
<tr>
<td>CHEK1</td>
<td>1111</td>
<td>BIRC3</td>
<td>330</td>
<td>NCAPG</td>
<td>64151</td>
<td>AURKB</td>
<td>9212</td>
</tr>
<tr>
<td>ZWINT</td>
<td>11130</td>
<td>BIRC5</td>
<td>332</td>
<td>SFTPB</td>
<td>6439</td>
<td>CACNA2D2</td>
<td>9254</td>
</tr>
<tr>
<td>OIP5</td>
<td>11339</td>
<td>IL16</td>
<td>3603</td>
<td>SLAMF1</td>
<td>6504</td>
<td>CD19</td>
<td>930</td>
</tr>
<tr>
<td>FCRL1</td>
<td>115350</td>
<td>KLRB1</td>
<td>3820</td>
<td>SLC18A2</td>
<td>6571</td>
<td>MS4A1</td>
<td>931</td>
</tr>
<tr>
<td>CCR6</td>
<td>1235</td>
<td>KIF11</td>
<td>3832</td>
<td>SPIB</td>
<td>6689</td>
<td>SIGLEC6</td>
<td>946</td>
</tr>
<tr>
<td>ADH1A</td>
<td>124</td>
<td>LY9</td>
<td>4063</td>
<td>AURKA</td>
<td>6790</td>
<td>KIF23</td>
<td>9493</td>
</tr>
<tr>
<td>ADH1B</td>
<td>125</td>
<td>MAD2L1</td>
<td>4085</td>
<td>BUB1B</td>
<td>701</td>
<td>CD40LG</td>
<td>959</td>
</tr>
<tr>
<td>COL4A3</td>
<td>1285</td>
<td>MAL</td>
<td>4118</td>
<td>TK1</td>
<td>7083</td>
<td>ESPL1</td>
<td>9700</td>
</tr>
<tr>
<td>CR2</td>
<td>1380</td>
<td>MKI67</td>
<td>4288</td>
<td>CLEC3B</td>
<td>7123</td>
<td>CD79A</td>
<td>973</td>
</tr>
<tr>
<td>CTSG</td>
<td>1511</td>
<td>MYBL2</td>
<td>4605</td>
<td>TOP2A</td>
<td>7153</td>
<td>CD79B</td>
<td>974</td>
</tr>
<tr>
<td>DNASE1L3</td>
<td>1776</td>
<td>NEK2</td>
<td>4751</td>
<td>TTK</td>
<td>7272</td>
<td>ACAP1</td>
<td>9744</td>
</tr>
<tr>
<td>FAM129C</td>
<td>199786</td>
<td>PAX5</td>
<td>5079</td>
<td>CCR2</td>
<td>729230</td>
<td>DLGAP5</td>
<td>9787</td>
</tr>
<tr>
<td>SKA1</td>
<td>220134</td>
<td>NUSAP1</td>
<td>51203</td>
<td>VIPR1</td>
<td>7433</td>
<td>CDK1</td>
<td>983</td>
</tr>
<tr>
<td>MS4A2</td>
<td>2206</td>
<td>GTSE1</td>
<td>51512</td>
<td>DSCC1</td>
<td>79075</td>
<td>MELK</td>
<td>9833</td>
</tr>
<tr>
<td>Fancer1</td>
<td>2208</td>
<td>PLK1</td>
<td>5347</td>
<td>ANKRD55</td>
<td>79722</td>
<td>CDC6</td>
<td>990</td>
</tr>
<tr>
<td>TMEM130</td>
<td>222865</td>
<td>POLE2</td>
<td>5427</td>
<td>CXor21</td>
<td>80231</td>
<td>CDC20</td>
<td>991</td>
</tr>
<tr>
<td>TPX2</td>
<td>22974</td>
<td>ERCC6L</td>
<td>54821</td>
<td>TRAF3IP3</td>
<td>80342</td>
<td>KIF14</td>
<td>9928</td>
</tr>
<tr>
<td>FOXM1</td>
<td>2305</td>
<td>PARPB</td>
<td>55010</td>
<td>TLR10</td>
<td>81793</td>
<td>CD302</td>
<td>9936</td>
</tr>
<tr>
<td>Ncaph</td>
<td>23397</td>
<td>CEP55</td>
<td>55165</td>
<td>CDC45</td>
<td>8318</td>
<td>CDC25C</td>
<td>995</td>
</tr>
<tr>
<td>TNFRSF13B</td>
<td>23495</td>
<td>FANCI</td>
<td>55215</td>
<td>ATP13A4</td>
<td>84239</td>
<td>HS3ST2</td>
<td>9956</td>
</tr>
<tr>
<td>LILRA4</td>
<td>23547</td>
<td>NEIL3</td>
<td>55247</td>
<td>FCRLA</td>
<td>84824</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIF4A</td>
<td>24137</td>
<td>HIURP</td>
<td>55355</td>
<td>CCNA2</td>
<td>890</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Supplementary Table 3. Distinct clinical characteristics between 2 LUAD subtypes.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Low risk ($n=210$)</th>
<th>High risk ($n=256$)</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>66.6±9.1</td>
<td>64.2±10.6</td>
<td>0.125</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>137 (65.2)</td>
<td>120 (46.9)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>73 (34.8)</td>
<td>136 (53.1)</td>
<td></td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>139 (66.2)</td>
<td>115 (44.9)</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>38 (18.1)</td>
<td>74 (28.9)</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>25 (11.9)</td>
<td>50 (19.5)</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>8 (3.8)</td>
<td>17 (6.6)</td>
<td></td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smoker</td>
<td>39 (18.6)</td>
<td>31 (12.1)</td>
<td></td>
</tr>
<tr>
<td>Never-smoker&gt;15 years</td>
<td>69 (32.9)</td>
<td>52 (20.3)</td>
<td></td>
</tr>
<tr>
<td>Non-smoker≤15 years</td>
<td>67 (31.9)</td>
<td>95 (37.1)</td>
<td></td>
</tr>
<tr>
<td>Current smoker</td>
<td>35 (16.7)</td>
<td>78 (30.5)</td>
<td></td>
</tr>
</tbody>
</table>

### Supplementary Table 4. Summary of TCGA and 9 validation datasets included in this study.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Platforms</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA</td>
<td>Illumina RNAseq HTSeq</td>
<td>502</td>
</tr>
<tr>
<td>GSE72094</td>
<td>Rosetta/Merck Human RSTA Custom Affymetrix 2.0 microarray</td>
<td>398</td>
</tr>
<tr>
<td>GSE68465</td>
<td>Affymetrix Human Genome U133A Array</td>
<td>442</td>
</tr>
<tr>
<td>GSE50081</td>
<td>Affymetrix Human Genome U133 Plus 2.0 Array</td>
<td>127</td>
</tr>
<tr>
<td>GSE42127</td>
<td>Illumina HumanWG-6 v3.0 expression beadchip</td>
<td>132</td>
</tr>
<tr>
<td>GSE41271</td>
<td>Illumina HumanWG-6 v3.0 expression beadchip</td>
<td>181</td>
</tr>
<tr>
<td>GSE31210</td>
<td>Affymetrix Human Genome U133 Plus 2.0 Array</td>
<td>226</td>
</tr>
<tr>
<td>GSE30219</td>
<td>Affymetrix Human Genome U133 Plus 2.0 Array</td>
<td>85</td>
</tr>
<tr>
<td>GSE13213</td>
<td>Agilent-014850 Whole Human Genome Microarray 4x44K G4112F</td>
<td>117</td>
</tr>
<tr>
<td>GSE11969</td>
<td>Agilent Homo sapiens 21.6K custom array</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2300</td>
</tr>
</tbody>
</table>
### Supplementary Table 5. Descriptions of clinical characteristics for LUAD patients in TCGA and 9 validation datasets.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>TCGA</th>
<th>GSE72094</th>
<th>GSE68465</th>
<th>GSE50081</th>
<th>GSE42127</th>
<th>GSE41271</th>
<th>GSE31210</th>
<th>GSE30219</th>
<th>GSE13213</th>
<th>GSE11969</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>502</td>
<td>398</td>
<td>422</td>
<td>127</td>
<td>132</td>
<td>181</td>
<td>226</td>
<td>85</td>
<td>117</td>
<td>90</td>
</tr>
<tr>
<td>Age (years)</td>
<td>65.3 ± 9.9</td>
<td>69.4 ± 9.5</td>
<td>64.5 ± 10.1</td>
<td>68.7 ± 9.7</td>
<td>65.8 ± 10.3</td>
<td>64.6 ± 10.4</td>
<td>59.6 ± 7.4</td>
<td>61.5 ± 9.3</td>
<td>60.7 ± 10.2</td>
<td>61.1 ± 9.8</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>209</td>
<td>174</td>
<td>219</td>
<td>65</td>
<td>67</td>
<td>91</td>
<td>105</td>
<td>66</td>
<td>60</td>
<td>47</td>
</tr>
<tr>
<td>Female</td>
<td>257</td>
<td>219</td>
<td>214</td>
<td>62</td>
<td>65</td>
<td>90</td>
<td>121</td>
<td>19</td>
<td>56</td>
<td>43</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>254</td>
<td>254</td>
<td></td>
<td>92</td>
<td>89</td>
<td>95</td>
<td>168</td>
<td></td>
<td>78</td>
<td>52</td>
</tr>
<tr>
<td>II</td>
<td>112</td>
<td>67</td>
<td>-</td>
<td>35</td>
<td>22</td>
<td>27</td>
<td>58</td>
<td>-</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>III</td>
<td>75</td>
<td>57</td>
<td>-</td>
<td>0</td>
<td>20</td>
<td>43</td>
<td>0</td>
<td>-</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>IV</td>
<td>25</td>
<td>15</td>
<td>-</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smoker</td>
<td>70</td>
<td>298</td>
<td>15</td>
<td>23</td>
<td>23</td>
<td>25</td>
<td>111</td>
<td></td>
<td>55</td>
<td>45</td>
</tr>
<tr>
<td>Ever smoker</td>
<td>0</td>
<td>30</td>
<td>94</td>
<td>56</td>
<td>-</td>
<td>156</td>
<td>115</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current smoker</td>
<td>113</td>
<td>-</td>
<td>16</td>
<td>36</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Non-smoker&gt;15 years</td>
<td>121</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Non-smoker≤ 15 years</td>
<td>162</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NA</td>
<td></td>
<td>65</td>
<td>90</td>
<td>12</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>