



OPEN

## A 13-gene signature to predict the prognosis and immunotherapy responses of lung squamous cell carcinoma

Qin Yang<sup>1,2,5</sup>, Han Gong<sup>1,5</sup>, Jing Liu<sup>1,4</sup>, Mao Ye<sup>3</sup>, Wen Zou<sup>1✉</sup> & Hui Li<sup>1,3✉</sup>

Lung squamous cell carcinoma (LUSC) comprises 20–30% of all lung cancers. Immunotherapy has significantly improved the prognosis of LUSC patients; however, only a small subset of patients responds to the treatment. Therefore, we aimed to develop a novel multi-gene signature associated with the immune phenotype of the tumor microenvironment for LUSC prognosis prediction. We stratified the LUSC patients from The Cancer Genome Atlas dataset into hot and cold tumor according to a combination of infiltration status of immune cells and PD-L1 expression level. Kaplan–Meier analysis showed that hot tumors were associated with shorter overall survival (OS). Enrichment analyses of differentially expressed genes (DEGs) between the hot and cold tumors suggested that hot tumors potentially have a higher immune response ratio to immunotherapy than cold tumors. Subsequently, hub genes based on the DEGs were identified and protein–protein interactions were constructed. Finally, we established an immune-related 13-gene signature based on the hub genes using the least absolute shrinkage and selection operator feature selection and multivariate cox regression analysis. This gene signature divided LUSC patients into high-risk and low-risk groups and the former inclined worse OS than the latter. Multivariate cox proportional hazard regression analysis showed that the risk model constructed by the 13 prognostic genes was an independent risk factor for prognosis. Receiver operating characteristic curve analysis showed a moderate predictive accuracy for 1-, 3- and 5-year OS. The 13-gene signature also performed well in four external cohorts (three LUSC and one melanoma cohorts) from Gene Expression Omnibus. Overall, in this study, we established a reliable immune-related 13-gene signature that can stratify and predict the prognosis of LUSC patients, which might serve clinical use of immunotherapy.

### Abbreviations

AUC	Area under the ROC curve
BP	Biological process
CC	Cellular component
DEGs	Differentially expressed genes
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
ICIs	Immune checkpoint inhibitors
ICMs	Immune checkpoint molecules
KEGG	Kyoto Encyclopedia of Genes
KRT	Keratin

<sup>1</sup>Department of Oncology, the Second Xiangya Hospital of Central South University, Molecular Biology Research Center and Center for Medical Genetics, School of Life Sciences, Central South University, Changsha 410000, Hunan, China. <sup>2</sup>School of Medical Technology, Shao Yang University, Shaoyang 422000, Hunan, China. <sup>3</sup>Molecular Science and Biomedicine Laboratory, State Key Laboratory for Chemo/Biosensing and Chemometrics, College of Biology, College of Chemistry and Chemical Engineering, Collaborative Innovation Center for Chemistry and Molecular Medicine, Hunan University, Changsha 410082, Hunan, China. <sup>4</sup>Hunan Province Key Laboratory of Basic and Applied Hematology, Central South University, Changsha 410011, China. <sup>5</sup>These authors contributed equally: Qin Yang and Han Gong. ✉email: zouwen29w@csu.edu.cn; lihuisience@163.com

LASSO	Least absolute shrinkage and selection operator
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MF	Molecular function
NES	Normalized enrichment scores
OS	Overall survival
PPI	Protein-protein interaction
ROC	Receiver operating characteristic
TCGA	The Cancer Genome Atlas
TME	Tumor microenvironment

Despite a recent decline in the incidence and death rate, lung cancer is still the world's leading cause of cancer death<sup>1</sup>. Over 80% of lung cancer patients are diagnosed with non-small cell lung cancer (NSCLC). Lung squamous cell carcinoma (LUSC) represents the second main NSCLC histotype<sup>2</sup> and is particularly challenging to treat because of the highly heterogeneous nature and wide range of mutations present<sup>3</sup>. Immune checkpoint molecules (ICMs) are key regulators in maintaining immune homeostasis. Modulating ICMs expression (such as upregulating PD-L1 expression) is a normal strategy for cancer cells to escape from host immunity<sup>4,5</sup>. Immune checkpoint inhibitors (ICIs) are blocking-antibodies targeted to ICMs. The ICIs-based immunotherapy has revolutionized the standard care of patients with LUSC, prolonged overall survival (OS) recently. But a large portion of patients still does not experience tumor shrinkage or extended survival<sup>2</sup>. The current major clinical determinants of LUSC prognosis are traditional AJCC/UICC-TNM stratification systems; however, various outcomes of LUSC patients with similar AJCC/UICC-TNM features indicate that new reliable prognostic markers with higher sensitivity and accuracy are in need. Reliable signatures can evaluate benefits from immunotherapies and conduct new biomarker-directed immunotherapies for LUSC patients; however, there are no definitive biomarkers for predicting the immunotherapy response of patients at present.

The tumor is not only an accumulation of neoplastic cells, but constitutes a tumor microenvironment (TME). The composition of the TME differs across patients with the same kind of cancer, which has been demonstrated to be a major determinant of tumor characteristics and patient outcomes<sup>6</sup>. Recently, some studies have showed that the immune fraction of the TME has prognostic value in cancer. For example, the classification of tumors based on their immune phenotype of TME is used to explain the clinical response to ICIs-based immunotherapy. Immunologically hot tumor with a higher level of immune infiltration is prone to benefit from immunotherapy, while patients with immunologically cold tumors are more likely to be resistant to ICIs-based immunotherapy<sup>7,8</sup>. PD-L1 expression on tumor and immune cells was used to predict the response of patients receiving PD-1-based immunotherapy<sup>9</sup>. Cytotoxic T cells and helper T cells are potential prognostic factors following resection in NSCLC patients<sup>10,11</sup>. With the remarkable results achieved in immunotherapy, insight knowledge of the current guidelines for tumor classification, prognostic marker and subsequent treatment by analyzing TME composition has become a pressing necessity<sup>12</sup>.

This study aims to establish a prognostic multi-gene signature associated with the immune phenotype of TME for LUSC patients. We divide LUSC patients from The Cancer Genome Atlas (TCGA) dataset into two groups, hot and cold tumors, based on the immune infiltrate scores and PD-L1 expression level. To explore the underlying mechanism, we analyze the differential expression genes (DEGs) between the hot and cold tumors using enrichment analyses. Moreover, we identify hub genes based on the DEGs and constructed protein-protein interaction (PPI) network in hot and cold tumors. After that, a 13-gene signature based on the hub genes is developed and then, validated in multiple independent datasets across different platforms (TCGA and Gene Expression Omnibus (GEO)) and cancers (LUSC and melanoma). Overall, we are the first to specifically classify LUSC patients according to a combination of infiltration status of immune cells and PD-L1 expression. This study provides a reliable immune-related 13-gene signature that has significant implications for the prediction of outcomes for LUSC patients, which might facilitate the clinical use of immunotherapy.

## Materials and methods

**RNA sequencing data acquisition.** Gene expression profile and corresponding clinical data of TCGA-LUSC were obtained from the “TCGA TARGET GTEx” cohort of UCSC XENA (<http://xena.ucsc.edu/>) and the Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>). The TCGA-LUSC dataset included 504 samples and the four databases (GSE30219, GSE12472, GSE157011 and GSE78220) from GEO included 606 samples. The clinical information of the TCGA-LUSC dataset was shown in Table 1.

The data collection and processing of this research complied with data policy of TCGA and GEO to protect human subjects. All experiments were performed in accordance with relevant guidelines and regulations.

**Identification of hot and cold tumors.** ImmuneCellAI (<http://bioinfo.life.hust.edu.cn/web/ImmucellAI/>) was used to estimate the abundance of 24 immune cells and infiltration scores from gene expression dataset<sup>13</sup>. The Immune infiltration level was used to combine with the PD-L1 expression level for dividing LUSC tumors into two groups: hot and cold tumors<sup>14</sup>. The hot tumors were of both the top 50% immune infiltration and the top 50% PD-L1 expression levels, while the others were defined as cold tumors. The Kaplan–Meier curve was performed for OS analysis of the hot versus cold tumors. OS was the length of time from the date of diagnosis or the start of treatment for a disease, such as cancer, that patients diagnosed with the disease were still alive.

Furthermore, our definition of the hot and cold tumors was validated by the expression of ICMs, ESTIMATE and CIBERSORT analyses. ICMs such as PD-L1 were used to predict clinical outcomes with ICIs-based immunotherapy<sup>15</sup>. ESTIMATE was used to calculate the immune score, stromal score, ESTIMATE score and

Clinical features	Count (%) (n = 504)
<b>Status</b>	
Alive	284 (56.3)
Dead	220 (43.7)
<b>Gender</b>	
Female	131 (26.0)
Male	373 (74.0)
<b>Age</b>	
≤ 65	190 (37.7)
> 65	305 (60.5)
Unknown	9 (1.8)
<b>Stage T</b>	
T1	114 (22.6)
T2	295 (58.5)
T3	71 (14.1)
T4	24 (4.8)
<b>Stage M</b>	
M0	414 (82.1)
M1	7 (1.4)
Unknown	83 (16.5)
<b>Stage N</b>	
N0	320 (63.5)
N1	133 (26.4)
N2	40 (7.9)
N3	5 (1.0)
Unknown	6 (1.2)
<b>Stage</b>	
Stage I	246 (48.8)
Stage II	165 (32.7)
Stage III	85 (16.9)
Stage IV	7 (1.4)
Unknown	1 (0.2)
<b>Neoplasm cancer status</b>	
Bronchial	10 (2.0)
L-lower	77 (15.3)
L-upper	137 (27.2)
R-lower	109 (21.6)
R-middle	18 (3.6)
R-upper	136 (27.0)
Unknown	17 (3.4)
<b>Smoking history</b>	
Current reformed smoker for < or = 15 years	252 (50.0)
Current reformed smoker for > 15 years	83 (16.5)
Current reformed smoker, duration not specified	5 (1.0)
Current smoker	134 (26.6)
Lifelong non-smoker	18 (3.6)
Unknown	12 (2.4)

**Table 1.** Clinical features of patients with lung squamous cell carcinoma in TCGA.

tumor purity<sup>16</sup>. Stromal and immune scores were used to predict the level of infiltrating stromal and immune cells. ESTIMATE score based on the immune score and stromal scores were further used to infer tumor purity in tumor tissue<sup>16</sup>. CIBERSORT algorithm was performed to calculate the abundance of 22 types of immune cell subsets in each sample<sup>17</sup>.

**Identification of differentially expressed genes (DEGs) and enrichment analysis.** The R package “DESeq2”<sup>18</sup> was applied to identify DEGs (p.adjust value < 0.05 and |log2FC| > 1) between the hot and cold tumors. Afterward, the DEGs were used to perform Gene Ontology (GO) and Kyoto Encyclopedia of Genes

(KEGG) analyses by using the R package “clusterProfiler”<sup>19</sup>. All genes were arranged into a ranked list according to the fold change and then used to conduct a GSEA analysis. For the target set of GSEA analysis requirements, we obtained hallmark gene sets (h.all.v7.4.entrez.gmt) from Molecular Signatures Database (MSigDB). Pathways with  $p$ -adjust value  $< 0.05$  were selected.

**Protein protein interaction (PPI) network and hub genes identification.** To further investigate the interactions between the DEGs, a PPI network was constructed using the STRING (<https://string-db.org/>) database and interaction scores  $> 0.4$  were considered statistically significant. The MCODE was a tool to detect densely connected regions in large protein–protein interaction networks that might represent molecular complexes<sup>20</sup>, which was used to extract key sub-networks and identify hub genes in the network. Enrichment analyses were performed on the hub genes. Subsequently, the CytoHubba<sup>21</sup> was used to obtain the top 10 hub genes in hot and cold tumors, respectively. All parameters were default values and all the above networks were visualized in Cytoscape software (v3.8.2).

**Screening of prognostic multi-gene signature.** The univariate Cox regression analysis was used to obtain prognostic genes in the hub genes. The prognostic 13-gene signature was established by the multivariate cox analysis and the least absolute shrinkage and selection operator (LASSO) analysis<sup>22</sup>. We evaluated the risk score of each patient by the below formula:

$$\text{Risk score} = \text{Coef}_{\text{gene1}} \times \text{Exp}_{\text{gene1}} + \text{Coef}_{\text{gene2}} \times \text{Exp}_{\text{gene2}} + \dots + \text{Coef}_{\text{gene13}} \times \text{Exp}_{\text{gene13}}$$

The Kaplan–Meier survival curve was used to validate the prognostic value of the 13-gene signature. Receiver operating characteristic (ROC) curve analysis was used to estimate the sensitivity and specificity of the 13-gene signature. Moreover, to demonstrate that the risk score was an independent prognostic factor, we conducted univariate and multivariate Cox regression analyses to examine the prognostic value of the risk score and other clinical indicators in the TCGA-LUSC patients.

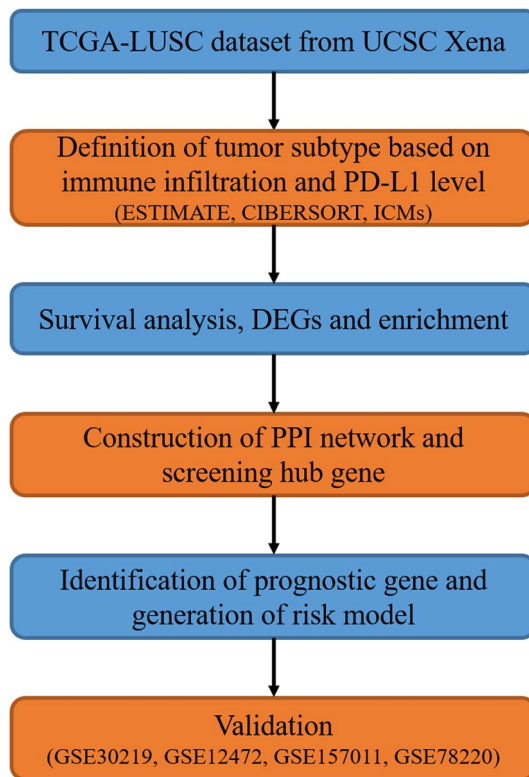
**Statistical analysis.** The Wilcoxon test was used to compare the differences in the proportion of immune cells, ICMs expression levels and ESTIMATE scores between the hot and cold tumors. For the OS analysis, the Kaplan–Meier curve and the two-sided log-rank test were performed. R package v4.0.2 was performed for all analyses and  $p < 0.05$  was considered as statistical significance.

## Results

**Stratification of hot and cold tumors.** To construct a multi-gene signature for predicting OS of LUSC patients, we designed and processed our study as shown in the flow chart (Fig. 1). Hot tumors are supposed to have a relatively higher immune infiltration and are thus more likely to respond to immunotherapy compared with cold tumors<sup>7,8</sup>. PD-L1 is a co-inhibitory ICM that contributes to the immune escape of cancer cells<sup>5</sup>. LUSC patients with an upregulated PD-L1 expression are more likely to benefit from immunotherapy<sup>23</sup>. Five hundred and four LUSC-TCGA tumor samples (Table 1) were divided into two groups: hot and cold tumors, when a combination of immune infiltration scores and the PD-L1 expression level was used as a cutoff. Tumors responding to immunotherapy had the top 50% immune infiltrates scores and the top 50% PD-L1 expression were referred to as ‘hot tumors’, whereas the rest tumors were ‘cold tumors’. Kaplan–Meier analysis demonstrated that patients with hot tumors had a significantly shorter OS than patients with cold tumors (Fig. 2a). We subsequently compared immune cell infiltration levels between the hot and cold tumors using ESTIMATE (Fig. 2b), the expression level of co-stimulatory ICMs and CIBERSORT (Fig. 2c,d). Most of the co-stimulatory ICMs were upregulated in hot tumors (Fig. 2c). Immune cells were more infiltrated in hot tumors than cold tumors (Fig. 2b–d). These results together indicated that hot tumors were more likely to respond to immunotherapy than cold tumors.

**Enrichment analyses of the DEGs.** 1203 DEGs (including 564 upregulated DEGs and 639 downregulated DEGs) were identified in the hot compared with cold tumors (Fig. 3a). To gain a functional understanding of the DEGs, we conducted GO (Fig. 3b) and KEGG (Fig. 3c) analyses on the 1203 DEGs. We also performed a GSEA analysis (Fig. 3d) based on the rank information of all genes. The GO biological process (BP) mainly included the proliferation and regulation of multi-immune cells (Fig. 3b). The most abundant GO molecular function (MF) was immune receptor activity (Fig. 3b). GO cellular component (CC) was enriched in ‘neuroactive ligand-receptor interaction’ and ‘metabolism of xenobiotics by cytochrome P450’ (Fig. 3b). Pathway enrichment analysis regarding KEGG focused on ‘cytokine-cytokine receptor interaction’, ‘cytokine signaling pathway’, ‘antigen processing and presentation’ and ‘natural killer cell-mediated cytotoxicity’ (Fig. 3c). The GSEA was enriched in the hallmark gene sets of ‘interferon  $\alpha/\gamma$  response’, ‘inflammatory response’, ‘IL5-STAT5 signaling’, ‘IL6-STAT3 signaling’ and ‘TNF- $\alpha$  signaling via NF-KB’ (Fig. 3d).

**Hub genes and protein–protein interactions (PPIs).** To further explore the functions of DEGs, we conducted a KEGG analysis in hot and cold tumors, respectively. When compared with the cold tumor, the upregulated DEGs in the hot tumors were mainly enriched in immune-related pathways (Fig. 4a) and the downregulated DEGs in the hot tumors were mainly enriched in metabolic pathways (Fig. 4b). To further explore the interaction between the DEGs, we constructed PPI networks by STRING in hot and cold tumors, respectively. Afterward, 337 hub genes based on the 1203 DEGs were identified through MCODE in Cytoscape software. The hub genes have relatively higher intro module connectivity and gene significance than the other genes and play key roles in pathways in the co-expression network. The top 10 hub genes in hot and cold tumors were filtered

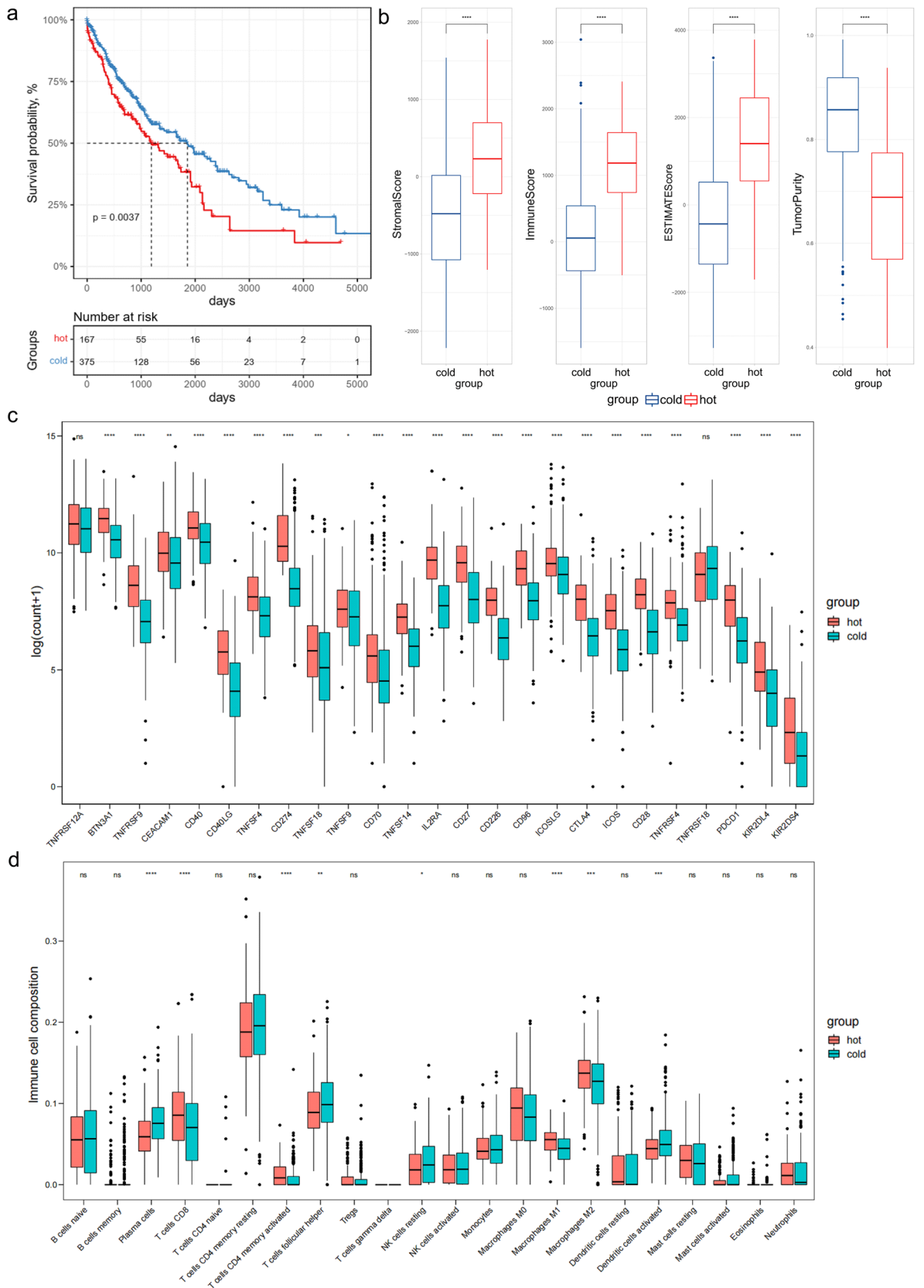


**Figure 1.** Flow chart of this study. The flow chart shows the strategy for developing and validating an immune-related 13-gene signature. This signature is constructed to predict immunotherapy responses and the overall survival of lung squamous cell carcinoma.

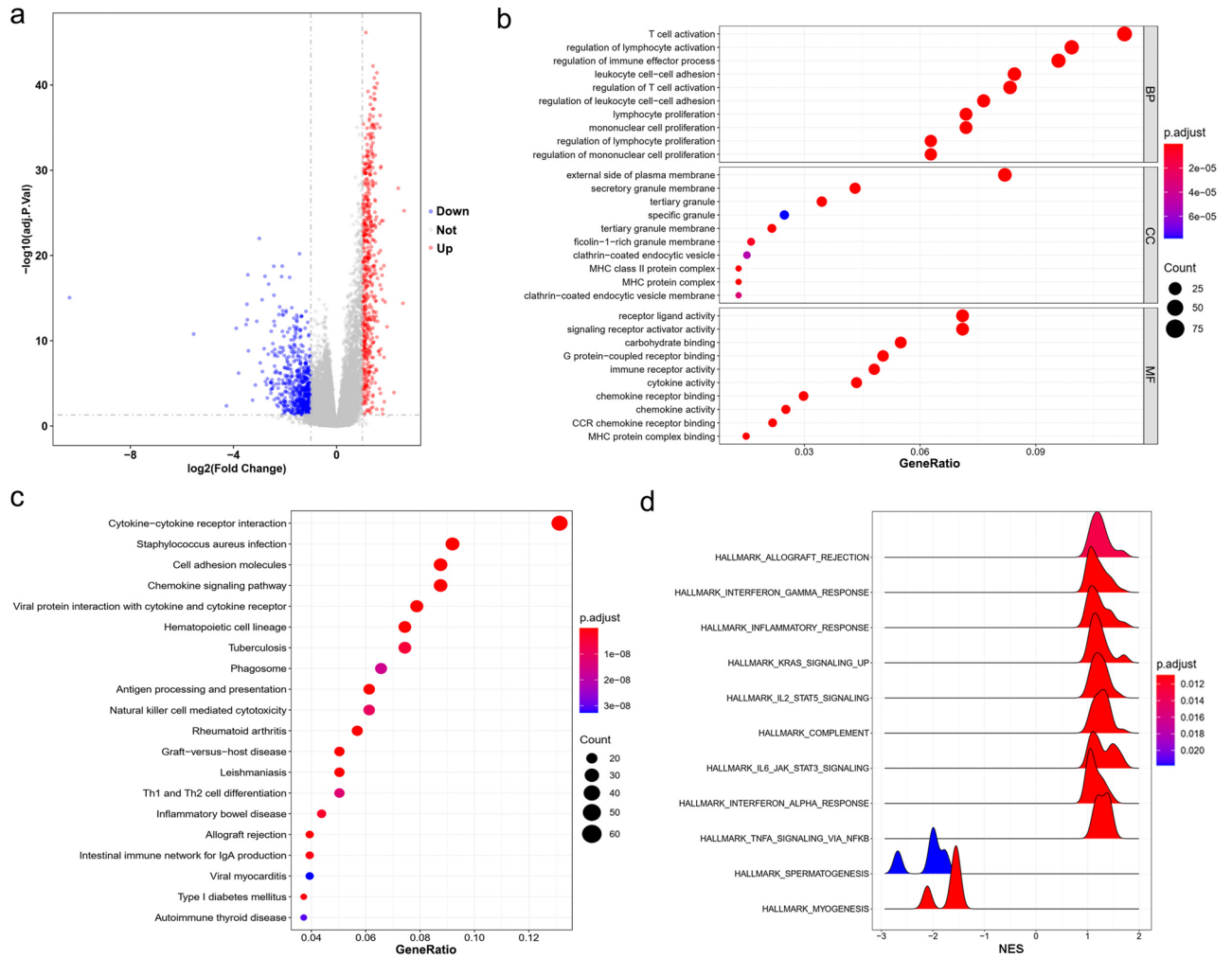
into the PPI network (Fig. 4c,d). The common feature of the top hub genes (for instance, IL17A, CD28, CD80 and CD40LG) in hot tumors was that they were involved in the immune activation process directly or indirectly (Fig. 4c). The top hub genes in cold tumors were keratin (KRT) family members, which were not so closely related to immune responses (Fig. 4d).

**Identification of an immune-related 13-genes signature.** To estimate the value of the 337 hub genes in predicting OS in LUSC, the TCGA-LUSC dataset was used as a training cohort. The univariate Cox regression analysis was used to obtain prognostic genes in the above hub genes. Subsequently, the LASSO and multivariate cox regression analyses were performed to identify prognostic genes with the strongest predicting ability in the training cohort (Fig. 5a,b). Finally, 13 prognostic genes (risk model) were identified (Fig. 5c) and the risk score was calculated by the following formula: risk score =  $(0.11865 \times \text{FGF4 expression}) + (0.06922 \times \text{FGL1 expression}) + (-0.13624 \times \text{LIM2 expression}) + (-0.08743 \times \text{NPY expression}) + (0.13426 \times \text{F13A1 expression}) + (-0.06918 \times \text{CDH12 expression}) + (-0.15824 \times \text{CD1E expression}) + (-0.06260 \times \text{OTX2 expression}) + (0.06185 \times \text{ADRA1D expression}) + (0.23177 \times \text{SAMMD9L expression}) + (0.07060 \times \text{ZFP42 expression}) + (0.07751 \times \text{GAGE2A expression}) + (-0.16373 \times \text{KLRC2 expression})$ . Functions of the 13 prognostic genes were showed in Supplementary Table S1.

By the median risk score, LUSC patients were divided into high- and low-risk groups and Kaplan–Meier curve showed that poor OS outcomes of LUSC patients were associated with the high-risk scores (Fig. 6a). According to the risk scores, the LUSC patients were ranked from left to right shown in the upper panel of Fig. 6b. The risk scores increased from left to right. OS distribution of each patient was shown in the lower panel of Fig. 6b where LUSC patients were ranked from left to right according to risk scores. A ROC curve was constructed to analyze the diagnostic accuracy of the 13-gene signature. It revealed that the 13-gene signature could serve as valuable biomarker for distinguishing between LUSC and control subjects (for 1-year, areas under the curve (AUCs) = 0.70; for 3-year, AUC = 0.76; for 5-year, AUC = 0.76) (Fig. 6c). To determine if the 13-gene signature was an independent prognostic marker, univariate and multivariate cox regression analyses were performed on the TCGA-LUSC dataset. The risk score of the 13-gene signature and other clinico-pathological factors, including gender, age, neoplasm cancer status, stage and smoking status were used as covariates in the cox regression analysis. We found a significant association between the 13-gene signature and OS in the TCGA dataset (HR = 1.5893,  $p < 0.0001$ ). Our results showed that this 13-gene signature was an independent risk factor for predicting the OS of LUSC patients (Fig. 6d). The detailed univariate and multivariate cox analyses of the 13-gene signature and other clinico-pathological factors were showed in Table 2. External four GEO-LUSC datasets (GSE30219, GSE12472, GSE157011 and GSE78220) were utilized to validate the prediction power of the 13-gene signature,



**Figure 2.** Stratification of hot and cold tumors. **(a)** The Kaplan–Meier curve shows the correlation between tumor stratification and overall survival. The p value is based on the log-rank test. **(b)** Immune score, stromal score, ESTIMATE score and tumor purity were obtained from ESTIMATE analysis in hot and cold tumors. **(c)** Expression of immune checkpoint molecules in hot and cold tumors. **(d)** Boxplot shows 22 immune cells infiltration obtained from CIBERSORT analysis in hot and cold tumors. \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ , \*\*\*\* $p < 0.0001$ , hot hot tumors, cold cold tumors.

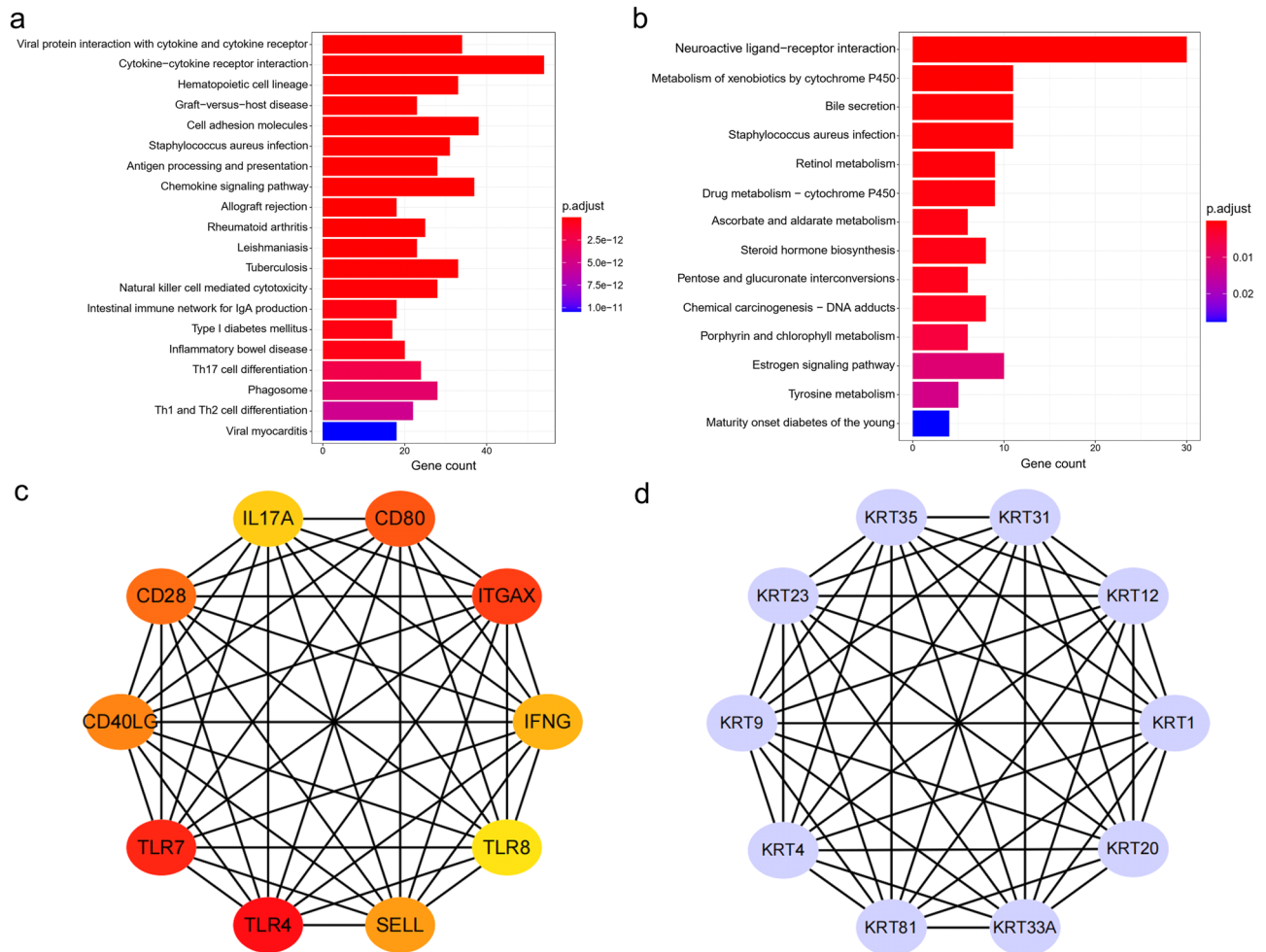


**Figure 3.** Enrichment analyses of differentially expressed genes (DEGs) between hot and cold tumors. (a) Volcano plot of the DEGs in the hot and cold tumors. Red plots (UP) represent significantly upregulated DEGs. Blue plots (Down) represent significantly downregulated DEGs. Grey plots (Not) represent DEGs with no significant difference. The vertical dashed line represents  $|\log_2\text{FC}| > 1$  (fold change > 2) and the horizontal dashed line represents  $p.\text{adjust value} = 0.05$ .  $p.\text{adjust value} < 0.05$  is considered significant. (b) Gene Ontology enrichment analysis of the top 10 enrichments for biological process, cellular component and molecular function. Enrichment levels are shown as a continuous variable from blue to red color. The bluer color represents the higher  $p.\text{adjust}$  value, and the redder color represents the lower  $p.\text{adjust}$  value. Enrichment counts are shown by dot sizes. The bigger size represents the higher enrichment count and the smaller size represents the lower enrichment count. (c) Kyoto Encyclopedia of Genes and Genomes analysis<sup>24–26</sup>. (d) Gene Set Enrichment Analysis is shown by normalized enrichment scores (NES) on X-axis and enriched pathways on Y-axis. The NES < 0 represents pathways enriched in cold tumors while the NES > 0 represents pathways enriched in hot tumors.

of which GSE78220 is a melanoma immunotherapy dataset. In line with the results in the training cohort (TCGA dataset), the Kaplan–Meier curve indicated that the risk scores could distinguish the patients well in the GEO datasets (Fig. 7); LUSC patients with low-risk scores demonstrated a significantly longer OS in the three validation cohorts. Based on these results, the 13-gene signature performed well in predicting OS of LUSC patients and could potentially guide the clinical management.

### Discussions

LUSC comprises about 20–30% of all lung cancers<sup>27</sup>. Its clinical outcome has been poor for the past decades, because of a limited treatment strategy. The situation has dramatically changed mainly with the clinical introduction of ICIs-based immunotherapy in recent years; however, it is quite clear that only a subgroup of LUSC patients achieves sustained benefits from ICIs-based immunotherapy. At present, various LUSC outcomes have been identified in patients with similar clinical and pathological features, suggesting that the current clinical prognostic factors used may be insufficient to consistently predict individual clinical outcomes. Predictive indicators are the most important to choose rational treatment. Identifying reliable prognostic markers with higher sensitivity and accuracy in LUSC is in urgent need. Molecular markers on tumor have been extensively

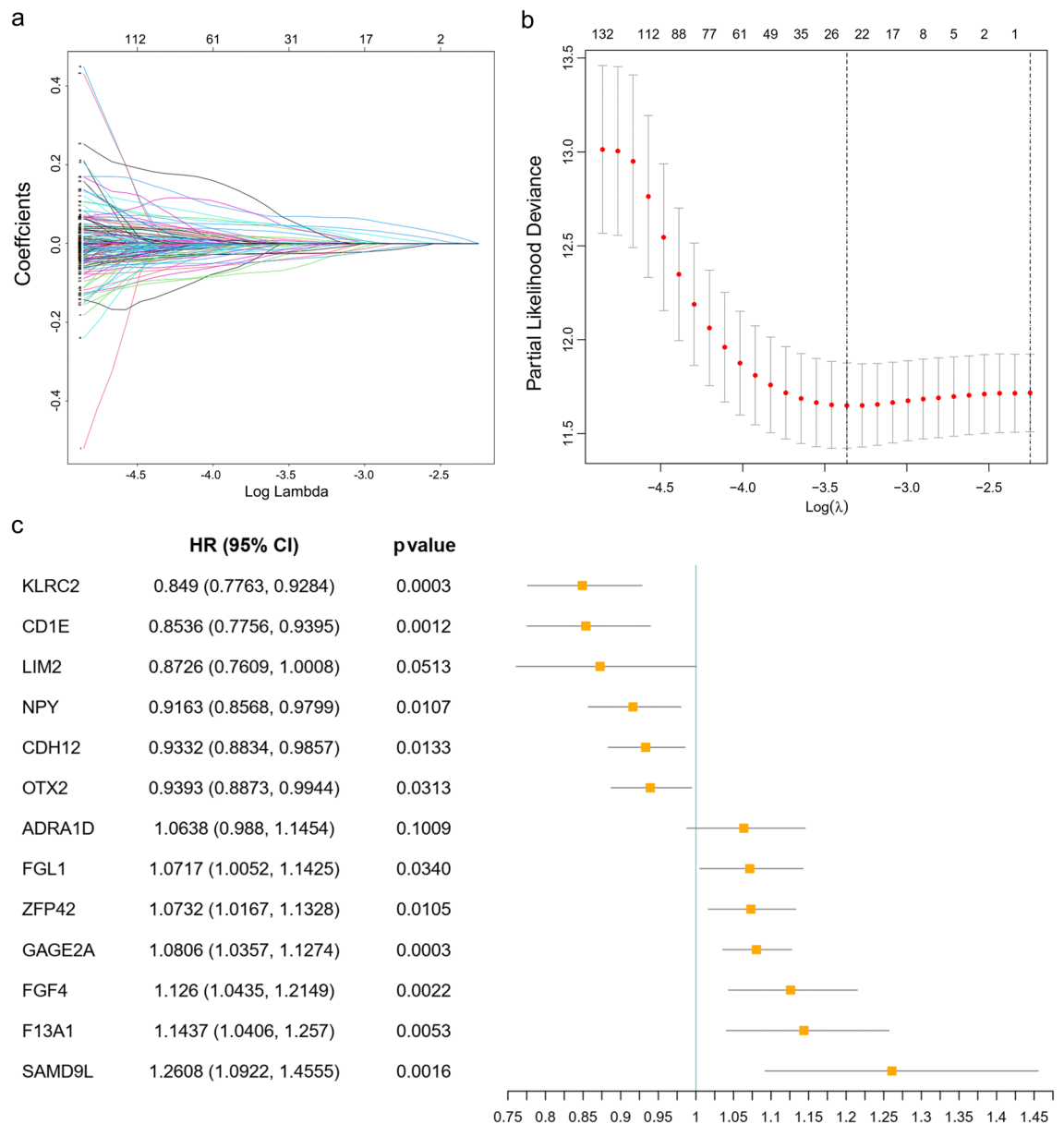


**Figure 4.** Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis<sup>24–26</sup> and protein–protein interaction (PPI) network. **(a)** KEGG analysis of the upregulated genes for hot tumors. Enrichment levels are shown as a continuous variable from blue to red color. The bluer color represents the higher p.adjust value, and the redder color represents the lower p.adjust value. P.adjust value < 0.05 is considered significant. **(b)** KEGG analysis of upregulated genes for cold tumors. **(c)** PPI network of top 10 hub genes obtained from CytoHubba analyses in hot tumors. Gene rankings are shown as colors. The redder color represents the higher gene ranking, and the yellow color represents the lower gene ranking. **(d)** PPI network of top 10 genes obtained from CytoHubba analysis in the cold tumors. The 10 genes are the same as the gene ranking.

investigated for prognosis and guidance of cancer therapy; much less for tumor-associated immune cells in TME. TME is an environment where tumors are considered as complex dynamic tissues with an important interplay of various cells including tumor-infiltrating immune cells, which is crucial for identifying effective biomarkers for predicting drug resistance and cancer progression<sup>28</sup>. Many studies have demonstrated that T cells are the major immune cells infiltrating tumors in TME and the degree of lymphocytic infiltration is positively associated with an absence of tumor metastases<sup>29</sup>. ICMs are indispensable for the full activation of T cells. PD-L1, as a major ICM, is expressed on the cell surface in tumor-associated immune cells and various cancer cells<sup>5</sup>. Although the expression of PD-L1 has been widely evaluated in ICI-based immunotherapies as a positive predictive marker, it is still an imperfect predictive biomarker<sup>2,30</sup>.

In our study, for the first time, LUSC patients distributed in hot and cold tumors were characterized by a combination of immune cell infiltration and PD-L1 expression associated with TME. Tumors responding to immunotherapy had a higher level of PD-L1 expression (top 50%) and immune infiltrates scores (top 50%) were referred to as ‘hot tumors’, whereas the rest of the tumors were ‘cold tumors’. Hot and cold tumors defined by our method predicted well in OS of LUSC patients. ESTIMATE, the expression level of co-stimulatory ICM, CIBERSORT and enrichment analyses all suggested that the hot tumors potentially had a higher immune response to immunotherapy than cold tumors, which further approved our stratification of tumors. Recently, this unofficial classification of tumors into two categories, ‘hot tumors’ and ‘cold tumors’, has been increasingly advocated. This immune-based, rather than tumor-based patient classification according to tumor immune infiltration, has shown a greater relative prognostic value than the traditional AJCC/UICC-TNM stratification system<sup>12,31,32</sup>. Different classifications of tumors represent various responses to immunotherapeutic options. Hot tumors are

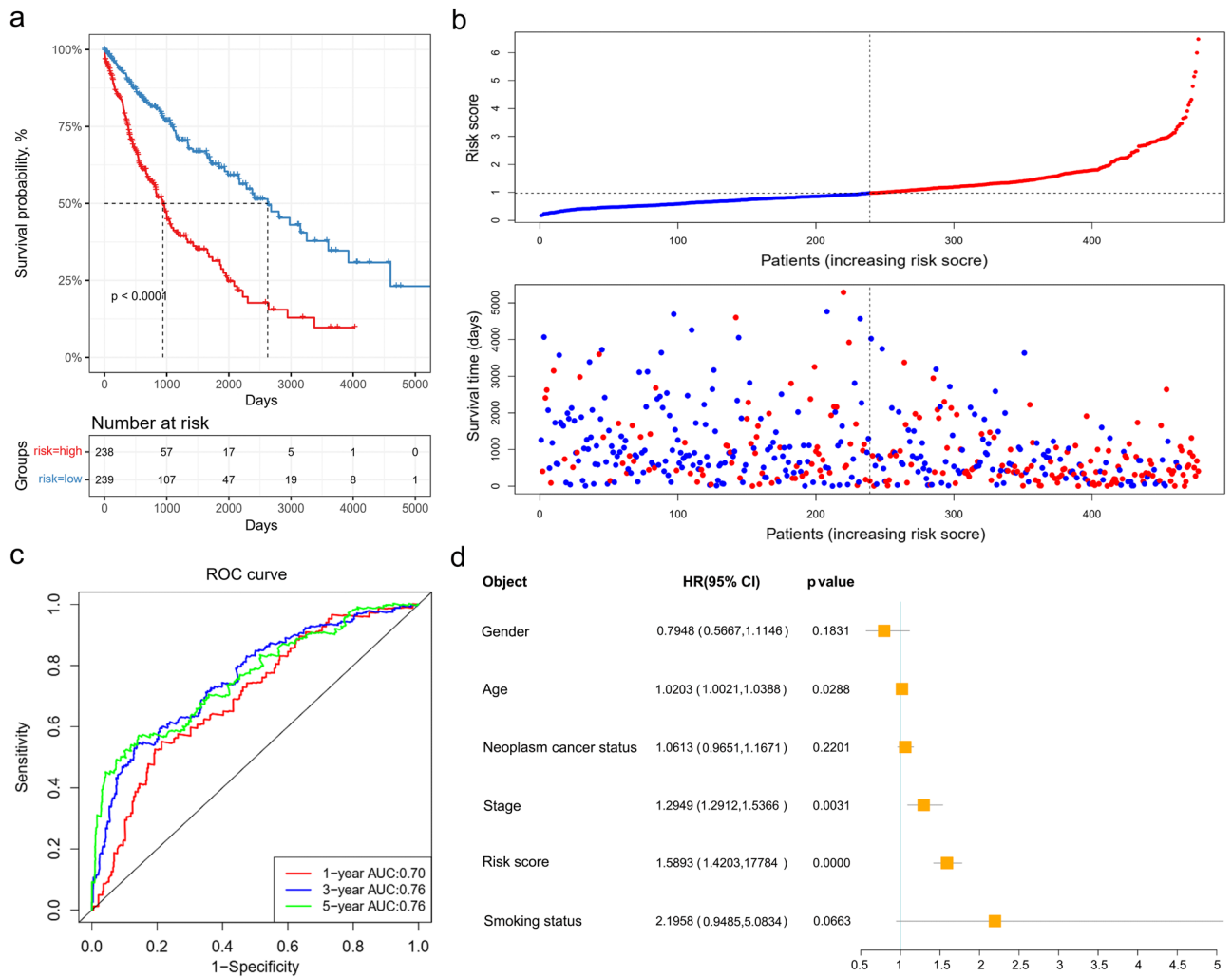




**Figure 5.** Identification of 13 prognostic genes in The Cancer Genome Atlas cohort. **(a,b)** Partial likelihood deviance coefficient profiles and Least absolute shrinkage and selection operator cox analysis of the hub genes. After 10 rounds of cross-validation, the optimal value of turning parameter is determined as 23. **(c)** Hazard ratios (HR) forest plot of the 13-gene signature constructed by the multivariate cox analysis. HR > 1 indicates that the gene is a risk factor. HR < 1 indicates that the gene is a protection factor.

more likely to benefit from immunotherapy. Our stratification of LUSC patients contribute to dichotomizing tumors and can ultimately contribute to converting cold tumors to hot tumor.

Hot tumors showed remarked differences in hub genes profile from the cold tumors. We found that the top hub genes of hot tumors comprised many immune-related genes (for instance, IL17A, CD28, CD80 and CD40LG). Co-stimulatory ICMs CD28, CD80 and CD40LG are secondary signal molecules in the T lymphocyte activation, which activate patients' anti-tumor immune responses, leading to increased efficacy of cancer immunotherapy<sup>5</sup>. CD28 is associated with an abundance of lymphocytes and longer OS in lung adenocarcinoma (LUAD)<sup>33</sup>. CD80 activates effector T cells via interacting with the receptors CD28 on the surface of the T cells. Upregulated CD80 predicts good prognosis in gastric adenocarcinoma<sup>34</sup> and oral squamous cell carcinoma<sup>35</sup>. Expression of CD40LG in the tumor-free lymph node is positively related to a good prognosis in oral squamous cell carcinoma<sup>35</sup>. We also observed that all top hub genes in cold tumors were keratin family members (for instance, KRT20, KRT12 and KRT4). Keratins are expressed in highly specific patterns correlated to the epithelial type and stage of cellular differentiation. Characteristic expression patterns of keratins are also observed in cancers<sup>36</sup>. Moreover, keratins are diagnostic and prognostic markers in epithelial cancers. For example, downregulated hub gene KRT20 indicates poor patient outcomes in colorectal cancer<sup>37</sup>, pancreatic adenocarcinomas<sup>38–40</sup>

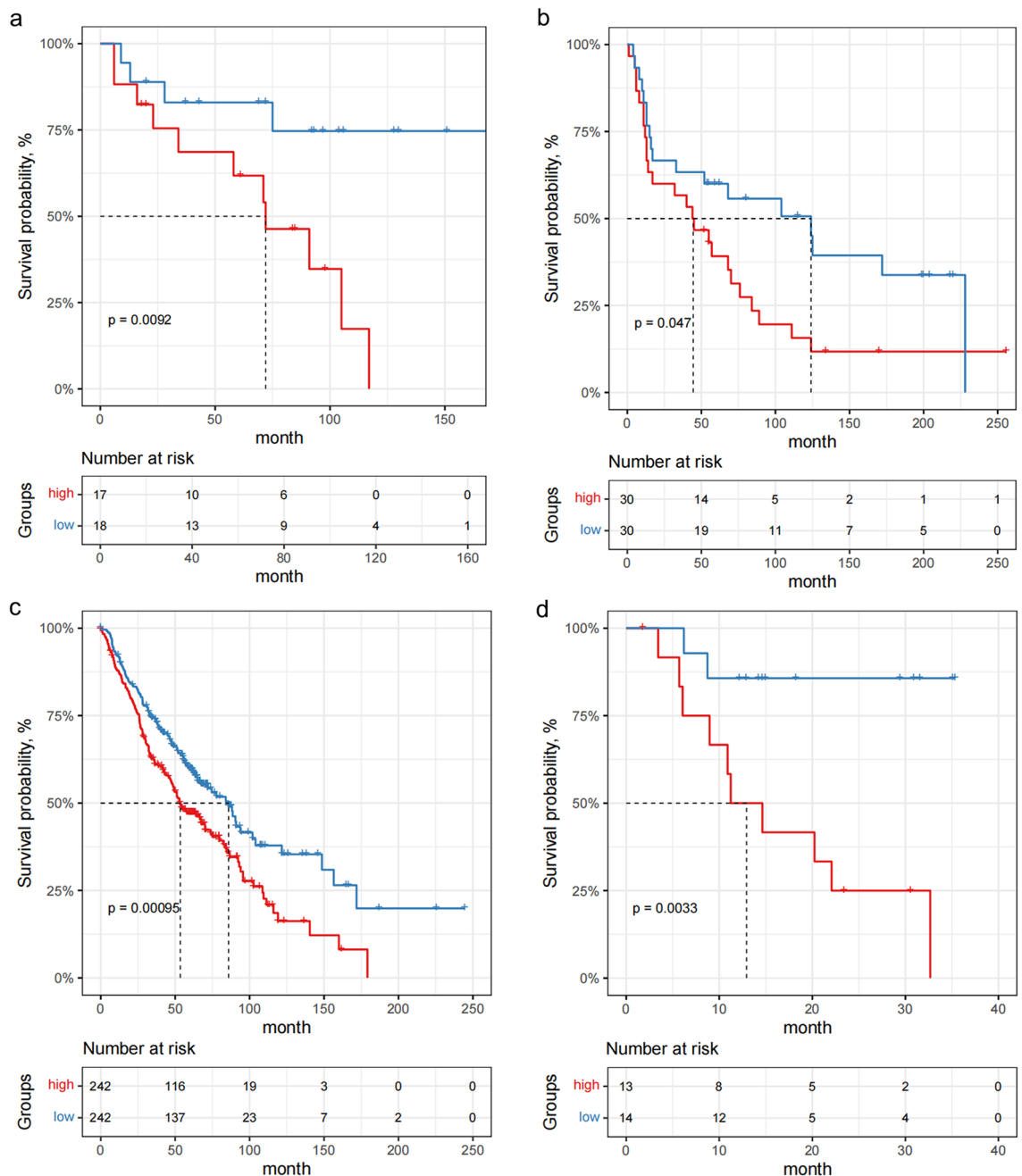


**Figure 6.** Validation of the 13-gene prognosis signature in The Cancer Genome Atlas cohort. (a) Kaplan–Meier curves of overall survival of the high- and low-risk groups. The p value is based on the log-rank test. (b) The distribution of risk scores, patient relapse status and survival time. In the upper panel, the red color represents the high-risk group and the blue color represents the low-risk group. In the lower panel, red dots represent dead patients and blue dots represent alive patients. (c) Receiver operating characteristic curve (ROC) for the 1-, 3-, and 5-year survival prediction using the 13-gene signature. AUC area under the ROC curve. (d) Multivariate cox analysis shows the hazard ratios of the 13-gene signature and other clinico-pathological factors with forest plots.

Variables	Univariate analysis			Multivariate analysis		
	HR	95% CI	p value	HR	95% CI	p value
Gender	0.824	0.824 (0.596, 1.140)	0.243	0.795	0.795 (0.567, 1.115)	0.183
Age	1.014	1.014 (0.997, 1.031)	0.097	1.020	1.020 (1.002, 1.039)	0.029
Neoplasm cancer status	1.082	1.082 (0.987, 1.187)	0.094	1.061	1.061 (0.965, 1.167)	0.220
Stage	1.297	1.297 (1.098, 1.53)	0.002	1.295	1.295 (1.091, 1.537)	0.003
Risk score	1.634	1.634 (1.467, 1.820)	3.96E-19	1.589	1.589 (1.420, 1.778)	6.67E-16
Smoking status	1.683	1.683 (0.743, 3.813)	0.212	2.196	2.196 (0.949, 5.083)	0.066

**Table 2.** Univariate and multivariate cox regression analyses of the prognosis-related factors.

and gastric cancer<sup>41</sup>. Soluble keratins in the circulation of NSCLC patients carry prognostic significance and are used to monitor tumor load and disease progression in clinical practice<sup>42,43</sup>. Cold tumors are the most challenging to eradicate and are invariably associated with a poor prognosis. Our results on the top hub genes in the cold tumors suggested a critical role of keratins in immunotherapeutic resistance. In line with this result, one widely accepted role of keratins is a protector of mechanical stability and epithelial cell integrity under a variety



**Figure 7.** Re-sampling analyses in Gene Expression Omnibus datasets (GSE30219, GSE12472, GSE157011 and GSE78220) implicate the reliability of the 13-gene signature for predicting overall survival of LUSC patients. **(a)** Kaplan–Meier survival curves for overall survival in the lung squamous cell carcinoma (LUSC) immunotherapy dataset GSE12472. The  $p$  value is based on the log-rank test. **(b)** Kaplan–Meier survival curves for overall survival in the LUSC immunotherapy dataset GSE31210. The  $p$  value is based on the log-rank test. **(c)** Kaplan–Meier survival curves for overall survival in the LUSC immunotherapy dataset GSE157011. The  $p$  value is based on the log-rank test. **(d)** Kaplan–Meier survival curves for overall survival in the melanoma immunotherapy dataset GSE78220. The  $p$  value is based on the log-rank test.

of stressful conditions including death receptor activation and drugs<sup>42,43</sup>. Further studies are required to identify the mechanisms of keratins explaining this possible decreased susceptibility and identifying prognostic markers in immunotherapy of LUSC.

A gene signature predicting the prognosis of a large cohort of cancer patients is of great significance, because the gene expression can capture the influence of the changes of multiple genes at the same time and summarize the prognosis of multiple ‘conventional’ risk factors into one risk score<sup>44,45</sup>. Although gene expressions currently are not involved in the standard diagnosis of LUSC, it is proved to be a comprehensive tool for predicting outcomes in many cancers. For instance, a 3-gene signature has been proved to be a comprehensive tool for leukemia diagnosis and classification due to its high accuracy in all clinically relevant leukemia sub-entity

predictions<sup>44</sup>. Zheng et al. have identified a 9-gene signature to predict OS in LUAD patients<sup>46</sup>. In this study, we applied multi-cox regression analysis and LASSO feature selection to screen a 13-gene signature among 337 hub genes. In TCGA and three GEO cohorts validation, the 13-gene signature significantly stratified patients into high- vs low-risk groups in terms of OS and remained as an independent prognostic factor in multivariate analysis. Among the 13 prognostic genes, CD1E, KLRC2 and GAGE2A are more relevant to tumor immunity. CD1E is an MHC class I-like molecule that presents antigens to T cells and thus regulates T cells participation in the immune response. CD1E can predict the efficacy of immunotherapy in patients with nonmuscle-invasive bladder cancer<sup>47</sup> and glioblastomas<sup>48</sup>. KLRC is expressed primarily in natural killer (NK) cells. Tumor infiltration of NK cells is correlated with the prolonged survival of cancer patients. Either acute exercise or in vitro expansion of KLRC+/NKG2A- NK cells can enhance the anti-tumor cytotoxicity of NK cells for immunotherapy<sup>49</sup>. In ovarian cancer, tumor-specific antigen GAGE2A can be used as an indicator for early diagnosis, efficacy evaluation and prognostic determination<sup>50</sup>.

In conclusion, for the first time, by dividing tumors into hot and cold tumors according to a combination of their immune infiltration and PD-L1 expression, this study proposed an immune-based rather than a tumor-based classification specifically for LUSC. Moreover, an immune-related 13-gene prognostic signature was developed and validated for prognosis prediction in LUSC through multi-step bioinformatics. This signature was strongly associated with OS in LUSC patients and might serve as a potential prognostic biomarker for clinical use of immunotherapy in the future. Prospective studies are needed to test the clinical utility of the signature for effective treatment strategies and personalized therapies of LUSC.

### Data availability

Raw RNA sequence data that support the findings of this study are available from the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) or TCGA (<https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>), respectively.

Received: 5 November 2021; Accepted: 29 July 2022

Published online: 11 August 2022

### References

1. Ferlay, J. C. *et al.* Cancer statistics for the year 2020: An overview. *Int. J. Cancer* <https://doi.org/10.1002/ijc.33588> (2021).
2. Yuan, H., Liu, J. & Zhang, J. The current landscape of immune checkpoint blockade in metastatic lung squamous cell carcinoma. *Molecules* **26**, 1392. <https://doi.org/10.3390/molecules26051392> (2021).
3. Santos, E. S. & Hart, L. Advanced squamous cell carcinoma of the lung: Current treatment approaches and the role of Afatinib. *Oncol. Targets Ther.* **13**, 9305–9321. <https://doi.org/10.2147/OTT.S250446> (2020).
4. Wagner, M., Jasek, M. & Karabon, L. Immune checkpoint molecules-inherited variations as markers for cancer risk. *Front. Immunol.* **11**, 606721. <https://doi.org/10.3389/fimmu.2020.606721> (2020).
5. Yang, Q., Cao, W., Wang, Z., Zhang, B. & Liu, J. Regulation of cancer immune escape: The roles of miRNAs in immune checkpoint proteins. *Cancer Lett.* **431**, 73–84. <https://doi.org/10.1016/j.canlet.2018.05.015> (2018).
6. Bremnes, R. M. *et al.* The role of tumor-infiltrating lymphocytes in development, progression, and prognosis of non-small cell lung cancer. *J. Thorac. Oncol.* **11**, 789–800. <https://doi.org/10.1016/j.jtho.2016.01.015> (2016).
7. Jiang, P. *et al.* Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* **24**, 1550–1558. <https://doi.org/10.1038/s41591-018-0136-1> (2018).
8. Mariathasan, S. *et al.* TGFbeta attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544–548. <https://doi.org/10.1038/nature25501> (2018).
9. Maleki Vareki, S. High and low mutational burden tumors versus immunologically hot and cold tumors and response to immune checkpoint inhibitors. *J. Immunother. Cancer* **6**, 157. <https://doi.org/10.1186/s40425-018-0479-7> (2018).
10. Al-Shibli, K. I. *et al.* Prognostic effect of epithelial and stromal lymphocyte infiltration in non-small cell lung cancer. *Clin. Cancer Res.* **14**, 5220–5227. <https://doi.org/10.1158/1078-0432.CCR-08-0133> (2008).
11. Wakabayashi, O. *et al.* CD4+ T cells in cancer stroma, not CD8+ T cells in cancer cell nests, are associated with favorable prognosis in human non-small cell lung cancers. *Cancer Sci.* **94**, 1003–1009. <https://doi.org/10.1111/j.1349-7006.2003.tb01392.x> (2003).
12. Galon, J. & Bruni, D. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat. Rev. Drug Discov.* **18**, 197–218. <https://doi.org/10.1038/s41573-018-0007-y> (2019).
13. Miao, Y. R. *et al.* ImmuCellAI: A unique method for comprehensive T-cell subsets abundance prediction and its application in cancer immunotherapy. *Adv. Sci.* **7**, 1902880. <https://doi.org/10.1002/adv.201902880> (2020).
14. Jiang, T. *et al.* Genomic landscape and its correlations with tumor mutational burden, PD-L1 expression, and immune cells infiltration in Chinese lung squamous cell carcinoma. *J. Hematol. Oncol.* **12**, 75. <https://doi.org/10.1186/s13045-019-0762-1> (2019).
15. Doroshow, D. B. *et al.* PD-L1 as a biomarker of response to immune-checkpoint inhibitors. *Nat. Rev. Clin. Oncol.* **18**, 345–362. <https://doi.org/10.1038/s41571-021-00473-5> (2021).
16. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612. <https://doi.org/10.1038/ncomms3612> (2013).
17. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457. <https://doi.org/10.1038/nmeth.3337> (2015).
18. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. <https://doi.org/10.1186/s13059-014-0550-8> (2014).
19. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287. <https://doi.org/10.1089/omi.2011.0118> (2012).
20. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **4**, 2. <https://doi.org/10.1186/1471-2105-4-2> (2003).
21. Chin, C. H. *et al.* cytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **8**(Suppl 4), S11. <https://doi.org/10.1186/1752-0509-8-S4-S11> (2014).
22. Wang, L. & Li, X. Identification of an energy metabolism-related gene signature in ovarian cancer prognosis. *Oncol. Rep.* **43**, 1755–1770. <https://doi.org/10.3892/or.2020.7548> (2020).
23. Herbst, R. S. *et al.* Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563–567. <https://doi.org/10.1038/nature14011> (2014).

24. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci. Publ. Protein Soc.* **28**, 1947–1951. <https://doi.org/10.1002/pro.3715> (2019).
25. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551. <https://doi.org/10.1093/nar/gkaa970> (2021).
26. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30. <https://doi.org/10.1093/nar/28.1.27> (2000).
27. Senoo, S., Ninomiya, K., Hotta, K. & Kiura, K. Recent treatment strategy for advanced squamous cell carcinoma of the lung in Japan. *Int. J. Clin. Oncol.* **24**, 461–467. <https://doi.org/10.1007/s10147-019-01424-y> (2019).
28. Mittal, V. *et al.* The microenvironment of lung cancer and therapeutic implications. *Adv. Exp. Med. Biol.* **890**, 75–110. [https://doi.org/10.1007/978-3-319-24932-2\\_5](https://doi.org/10.1007/978-3-319-24932-2_5) (2016).
29. Sokratous, G., Polyzoidis, S. & Ashkan, K. Immune infiltration of tumor microenvironment following immunotherapy for glioblastoma multiforme. *Hum. Vaccin. Immunother.* **13**, 2575–2582. <https://doi.org/10.1080/21645515.2017.1303582> (2017).
30. Lantuejoul, S. *et al.* PD-L1 testing for lung cancer in 2019: Perspective from the IASLC pathology committee. *J. Thorac. Oncol.* **15**, 499–519. <https://doi.org/10.1016/j.jtho.2019.12.107> (2020).
31. Pages, F. *et al.* International validation of the consensus Immunoscore for the classification of colon cancer: A prognostic and accuracy study. *Lancet* **391**, 2128–2139. [https://doi.org/10.1016/S0140-6736\(18\)30789-X](https://doi.org/10.1016/S0140-6736(18)30789-X) (2018).
32. Galon, J. & Lanzi, A. Immunoscore and its introduction in clinical practice. *Q. J. Nucl. Med. Mol. Imaging* **64**, 152–161. <https://doi.org/10.23736/S1824-4785.20.03249-5> (2020).
33. Sun, D. *et al.* The role of CD28 in the prognosis of young lung adenocarcinoma patients. *BMC Cancer* **20**, 910. <https://doi.org/10.1186/s12885-020-07412-0> (2020).
34. Feng, X. Y. *et al.* Low expression of CD80 predicts for poor prognosis in patients with gastric adenocarcinoma. *Future Oncol.* **15**, 473–483. <https://doi.org/10.2217/fon-2018-0420> (2019).
35. Rah, Y. C. *et al.* Low expression of CD40L in tumor-free lymph node of oral cavity cancer related with poor prognosis. *Int. J. Clin. Oncol.* **23**, 851–859. <https://doi.org/10.1007/s10147-018-1294-3> (2018).
36. Jacob, J. T., Coulombe, P. A., Kwan, R. & Omary, M. B. Types I and II Keratin Intermediate Filaments. *Cold Spring Harb. Perspect. Biol.* **10**, a018275. <https://doi.org/10.1101/cshperspect.a018275> (2018).
37. Knosel, T. *et al.* Cytokeratin profiles identify diagnostic signatures in colorectal cancer using multiplex analysis of tissue microarrays. *Cell. Oncol.* **28**, 167–175. <https://doi.org/10.1155/2006/354295> (2006).
38. Soeth, E. *et al.* Detection of tumor cell dissemination in pancreatic ductal carcinoma patients by CK 20 RT-PCR indicates poor survival. *J. Cancer Res. Clin. Oncol.* **131**, 669–676. <https://doi.org/10.1007/s00432-005-0008-1> (2005).
39. Matros, E. *et al.* Cytokeratin 20 expression identifies a subtype of pancreatic adenocarcinoma with decreased overall survival. *Cancer* **106**, 693–702. <https://doi.org/10.1002/cncr.21609> (2006).
40. Schmitz-Winnenthal, F. H. *et al.* Expression of cytokeratin-20 in pancreatic cancer: An indicator of poor outcome after R0 resection. *Surgery* **139**, 104–108. <https://doi.org/10.1016/j.surg.2005.06.058> (2006).
41. Katsuragi, K. *et al.* Prognostic impact of PCR-based identification of isolated tumour cells in the peritoneal lavage fluid of gastric cancer patients who underwent a curative R0 resection. *Br. J. Cancer* **97**, 550–556. <https://doi.org/10.1038/sj.bjc.6603909> (2007).
42. Moll, R., Divo, M. & Langbein, L. The human keratins: Biology and pathology. *Histochem. Cell Biol.* **129**, 705–733. <https://doi.org/10.1007/s00418-008-0435-6> (2008).
43. Karantza, V. Keratins in health and cancer: More than mere epithelial cell markers. *Oncogene* **30**, 127–138. <https://doi.org/10.1038/onc.2010.456> (2011).
44. Zhu, X. *et al.* A three-gene signature might predict prognosis in patients with acute myeloid leukemia. *Biosci. Rep.* <https://doi.org/10.1042/BSR20193808> (2020).
45. Wagner, S. *et al.* A parsimonious 3-gene signature predicts clinical outcomes in an acute myeloid leukemia multicohort study. *Blood Adv.* **3**(BSR20193808), 1330–1346. <https://doi.org/10.1182/bloodadvances.2018030726> (2019).
46. Zheng, Y. *et al.* A novel immune-related prognostic model for response to immunotherapy and survival in patients with lung adenocarcinoma. *Front. Cell Dev. Biol.* **9**, 651406. <https://doi.org/10.3389/fcell.2021.651406> (2021).
47. Videira, P. A. *et al.* Efficacy of bacille Calmette-Guerin immunotherapy predicted by expression of antigen-presenting molecules and chemokines. *Urology* **74**, 944–950. <https://doi.org/10.1016/j.urology.2009.02.053> (2009).
48. Zhang, H. & Chen, Y. Identification of glioblastoma immune subtypes and immune landscape based on a large cohort. *Hereditas* **158**, 30. <https://doi.org/10.1186/s41065-021-00193-x> (2021).
49. Bigley, A. B. & Simpson, R. J. NK cells and exercise: Implications for cancer immunotherapy and survivorship. *Discov. Med.* **19**, 433–445 (2015).
50. Zhang, S., Zhou, X., Yu, H. & Yu, Y. Expression of tumor-specific antigen MAGE, GAGE and BAGE in ovarian cancer tissues and cell lines. *BMC Cancer* **10**, 163. <https://doi.org/10.1186/1471-2407-10-163> (2010).

## Acknowledgements

This study was funded by the National Natural Science Foundation of China [Grant numbers 81770107, 82003286]; Natural Science Foundation of Hunan Province [Grant number 2020JJ4560, 2021JJ40054]; Scientific Research Foundation of Hunan Provincial Education Department [Grant number 20B528]; Guidance Science and Technology Program of Shaoyang City [Grant number 2019ZD07]; the fellowship of China Postdoctoral Science Foundation [Grant numbers 2020M672474, 2021T140195]; and the Changsha Municipal Natural Science Foundation [Grant number kq20140421].

## Author contributions

Q.Y., J.L. and M.Y. performed the literature search and wrote the manuscript. H.G. performed the analyses. H.L. and W.Z. had the idea for the article and critically revised the manuscript. All authors reviewed and approved the final manuscript.

## 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/s41598-022-17735-6>.

**Correspondence** and requests for materials should be addressed to W.Z. or H.L.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022