RESEARCH

Characteristic of molecular subtypes in lung adenocarcinoma based on m6A RNA methylation modification and immune microenvironment

Hao Zhou[1†], Miaosen Zheng[2†], Muqi Shi[3], Jinjie Wang[1], Zhanghao Huang[1], Haijian Zhang[4], Youlang Zhou[4*] and Jiahai Shi[1*]

Abstract

Background: Lung adenocarcinoma (LUAD) is a major subtype of lung cancer and closely associated with poor prognosis. N6-methyladenosine (m6A), one of the most predominant modifications in mRNAs, is found to participate in tumorigenesis. However, the potential function of m6A RNA methylation in the tumor immune microenvironment is still murky.

Methods: The gene expression profile cohort and its corresponding clinical data of LUAD patients were downloaded from TCGA database and GEO database. Based on the expression of 21 m6A regulators, we identified two distinct subgroups by consensus clustering. The single-sample gene-set enrichment analysis (ssGSEA) algorithm was conducted to quantify the relative abundance of the fraction of 28 immune cell types. The prognostic model was constructed by Lasso Cox regression. Survival analysis and receiver operating characteristic (ROC) curves were used to evaluate the prognostic model.

Result: Consensus classification separated the patients into two clusters (clusters 1 and 2). Those patients in cluster 1 showed a better prognosis and were related to higher immune scores and more immune cell infiltration. Subsequently, 457 differentially expressed genes (DEGs) between the two clusters were identified, and then a seven-gene prognostic model was constricted. The survival analysis showed poor prognosis in patients with high-risk score. The ROC curve confirmed the predictive accuracy of this prognostic risk signature. Besides, further analysis indicated that there were significant differences between the high-risk and low-risk groups in stages, status, clustering subtypes, and immunoscore. Low-risk group was related to higher immune score, more immune cell infiltration, and lower clinical stages. Moreover, multivariate analysis revealed that this prognostic model might be a powerful prognostic predictor for LUAD. Ultimately, the efficacy of this prognostic model was successfully validated in several external cohorts (GSE30219, GSE50081 and GSE72094).

---

* Correspondence: zhouyoulang@ntu.edu.cn; sjh@ntu.edu.cn
[†]Hao Zhou and Miaosen Zheng contributed equally to this work.
[4]Research Center of Clinical Medicine, Affiliated Hospital of Nantong University, Nantong 226001, Jiangsu, China
[1]Department of Thoracic Surgery, Affiliated Hospital of Nantong University and Medical School of Nantong University, Nantong 226001, Jiangsu, China
Full list of author information is available at the end of the article

Conclusion: Our study provides a robust signature for predicting patients' prognosis, which might be helpful for therapeutic strategies discovery of LUAD.

Keywords: Lung adenocarcinoma, m6A RNA methylation, Immune microenvironment, Prognostic model

Introduction

Lung cancer is one of the common cancers worldwide, leading to high mortality every year [1]. Non-small cell lung cancer (NSCLC) accounts for 85% of cases, and lung adenocarcinoma (LUAD) is the most prevalent subtype [2, 3]. Despite great improvements in diagnostic and therapeutic techniques, the prognosis for LUAD patients is still poor [4]. Therefore, identifying treatment targets and effective predictors to improve the prognosis of LUAD patients is critical.

N6-methyladenosine (m6A) is one of the most common mRNA internal modifications in eukaryotic organisms and methylation modificatied at the N6 position of adenosine residues in RNA [5, 6]. The modification of RNA m6A is involved in many essential biological processes including gene expression, immunomodulation and cancers [7]. In addition, RNA modifications, especially m6A modification, has been proved to be necessary for tumor development [8]. The tumor microenvironment (TME) is important in the formation, development, and treatment of tumors which contains tumor cells, immune cells and stromal cells [9]. Tumor infiltrating immune cells are an important component of the complex microenvironment. Currently, TME has been the hot pot of the tumor field because research suggested that TME immune cells are closely related to the prognosis and malignancy of tumors [10]. Hence, exploring the relationship between m6A RNA methylation and immune microenvironment is beneficial to improve the prognosis accuracy of patients with LUAD.

In this study, we assessed the expression of 21 m6A regulatory factors from TCGA database, and the TCGA patients were divided into two clusters according to the expression of these genes. Then, a risk prognostic signature was established on the base of the differentially expressed genes (DEGs) between the two clusters. We analyzed and compared the different immune cell infiltration and clinical outcomes between the high-risk group and low-risk group. Furthermore, this prognostic model showed pretty good predictive accuracy compared with other clinical factors. Importantly, we validated the prognostic model in three independent external cohorts (GSE30219, GSE50081 and GSE72094). These results discovered novel insight for the diagnosis and treatment of LUAD by using bioinformatics tools.

Materials and methods

Data source

The TCGA database was used to obtaine the gene expression profile cohort and its corresponding clinical data of LUAD patients, including the FPKM value of gene expression of 535 LUAD samples and 59 normal samples. Four hundred ninety-four samples with complete survival information were used for subsequent analysis. Because a considerable number of patients lack the clinical information of M classification, in order to ensure the number of samples, we did not include M classification in the clinical correlation analysis. The normalized matrix files of the three cohorts (GSE30219, GSE50081 and GSE72094) from the GEO database were downloaded for the validation data sets.

Consensus cluster analysis for 21 m6A regulators

In this study, we included 21 m6A regulators including 8 writers (METTL3, METTL14, RBM15, RBM15B, WTAP, VIRMA, CBLL1, ZC3H13), 2 erasers (ALKBH5, FTO) and 11 readers (YTHDC1, YTHDC2, YTHDF1, YTHDF2, YTHDF3, IGF2BP1, HNRNPA2B1, HNRNPC, FMR1, LRPPRC, ELAVL1) [11]. Based on the expression of 21 m6A regulators, we performed consensus classification to identify different m6A modification patterns. The patients were divided into different subtypes using the R package "ConsensusClusterPlus" for further analysis. To ensure the stability of classification, 1000 iterations, and a resample rate of 80% were conducted. The cumulative distribution function (CDF) curve was used to determine the clustering number [12].

Inference of tumor microenvironment and immune cells

The immune score was calculated by applying the ESTIMATE algorithm to each patient via the "estimate" R package [13]. The single-sample gene-set enrichment analysis (ssGSEA) algorithm was conducted to quantify the relative abundance of the fraction of 28 immune cell types [14].

Differential gene expression and functional analyses

Wilcoxon test conducted by R software was performed to identify the differentially expressed genes between cluster 1 and cluster 2. The cut-off criteria was | logFC | > 1 and FDR < 0.05. Gene ontology (GO) enrichment analysis by clusterProfiler R package was performed to know the potential biological processes associated with

Zhou *et al. BMC Cancer*   (2021) 21:938

Page 3 of 14

the DEGs [15]. The *P*-value adjusted < 0.05 was regarded significant.

Risk assessment model construction

Firstly, 457 DEGs were identified between two clusters for our data screening. The DEGs associated with overall survival (OS) were screened via univariate Cox regression analysis. Then, Lasso Cox regression was used to selected the most striking markers to establish the prognostic model. The optimal value of the penalty parameter λ was defined by 10-fold cross validation [16]. The risk score was obtained after multiplying the expression level of each gene by its coefficient obtained from the LASSO Cox regression. Then, the patients were divided into high-risk and low-risk groups according to the optimal cut-off value.

Gene set enrichment analysis

GSEA was used to investigate potential biological functions between different risk groups of LUAD patients [17]. The number of random sample permutations was performed 1000 times. The significance was based on the false discovery rate (FDR) < 0.05 and *P*-value < 0.05.

Statistical analysis

All data analyses were conducted using R language (version 3.6.3). Kaplan-Meier method was used to draw survival curves. The receiver operating characteristic (ROC) curve was plotted with the SurvivalROC R package. The Wilcoxon test was used to compare the two groups' differences. The correlation tests were conducted by Pearson correlation analysis. Univariate and multivariate analyses were conducted to determine whether the prognostic model was independently variable when integrated with other clinical factors. Statistical *P* < 0.05 was considered significantly.

Results

The different expression of m6A RNA methylation regulatory molecules and consensus clustering analysis

In order to explore the potential role of m6A regulatory factors in the occurrence and development of LUAD, we systematically analyzed the expression of 21 m6A regulatory factors between LUAD and adjacent normal tissues. The different expression values of these genes between LUAD and normal tissues were observed in Fig. 1a. Figure 1b showed the important co-expression patterns between all the 21 m6A regulatory factors. The results indicated that YTHDF3 was strongest positively correlated with VIRMA ($r = 0.75$). RBM15 was positively associated with YTHDC2 ($r = 0.65$). METTL14 was positively associated with YTHDC1 ($r = 0.64$). Instead, FTO was negatively associated with HNRNPC ($r = -0.27$). Then, we performed a consensus classification to determine
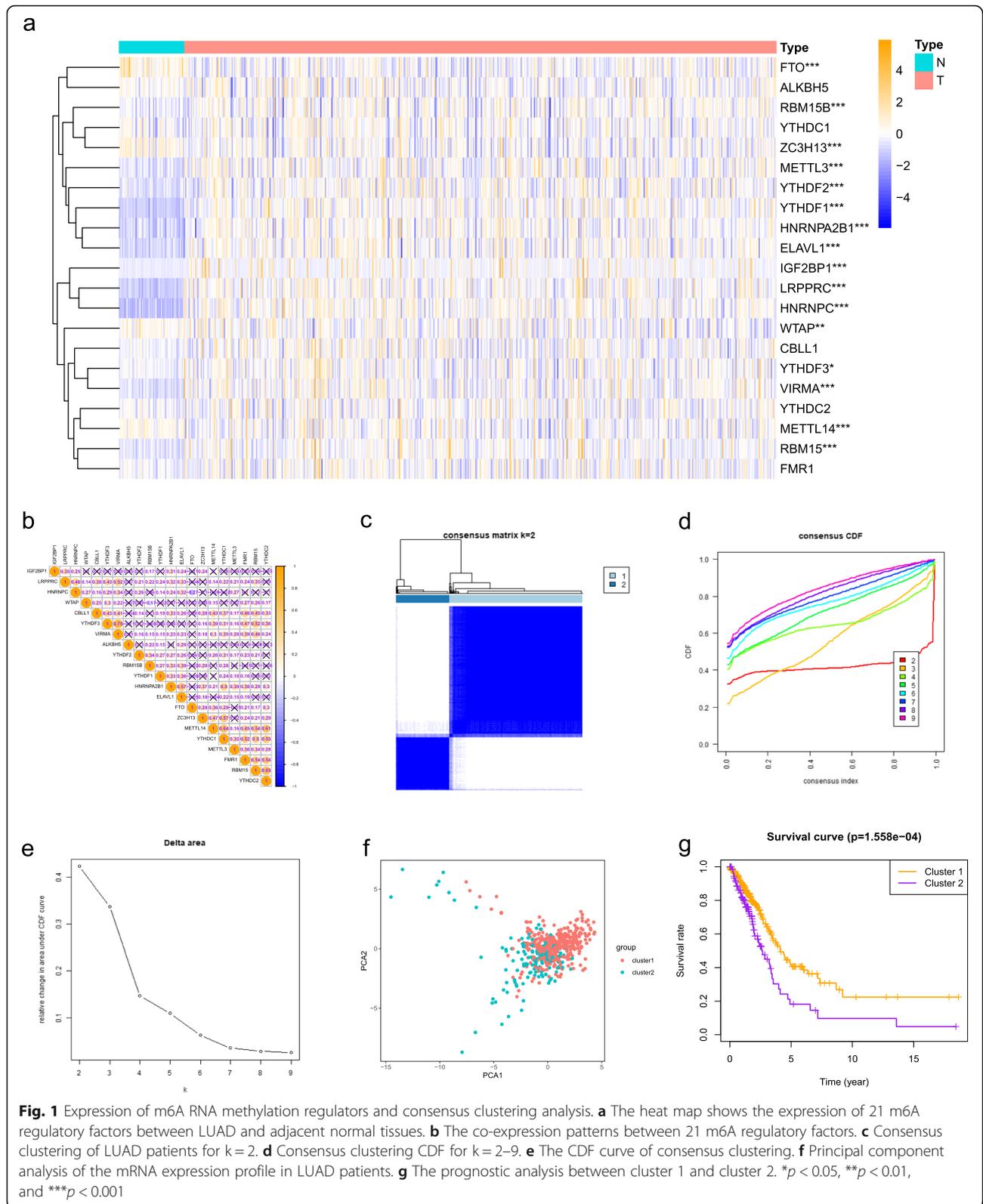
the optimal number of clusters by looking for a suitable k value. The k = 2 was demonstrated to have the best clustering stability (Fig. 1c-e). The LUAD patient cohort was separated into two clusters, namely cluster 1 and cluster 2. In addition, PCA shows two distinct distribution patterns, which shows that the classification generated by consensus clustering analysis is effective (Fig. 1f). The prognostic analysis showed that the OS of patients with LUAD in cluster 2 was obviously shorter than that in cluster 1 (Fig. 1g).
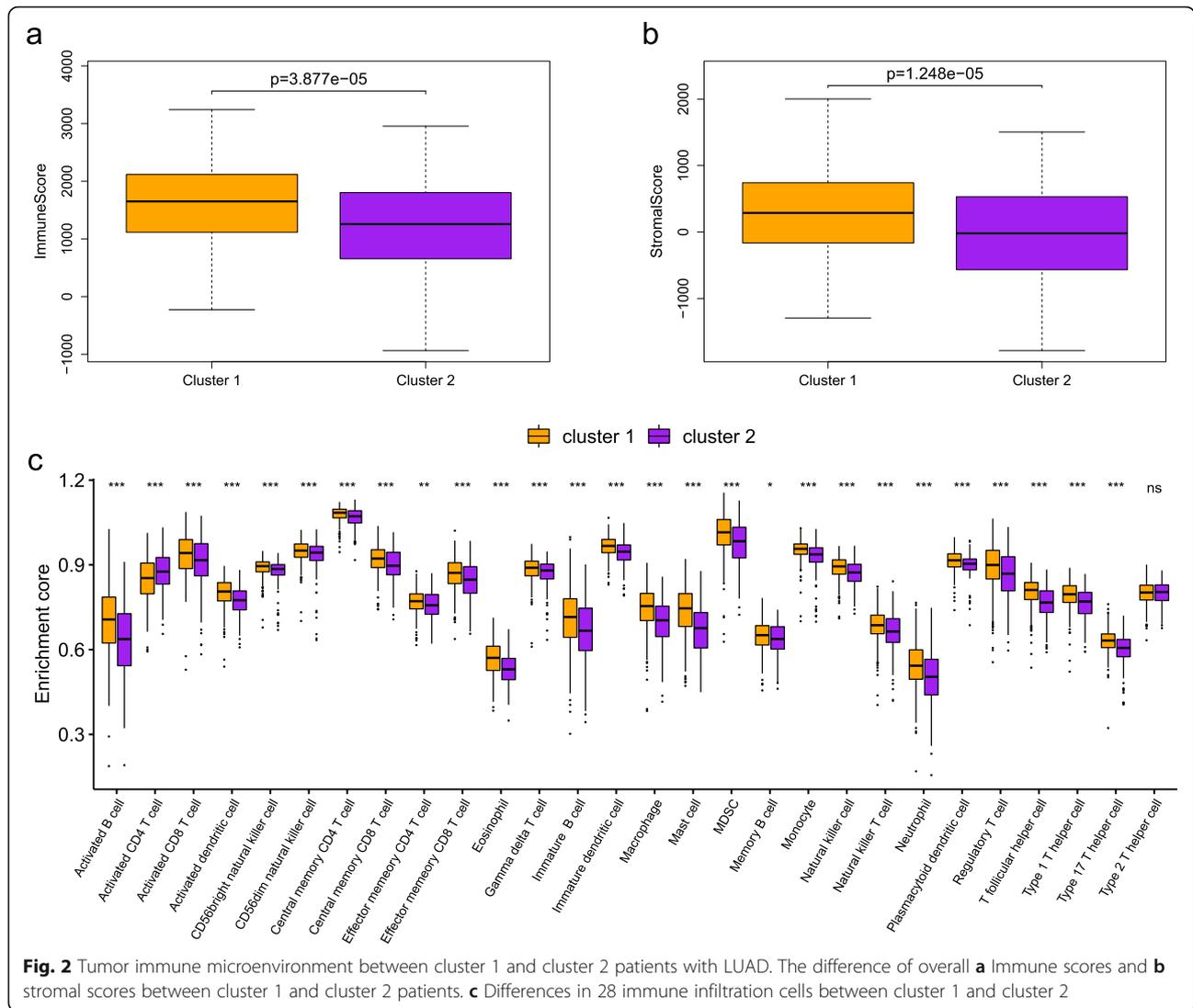
Tumor immune microenvironment of the two clusters of LUAD patients

A piece of increasing evidence indicated the close relationship between the malignant degree of cancers and the tumor immune microenvironment. Considering the obvious differences of the OS in two clusters of LUAD patients, we speculated that tumor immune microenvironment might be an important contributor to the LUAD progress. Then we explored the differences of immune infiltration to distinguish the patients in two clusters. Utilizing the ESTIMATE algorithm, we found that the immune score of cluster 1 was much higher than that of cluster 2 (Fig. 2a). And the stromal score presented the same tendency (Fig. 2b). Based on the ssGSEA, we evaluated the proportion difference of 28 immune cell types in the two clusters (Table S1). As shown in Fig. 2c, the two clusters had significantly different characteristics of immune cell infiltration. In addition, the degree of immune cell infiltration in cluster 1 was significantly higher than that in cluster 2. Overall, these findings demonstrated that the presence of different immune cell populations may influence the OS of the patients with LUAD.

Identification of differentially expressed genes and functional analyses

Four hundred fifty-seven DEGs were identified to further explore the differences between the two m6A modification patterns (Table S2). Among them, 168 DEGs were up-regulated in cluster 1 and 289 DEGs were up-regulated in cluster 2. Figure 3a-b showed the heat map and volcano plot. For better understanding the biological functions of the two clusters based on DEGs, GO enrichment analysis was performed by clusterProfiler R package. And it was found that overexpression of genes in cluster 1 mainly enriched in immune-related biological processes, such as humoral immune response and regulation of adaptive immune response (Fig. 3c and Table S3). The genes overexpressed in cluster 2 were found to be mainly enriched in many biological processes related to the cell cycle (Fig. 3d and Table S4).

Zhou et al. BMC Cancer (2021) 21:938

Page 4 of 14

Fig. 1 Expression of m6A RNA methylation regulators and consensus clustering analysis. **a** The heat map shows the expression of 21 m6A regulatory factors between LUAD and adjacent normal tissues. **b** The co-expression patterns between 21 m6A regulatory factors. **c** Consensus clustering of LUAD patients for k = 2. **d** Consensus clustering CDF for k = 2–9. **e** The CDF curve of consensus clustering. **f** Principal component analysis of the mRNA expression profile in LUAD patients. **g** The prognostic analysis between cluster 1 and cluster 2. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$

Zhou et al. BMC Cancer   (2021) 21:938

Page 5 of 14

Fig. 2 Tumor immune microenvironment between cluster 1 and cluster 2 patients with LUAD. The difference of overall **a** Immune scores and **b** stromal scores between cluster 1 and cluster 2 patients. **c** Differences in 28 immune infiltration cells between cluster 1 and cluster 2
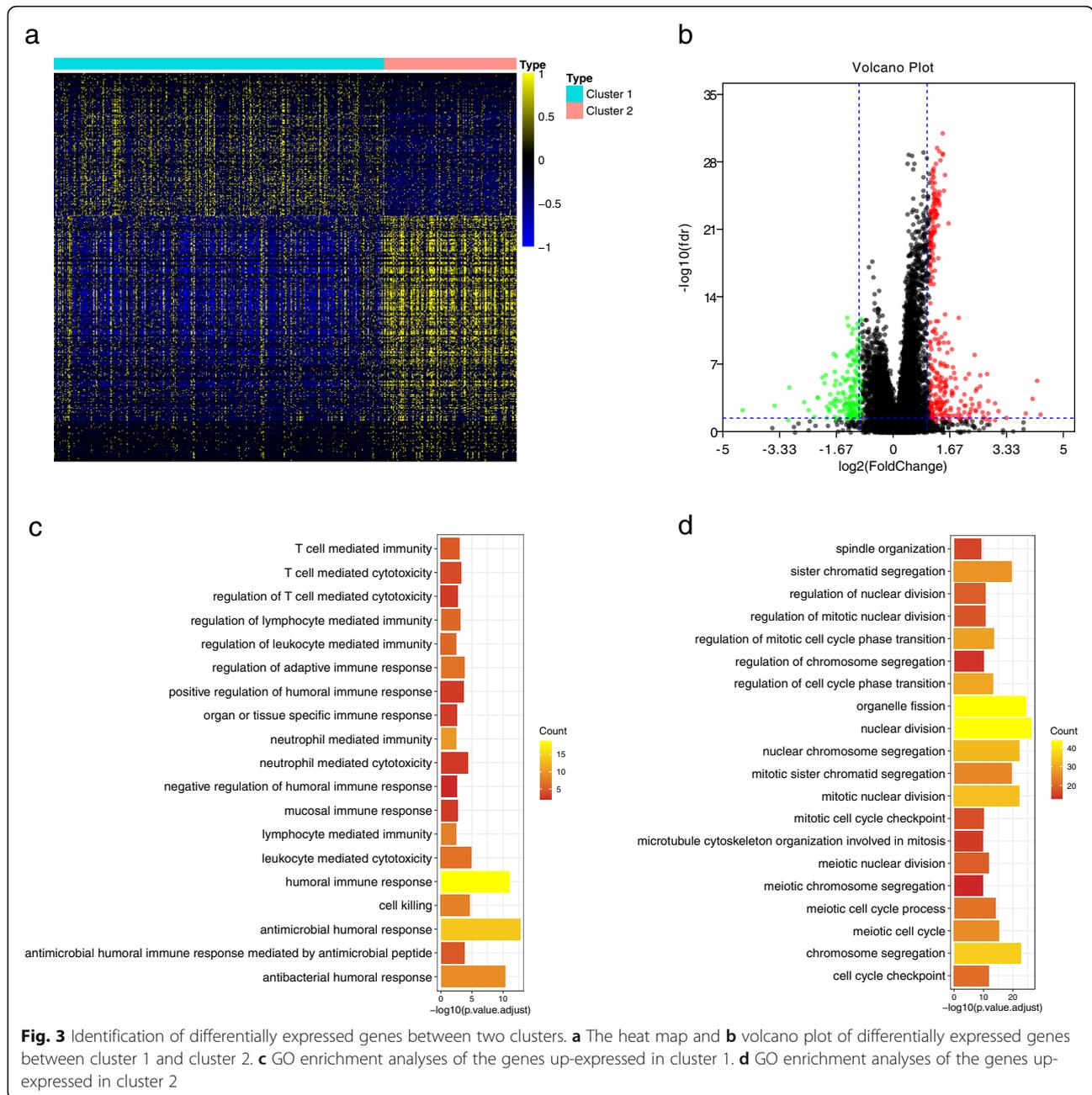
Construction of a seven-gene prognostic model

With univariate Cox regression analysis, 24 prognostic-related genes based on the DEGs were identified (Fig. 4a). Then, Lasso Cox regression was conducted to determine the key genes with the best prognostic value by reducing the dimensionality and calculate the relative coefficients of the genes (Fig. 4b-c). Ultimately, seven optimal genes including CLEC3B, TENM3, IGF2BP1, E2F7, ANLN, ANKRD18B and FBN2 were chosen to construct the prognostic model for LUAD. The coefficients for each gene were listed in Table S5. The risk score for each patient was acquired by multiplying the each gene's expression level by its coefficient. Then, according to the optimal cut-off value, the patients were divided into high- and low-risk groups (Fig. 4d). One hundred forty-eight patients were in the high-risk group and 346 patients were in the low-risk group. Besides, the survival

analysis in Fig. 4e revealed the worse prognosis in patients with a high-risk score. The sensitivity of the prognostic model was assessed by ROC curve (Fig. 4f). And the AUC result of this risk score model was 0.703. Compared with other clinical parameters, our prognostic risk signature had pretty good predictive accuracy. The risk score distribution and each patient's survival status were shown in Fig. 4g. The tendency of seven hub genes' expression in high-risk and low-risk groups was shown in the heat map of Fig. 4h.

Correlation analysis was performed between the risk model and immune cell infiltration

To explore the difference of tumor immune microenvironment between the two groups with different degrees of risk, the infiltration levels of 28 immune cell types were evaluated based on ssGSEA (Fig. 5a). The result

Fig. 3 Identification of differentially expressed genes between two clusters. **a** The heat map and **b** volcano plot of differentially expressed genes between cluster 1 and cluster 2. **c** GO enrichment analyses of the genes up-expressed in cluster 1. **d** GO enrichment analyses of the genes up-expressed in cluster 2

suggested that high immune cell infiltration in a significant number of low-risk samples. Then, we examined the distribution of stromal and immune scores between the two risk groups of patients using the ESTIMATE algorithm. Comparing with the high-risk group, the low-risk group presented higher immune and stromal scores (Fig. 5b-c). In addition, PD-L1, PD-1, PD-L2, and LAG3 were assessed to know these different expression of immune checkpoint molecules in different groups (Fig. 5d-g). When compared to patients in low-risk group, high-risk group patients had higher expression of immune

checkpoint molecules. This indicated that the prognosis model may have a potential role in predicting patients' response to anti-checkpoint immunotherapy. Based on the TCGA cohort, we further analyzed the distribution differences of somatic genomic mutation between high-risk group and low-risk group. We found that high-risk group presented more extensive tumor mutation burden than low-risk group (Fig. 5h-i). In addition, we analyzed the difference in miRNA of the two groups based on the risk score. A total of 53 differentially expressed miRNAs were identified with the cut-off criteria | logFC | > 1 and
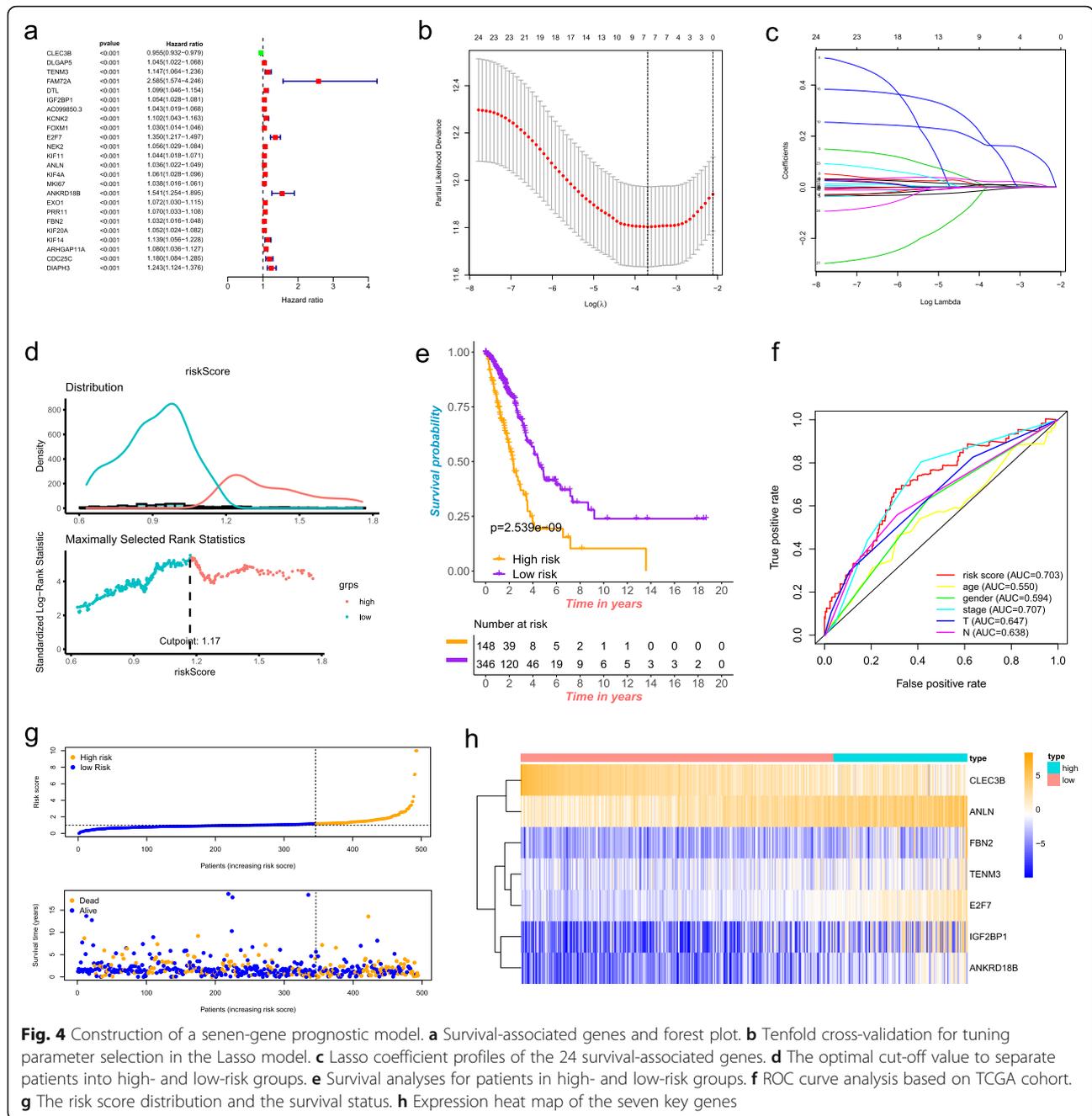
Fig. 4 Construction of a senen-gene prognostic model. **a** Survival-associated genes and forest plot. **b** Tenfold cross-validation for tuning parameter selection in the Lasso model. **c** Lasso coefficient profiles of the 24 survival-associated genes. **d** The optimal cut-off value to separate patients into high- and low-risk groups. **e** Survival analyses for patients in high- and low-risk groups. **f** ROC curve analysis based on TCGA cohort. **g** The risk score distribution and the survival status. **h** Expression heat map of the seven key genes

FDR < 0.05 based on the TCGA cohort. The results were shown in Table S6 and visualized in the form of volcano plot in Fig. S1.

Correlation assessment between risk signature and clinical features

We further analyzed the relationship between the risk score model and clinical features. Figure 6a summarized the expression levels of the seven key genes as a heat map. Almost all these genes expressed highly in the high-risk group while CLEC3B highly expressed in the low-risk group. In addition, CLEC3B was also highly expressed in normal samples and related to better OS while the rest six genes were highly expressed in tumor samples and associated with worse OS (Fig. S2-S3). The correlation between the seven key genes was showed in Fig. 6b. The high-risk and low-risk groups have a significant difference in terms of stages, status, clustering subtypes, and immunoscore (Fig. 6a). Compared to the cluster 2, cluster 1 had a lower risk score (Fig. 6c). This was also in line with our previous result that patients in cluster 1 had a better OS. Moreover, compared to the
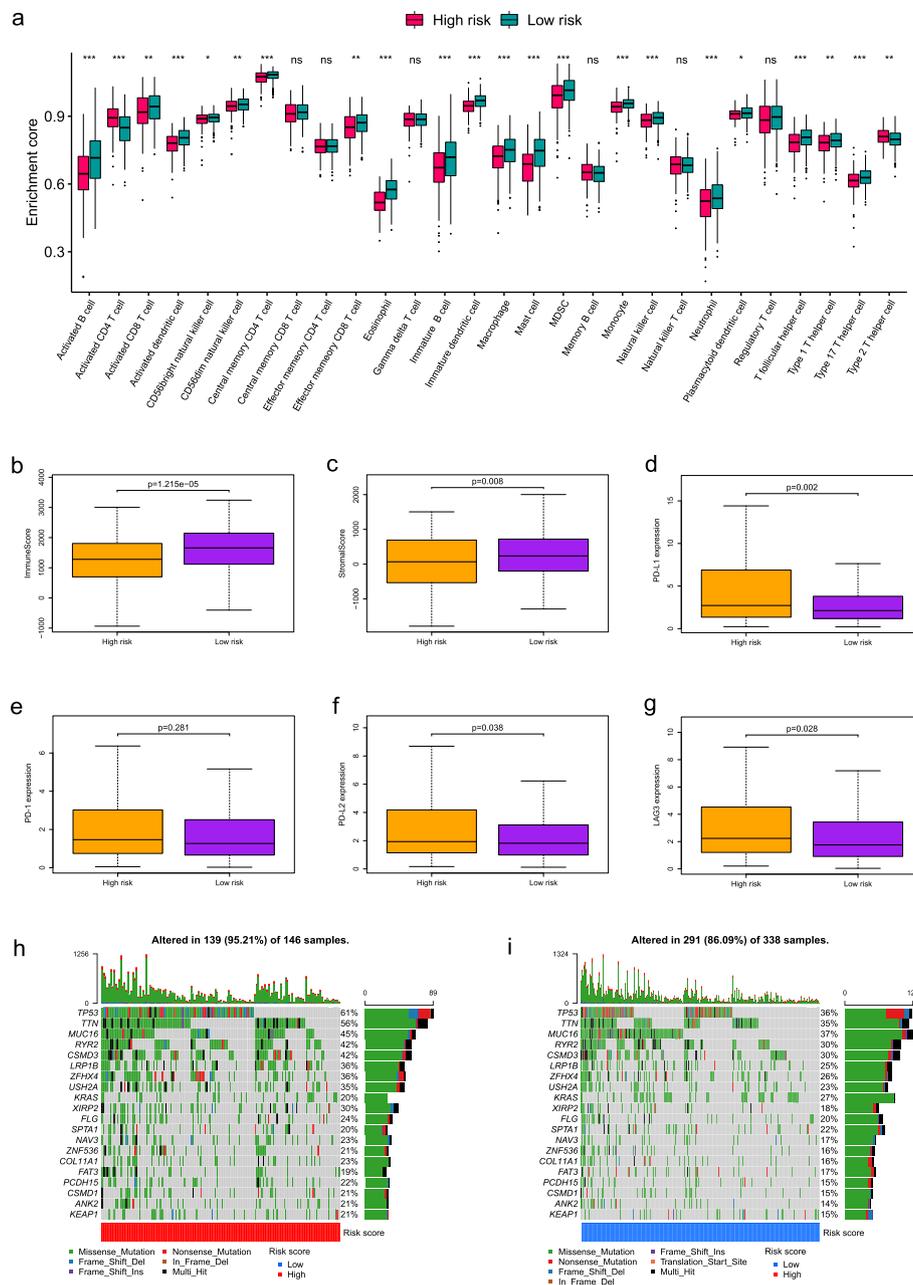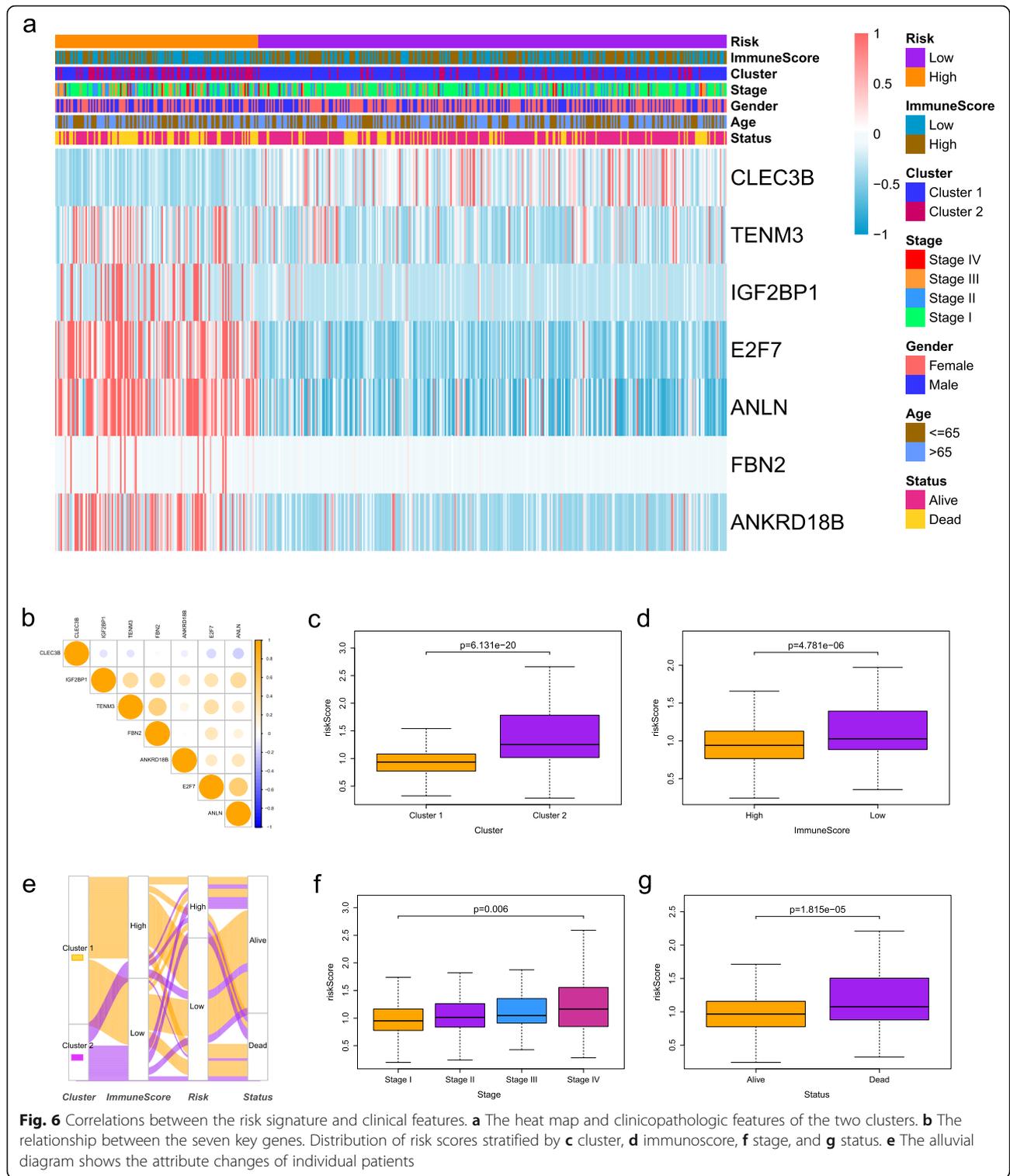
Fig. 5 Correlation analysis of the prognostic model and immune cell infiltration. **a** Differences in 28 immune infiltration cells between high- and low-risk groups. **b** Immune scores and **c** stromal scores between high- and low-risk patients. The expression level of immune checkpoint molecules including **d** PD-L1, **e** PD-1, **f** PD-L2, and **g** LAG3 between high- and low-risk patients. The waterfall plot showing the differences of tumor somatic genomic mutation between **h** high-risk group and **i** low-risk group

low immunescore group, a lower risk sore in the high immunescore group was observed in Fig. 6d. With the clinical-stage increased, the risk score also increased (Fig. 6f). Figure 6g revealed a slightly better status in the low-risk group patients. These findings demonstrated that the prognostic risk signature was closely related to the malignant degree of tumor. And the attribute
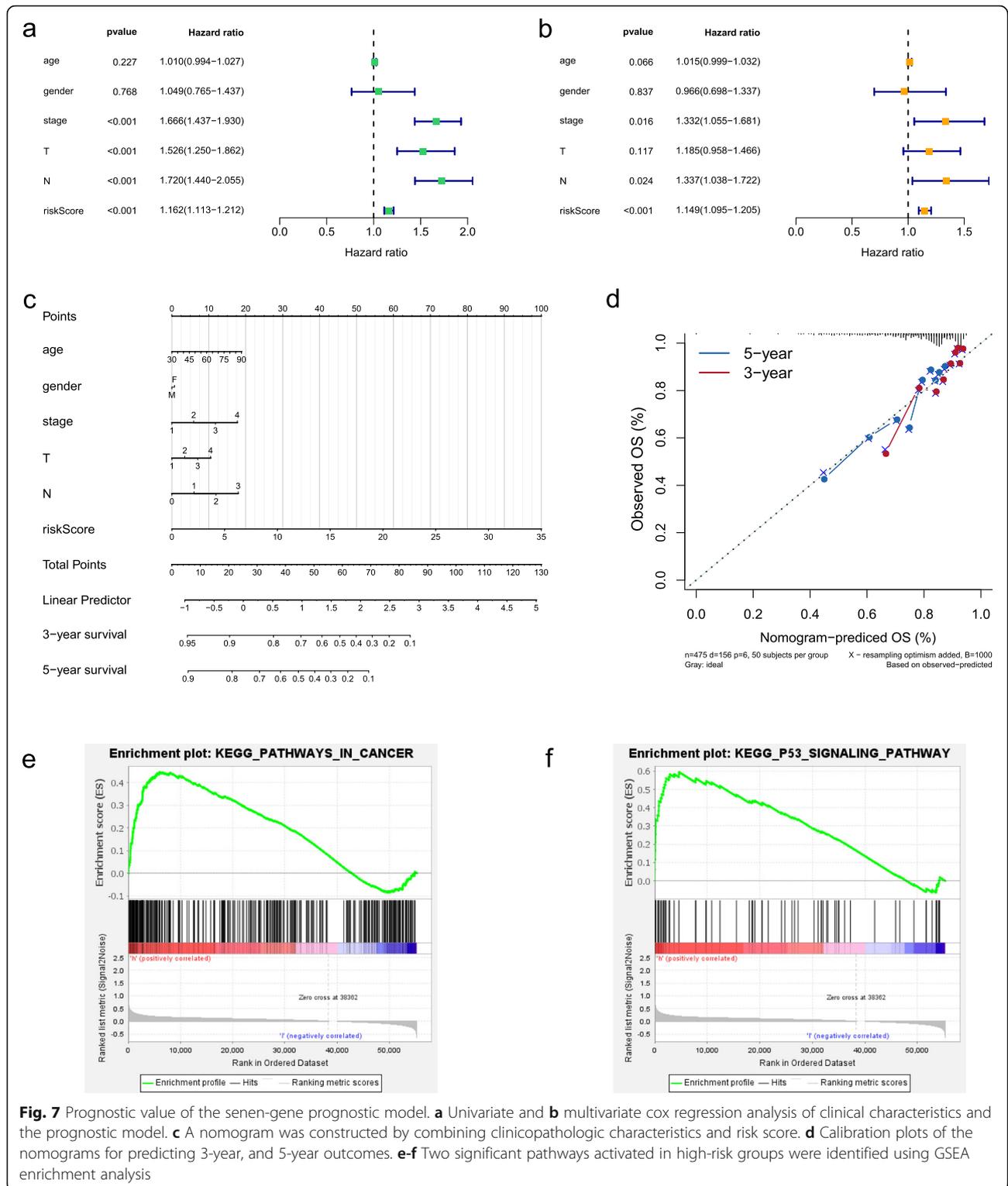
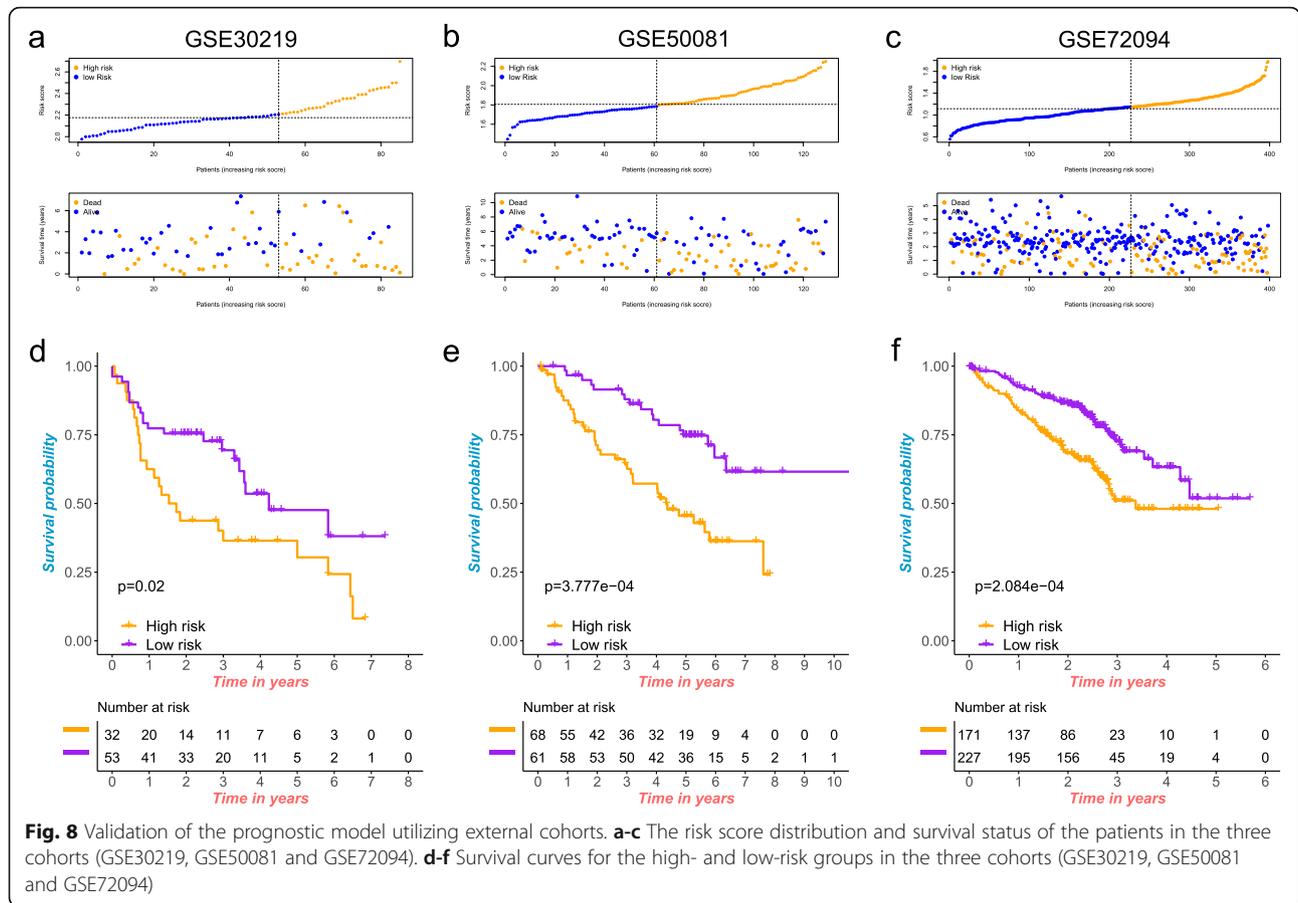changes of individual patients can be clearly presented by an alluvial diagram (Fig. 6e).

Considering other clinical features in TCGA, whether the risk score for OS is an independent indicator needs further exploration. As the univariate analysis shown in Fig. 7a, the clinical stage, T stage, N stage, and risk score were related to poorer survival. The multivariate analysis

Fig. 6 Correlations between the risk signature and clinical features. **a** The heat map and clinicopathologic features of the two clusters. **b** The relationship between the seven key genes. Distribution of risk scores stratified by **c** cluster, **d** immunoscore, **f** stage, and **g** status. **e** The alluvial diagram shows the attribute changes of individual patients

indicated that the risk score could be considered as an independent prognostic factor for patients with LUAD (Fig. 7b). To facilitate the utilization of risk score, a nomogram was plotted considering risk score and other clinical factors (Fig. 7c). Calibration plots for 3-year and 5-year OS were used to visualize the performances of the nomograms (Fig. 7d). To further explore the biological pathways associated with the risk signature, we

Fig. 7 Prognostic value of the senen-gene prognostic model. **a** Univariate and **b** multivariate cox regression analysis of clinical characteristics and the prognostic model. **c** A nomogram was constructed by combining clinicopathologic characteristics and risk score. **d** Calibration plots of the nomograms for predicting 3-year, and 5-year outcomes. **e-f** Two significant pathways activated in high-risk groups were identified using GSEA enrichment analysis

Fig. 8 Validation of the prognostic model utilizing external cohorts. **a-c** The risk score distribution and survival status of the patients in the three cohorts (GSE30219, GSE50081 and GSE72094). **d-f** Survival curves for the high- and low-risk groups in the three cohorts (GSE30219, GSE50081 and GSE72094)

performed GSEA enrichment analysis and found that pathways in cancer and p53 signaling pathway were activated in high-risk groups (Fig. 7e-f).

Validation of the prognostic signature by the GEO database

To determine the prognostic potential of the seven-gene signature in other datasets, Three GEO data sets (GSE30219, GSE50081 and GSE72094) were adapted as independent external validation. The same formula was used to calculate the risk score for each patient in the GEO cohorts. Then, according to the optimal cut-off value, we divided LUAD patients into high-risk group and low-risk group. The patients' risk score distribution and survival status were shown in Fig. 8a-c. The survival curves revealed that patients in the high-risk group have a shorter survival time than patients in the low-risk group (Fig. 8d-f). Therefore, our risk signature can well distinguish patients according to the risk score. It is helpful for prognosis prediction and treatment of LUAD patients.

Discussion

Lung cancer is a global public health challenge with its high mortality [18]. LUAD, a fatal malignancy associated with poor prognosis and high mortality rates, accounts for more than 40% of the total incidence of lung cancer [19, 20]. Although the diagnosis and treatment of LUAD have made great progress, the effective diagnosis and prognosis prediction of LUAD patients is still a major clinical challenge [21]. Therefore, identifying key molecules and constructing a prediction model with high stability and effectiveness are conducive to the implementation of precise treatment and improve the prognosis of patients.

N6-methyladenosine (m6A) is the most abundant internal modification of eukaryotic mRNA. Almost every stage of mRNA metabolism is affected by m6A mRNA methylation [22]. In addition, new evidence suggests that m6A RNA methylation plays a vital role in tumorigenesis and development [23]. FTO and METTL3 have been reported as potential targets for the diagnosis and treatment of LUAD patients. It is reported that FTO facilitates LUAD cell progression by activating cell migration through m6A demethylation [24]. High expression of

Zhou *et al. BMC Cancer* (2021) 21:938

Page 12 of 14

METTL3 in LUAD is believed to promote the growth and invasion of cancer cells [25]. However, there are few studies on the relationship between m6A related genes and LUAD. In this study, the expression of 21 m6A regulatory factors from LUAD and adjacent normal tissues were systematically analyzed. We observed the different expression levels of these genes between LUAD and normal tissues. Based on the expression of 21 m6A regulatory factors, TCGA patients were divided into two clusters utilizing consensus classification. The prognosis of patients in cluster 1 was better than that of patients in cluster 2. Given that the patients' prognosis was probably related to tumor immune microenvironment, the differences of immune infiltration between two clusters were analyzed. We found that patients in cluster 1 had a higher immune score and immune cell infiltration. Previous studies indicated that the immune-inflamed phenotype shows the infiltration of massive amounts of immune cells in the tumor microenvironment, and correlated with better prognosis [11, 26]. To further explore the difference between the two clusters, we identified 457 DEGs and analyzed their biological functions by GO enrichment analyses. Interestingly, the immune-related biological processes were mainly enriched for the up-expressed genes in cluster 1. Then, after selecting seven survival-associated DEGs by Lasso Cox regression, a reliable prognostic model was successfully established. The survival analysis demonstrated patients with a high-risk score have a worse prognosis. The ROC curve confirmed the predictive accuracy of this prognostic risk signature. In addition, higher immune scores and immune cell infiltration were founded in the low-risk group. In recent years, immune checkpoint inhibitors have attracted much attention due to their promising application in the immunotherapy of cancer [27]. Therefore, expressions of immune checkpoint molecules like PD-L1, PD-1, PD-L2, and LAG3 in different groups were examined. And it was found that there existed a positive correlation between risk score and the expressions of immune checkpoint molecules. Therefore, these results demonstrated that the prognosis model might have a potential role in predicting the clinical response of immunotherapy.

Among the seven key molecules (CLEC3B, TENM3, IGF2BP1, E2F7, ANLN, ANKRD18B, and FBN2), only CLEC3B highly expressed in the low-risk group while other genes highly expressed in the high-risk group. CLEC3B, C-type lectin domain family 3 member B, is a member of the C-type lectin superfamily [28]. It encodes tetranectin, a plasminogen kringle-4-binding protein in cells [29]. It is reported that CLEC3B was down-regulated in several tumors and considered as a tumor suppressor in oral squamous cell carcinoma [30]. A previous study demonstrated that the expression of CLEC3B is correlated with the level of immune infiltration in lung cancer, and it is promising to be the important marker for the early diagnosis of lung cancer [31]. The protein encoded by TENM3 gene belongs to the teneurin family and is involved in tumorigenesis and drug resistance [32]. It is reported that TENM3 was up-regulated in tumor tissues, and it may function as an oncogenic gene in esophageal cancer [33]. However, the role of TENM3 in lung cancer is not yet clear. IGF2BP1 is an RNA-binding protein that participates in tumor progression, tumor cell proliferation and growth [34, 35]. The let-7 family exerts its role in suppressing the migration and growth of tumor cells by inhibiting the expression of IGF2BP1 [36]. Previous studies showed the up-regulation of IGF2BP1 in LUAD, which affects the progression of the disease [37, 38]. Furthermore, high expression of IGF2BP1 was associated with poor OS in LUAD [37]. E2F7 is a member of the E2F transcription factors family. It is reported that the mammalian E2F transcription factors play vital roles in the cell cycle, so they are closely related to cancer [39]. In addition, E2F7 has been found up-regulated in various malignant tumors, such as acute myeloid leukaemia and cutaneous squamous cell carcinomas [40, 41]. Knockdown of E2F7 can repress cell growth in endometrial carcinoma [42]. However, how E2F7 participates in LUAD is still unknown. The ANLN gene encodes an actin-binding protein that contributes to cell growth and migration [43]. The expression of ANLN is up-regulated in a variety of types of tumors, including lung cancer, and the development of cancers is related to the expression level of ANLN [44]. Previous studies indicated the vital role of ANLN in cell proliferation, and lacking ANLN reduced cell migration and invasion [45, 46]. ANLN has been reported to participate in the metastasis of LUAD by promoting the EMT of tumor cells [47]. ANKRD18B is a member of ANKRD family that functions in the occurrence of cancer, evidence that over-expression of ANKRD18B suppressed the growth of lung cancer cells has been reported [48]. FBN2 encoded the protein which belongs to the connective tissue microfibrils and participates in elastic fiber assembly. Nevertheless, there is no doubt to determine the impact of FBN2 in LUAD in the future.

Nevertheless, our study had a few limitations to be considered. Firstly, further experiments are necessary to verify our results because our study was only based on public databases. Secondly, larger sample size is necessary to confirm the predictive ability of our prognostic models. Thirdly, the biological roles of the seven key genes in LUAD require further experimental validation.

Conclusions
In summary, based on the differentially expressed genes between the two clusters, a reliable prognostic model was established and identified as an independent

prognostic predictor for LUAD. Furthermore, this risk signature could also be considered as a predictor of increased immune cell infiltration, proving its potential role in the tumor immune microenvironment. Importantly, the prognostic value of this risk signature was successfully validated in independent external cohorts (GSE30219, GSE50081 and GSE72094). Our current study provides a robust prognostic model to predict the prognosis of LUAD patients, which may provide significant guidance for the diagnosis and treatment of LUAD.

Abbreviations
LUAD: Lung adenocarcinoma; m6A: N6-methyladenosine; DEG: Differentially expressed gene; NSCLC: Non-small cell lung cancer; TME: Tumor microenvironment; CDF: Cumulative distribution function; ssGSEA: single-sample gene-set enrichment analysis; GO: Gene ontology; OS: Overall survival; FDR: False discovery rate; ROC: Receiver operating characteristic; GEO: Gene Expression Omnibus; LASSO: Least absolute shrinkage and selection operator; TCGA: The Cancer Genome Atlas

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12885-021-08655-1.

Additional file 1: Table S1. Estimating relative abundance of 28 different immune cell types infiltration by the Single-Sample Gene-Set Enrichment Analysis (ssGSEA). **Table S2.** The differentially expressed genes between cluster 1 and cluster 2. **Table S3.** Gene Ontology (GO) enrichment analyses for differentially expressed genes up-regulated in cluster 1. **Table S4.** Gene Ontology (GO) enrichment analyses for differentially expressed genes up-regulated in cluster 2. **Table S5.** Regression coefficients of the genes in the prognostic model. **Table S6.** The differentially expressed miRNAs between high risk group and low risk group.

Additional file 2: Figure S1. The volcano plot of differentially expressed miRNAs between high risk group and low risk group.

Additional file 3: Figure S2. Survival analyses for the seven key genes between LUAD and adjacent normal samples. (a) CLEC3B, (b) TENM3, (c) IGF2BP1, (d) E2F7, (e) ANLN, (f) ANKRD18B, (g) FBN2.

Additional file 4: Figure S3. Paired sample expression analyses of the seven key genes. (a) CLEC3B, (b) TENM3, (c) IGF2BP1, (d) E2F7, (e) ANLN, (f) ANKRD18B, (g) FBN2.

Acknowledgments
Not applicable.

Authors' contributions
H.Z. (Hao Zhou) and M.Z. conducted the formal analysis and drafted the manuscript; M.S., J.W., and Z.H. performed the project administration; H.Z. (Hao Zhou) and H.Z. (Haijian Zhang) participated in software and data analysis; Z.H. (Hao Zhou) and Y.Z. contributed to writing, reviewing, and editing the article; and J.S. revised the manuscript. All authors read and approved the final manuscript.

Availability of data and materials
The datasets downloaded for supporting the results of this article are publicly available at the TCGA (https://portal.gdc.cancer.gov/) and GEO (https://www.ncbi.nlm.nih.gov/geo/).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no conflicts of interest.

Author details
[1]Department of Thoracic Surgery, Affiliated Hospital of Nantong University and Medical School of Nantong University, Nantong 226001, Jiangsu, China. [2]Department of Pathology, Affiliated Hospital of Nantong University and Medical School of Nantong University, Nantong 226001, Jiangsu, China. [3]Medical School of Nantong University, Nantong 226001, Jiangsu, China. [4]Research Center of Clinical Medicine, Affiliated Hospital of Nantong University, Nantong 226001, Jiangsu, China.

References
1. Cassim S, Chepulis L, Keenan R, Kidd J, Firth M, et al. Patient and carer perceived barriers to early presentation and diagnosis of lung cancer: a systematic review. BMC Cancer. 2019;19(1):25.
2. Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature. 2018;553(7689):446–54. https://doi.org/10.1038/nature25183.
3. Saad MI, Alhayyani S, McLeod L, Yu L, Alanazi M, et al. ADAM17 selectively activates the IL-6 trans-signaling/ERK MAPK axis in KRAS-addicted lung cancer. EMBO Mol Med. 2019;11(4):e9976.
4. Dolly SO, Collins DC, Sundar R, Popat S, Yap TA. Advances in the development of molecularly targeted agents in non-small-cell lung cancer. Drugs. 2017;77(8):813–27. https://doi.org/10.1007/s40265-017-0732-2.
5. Zhang P, He Q, Lei Y, Li Y, Wen X, et al. m (6) A-mediated ZNF750 repression facilitates nasopharyngeal carcinoma progression. Cell Death Dis. 2018;9(12):1169.
6. Baquero-Perez B, Antanaviciute A, Yonchev ID, Carr IM, Wilson SA, et al. The Tudor SND1 protein is an m (6) A RNA reader essential for replication of Kaposi's sarcoma-associated herpesvirus. Elife. 2019;8:e47261.
7. Fang J, Hu M, Sun Y, Zhou S, Li H. Expression profile analysis of m6A RNA methylation regulators indicates they are immune signature associated and can predict survival in kidney renal cell carcinoma. DNA Cell Biol. 2020;39(12):2194–211. https://doi.org/10.1089/dna.2020.5767.
8. Dong Z, Cui H. The emerging roles of RNA modifications in glioblastoma. Cancers (Basel). 2020;12(3):736.
9. Do HTT, Lee CH, Cho J. Chemokines and their receptors: multifaceted roles in cancer progression and potential value as cancer prognostic markers. Cancers (Basel). 2020;12(2):287.
10. Hanahan D, Coussens LM. Accessories to the crime: functions of cells recruited to the tumor microenvironment. Cancer Cell. 2012;21(3):309–22. https://doi.org/10.1016/j.ccr.2012.02.022.
11. Zhang B, Wu Q, Li B, Wang D, Wang L, et al. m (6) A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer. Mol Cancer. 2020;19(1):53.
12. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26(12):1572–3. https://doi.org/10.1093/bioinformatics/btq170.
13. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4(1):2612. https://doi.org/10.1038/ncomms3612.
14. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462(7269):108–12. https://doi.org/10.1038/nature08460.
15. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7. https://doi.org/10.1089/omi.2011.0118.

Zhou *et al. BMC Cancer*   (2021) 21:938

Page 14 of 14

16. Gao J, Kwan PW, Shi D. Sparse kernel learning with LASSO and Bayesian inference algorithm. Neural Netw. 2010;23(2):257–64. https://doi.org/10.1016/j.neunet.2009.07.001.

17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43): 15545–50. https://doi.org/10.1073/pnas.0506580102.

18. Hirsch FR, Scagliotti GV, Mulshine JL, Kwon R, Curran WJ Jr, et al. Lung cancer: current therapies and new targeted treatments. Lancet. 2017; 389(10066):299–311. https://doi.org/10.1016/S0140-6736(16)30958-8.

19. Perez-Ramirez C, Canadas-Garre M, Robles AI, Molina MA, Faus-Dader MJ, et al. Liquid biopsy in early stage lung cancer. Transl Lung Cancer Res. 2016; 5(5):517–24. https://doi.org/10.21037/tlcr.2016.10.15.

20. Shi J, Hua X, Zhu B, Ravichandran S, Wang M, et al. Somatic genomics and clinical features of lung adenocarcinoma: a retrospective study. PLoS Med. 2016;13(12):e1002162.

21. Park CK, Cho HJ, Choi YD, Oh IJ, Kim YC. A phase II trial of osimertinib in the second-line treatment of non-small cell lung cancer with the EGFR T790M mutation, detected from circulating tumor DNA: LiquidLung-O-cohort 2. Cancer Res Treat. 2019;51(2):777–87. https://doi.org/10.4143/crt.2018.387.

22. Zhang F, Zhang YC, Liao JY, Yu Y, Zhou YF, et al. The subunit of RNA N6-methyladenosine methyltransferase OsFIP regulates early degeneration of microspores in rice. PLoS Genet. 2019;15(5):e1008120.

23. Pan Y, Ma P, Liu Y, Li W, Shu Y. Multiple functions of m (6) A RNA methylation in cancer. J Hematol Oncol. 2018;11(1):48.

24. Ding Y, Qi N, Wang K, Huang Y, Liao J, Wang H, et al. FTO facilitates lung adenocarcinoma cell progression by activating cell migration through mRNA demethylation. Onco Targets Ther. 2020;13:1461–70. https://doi.org/10.2147/OTT.S231914.

25. Du M, Zhang Y, Mao Y, Mou J, Zhao J, et al. MiR-33a suppresses proliferation of NSCLC cells via targeting METTL3 mRNA. Biochem Biophys Res Commun. 2017;482(4):582–9. https://doi.org/10.1016/j.bbrc.2016.11.077.

26. Zeng D, Li M, Zhou R, Zhang J, Sun H, et al. Tumor microenvironment characterization in gastric cancer identifies prognostic and immunotherapeutically relevant gene signatures. Cancer Immunol Res. 2019;7(5):737–50. https://doi.org/10.1158/2326-6066.CIR-18-0436.

27. Wang D, DuBois RN. Immunosuppression associated with chronic inflammation in the tumor microenvironment. Carcinogenesis. 2015;36(10): 1085–93. https://doi.org/10.1093/carcin/bgv123.

28. Tanisawa K, Arai Y, Hirose N, Shimokata H, Yamada Y, et al. Exome-wide association study identifies CLEC3B missense variant p.S106G as being associated with extreme longevity in east Asian populations. J Gerontol A Biol Sci Med Sci. 2017;72(3):309–18.

29. Dai W, Wang Y, Yang T, Wang J, Wu W, et al. Downregulation of exosomal CLEC3B in hepatocellular carcinoma promotes metastasis and angiogenesis via AMPK and VEGF signals. Cell Commun Signal. 2019;17(1):113.

30. Cao R, Wu Q, Li Q, Yao M, Zhou H. A 3-mRNA-based prognostic signature of survival in oral squamous cell carcinoma. PeerJ. 2019;7:e7360. https://doi.org/10.7717/peerj.7360.

31. Sun J, Xie T, Jamal M, Tu Z, Li X, Wu Y, et al. CLEC3B as a potential diagnostic and prognostic biomarker in lung cancer and association with the immune microenvironment. Cancer Cell Int. 2020;20(1):106. https://doi.org/10.1186/s12935-020-01183-1.

32. Ziegler A, Corvalan A, Roa I, Branes JA, Wollscheid B. Teneurin protein family: an emerging role in human tumorigenesis and drug resistance. Cancer Lett. 2012;326(1):1–7. https://doi.org/10.1016/j.canlet.2012.07.021.

33. Li XC, Wang MY, Yang M, Dai HJ, Zhang BF, et al. A mutational signature associated with alcohol consumption and prognostically significantly mutated driver genes in esophageal squamous cell carcinoma. Ann Oncol. 2018;29(4):938–44. https://doi.org/10.1093/annonc/mdy011.

34. Li Y, Ge D, Gu J, Xu F, Zhu Q, et al. A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies. BMC Cancer. 2019;19(1):886.

35. Bell JL, Wachter K, Muhleck B, Pazaitis N, Kohn M, et al. Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? Cell Mol Life Sci. 2013;70(15):2657–75. https://doi.org/10.1007/s00018-012-1186-z.

36. Busch B, Bley N, Muller S, Glass M, Misiak D, et al. The oncogenic triangle of HMGA2, LIN28B and IGF2BP1 antagonizes tumor-suppressive actions of the let-7 family. Nucleic Acids Res. 2016;44(8):3845–64. https://doi.org/10.1093/nar/gkw099.

37. Shi R, Yu X, Wang Y, Sun J, Sun Q, et al. Expression profile, clinical significance, and biological function of insulin-like growth factor 2 messenger RNA-binding proteins in non-small cell lung cancer. Tumour Biol. 2017;39(4):1010428317695928.

38. Kato T, Hayama S, Yamabuki T, Ishikawa N, Miyamoto M, et al. Increased expression of insulin-like growth factor-II messenger RNA-binding protein 1 is associated with tumor progression in patients with lung cancer. Clin Cancer Res. 2007;13(2 Pt 1):434–42. https://doi.org/10.1158/1078-0432.CCR-06-1297.

39. Evangelou K, Havaki S, Kotsinas A. E2F transcription factors and digestive system malignancies: how much do we know? World J Gastroenterol. 2014; 20(29):10212–6. https://doi.org/10.3748/wjg.v20.i29.10212.

40. Salvatori B, Iosue I, Mangiavacchi A, Loddo G, Padula F, et al. The microRNA-26a target E2F7 sustains cell proliferation and inhibits monocytic differentiation of acute myeloid leukemia cells. Cell Death Dis. 2012;3:e413. https://doi.org/10.1038/cddis.2012.151.

41. Endo-Munoz L, Dahler A, Teakle N, Rickwood D, Hazar-Rethinam M, et al. E2F7 can regulate proliferation, differentiation, and apoptotic responses in human keratinocytes: implications for cutaneous squamous cell carcinoma formation. Cancer Res. 2009;69(5):1800–8. https://doi.org/10.1158/0008-5472.CAN-08-2725.

42. Li Q, Qiu XM, Li QH, Wang XY, Li L, et al. MicroRNA-424 may function as a tumor suppressor in endometrial carcinoma cells by targeting E2F7. Oncol Rep. 2015;33(5):2354–60. https://doi.org/10.3892/or.2015.3812.

43. Tuan NM, Lee CH. Role of anillin in tumour: from a prognostic biomarker to a novel target. Cancers (Basel). 2020;12(6):1600.

44. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015; 347(6220):1260419.

45. Suzuki C, Daigo Y, Ishikawa N, Kato T, Hayama S, et al. ANLN plays a critical role in human lung carcinogenesis through the activation of RHOA and by involvement in the phosphoinositide 3-kinase/AKT pathway. Cancer Res. 2005;65(24):11314–25. https://doi.org/10.1158/0008-5472.CAN-05-1507.

46. Wang A, Dai H, Gong Y, Zhang C, Shu J, et al. ANLN-induced EZH2 upregulation promotes pancreatic cancer progression by mediating miR-218-5p/LASP1 signaling axis. J Exp Clin Cancer Res. 2019;38(1):347.

47. Xu J, Zheng H, Yuan S, Zhou B, Zhao W, et al. Overexpression of ANLN in lung adenocarcinoma is associated with metastasis. Thorac Cancer. 2019; 10(8):1702–9. https://doi.org/10.1111/1759-7714.13135.

48. Liu WB, Han F, Jiang X, Yin L, Chen HQ, et al. Epigenetic regulation of ANKRD18B in lung cancer. Mol Carcinog. 2015;54(4):312–21. https://doi.org/10.1002/mc.22101.

Publisher's Note