scientific reports

Check for updates

OPEN

A prognostic model of non small cell lung cancer based on TCGA and ImmPort databases

Dongliang Yang[1,3], Xiaobin Ma[2,3] & Peng Song[2]✉

Bioinformatics methods are used to construct an immune gene prognosis assessment model for patients with non-small cell lung cancer (NSCLC), and to screen biomarkers that affect the occurrence and prognosis of NSCLC. The transcriptomic data and clinicopathological data of NSCLC and cancer-adjacent normal tissues were downloaded from the Cancer Genome Atlas (TCGA) database and the immune-related genes were obtained from the IMMPORT database (http://www.immport.org/); then, the differentially expressed immune genes were screened out. Based on these genes, an immune gene prognosis model was constructed. The Cox proportional hazards regression model was used for univariate and multivariate analyses. Further, the correlations among the risk score, clinicopathological characteristics, tumor microenvironment, and the prognosis of NSCLC were analyzed. A total of 193 differentially expressed immune genes related to NSCLC were screened based on the "wilcox.test" in R language, and Cox single factor analysis showed that 19 differentially expressed immune genes were associated with the prognosis of NSCLC ($P < 0.05$). After including 19 differentially expressed immune genes with $P < 0.05$ into the Cox multivariate analysis, an immune gene prognosis model of NSCLC was constructed (it included 13 differentially expressed immune genes). Based on the risk score, the samples were divided into the high-risk and low-risk groups. The Kaplan–Meier survival curve results showed that the 5-year overall survival rate in the high-risk group was 32.4%, and the 5-year overall survival rate in the low-risk group was 53.7%. The receiver operating characteristic model curve confirmed that the prediction model had a certain accuracy (AUC = 0.673). After incorporating multiple variables into the Cox regression analysis, the results showed that the immune gene prognostic risk score was an independent predictor of the prognosis of NSCLC patients. There was a certain correlation between the risk score and degree of neutrophil infiltration in the tumor microenvironment. The NSCLC immune gene prognosis assessment model was constructed based on bioinformatics methods, and it can be used to calculate the prognostic risk score of NSCLC patients. Further, this model is expected to provide help for clinical judgment of the prognosis of NSCLC patients.

Globally, due to high-risk factors, such as smoking, radon, occupational exposure, traffic exhaust, and air pollution, lung cancer has become the leading cause of cancer-related deaths, and it is also a major global health problem that is currently attracting widespread attention[1]. Non-small cell lung cancer (NSCLC) accounts for 85% of lung cancer diagnoses. Approximately 50% of NSCLC patients are in stage IV when they are detected, and their 5-year survival rate is less than 10%[2]. In recent years, immune checkpoint inhibitors (ICIs) targeting programmed cell death 1 (PD-1) or its ligands (PD-L1) have been developed, which has caused significant progress in the treatment and overall management of locally advanced and advanced NSCLC[3]. The role of abnormal expression of tumor immune-related genomes in tumor immune evasion has become a new direction in tumor research. Abnormal immune genomes have an important impact on patients with ovarian cancer, gastric cancer, liver cancer, and kidney cancer[4]. However, there is no relevant report on how an abnormal genome affects NSCLC. In addition, a variety of molecular markers are used to predict the prognosis of NSCLC, but they have not yet been widely recognized. Therefore, it is necessary to explore the genes related to the prognosis of NSCLC at the molecular level, and the construction of genetic models related to the prognosis of NSCLC has a strong clinical significance. Therefore, this study is based on the TCGA and ImmPort data sets to explore immune gene expression and immune cell differential analysis, and combine its clinicopathological characteristics and

[1]Department of General Education, Cangzhou Medical College, Cangzhou 061001, China. [2]Department of Respiratory Medicine, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan 252200, China. [3]These authors contributed equally: Dongliang Yang and Xiaobin Ma. ✉email: songpeitong@163.com

immune gene characteristics to construct a prognostic model of NSCLC, it is expected to have certain guiding significance for the treatment of NSCLC.

Materials and methods

Data download. The transcriptomic data and clinical data of NSCLC patients in TCGA-LUAD and TCGA-LUSC were downloaded through the Genome Data General Database (GDC) data portal, and 929 clinical data were obtained. Immune gene data were downloaded through the ImmPort data portal, and 2498 immune-related genes were obtained. The transcription factor data were downloaded from the Cistorm website.

Differential expression analysis. *Differential gene expression analysis.* The Wilcoxon test in R software was used to analyze the differences in all transcriptomic data, and to screen out genes with significant differences in expression between normal tissues and tumor tissues. The screening criteria were $|\log FC| > 1$, $FDR < 0.05$.

Analysis of immune gene differences. Differential genes were combined with the acquired immune gene data and analyzed in R software to screen out differential immune genes from all differential genes.

Establishment and evaluation of the immune gene prognosis model. The expression of immune genes was combined with survival time and survival status, survival analysis of differential immune genes and clinical survival time was conducted, and immune genes that could affect the prognosis of NSCLC were determined. Based on these genes, an immune gene prognostic model was constructed. Receiver operating characteristic (ROC) and risk scoring curves were drawn in R software to evaluate the effectiveness of this model.

Correlation analysis between risk score and immune cells. Immune genes related to the prognosis were combined with clinical data, and the patient's risk score was calculated based on the immune gene prognosis model. Correlation analysis was performed between risk score and immune cells infiltrated by the tumor microenvironment (immune cell data were downloaded from the TIMER immune cell infiltration database).

Statistical methods. R 3.6.0 software (https://mirrors.tuna.tsinghua.edu.cn/CRAN/) was used for statistical analysis and graph drawing. The wilcox.test was used to screen differential genes. The "ggplot" package was used for graph drawing, and the "survival" package was used for single-factor and multi-factor Cox proportional regression model screening and to establish the multiple gene prognosis model. The "survival ROC" package was used to calculate the ROC curve to evaluate the effectiveness of the model and the area under the curve (AUC). The statistical inference level was set at two-sided $\alpha = 0.05$.

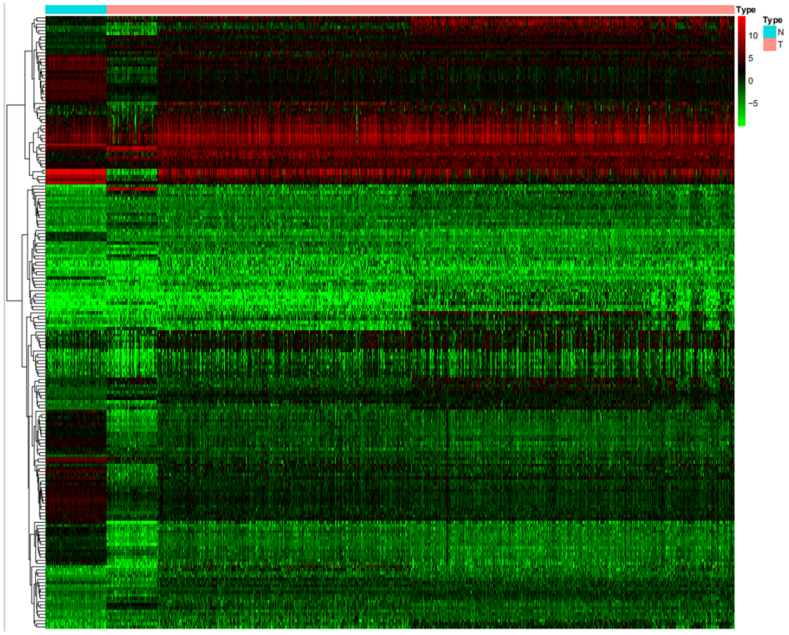
Results

Screening of differentially expressed immune genes. Data of the TCGA database containing 1128 non-small cell lung cancer samples and 110 normal tissues were downloaded. Differential expression analysis screened a total of 2875 differential genes ($FDR < 0.01$, $|\log 2FC| > 1$), of which 2317 differential genes were highly expressed and expression of 557 differential genes was low. A total of 2498 tumor-related immune genes were downloaded from the ImmPort database. In the R language, immune genes and all differentially expressed genes were intersected and a total of 193 differential immune genes related to NSCLC were obtained, of which 121 differential immune genes were highly expressed and 72 differentially expressed genes were lowly expressed. The R-ggplot package (version: 3.3.5) was used to draw a heat map (Fig. 1A), and the R-pheatmap package (version: 1.0.12) was used to draw a volcano map (Fig. 1B).

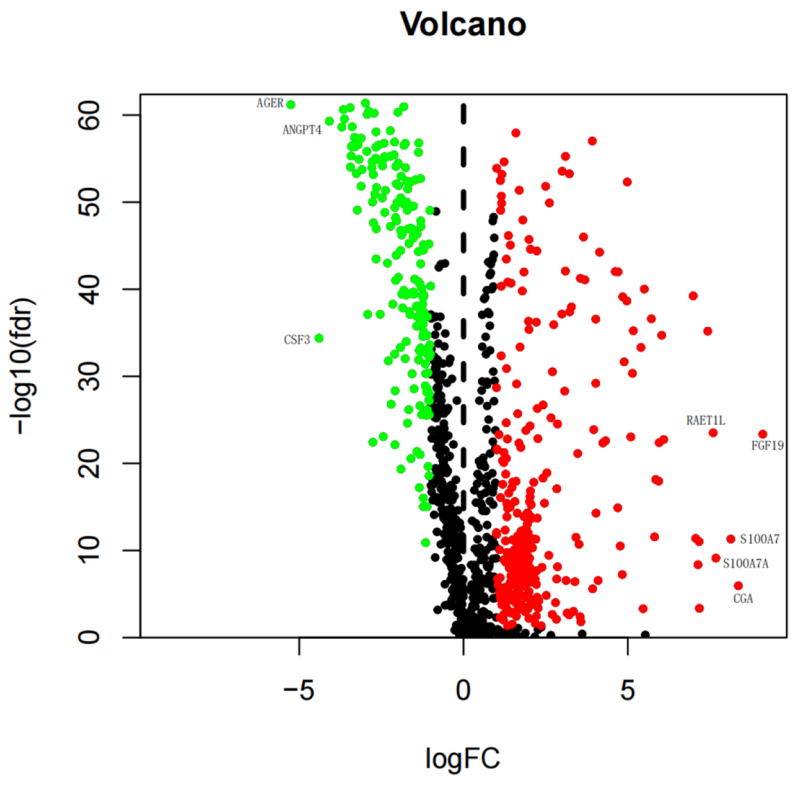
Transcriptional regulatory network mapping. A total of 318 transcription factors (TFs) downloaded from the Cistrome Cancer database and 2875 differentially expressed genes were crossed to obtain 83 differentially expressed TFs, of which 50 TFs were up-regulated and 33 TFs were down-regulated (Fig. 2A, B). The correlation between immune-related genes and differentially expressed TFs was further analyzed by Pearson correlation test, and the intersection group with correlation coefficient > 0.4 and $P < 0.01$ was screened out. Cytoscape was used to draw the immune factors and transcriptional gene regulatory network (Fig. 3). VIPR1 (low-risk immune genes) is negatively regulated by NCAPG, MYBL2, CENPA and positively regulated by ERG, EPAS1, TCF21. SHC3 is negatively regulated by CENPA and positively regulated by TCF21. All the other genes were positively regulated.

Establishment of an immune gene prognostic model for NSCLC. Univariate regression analysis was performed on 193 differential immune genes related to NSCLC. The results showed that 19 differential immune genes were significantly related to the overall survival rate of NSCLC ($P < 0.05$) (Fig. 4).

These 19 differential immune genes with $P < 0.05$ were selected for the Cox multivariate analysis, and the NSCLC prognosis model containing 13 differential immune genes was obtained: Risk Score = MMP12 × 0.0022 + PLAU × 0.0023 + S100P × 0.003 + CRABP1 × 0.0036 + RBP2 × 0.0531 + LTB4R × (-0.0255) + RNASE7 × 0.0174 + IGLV4-3 × 0.0017 + IL33 × (-0.014) + INHA × 0.0053 + FGFR4 × 0.0443 + SHC3 × (-0.1775) + HNF4G × 0.0516. From the risk score, LTB4R, IL33 and SHC3 are the immune genes that are beneficial for the prognosis of NSCLC. Using the median of Risk Score (RS) (0.9506237) as the boundary value, the RS was divided into high risk and low risk groups, and patients were sorted according to risk scores from low to high. The risk score curves (Fig. 5) and survival heat maps (Fig. 6A) were drawn. The patient was used as the abscissa to plot the RS and survival time. It was noted that as the RS score increased, the immune gene expression content increased (Fig. 6B); and as the RS



A



B

Figure 1. A heat map (**A**) and a volcano map (**B**) of differential immune genes. The heat map abscissa represents the sample: the blue area represents normal tissue and the red area represents tumor tissue; the ordinate represents the gene. On the volcano map, the green area represents the downregulated differential genes and the red area represents the upregulated differential genes.

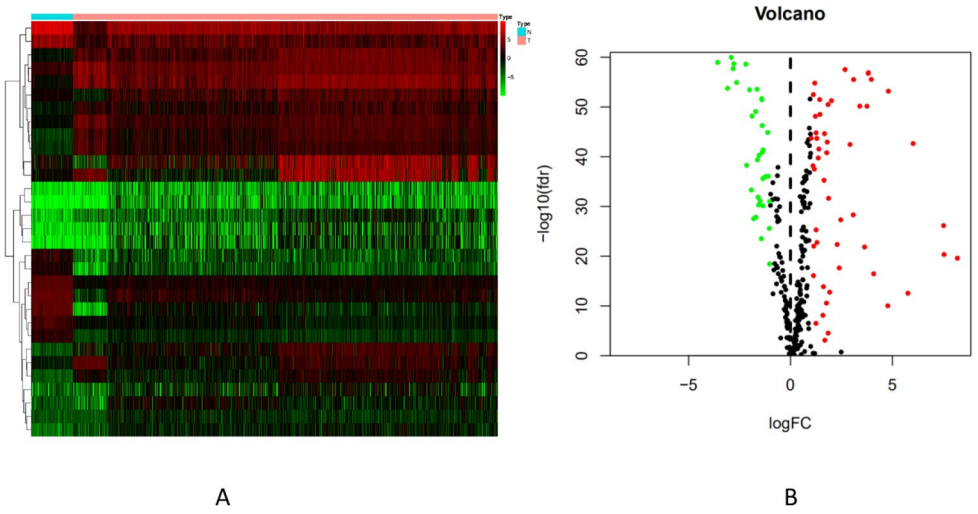


Figure 2. Heat map and volcano map of differentially expressed TFs of non-small cell lung cancer. **A** heat map of differentially expressed TFs of non-small cell lung cancer, red represents high expression, blue represents low expression; **B** volcano map of TFs of non-small cell lung cancer. The X-axis is log FC, and the larger the absolute value is, the larger the corrected *P* value is, indicating the larger the multiple of the difference is. The Y-axis is the corrected *P* value, and the larger the logarithm of log10 is, indicating the more significant the difference is.

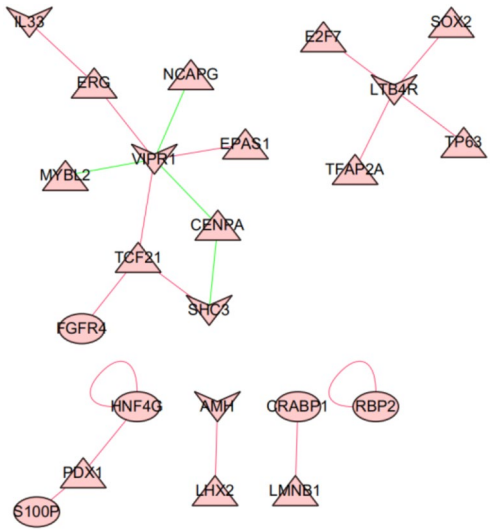


Figure 3. Transcriptional factors and immune gene regulatory network (Triangles represent transcription factors, circles represent high-risk immune genes, and cones represent low-risk immune genes; The red line represents positive regulation, and the blue line represents negative regulation).

increased, the patient's survival time shortened and the number of deaths increased significantly (Fig. 6C). Use the R package "princomp" to perform principal component analysis (PCA) on 13 immune genes, it can be seen that the high-risk genome and the low-risk genome are clearly divided into two discontinuous groups (Fig. 7).

COX survival analysis and prognostic model evaluation. After using the R-survival package to perform COX survival analysis in the high-risk and low-risk groups, the results showed that the 5-year survival rate in the high-risk group was 32.4%, and the 5-year survival rate in the low-risk group was 53.7%. The difference was statistically significant ($P < 0.01$). In order to further verify the accuracy of the prognostic evaluation model, we used the R-survival ROC package to draw the model ROC curve (Fig. 8), and the results showed an AUC = 67.3%. This finding suggested that the risk assessment model had better sensitivity and specificity in predicting the prognosis of NSCLC.

In order to verify whether the machine learning modeling algorithm is better than the traditional COX regression analysis, We use the decision tree algorithm in machine learing to build a new model. Using the rpart package of R version 4.0.2 to model the original data decision tree (Fig. 9), it was found that the area under the ROC curve of the decision tree model was 0.601 (Fig. 10), which was lower than the model established by

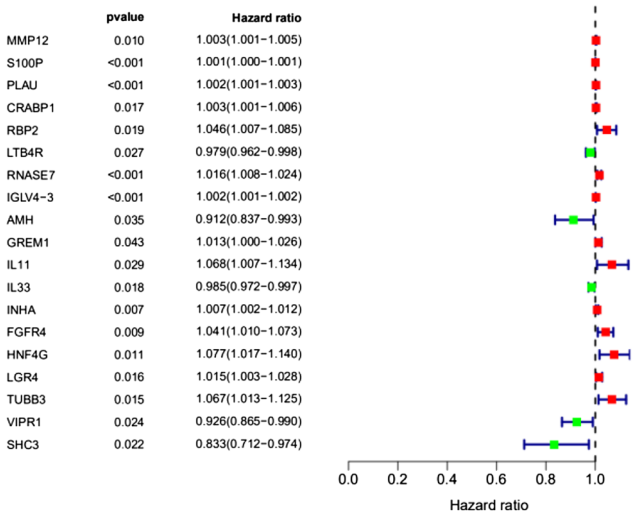


Figure 4. Forest map of 19 differentially expressed immune genes in the univariate Cox regression model.

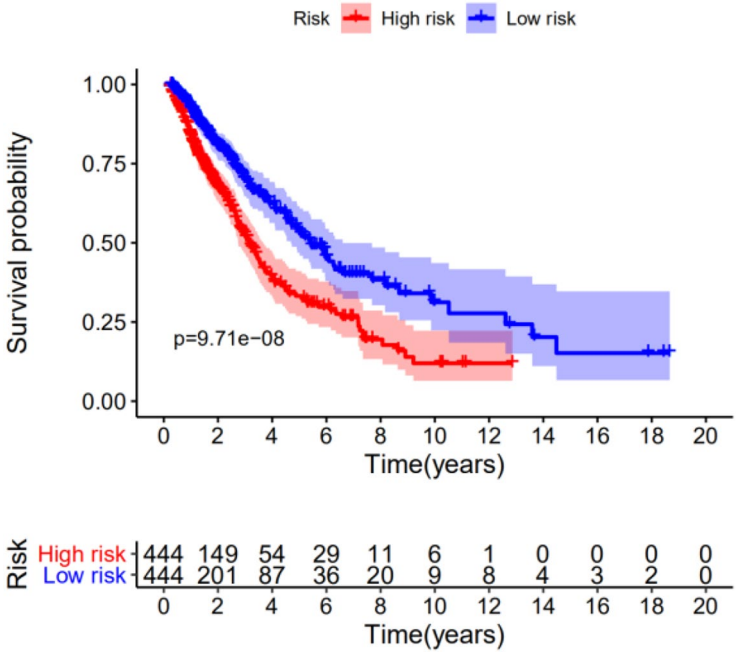


Figure 5. Kaplan–Meier survival analysis of non-small-cell lung cancer patients by risk stratification.

the original cox regression area under the curve. Therefore, the model established by cox proportional hazard regression is the final predictive model.

An independent prognostic factor in NSCLC. In order to further verify whether the prediction model can independently predict the prognosis of NSCLC, various parameters in the clinical data downloaded from the TCGA database were used as independent variables, and the patient's survival time was used as the dependent variable to perform Cox factor regression analysis. The results suggested that the risk score was an independent risk factor affecting the prognosis of NSCLC ($P<0.05$) (Fig. 11).

Validation of the prognostic value of risk score using the GEO datasets. Independent validations were conducted using the GEO datasets to further test the prognostic value of risk score. The same method was used to generate a risk score for risk stratification of NSCLC patients in GSE68465 and GSE101929. In order to further verify the accuracy of the prognostic evaluation model in GEO database, we used the R-survival ROC package to draw the model ROC curve (Figs. 12 and 13), and the results showed an AUC = 71.2% (GSE68465)

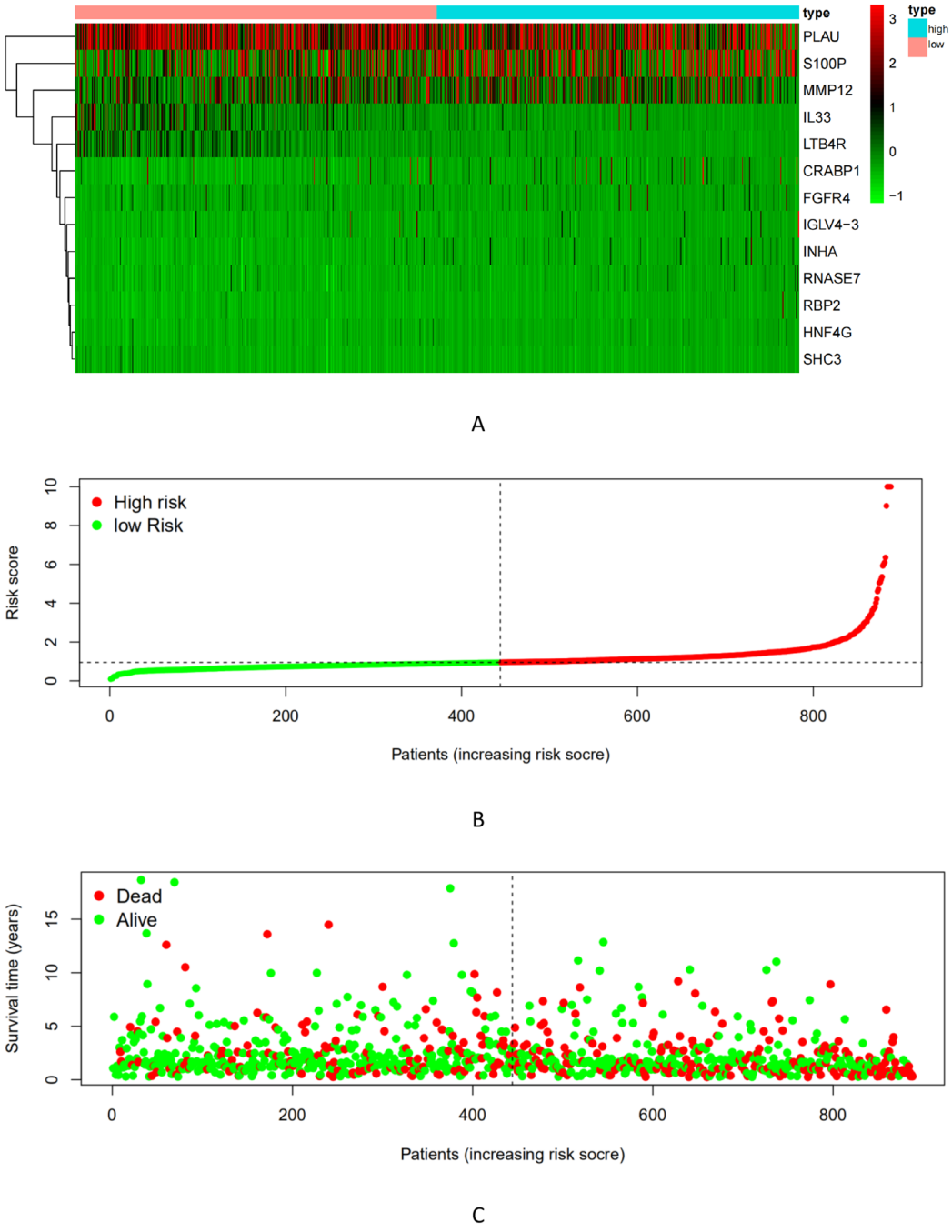


Figure 6. Risk score curve and survival heat map. (**A**) survival heat map, with the increase of risk score, the expression of immune genes increased; (**B**) risk score curve, from left to right, the patient's risk score increased gradually; (**C**) point of survival chart (With the increase of patients'risk value, more patients died).

/ 65.4% (GSE101929). This result can confirm that the 13-gene NSCLC prediction model still performs well in the GEO database.

The relationship between risk score and immune cell infiltration. We also analyzed the relationship between risk score and immune cell infiltration in the tumor microenvironment. The results showed that the degree of neutrophil infiltration had a certain correlation with the risk score ($P = 0.089$), but there was no statistically significant difference. There was no correlation with B cells, CD4 + T cells, CD8 + T cells, dendritic cells, and macrophages (Fig. 14).

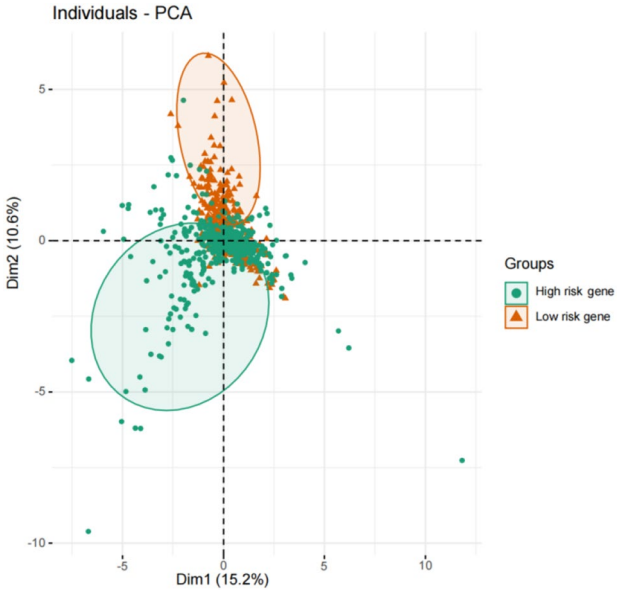


Figure 7. Principal component analysis plot using expression values at 13 selected immune genes.

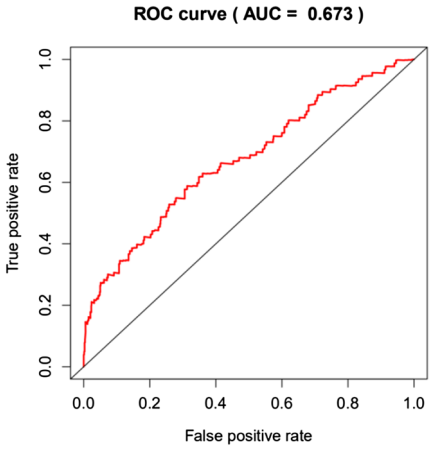


Figure 8. ROC curve of multivariate Cox analysis model.

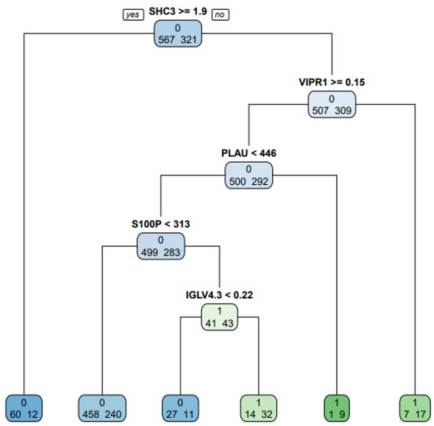


Figure 9. Immune gene prediction model (decision tree algorithm).

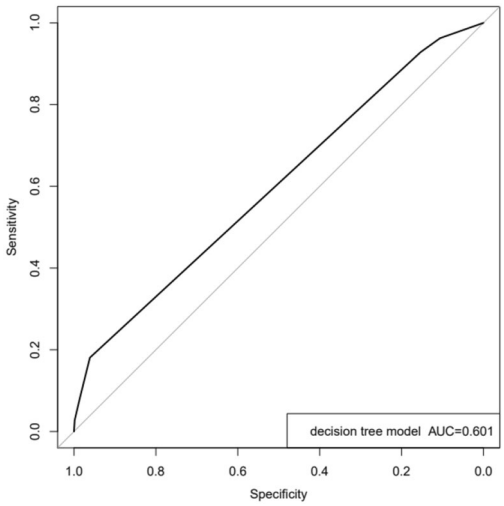


Figure 10. ROC curve of decision tree algorithm model.

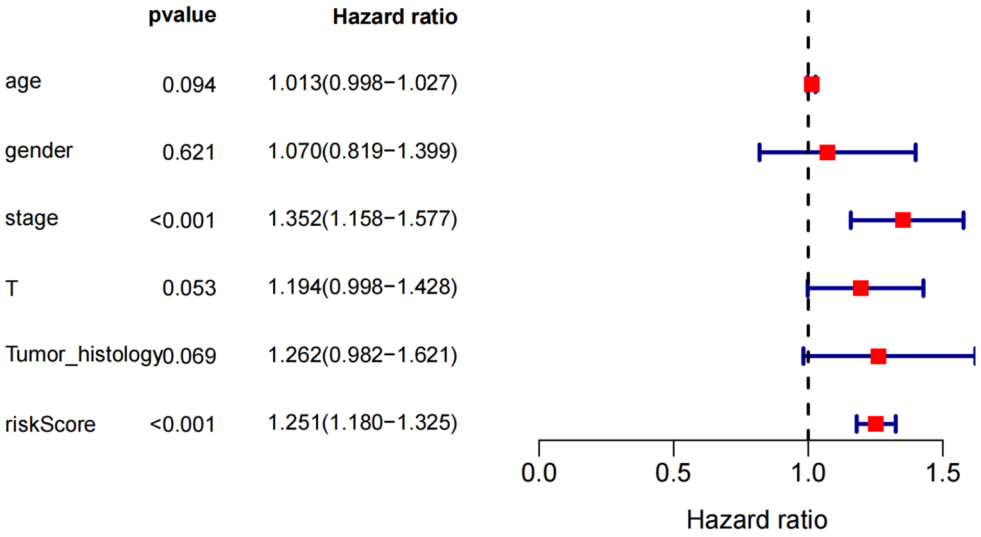


Figure 11. Cox multivariate regression analysis.

Discussion

In the past decade, it has been recognized that the occurrence and development of tumors should not only be attributed to the internal genetic background of cancer cells, but it is also related to the interaction of various systems in the body[5], especially the immune system[6]. Immune-related cells and factors are involved in the entire process of tumorigenesis, proliferation, and development[7,8]. Therefore, it is necessary to explore the characteristics of immune-related molecules and evaluate the function of immune genes in lung cancer[9]. In this study, we sorted out 2498 immune-related genes in TCGA mRNA in 929 NSCLC patients and further performed COX univariate analysis of 193 immune-related genes, and we found that 19 differential genes were significantly related to the prognosis of NSCLC patients. COX multivariate analysis yielded a NSCLC multivariate prognostic risk model containing 13 differential genes, and the effectiveness of the model was verified by Kaplan–Meier and ROC curves. Among these genes, MMP12, PLAU, S100P, CRABP1, RBP2, RNASE7, IGLV4-3, INHA, FGFR4, and HNF4G may be immune-related genes that promote tumorigenesis; and LTB4R, IL33, and SHC3 may be immune-related genes that inhibit tumorigenesis. The immune gene risk score and clinicopathological characteristics were included in the univariate and multivariate Cox regression analyses of the prognosis of NSCLC. The results suggest that the immune gene risk score is an independent predictor of the prognosis of NSCLC. In view of their important role in the prognostic evaluation of NSCLC, these genes play an important role in the occurrence and development of NSCLC, and they may become new targets for precision treatment of NSCLC, which are worthy of an in-depth study[10].

More and more evidence shows that abnormally expressed genes can be used as prognostic markers for NSCLC[11]. SD DEr have identified and verified 15 gene characteristics (ATP1B1, TRIM14, FAM64A, FOSL2,

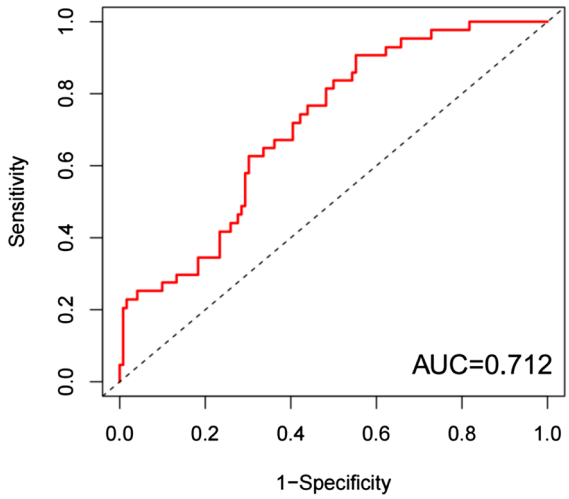


Figure 12. ROC curve of multivariate Cox analysis model in GSE68465 database.

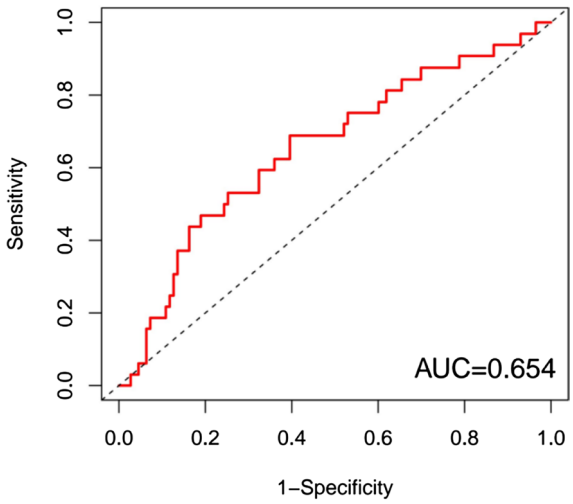


Figure 13. ROC curve of multivariate Cox analysis model in GSE101929 database.

HEXIM1, MB, L1CAM, UMPS, EDN3, STMN2, MYT1L, IKBKAP, MLANA, MDM2, ZNF236) that affect the prognosis of NSCLC[12]. Shukla et al. divided the TCGA RNA sequencing data into training and validation cohorts, based on 4 gene characteristics (RHOV, CD109, FRRS1 and long non-coding RNA (lncRNA) genes LINC00941) divide LUAD patients into high-risk and low-risk survival groups[13]. In another study, 20 gene characteristics based on TCGA data can predict the OS of NSCLC, combined with a comprehensive analysis of differentially expressed genes in the GEO data set (GSE85841), including four of FUT4, SLC25A42, IGFBP1 and KLHDC8B Genes can predict OS (AUC of prognostic score 20 genes = 0.615, AUC of prognostic score 4 genes = 0.5731)[14]. Recently, Xie et al. used DE genes in the TCGA and GEO datasets to construct a weighted gene co-expression network, and found that 6 gene features (RRAGB, RSPH9, RPS6KL1, RXFP1, RRM2, and RTL1) can be used for prognostic stratification of lung adenocarcinoma (the area under ROC curve (AUC) was 0.776 in predicting the 10-year survival of NSCLC patients)[15]. Our data shows that the sensitivity of the prediction model is 0.710, the specificity is 0.687, the AUC of prognostic score = 0.673. The prediction accuracy is higher than the data of Zhao K et al., and is similar to the prediction effectiveness of the model of Xie et al.

LTB4R is the first discovered protective gene for NSCLC, and IGLV4-3 is the first discovered harmful gene for NSCLC. The other 11 differential immune-related genes in this prognostic assessment model have been rarely reported in NSCLC. MMP12 is one of the zinc-dependent proteolytic enzymes, which plays a vital role in all aspects of tumor progression (such as tumor angiogenesis and metastasis)[16,17]. Klupp et al. have reported that serum MMP12 levels are a negative prognostic marker in colon cancer patients[18]. In addition, MMP12 polymorphisms are associated with a higher risk of lung cancer[19–21]. Lv FZ et al. found that high expression of MMP12 is related to pathological staging and tumor metastasis of lung adenocarcinoma, indicating that MMP12 may be a promising target for the treatment of lung adenocarcinoma[22]. According to the reports, RNASE7 and PLAU are

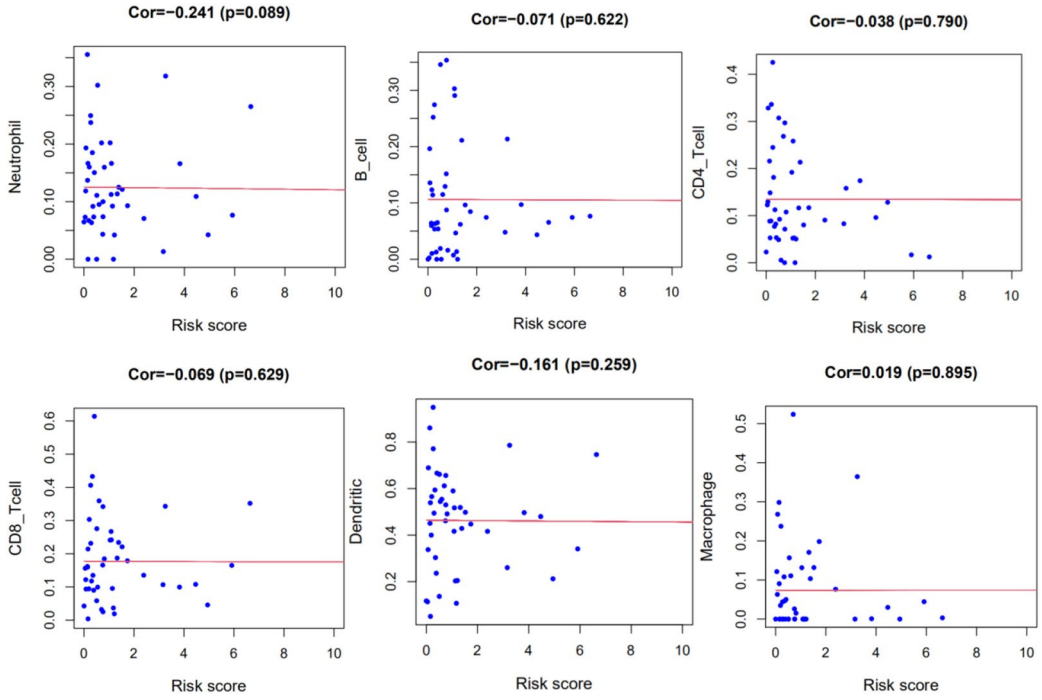


Figure 14. Correlation analysis between risk score and immune cell infiltration in tumor microenvironment.

unfavorable prognostic factors for NSCLC[23,24]. S100P is a pleiotropic tumor-promoting factor. According to the reports, in addition to promoting tumor migration, invasion, and metastasis, S100P also enhances cell proliferation by up-regulating cyclin D1 and CDK2[25] and confers chemoresistance by binding and inactivating p53[26,27]. Some studies have shown that CRABP1 expression is abnormal in NSCLC and is significantly associated with distant lymph node metastasis. A total of 42% of NSCLC samples have shown elevated CRABP1mRNA levels, which may be related to the transfer of NSCLC[28]. RBP2 is overexpressed in human lung cancer tissues and is necessary for lung cancer cell proliferation, movement, migration, invasion, and metastasis. These capabilities have been further proved to be regulated by the deethylase and DNA binding activity of RBP2. RBP2 directly binds to the integrin b1 (ITGB1) promoter and is involved in tumor migration and invasion[29]. Another study showed that RBP2 regulates the expression of n-cadherin and snails by activating Akt signaling[30]. In addition, ITGB1 and Akt signaling are significantly related to tumor angiogenesis[31,32]. These results also indicate that RBP2 promotes tumor angiogenesis[33,34]. Genetic polymorphisms and abnormal levels of IL33 are closely related to lung cancer[35,36]. Mei LJ also demonstrated the protective effect of IL-33 alleles on lung cancer[37]. Wang JJ conducted a comprehensive review, meta-analysis, and evaluation of the strength of evidence on published studies on lung cancer candidate genes. Among these studies, 2910 gene variants in 754 different genes or chromosomal loci were eligible for inclusion. A major meta-analysis of 246 variants of 138 different genes found that FGFR4rs351855 is significantly associated with the cumulative epidemiological sensitivity of lung cancer[38]. Li R et al. identified SHC3 and IL33 immune genes as independent prognostic factors for predicting the survival of NSCLC patients[39]. Hepatocyte nuclear factor 4 (HNF4) belongs to the orphan nuclear receptor superfamily[40]. Compared with the adjacent normal lung tissue, the expression of HNF4G is significantly up-regulated in lung cancer tissues. The expression level of HNF4G is related to the tumor size and overall survival rate. Genome set enrichment analysis and biological function determination have proved that HNF4G can exert a carcinogenic effect by promoting cell proliferation and inhibiting cell apoptosis[41]. The immune-related genes in this prognostic gene model are closely related to the occurrence and development of NSCLC. Related immune genes can be used as specific molecular markers for early diagnosis of NSCLC, and they can also be used as indicators for prognostic evaluation.

This study suggests that the degree of neutrophil infiltration had a certain correlation with the risk score, but there was no statistically significant difference. There was no correlation with B cells, CD4 + T cells, CD8 + T cells, dendritic cells, and macrophages. The cytoplasm of neutrophils contains a large number of neutral fine particles that are neither basophilic nor acidophilic. Most of these particles are lysosomes, containing peroxidase, lysozyme, alkaline phosphatase and acid hydrolase, etc. which are related to the phagocytic and digestive functions of cells. Neutrophil count is a representative indicator of systemic inflammation, and its increase is associated with the poor prognosis of many cancer[42]. In the tumor microenvironment, neutrophils can be manipulated, including in the early stages of the differentiation process, to develop different phenotypes and functional polarization states, thereby inducing anti-tumor or pro-tumor effects[43]. In the pro-inflammatory state, it will rapidly increase the production of neutrophils and release immature or poorly differentiated neutrophils. The recruitment of these immature neutrophils into the tumor matrix can inhibit cell apoptosis, promote metastasis and angiogenesis leading to tumorigenicity[44]. Fred Hutchinson's researchers found that neutrophils

in tumors can continuously produce substances that inhibit the activity of T cells, which affects the efficacy of immune checkpoint inhibitors against tumors[45]. Recently, the neutrophil to lymphocyte ratio (NLR) has been the most extensively studied in solid tumors. Koung Jin et al. retrospectively analyzed the relevant data of 54 patients with non-small cell lung cancer treated with PD-1 inhibitors. Multivariate analysis showed that higher NLR after treatment was an independent prognostic factor for shorter PFS and OS[46]. Relevant studies have shown that neutrophils can be used as a predictor of immunotherapy response and can help make clinical decisions in specific situations. This study evaluated the correlation between the risk score of the prediction model and the penetration of six types of immune cells in the tumor microenvironment, which may provide an important reference for monitoring the status of the tumor microenvironment to guide immunotherapy.

However, this study has certain limitations. First, verifying the capabilities of the predictive model still requires a large amount of evidence-based medical evidence from multiple centers. Second, the prognostic evaluation model is based on the results of RNA sequencing analysis in the TCGA database. Third, there is a lack of clinical, cellular, and animal functional tests; hence, the reliability of data analysis results needs further verification.

Received: 28 June 2021; Accepted: 15 December 2021
Published online: 10 January 2022

References

1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359–E386 (2015).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J Clin* **68**, 7–30 (2018).
3. Reck, M. & Rabe, K. F. Precision diagnosis and treatment for advanced non-small-cell lung cancer. *N Engl J Med* **377**, 849–861 (2017).
4. Ohyama, K. *et al.* Immune complexome analysis reveals the specific and frequent presence of immune complex antigens in lung cancer patients: a pilot study. *Int J Cancer* **140**, 370–380 (2017).
5. Ladbury, C. J. *et al.* Impact of radiation dose to the host immune system on tumor control and survival for stage III non-small cell lung cancer treated with definitive radiation therapy. *Int J Radiat Oncol Biol Phys* **105**, 346–355 (2019).
6. Mendes, F. *et al.* Lung cancer: the immune system and radiation. *Br J Biomed Sci* **72**, 78–84 (2015).
7. Tian, H. *et al.* Effects of siRNAs targeting CD97 immune epitopes on biological behavior in breast cancer cell line MDA-MB231. *Zhejiang Da Xue Xue Bao Yi Xue Ban* **46**, 341–348 (2017).
8. Giatromanolaki, A. *et al.* Programmed death-1 receptor (PD-1) and PD-ligand-1 (PD-L1) expression in non-small cell lung cancer and the immune-suppressive effect of anaerobic glycolysis. *Med Oncol* **36**, 76 (2019).
9. Zhang, Y. *et al.* Systemic immune-inflammation index is a promising noninvasive marker to predict survival of lung cancer: a meta-analysis. *Medicine (Baltimore)* **98**, e13788 (2019).
10. Sun, L. *et al.* Analysis of expression differences of immune genes in non-small cell lung cancer based on TCGA and ImmPort data sets and the application of a prognostic model. *Ann Transl Med* **8**(8), 550 (2020).
11. Giatromanolaki, A. *et al.* Increased expression of transcription factor EB (TFEB) is associated with autophagy, migratory phenotype and poor prognosis in non-small cell lung cancer. *Lung Cancer* **90**(1), 98–105 (2015).
12. Der, S. D. *et al.* Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *J Thorac Oncol* **9**(1), 59–64 (2014).
13. Shukla S, Evans JR, Malik R, et al. Development of a RNA-seq based prognostic signature in lung adenocarcinoma. J Natl Cancer Inst 2016 ;109(1):djw200.
14. Zhao, K., Li, Z. & Tian, H. Twenty-gene-based prognostic model predicts lung adenocarcinoma survival. *Onco Targets Ther* **11**, 3415–3424 (2018).
15. Xie, H. & Xie, C. A six-gene signature predicts survival of adenocarcinoma type of non-small-cell lung cancer patients: a comprehensive study based on integrated analysis and weighted gene coexpression network. *Biomed Res Int* **2019**, 4250613 (2019).
16. Chambers, A. F. & Matrisian, L. M. Changing views of the role of matrix metalloproteinases in metastasis. *J Natl Cancer Inst* **89**(17), 1260–1270 (1997).
17. Wang, T. *et al.* Screening of tumor-associated antigens based on Oncomine database and evaluation of diagnostic value of autoantibodies in lung cancer. *Clin Immunol.* **210**, 108262 (2020).
18. Klupp, F. *et al.* Serum MMP7, MMP10 and MMP12 level as negative prognostic markers in colon cancer patients. *BMC Cancer* **16**, 494 (2016).
19. Su, L. *et al.* Genotypes and haplotypes of matrix metalloproteinase 1, 3 and 12 genes and the risk of lung cancer. *Carcinogenesis* **27**(5), 1024–1029 (2006).
20. Bradbury, P. A. *et al.* Matrix metalloproteinase 1, 3 and 12 polymorphisms and esophageal adenocarcinoma risk and prognosis. *Carcinogenesis* **30**(5), 793–798 (2009).
21. Sarkar, S. *et al.* Tenascin-C stimulates glioma cell invasion through matrix metalloproteinase-12. *Cancer Res* **66**(24), 11771–11780 (2006).
22. Lv, F. Z. *et al.* Knockdown of MMP12 inhibits the growth and invasion of lung adenocarcinoma cells. *Int J Immunopathol Pharmacol* **28**(1), 77–84 (2015).
23. Kowalczuk, O. *et al.* CXCL5 as a potential novel prognostic factor in early stage non-small cell lung cancer: results of a study of expression levels of 23 genes. *Tumor Biol* **35**(5), 4619–4628 (2014).
24. Zhang, J. *et al.* Establishment of the prognostic index of lung squamous cell carcinoma based on immunogenomic landscape analysis. *Cancer Cell Int* **20**, 330 (2020).
25. Kim, J. K. *et al.* Targeted disruption of S100P suppresses tumor cell growth by down-regulation of cyclin D1 and CDK2 in human hepatocellular carcinoma. *Int J Oncol* **35**, 1257–1264 (2009).
26. Gibadulinova, A. *et al.* Cancer-associated S100P protein binds and inactivatesp53, permits therapy-induced senescence and supports chemoresistance. *Oncotarget* **7**, 22508–22522 (2016).
27. Tan, B. S. *et al.* LncRNA NORAD is repressed by the YAP pathway and suppresses lung and breast cancer metastasis by sequestering S100P. *Oncogene* **38**(28), 5612–5626 (2019).
28. Favorskaya, I. *et al.* Expression and clinical significance of CRABP1 and CRABP2 in non-small cell lung cancer. *Tumour Biol* **35**(10), 10295–10300 (2014).
29. Teng, Y. C. *et al.* Histone demethylase RBP2 promotes lung tumorigenesis and cancer metastasis. *Cancer Res* **73**(15), 4711–4721 (2013).
30. Wang, S. *et al.* RBP2 induces epithelial-mesenchymal transition in non-small cell lung cancer. *PLoS One* **8**, e84735 (2013).

31. Bolas, G. *et al.* Inhibitory effects of recombinant RTS-jerdostatin on integrin alpha1beta1 function during adhesion, migration and proliferation of rat aortic smooth muscle cells and angiogenesis. *Toxicon* **79**, 45–54 (2014).
32. Belaiba, R. S. *et al.* Hypoxia up-regulates hypoxia-inducible factor-1alpha transcription by involving phosphatidylinositol 3-kinase and nuclear factor kappaB in pulmonary artery smooth muscle cells. *Mol Biol Cell* **18**, 4691–4697 (2007).
33. Jahangiri, A., Aghi, M. K. & Carbonell, W. S. beta1 integrin: Critical path to antiangiogenic therapy resistance and beyond. *Cancer Res* **74**, 3–7 (2014).
34. Qi, L. *et al.* Retinoblastoma binding protein 2 (RBP2) promotes HIF-1α-VEGF-induced angiogenesis of non-small cell lung cancer via the Akt pathway. *PLoS One* **9**(8), e106032 (2014).
35. Xia, J. *et al.* Increased IL-33 expression in chronic obstructive pulmonary disease. *Am J Physiol Lung Cell Mol Physiol* **308**(7), L619–L627 (2015).
36. Koca, S. S. *et al.* Serum IL-33 level and IL-33 gene polymorphisms in Behçet's disease. *Rheumatol Int* **35**(3), 471–477 (2015).
37. Mei, L. *et al.* Association between ADRB2, IL33, and IL2RB gene polymorphisms and lung cancer risk in a Chinese Han population. *Int Immunopharmacol* **77**, 105930 (2019).
38. Wang, J. *et al.* Genetic predisposition to lung cancer: comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. *Sci Rep* **7**(1), 8371 (2017).
39. Li, R. *et al.* Identification and validation of the prognostic value of immune-related genes in non-small cell lung cancer. *Am J Transl Res* **12**(9), 5844–5865 (2020).
40. Bertrand, S. *et al.* Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. *Mol Biol Evol* **21**, 1923–1937 (2004).
41. Wang, J. *et al.* Expression of HNF4G and its potential functions in lung cancer. *Oncotarget* **9**(26), 18018–18028 (2017).
42. Horne, Z. D. *et al.* Increased levels of tumor-infiltrating lymphocytes are associated with improved recurrence-free survival in stage 1A nonesmall-cell lung cancer. *J Surg Res* **171**, 1–5 (2011).
43. Sionov, R. V., Fridlender, Z. G. & Granot, Z. The multifaceted roles neutrophils play in the tumor microenvironment. *Cancer Microenviron* **8**(3), 125–158 (2015).
44. Sun, Z. & Yang, P. Role of imbalance between neutrophil elastase and alpha 1-antitrypsin in cancer development and progression. *Lancet Oncol* **5**, 182–190 (2004).
45. Kargl, J. *et al.* Neutrophil content predicts lymphocyte depletion and anti-PD1 treatment failure in NSCLC. *JCI Insight.* **4**(24), e130850 (2019).
46. Suh, K. J. *et al.* Post-treatment neutrophil-to-lymphocyte ratio at week 6 is prognostic in patients with advanced non-small cell lung cancers treated with anti-PD-1 antibody. *Cancer Immunol Immunother.* **67**(3), 459–470 (2018).

Acknowledgements

This study was supported by the Natural Science Foundation of Shandong Province, China (No. ZR2021QH334).

Author contributions

D.L.Y. and X.B.M. analyzed and visualized the results. P.S. wrote and revised the manuscript. All authors have read and approved the content, and agree to submit it for consideration for publication in your journal.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-04268-7.

Correspondence and requests for materials should be addressed to P.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.