Identification of seven-gene signature for prediction of lung squamous cell carcinoma

Zhe Wang1,* Zhongmiao Wang1,* Xing Niu2 Jie Liu3 Zhunling Wang2 Lijie Chen4 Baoli Qin1

1Department of Gastrointestinal Oncology, Cancer Hospital of China Medical University, Shenyang 110042, Liaoning Province, People’s Republic of China; 2Department of Second Clinical College, Shengjing Hospital affiliated to China Medical University, Shenyang 110004, Liaoning Province, People’s Republic of China; 3Science Experiment Center of China Medical University, China Medical University, Shenyang 110012, Liaoning Province, People’s Republic of China; 4Department of Third Clinical College, China Medical University, Shenyang 110122, Liaoning Province, People’s Republic of China

*These authors contributed equally to this work

Background and aim: Lung squamous cell carcinoma (LUSC), is a pathological subtype of lung cancer, accounting for 30% of the lung cancers. A reliable model was constructed, based on the whole gene expression profiles, to predict the prognosis of patients with LUSC.

Methods: The RNA-Seq data of LUSC was downloaded from the TCGA database, and differentially expressed genes (p<0.05, |log2fold change| >1) were screened out. By univariate and multivariate Cox regression analysis, we identified seven prognosis-related genes.

Then, we established a risk score staging system to predict the prognosis of patients with LUSC. Compared with other clinical parameters, the risk score was an independent prognostic factor and had a better performance in predicting prognosis. Finally, GSEA analysis was carried out to determine the enrichment pathway significantly. The risk score models were established by Cox proportional hazard regression analysis; the ROC curve was applied to test the performance of risk score model. All the statistical analysis was accomplished by R packages.

Results: In this study, a model was constructed to predict prognosis, which contains seven genes: CSRNP1, CLEC18B, MIR27A, AC130456.4, DEFA6, ARL14EPL, and ZFP42. Based on the model, the risk score of each patient was calculated with LUSC (hazard ratio [HR] =2.673, 95% CI=1.871–3.525). It was found that the risk score can distinguish high-risk and low-risk groups in prognosis of LUSC patients, independently. Furthermore, the model was validated by ROC curves in the testing dataset and the whole dataset. Lastly, by gene set enrichment analysis (GSEA), we showed the main enrichment pathways were DNA damage stimulus, DNA repair, and DNA replication. It was suggested that the risk score may provide a new and reliable method for prognosis prediction.

Conclusion: The results of this study suggested that the risk score based on seven-genes could indicate a promising and independent prognostic biomarker for LUSC patients.

Keywords: lung squamous cell carcinoma, prognosis, gene set enrichment analysis, Cox regression model, risk score

Introduction

Lung cancer is the most commonly diagnosed cancer (11.6% of the total cases) and the leading cause of cancer death, which made it a big concern for human public health worldwide.1 About 80% of primary lung cancers are non-small cell lung cancer (NSCLC), which can be further classified into adenocarcinoma and squamous cell carcinoma.2 Lung squamous cell carcinoma (LUSC) comprises 30% of all lung cancers.3,4 Primary treatments of advanced lung squamous cell carcinoma are surgery, radiotherapy, and chemotherapy. Although molecularly targeted agents and immunotherapy have played a major role in the treatment of lung adenocarcinoma, there...
have been few approved targeted therapies and effective chemotherapeutic options beyond the first line of therapy for LUSC. However, its 5-year survival rate is still less than 15%. Thus, it is important to find effective and reliable predictors for early detection.

Previous studies have identified the correlation between aberrant expressed genes and human cancers, and revealed promise of these genes as biomarkers in predicting patients’ survival. Similarly, the abnormality of RNA has also been reported as having impact on the survival. Overexpressed HUWE1 was found to be associated with poor patient outcome. VPS9D1-AS1 serves as a promising biomarker to predict the prognosis of NSCLC. These studies demonstrate that changes in single gene expression and in RNA expression are meaningful for predicting the prognosis of lung cancer. However, there is no independent element which can predict the prognosis of lung cancer based on a multi-gene model.

In this research, we presented the risk score based on seven genes as a dependent prognostic biomarker for LUSC; the risk score contained all genes information which was included in the model. Additionally, we found that these genes are mainly enriched in the DNA damage stimulus pathway, DNA repair pathway, and DNA replication pathway. In summary, we identified a new element (risk score) to predict the prognosis of LUSC patients. Compared to the existing studies of predictor, our work offered a new risk model as an independent prognostic biomarker in risk stratification for LUSC patients.

Materials and methods

Data source and preprocessing
Gene expression profiles and the clinical dataset of patients with LUSC were downloaded from the TCGA data portal. The gene expression profiles were chosen based on the criteria as follows: median expression in LUSC >0 and in adjacent-normal tissue >0. Then, all the genes were normalized by DESeq2 package. Clinical information, including the total number of patients (n=551), sex, age, TNM stage, T stage, N stage, M stage, survival time, and survival state were included in this research. A total of 551 LUSC patients were randomly divided into a training data set (n=363) and a testing data set (n=188).

Identification of prognosis-associated genes
Differential expression genes (DEGs) were chosen according to the criteria of P<0.05, |log2fold change| >1 in the training dataset using the R package edge, and analysed by univariate analysis. Then, 24 genes (p<0.001) were selected to build a Cox proportional hazards regression model. Finally, the model was carried out by stepwise regression method. Therefore, seven independent genes were chosen via the Cox regression model (Table 1). Based on this independent prognostic model, the survival risk of each LUSC patient was evaluated as follows:

Risk score = (−0.1311) × expression of CSRNP1 + 0.1390 × expression of MIR27A + (−0.1951) × expression of CLEC18B + 0.1708 × expression of AC130456.4 + 0.1702 × expression of DEFA6 + 0.1821 × expression of ARL14EPL + 0.1135 × expression of ZFP42

K-M survival analysis
The risk score of each LUSC sample was calculated according to the risk score formula, and all the LUSC patients were divided into low-risk and high-risk groups using the median value of the risk score as the cut-off value. Then the K-M survival curves of high-risk and low-risk groups were mapped using survival package. Lastly, the follow-up time of patients was shown in a scatter plot.

Correlation analysis between risk score system and clinical factors
All patients were grouped by age, gender, race, smoking status, and stage. The risk score between the two groups was calculated with the purpose of observing the difference between each feature.

Gene set enrichment analysis
Gene Set Enrichment Analysis (GSEA) was carried out to analyze the differences between two groups: high-risk

<table>
<thead>
<tr>
<th>Genes</th>
<th>Coefficient</th>
<th>HR</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSRNP1</td>
<td>−0.1311</td>
<td>0.8772</td>
<td>0.1348</td>
</tr>
<tr>
<td>MiR27A</td>
<td>0.1390</td>
<td>1.1492</td>
<td>0.0794</td>
</tr>
<tr>
<td>CLEC18B</td>
<td>−0.1951</td>
<td>0.8228</td>
<td>0.0014</td>
</tr>
<tr>
<td>AC130456.4</td>
<td>0.1708</td>
<td>1.1862</td>
<td>0.0343</td>
</tr>
<tr>
<td>DEFA6</td>
<td>0.1702</td>
<td>1.1855</td>
<td>0.0033</td>
</tr>
<tr>
<td>ARL14EPL</td>
<td>0.1821</td>
<td>1.1997</td>
<td>0.0245</td>
</tr>
<tr>
<td>ZFP42</td>
<td>0.1135</td>
<td>1.1202</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

Abbreviation: HR, hazard ratio.
Statistical analysis
We investigated DEGs using edgeR package, the principle of which is fitting generalized linear models and using likelihood ratio tests to identify DEGs. Next, we used the Cox regression model to do the multivariable survival analysis. Based on the median value of risk score, we divide the population into high-risk and low-risk groups. To further confirm the model, we used the Kaplan-Meier method to analyze the correlation between risk score and overall survival, and the log-rank test to compare survival curves. We used R software version 3.5.1 and the “survival” package to do the receiver operating characteristic (ROC) curve analysis. Moreover, we conducted 3-fold cross-validation to further confirm the efficacy of the model. The area under the curve (AUC) was also calculated. We compared two groups in different clinical parameters (age, gender, race, smoking status, and stage) using the t-test for continuous variables and $\chi^2$ test for categorical variables. With regard to both log-rank test and Cox proportional hazards regression model, the significance was set at $p<0.05$.

Results
Identification of prognostic gene signature
To identify new genetic prognostic markers for LUSC, we screened 24 prognostic related differential expression genes using univariate analysis. These prognostic significant genes were included in the Cox multivariate regression model, and the seven-gene prognosis prediction model (Table 1) was finally determined by stepwise regression.

Based on the model, the risk score of each sample was obtained, and, based on the median value of the risk score, LUSC patients were divided into high-risk group and low-risk group. In the training set, we found that the survival curves of the high-low risk group were significantly different, and the low-risk group had a better survival period (Figure 1A). The training population were ranked according to the risk score from low to high, and we showed the population follow-up time and genes heat map by the ranking (Figures 1C–E). Notably, the gene expression trend of CLEC18B is opposite to the others (Figure 1E).

Validation of the seven-gene signature in predicting survival using Kaplan–Meier curves
In order to further prove the accuracy of the model and the classification effect of the risk score, we further applied the K-M analysis to display the survival curve of the high-risk group and the low-risk group in the validation dataset and the whole dataset. We found the survival of the two groups is significantly different in each dataset ($p<0.05$) (Figures 3A and 4A). This result further demonstrated the effective performance of the model for classifying patients into two groups with different prognoses. Meanwhile, we showed the follow-up time of the validation dataset and the whole dataset by ascending ranking with risk score (Figures 3B and 4B). Consistently, the seven-gene expression heatmap sorted by risk score was also shown in the figures (Figures 3C and D and 4C and D).

The seven-gene signature is an independent prognostic factor of survival
To compare the risk score with conventional clinical factors, we performed univariate and multivariate Cox hazard regression analysis to evaluate the importance of these indicators in the patients’ prognosis, which included risk score, age, sex, and TNM stage. We found that the risk score was an independent prognostic factor (Figure 5) (hazard ratio [HR]=2.673; 95% confidence interval [CI]=1.871–3.525; $p<0.05$). This result suggested the risk score was a robust indicator in predicting the prognosis of the LUSC cohort.

Clinical staging is reported relating to prognosis in NSCLC. However, when it was analyzed together with the risk score in our research, the result in the model did not show significance. This was because each factor in the model was assigned a weight, and this made the function of the staging non-significant. These results suggest that the risk score is a more robust element in predicting the prognosis in the LUSC cohort. Furthermore, it is noteworthy that the seven-gene signature is competitive for survival prediction compared with clinical parameters.
By Cox regression analysis, we showed that the risk score is an independent prognostic factor. However, whether there is a difference in risk score for gender, age, race, stage, and smoking-pack years is unknown. In order to explore this issue, we tested the risk scores in different groups dividing by the parameters (gender, age, race, stage, and smoking-pack years), but there were no difference found (Figure 6).

Figure 1 Outcomes of the seven-gene model in the training cohort. (A) Kaplan-Meier survival curve demonstrates the differences between the high-risk group and low-risk group about survival rate in the training sample. (B) ROC curve shows the performance of the model. (C) The distribution of follow-up time in the training sample. (D) The distributions of the risk scores in the training sample. (E) The distributions of the seven-gene expression profiles of patients in the training sample.
Identification of the eight-gene signature-related biological processes and pathways

Based on the risk scores of the seven-gene expression, the population was divided into a high-risk group and a low-risk group, and there were significant differences in the prognosis between the two groups. To explore the mechanism of the seven-gene model, we examined the whole gene expression profiles by GSEA analysis. The GSEA analysis revealed that the mainly enrichment pathways were DNA damage stimulus pathway, DNA repair pathway, and DNA replication pathway, and the results suggested that the poor prognosis is related to the DNA damage repair pathway (Figure 7).

Discussion

The prediction of prognosis is of great significance for treatment choice. A large number of studies have explored
prognosis-related biomarkers, and gene expression profiles have been found to play an important role in prognosis.\textsuperscript{14-16}

In this research, we obtained a seven-gene prognostic model by Cox regression analysis and calculated the risk score of the LUSC patients. Our results suggested that the risk score is an independent prognostic factor. Furthermore, we explored the mechanism of the seven-gene predicting model by GSEA analysis. These findings suggested that the patient’s prognosis and molecular genetic background are inter-related, and prognostic assessment, including genetic background, will be more accurate. To the best of our knowledge, this is the first time that the seven-gene model was identified, and was found to be a candidate prognostic biomarker in LUSC.

Many studies have found that changes in gene expression affect prognosis, and there are reports that combine this analysis to predict prognosis of lung adenocarcinoma, kidney cancer, and pancreatic cancer.\textsuperscript{17-19} Multi-gene prediction of LUSC has not been reported. In our study; we obtained the prognosis-related seven-gene model by the Cox hazard regression analysis. Some of the seven genes (CSRNPI, CLEC18B, MIR27A, AC130456.4, DEFA6, ARL14EPL, and ZFP42) were reported in relation to tumorigenesis. CSRNP1 has been reported as a hub gene correlating to the prognosis of liver cancer.\textsuperscript{20} Similarly, it has been reported to be involved in the promotion of carcinogenesis through the eQTL mechanism in renal cancer.\textsuperscript{21} TGF-β1 could induce CSRNP1 expression, while SMAD3 was activated.\textsuperscript{22} A meta-analysis reported that the rs13275170 locus of DEFA6 increased 1.3-fold in gastric cancer, suggesting that DEFA6 is associated with tumorgenesis.\textsuperscript{22} ZFP42 is an oncogene, and a study reported a combined methylation model can accurately predict the occurrence of lung cancer, including SOX17, TAC1, HOXA7, CDO1, HOXA9, and ZFP42, which highly suggests the role of ZFP42 in the development of lung cancer.\textsuperscript{23} In this seven-gene model, MIR27A, AC130456.4, and ARL14EPL have not been reported in the relationship with tumors. It was interesting to find the increasing expression of CLEC18B has a negative correlation with the risk score in the heatmap. CLEC18B is a member of the CLEC superfamily and expressed in a variety of cancer cells. It was reported as strongly negatively correlated with total survival in GBM,

\textbf{Figure 3} Outcomes of seven-gene model in the all samples. (A) Kaplan-Meier survival curve illustrates the differences between the high-risk group and low-risk group about survival rate in all samples. (B) The distribution of follow-up time in the all samples. (C) The distributions of the risk scores in the all samples. (D) The distributions of the seven-gene expression profiles of patients in the all samples.
and it was found to promote proliferation, invasion, and migration in GBM tumor cells. The high expression of CLEC18B can activate the activity of Wnt/β-catenin signaling pathway. Our study has proven that the genes in the model are related to the DNA damage pathway. In addition, Wnt signaling and DNA damage do have crosstalk. Referring to ZFP42, there has been little known about the ZFP42 gene in lung cancer. In summary, the seven genes have not been reported to predict the

**Figure 4** The effects of clinical factors and risk score in all samples. The clinical importance of clinical factor and risk score.

**Abbreviations:** CI, confidence interval; HR, hazard ratio.

**Figure 5** The association between risk score and clinical factors (age (A), gender (B) race (C), smoking-pack years (D), stage (E)) in all samples.
prognosis of lung cancer individually. We obtained the risk score of the sample by the combined prediction model of the seven genes, and verified the risk score was an independent prognostic factor.

In order to understand the mechanism of prognosis prediction of this gene set. We performed GSEA analysis; GSEA is a computational method to evaluate whether a defined set of genes shows statistically significant, concordant differences between two biological states. We found the associated biological signaling pathways were the DNA damage stimulus pathway, DNA repair pathway, and DNA replication pathway. Consistent with previous reports, DNA damage repair genes affect prognosis of tumor patients. These results also suggested that poor prognosis may be associated with abnormal DNA damage pathways. However, the risk score as a comprehensive evaluation indicator reducing the genetic dimension can independently predict the LUSC patient’s prognosis.

The model built by this study was validated by an internal dataset. However, it also required more clinical data validation to reinforce the conclusion. Taken together, this study suggests that the integrated genetic information included with the clinical evaluation parameters to analyze prognosis will lead to more accurate predictions and provide more accurate treatment options.

**Conclusion**

Our study proposed a novel seven-gene model to predict the prognosis of patients with LUSC. We also further validated the feasibility of this model. In addition, genes in this model were mainly associated with the DNA damage stimulus, DNA repair, and DNA replication. Therefore, our study provided a reliable way to analyze the prognosis of patients with lung squamous cancer.

**Abbreviation list**

SCLC, small-cell lung cancer; NSCLC, non-small cell lung cancer; LUSC, lung squamous carcinoma; ROC, receiver operating characteristic; GSEA, Gene Set Enrichment Analysis; TCGA, The Cancer Genome Atlas; CI, confidence interval; HR, hazard ratio; CSRNP-1, cysteine-serine-rich nuclear protein-1.
Availability of data and material
The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgment
This work was funded by Science and Technology Project of Shenyang (F16-102-4-00).

Author contributions
All authors contributed to data analysis, drafting and revising the article, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Disclosure
The authors report no conflicts of interest in this work.

References