Comprehensive Analysis of WGCNA - Derived Cancer Associated Fibroblasts Model For Prognosis, Immune Features, and Candidate Drug Development in LUSC

Rui Sun  
China Medical University

Wang Jian ( wangjian_cmu2023@163.com )  
China Medical University

Article

Keywords: Lung Cancer, WGCNA, prognosis, immune feature

Posted Date: August 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3275724/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Cancer-associated fibroblasts (CAFs) directly affect the behavior of surrounding cells and reshape extracellular matrix (ECM) in tumor microenvironment (TME) via cell-cell contact, releasing regulatory factors. This study aimed to explore stromal CAF-related genes for prognostic prediction and therapeutic response in LUSC. We downloaded mRNA expression and clinical information of 243 LUSC cases from Gene Expression Omnibus (GEO) and 504 cases from The Cancer Genome Atlas (TCGA) databases. weighted gene co-expression network analysis (WGCNA) was performed to identify the key gene module. The protein-protein interaction (PPI) network and machine learning methodology were used to construct a prognostic model. The risk score was involved in 5 genes (COL1A2, COL4A1, COL5A1, MMP2, FN1). In addition, a series of methods based on bioinformatics were used and the results indicated the cases in high risk group suffered less survival time, weaker immune response and higher likely to respond to chemotherapeutic agents. Subsequently, we characterized prognostic model by single-cell sequencing and immunohistochemistry. This five-gene prognostic CAF signature may be a potential biomarker for guiding anti-CAFs therapy and a prognostic clue related to CAF for LUSC patients.

BACKGROUND

Cancer associated fibroblasts (CAFs) are fibroblast populations discovered in primary and metastatic tumors. As a highly heterogeneous and substantial section in tumor microenvironment (TME), CAFs are involved in tumor initiation, progression and metastasis\[1\]. Researchers have recently found that non-small cell lung cancer (NSCLC) CAFs also contain diverse molecular features in different subtypes\[2\]. The mortality of lung cancer is high, and annual deaths equal to the combination of prostate, breast, colon and rectum cancers\[3\]. CAFs play an important role in synchronizing angiogenesis and degradation of the basement membrane through the expression of numerous extracellular matrix (ECM) molecules and growth factors\[4\]. As a consequence, CAFs in NSCLC may affect the regulation of matrix degradation, angiogenesis, invasion, cell growth, survival and could be a new prognostic indicator\[5\]. In lung adenocarcinoma, several genes relevant to CAFs have identified to be better prognosis factor, including podoplanin\[6\], carbonic anhydrase IX\[7\]. While for squamous cell carcinomas patients the CAFs related markers were different, like MMP2 is a significant unfavorable prognostic factor\[8\]. In triple-negative breast cancers (TNBC), CAFs promoted an immunosuppressive environment by secreting CXCL12 to attract CD4+CD25+ T lymphocytes and enhancing the regulatory T cell capacity to inhibit T effector proliferation\[9\]. The activation of CAFs is the key event in partial epithelial-mesenchymal transition (EMT) in pancreatic cancer\[10\]. CAFs-associated paracrine signaling have been confirmed to be involved in drug resistance in epithelial ovarian cancer\[11\]. In the previous studies, α-smooth muscle actin (α-SMA)\[12\], vimentin (vim)\[13\], and fibroblast activation protein (FAP)\[14\] have been considered to be markers of CAFs. Due to the diversity in the sources, differences in protein expression and various function of CAFs, the marker model of CAFs in specific cancer remains to be studied. Weighted gene co-expression network analysis (WGCNA) is a novel algorithm to identify the complex relationships between gene expression
profiles and phenotypes\textsuperscript{[15]}. This is a powerful method to explore the clinical biomarkers and therapeutic targets through characterization of gene expression signature. WGCNA has been conducted to dig potential diagnostic and prognostic biomarkers in stomach adenocarcinoma\textsuperscript{[16]}, ovarian cancer\textsuperscript{[17]}, bladder cancer\textsuperscript{[18]}, and cervical cancer\textsuperscript{[19]}. In this study, we use WGCNA to explore potential biomarkers for LUSC using data from TCGA and GEO databases. We analysed the modules most relevant to CAF infiltration and find five hub genes, i.e., COL1A2, COL5A1, COL4A1. MMP2, FN1 are potential prognosis biomarkers for LUSC patients. Further, we used single cell sequencing and immunohistochemistry to characterize markers, suggesting these genes may be candidate prognostic biomarkers.

**METHODS**

**Datasets, Data Download and processing**

The RNA sequencing (RNA-seq) data with complete follow-up information of 504 samples whose diagnosis was lung squamous cell carcinoma (LUSC) were downloaded from the Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). Clinical and follow-up information was obtained and samples without survival follow-up information were rolled out. We divided RNA-seq data into mRNA and lncRNA. The simple nucleotide variation data is also acquired from TCGA database. We calculated the tumor mutation burden (TME), number of mutated bases per million bases, via the tmbl algorithm in the “maftools” package\textsuperscript{[20]}. Transforming the raw data above by log2 [transcripts per million (TPM) + 1] was performed. We downloaded GSE157010 dataset from Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), 243 cases whose diagnosis is LUSC with full follow-up information on the platform (GPL570). CAF markers were reported by researches before.

**Cancer-Assosicated Fibroblasts and Stromal Analysis**

CAF is a sector of stromal cell, four methods were launched to calculate CAF scores in both TCGA and GEO data, including Estimate the Proportion of Immune and Cancer Cells (EPIC) algorithm\textsuperscript{[21]}, microenvironment cell populations-counter (MCP-counter) algorithm\textsuperscript{[22]}, xCell algorithm\textsuperscript{[23]}, and Estimation of Stromal and Immune cells in Malignant Tumor tissues using Expression data (ESTIMATE) algorithm\textsuperscript{[24]}. We calculated CAF abundance by EPIF in WGCNA and defined Stromal score as phenotypic data. Tumor Immune Dysfunction and Exclusion (TIDE) scores were processed on TIDE database (http://tide.dfci.harvard.edu/)\textsuperscript{[25]}.

**Weighted Gene Co-Expression Network Analysis**

The expression profiles of TCGA - LUSC cohort and GSE157010 were used to screen module genes significantly associated with CAF scores by “WGCNA“ package in R. Weighted Gene Co-Expression Network Analysis (WGCNA) was a package to identify groups of genes sharing similar expression patterns (soft threshold (β) value = 3). We chose the highest positive correlation among the modules and CAF scores (EPIC) for the further analysis, calculating by Pearson correlation coefficient method. Then, we measured gene significance (GS) for each gene's traits and module membership (MM).
Protein-Protein Interaction Network and Hub genes

The protein-protein interaction (PPI) network was analyzed using STRING database[26] (http://string-db.org)[26]. The network was visualized by Cytoscape 3.9.1 software[27] and calculated top 10 hub genes by Maximal Clique Centrality (MCC) method.

Construction of Prognostic Model

TCGA - LUSC cohort was the training cohort due to it contained more simples, and the other was the test cohort. In the least absolute shrinkage and selection operator (LASSO) regression analysis, the top 10 hub genes identified using MCC were subjected to multicariated Cox regression to obtain the coefficients. The LUSC patients in both cohorts were divided into high- and low-risk groups by their respective thresholds[29].

Enumeration of immune cells from gene expression profiles

To quality the relative abundance of 22 tumor-associated Leukocyte (TAL) subsets in groups, we performed Cell type Identification By Estimating Relative Subsets Of known RNA Transcripts (CIBERSORT) method[30].

Chemotherapy Response Predictions

We used “oncoPredict” package to predict IC50 value of different risk groups based on gene expression level[31].

Immunohistochemistry

Tumor tissue from LUSC patients was fixed in 10% formalin at 4°C overnight and embed in paraffin. 5-µm sections were incubated with primary and secondary antibodies (Abcepta Biotech Ltd. Co.) successively[32].

Single-Cell Sequencing Analysis

Single-cell RNA-sequencing (scRNA-seq) data (GSE127465) was analyzed on the Tumor Immune Single Cell Hub (TISCH) database (http://tisch.comp-genomics.org)[33].

Statistical Analysis

We performed all statistical analysis by R software (V.4.2.1). The overall survival comparisons were calculated through the Kaplan-Meier analysis with the log-rank test. P < 0.05 was considered statistically significant.

RESULTS
Cancer-Associated Fibroblasts and Stromal Scores are relevant to the prognostic of lung squamous cell carcinoma

CAFs have been confirmed to be a critical section of stromal cells and are able to recruit immunosuppressive cells in tumor immune microenvironment (TIME).

We performed EPIC, xCell, MCP-counter algorithm to count the level of CAFs infiltration and ESTIMATE algorithm to measure stromal score in GEO cohort (n = 243) and TCGA - LUSC cohort (n = 504). Based on the scores before, the GEO cohort and TCGA - LUSC cohort are divided into high and low CAF infiltration groups compared to stromal cells by the cutoff values, respectively. In TCGA - LUSC cohort, the results from four algorithms depicted that patients with high CAF abundance had shorter overall survival (OS) (Fig. 1A), indicating CAF abundance could be a potential biomarker for the prognosis of LUSC. The GEO cohort is inconsistent to the TCGA cohort (Fig. 1B). The clinical details of the patients were summarized in Table 1.
Table 1
Clinical characteristics of the colon cancer patients used in this study

<table>
<thead>
<tr>
<th></th>
<th>TCGA cohort</th>
<th>GSE150010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>504</td>
<td>235</td>
</tr>
<tr>
<td>Age in years (median,range)</td>
<td>67 (39,90)</td>
<td>68 (46,89)</td>
</tr>
<tr>
<td>Gender (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>131 (26.0%)</td>
<td>82 (34.9%)</td>
</tr>
<tr>
<td>Male</td>
<td>373 (74.0%)</td>
<td>153 (65.1%)</td>
</tr>
<tr>
<td>Stage (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>245 (48.6%)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>163 (32.3%)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>85 (16.9%)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>7 (13.9%)</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>114 (22.6%)</td>
<td>63 (26.8%)</td>
<td></td>
</tr>
<tr>
<td>295 (58.5%)</td>
<td>139 (59.1%)</td>
<td></td>
</tr>
<tr>
<td>71 (14.1%)</td>
<td>31 (13.2%)</td>
<td></td>
</tr>
<tr>
<td>24 (4.8%)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M0</td>
<td>414 (82.1%)</td>
<td>NA</td>
</tr>
<tr>
<td>M1</td>
<td>7 (1.4%)</td>
<td>NA</td>
</tr>
<tr>
<td>MX</td>
<td>79 (15.7%)</td>
<td>NA</td>
</tr>
<tr>
<td>Number of patients</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N0</td>
<td>320 (63.5%)</td>
<td>NA</td>
</tr>
<tr>
<td>N1</td>
<td>133 (26.4%)</td>
<td>NA</td>
</tr>
<tr>
<td>N2</td>
<td>45 (8.9%)</td>
<td>NA</td>
</tr>
<tr>
<td>NX</td>
<td>6 (1.2%)</td>
<td>NA</td>
</tr>
</tbody>
</table>

NA, not applicable; TCGA, The Cancer Genome Atlas
The Kaplan-Meier analysis of (A) TCGA cohort and (B) GEO cohort, including CAF_EPIC, Fibroblasts_MCPcounter, Fibroblasts_xCell, TIDE score and StromalScore_estimate.

Scale independence and mean connectivity in (C) TCGA cohort and (G) GEO cohort.

Module trait relationship of (D) TCGA cohort and (H) GEO cohort. Topographical overlap matrix (TOM) of the co-expression network of 1000 randomly chose genes in (E) TCGA cohort and (I) GEO cohort. Scatterplot of MS and GS of CAFs_EPIC in (F) TCGA cohort and (J) GEO cohort.

Co-Expression Network of Cancer-Associated Fibroblast Scores

WGCNA was conducted to identify the significant modules in GEO and TCGA - LUSC cohorts. The soft threshold powers were 3 in both cohorts (Fig. 1C, G). Subsequently, we performed dynamic module with the minimum 30 genes (Fig. 1D, H). TCGA - LUSC cohort was clustered into 10 co-expression modules and the correlation between blue module and CAFs_EPIC is strongest (Cor = 0.69, P = 5e-71), while the turquoise module module was strongest correlated with StromalScore (Cor = 0.8, P = 1e-113) (Fig. 1E). The positive correlation with CAFs_EPIC was counted between Module membership (MM) and gene significance (GS) (Cor = 0.88, P = 1e-200) (Fig. 1F). For GEO cohort, the number of clustered modules was 14 and the blue module was highest positive correlated with CAFs_EPIC score (Cor = 0.87, P = 2e-74), as well as the StromalScore (Cor = 0.85, P = 5e-66) (Fig. 1I). The positive correlation with CAFs_EPIC was shown between MM and GS (Cor = 0.95, P = 1e-200) (Fig. 1J).

Prognosis-Related CAFs-associated Genes Risk Score Model Construction and Validation

We overlapped the gene lists in the key modules of TCGA cohort and GEO cohort(Fig. 2A), and calculated the PPI network(Fig. 2B) to obtain the key genes’ interaction network(Fig. 2E). Top pathway from KEGG datasets are extracellular matrix organization, collagen-containing extracellular matrix, and extracellular matrix structural constituent(Fig. 2C) and focal adhesion, ECM-receptor interaction and PI3K-Akt signaling pathway from GO datasets(Fig. 2D). We calculated λ value from the top 10 hub genes by LASSO regression analysis(Fig. 2F). Then, 5 genes were included in the risk score formula: CAF risk score = (-0.1737 * ExpCOL1A2) + (0.1365 * ExpCOL5A1) + (0.0705 * ExpFN1) + (0.0555 * ExpMMP2) + (0.0168 * ExpCOL4A1). The LUSC patients in both cohorts were separated into groups by the cutoff of riskScore for each cohort (TCGA - LUSC cohort = 0.4567; GEO cohort = 0.4564) as the threshold value. The analysis revealed a worse OS in high - risk groups than the other groups (Fig. 2G). To test the risk score, a nomogram which integrated the clinical factors and risk score was built to predict survival prognosis of TCGA - LUSC cohort. The nomogram showed that the risk score has great value in predicting prognosis and were able to be used to predict the 1-, 3-, 5-year survival condition of each patient. (Fig. 2H).

Prognosis-Relevant CAFs-associated Genes Risk Score Was Correlated With CAFs Markers
As the major producers of ECM components and secreted cytokines, CAFs exhibited an immunosuppressive function and strongly influenced various immune cell types\cite{34}. CAFs have been reported as a tumor promotor via the secretion of CXCL12, VEGFA, LIF, IGF1, HGF, IL-6, BMP4, and WNT5a\cite{35-40}. The correlation between risk score and CAF scores above was calculated by spearman correlation analysis. In both cohorts, the correlation between risk score and CAF abundance and stromal score were strongly positive (Fig. 3A-B). Apart from the genes which were counted in risk score, the other CAF marker genes expression level was also higher in high-risk group in two cohorts (Fig. 3C-D). In addition, the 5 hub genes were high and positively correlated with the gene sets (Fig. 3E-F).

**CAFs Associated Risk Score Plays A Role In Immune Microenvironment**

CAFs have been reported to exhibit a strong immunomodulatory ability by regulating the abundance of immune cells and anti-tumor function in TME\cite{41-43}. Patients with high risk score and TIDE score have worse response to immunotherapy (Fig. 4A), and greater possibility of immune escape (Fig. 4B). The area under the cover (AUC) of receiver operating characteristic curve (ROC) represented the reliability of immunotherapy effect verification with risk score, was 0.698 (95% CI: 0.623–0.763) in TCGA - LUSC cohort (Fig. 4C). The results above were verified in GEO cohort (Fig. 4D-F). Additionally, CIBERSORT has been considered as an approach to estimate the relevance between survival and abundance of diverse TAL subsets. We conducted CIBERSORT to reveal immune cell fractions between groups of LUSC patients in both cohorts with different risk score. In TCGA - LUSC cohort, the abundance of CD4+ memory resting T cells, Macrophages 0 (M0) were higher in high-risk group, and CD8+ T cells, follicular helper T cells were on the contrary (Fig. 4G), as well as GEO cohort (Fig. 4H).

**Gene Set Enrichment Analysis Enrichment in high-risk Group**

Gene set enrichment analysis enrichment of the two cohorts revealed the aberrant activation of multiple signaling pathways in high-risk group. In TCGA cohort, these included apical junction, epithelial mesenchymal transition (EMT), inflammatory response, TNFα signaling via NFκB pathway based on cancer hallmark gene set (Fig. 5A) and coagulation, epithelial mesenchymal transition (EMT), inflammatory response, Kras signaling up and TNFα signaling via NFκB pathway in GEO cohort (Fig. 5B). As to KEGG gene set, these were cell adhesion molecules cams, cytokine-cytokine receptor interaction, ECM receptor interaction, focal adhesion, pathway in cancer from TCGA cohort (Fig. 5C) and complement and complement cascades, cytokine-cytokine receptor interaction, ECM receptor interaction, focal adhesion and leishmania infection from GEO cohort (Fig. 5D). Previous research has demonstrated that CAFs-derived chemokine CCL5 enhanced Hepatocellular carcinoma (HCC) EMT process\cite{44} via hypoxia-inducible factor 1 alpha (HIF1α)/ zinc finger enhancer-binding protein 1 (ZEB1) axis. Furthermore, the ssGSEA score showed the similar results (Fig. 5E-H).
Estimation of Mutational Load and sensitivity to Anteoeplastic Drugs

To further explore the mechanisms for the difference in survival of patients from two groups, we analyze TMB between them and construct waterfall plots to demonstrate the tumor mutational burden (TME) details in groups (Fig. 6A,B). Some genes displayed higher mutation frequencies in the low-risk group, including TP53, TTN, CSMD3, MUC16, LPR1B. Immunotherapy and chemotherapy after surgery for LUSC patients are critical and the mutation of tumor protein p53 (TP53) is related to the efficacy of EGFR tyrosine kinase inhibitors (TKIs)\[^{45}\]. However, we discovered no significant difference between the risk score and TBM (Fig. 6C). Nevertheless, solely TIDE score of CAFs was correlated to TMB (Fig. 6D). Seventy-six potential small molecule drugs calculated by oncoPredict package that patients with different risk score had different sensitivity. Patients in high-risk group may be more sensitive to 5-Fluorouracil, Afatinib, Afuresertib, Cisplatin, Erlotinib, Gefitinib, Ibrutinib, Ipatasertib, Nilotinib, Osimertinib, Savolitinib, Staurosporine, Tamoxifen, Temozolomide, Uprosertib (Fig. 6E), and the other group of patients may benefit more from Entospletinib.

Validation in CCLE Database, scRNA-Seq AND IHC STAINING

We performed Wilcoxon test in Cancer Cell Line Encyclopedia (CCLE) dataset by “limma” package to compare the expression level of 5 genes related to CAF risk score between fibroblasts and lung in two cohorts. Genes were over expressed in fibroblast (Fig. 7A) and down regulated in lung (Fig. 7B). GSE 127465 was annotated the scRNA-seq into 12 clusters (Fig. 7C). The differential analysis results demonstrated that COL1A2, COL5A1, MMP2, and FN1 expression level were high in fibroblasts, while COL5A1 was over expressed in endothelial cells (Fig. 7D). In addition, the GSEA results of single-cell cohort was consistent with the GSEA above that upresulated genes of fibroblasts were enriched in EMT, myogenesis, angiogenesis, and coagulation (Fig. 7E). The protein expressing level of CAF risk score related genes are demonstrated in both IHC staining of LUSC tumor tissue and results downloaded from HAP database (https://www.proteinatlas.org/). COL1A2, COL5A1, MMP2, and FN1 were expressed in tumor stroma. (Fig. 8).

DISCUSSION

CAFs are part of tumor stroma and play a critical role in cancer progression and has a strong heterogeneity in different cancers. CAFs, as an important section of tumor microenvironment are even affected in response to treatment pressure and/or disease progression\[^{46}\]. Meanwhile, for lung cancer, LUAD and LUSC appear different characters at the molecular, pathological, and clinical level\[^{47}\]. We observed that CAFs and stromal cells abundances were relevant to immunotherapy efficacy and prognosis in LUSC patients. This research is the first exploratory study on CAFs associated biomarkers in LUSC patients with a larger sample and utilizing WGCNA as a method to find gene module. A 5-gene
progenostic (COL1A2, COL1A2, COL5A1, FN1, MMP2, COL4A1) model was calculated and validated through PPI network and LASSO regression and algorithms. Patients with high CAFs risk score were suffered poor effects of immunotherapy and multiple chemical drugs, and higher infiltration level of macrophage M0. Consistently, CAFs risk score were correlated with angiogenesis, ECM, and oxidative phosphorylation. CAFs are involved in “Reverse Warburg effect”. In tumor cells, studies have revealed that oxidative phosphorylation and aerobic glycolysis are not completely mutually exclusive. CAFs derived from stromal cells markedly affects multiple biological process of cancer\(^{[35]}\) in the process of energy supply and competition. CAFs undergo aerobic glycolysis and produce high-energy substance to feed cancer cells due to the oxidative stress in the microenvironment caused by hydrogen peroxide (H\(_2\)O\(_2\)) secreted by tumor. Therefore, mitochondria in cancer cells produce large amount of ATP by oxidative phosphorylation in the presence of anaerobic glycolysis\(^{[48]}\). More studies is still needed to explore the elucidated the above crosstalk in LUSC.

We ameliorated the algorithms to screen hub CAF markers on the basis of traditional differential gene expression (DEG) method in LUSC samples. We used WGCNA to explore the important gene module, and then used PPI for hub genes in the modules. To ensure the robustness of the prognosis model, we calculated the risk score by TCGA - LUSC cohort and validated it by GSE157010 and GSE127465. Based on the above approach, we are convinced that out model closely represents the extent of CAF infiltration in LUSC tumor and provide evidence for clinical prognosis prediction. In addition, the genes that we concluded in the CAF-associated prognostic model were highly expressed in the fibroblasts of LUSC, both at staining of proteins and single-cell levels.

The five genes involved in the risk model have been reported playing functions in the TME for cancer progression, migration, and invasion. During carcinogenesis, ECM is remodeled by various cytokines and collagen from stromal cells and tumor cells. As a critical component of ECM, many collagen proteins come from both fibroblasts and tumor cells and play a function in EMT, invasion and metastasis. CANCER-derived Type collagen is relevant to overall survival and cancer cell differentiation in lung cancer\(^{[49]}\). The pro-tumoral phenotype CAFs can be induced increased expression of \(\alpha\)-smooth muscle actin (\(\alpha\)-SMA) and Collagen alpha-2 (I) chain (COL1A2)\(^{[50]}\). COL1A2 is involved in a predictive logistic regression model for liver and lung metastasis of colorectal cancer\(^{[51]}\). COL4A1 is considered as a biomarker of immune infiltration in NSCLC\(^{[52]}\) and has confirmed to be one of the hub genes to identity the tumor and adjacent non-cancerous tissues in LUSC samples\(^{[53]}\). COL4A1 as involved as type IV collagen to deposit in tumor environment and activate the integrin signaling pathway, which is responsible for high metastatic tendency and more lung modules due to low cell elasticity\(^{[54]}\). COL5A1 gene codifies for the \(\alpha1\)-helix of collagen type V, is one of the prognostic markers (COL5A1, ALDH2, KIF20A, ADH1B, SDC1, VCAN) in malignant pleural mesothelioma\(^{[55]}\) from glycolysis-related pathway gene sets. Moreover, COL5A1 is identified as hub gene in the process to explore the common pathogenesis of lung adenocarcinoma (LUAD) and LUSC\(^{[56]}\). Fibronectin (FN) is one element of ECM and forms a fibrillar array that provides an obligate scaffold for the deposition of other matrix proteins and binding sites for functionalization by soluble factors in the tumor microenvironment. Compared to normal fibroblasts, FN1 produced and
organized by CAFs guides the cancer cells to migrate directionally\textsuperscript{[57]}. FN1 is also considered as an indicator of CAF abundance to protect the prognosis for oral squamous cell carcinoma (OSCC) patients along with TGFB2, TGFBRII, and TGFBIII. In addition, fibronectin regulates anoikis\textsuperscript{[58]}, cell migration and invasion\textsuperscript{[59]} in lung cancer\textsuperscript{[60]}. Matrix metalloproteinases (MMPs) are a family of zinc-dependent extracellular matrix (ECM) remodeling endopeptidases that have the capacity to degrade almost every component of the ECM. Among the MMPs, matrix metalloproteinase – 2 (MMP-2) activation is related to tumor progression and invasion\textsuperscript{[61, 62]} and may be a valuable prognosis variable for lung cancer patients\textsuperscript{[61, 63]}.

In this study, a CAF risk model for LUSC patients has been established and further defined the immunophenotype in TME through the infiltration and characteristics of CAF. Subsequently, we demonstrated that the risk score is a predictor for efficacy of chemotherapy and immunotherapy and survival for LUSC patients. Moreover, this risk model may help to explore the new drug and treatment strategy based on the biomarker of CAF in the future.

**CONCLUSIONS**

This study deepens our knowledge of CAF - associated genes in LUSC. Notably, we discovered that CAF infiltration level might be a biomarker for the response to drug and prognosis of LUSC tumors.

**Abbreviations**

- CAFs: Cancer-associated fibroblasts
- ECM: Extracellular matrix
- GEO: Gene Expression Omnibus
- TCGA: The Cancer Genome Atlas
- WGCNA: Weighted gene co-expression network analysis
- PPI: Protein-protein interaction
- TME: Tumor microenvironment
- NSCLC: Non-small cell lung cancer
- TNBC: Triple-negative breast cancers
EMT
Epithelial-mesenchymal transition
α-SMA
α-smooth muscle actin
vim
Vimentin
FAP
fibroblast activation protein
LUSC
lung squamous cell carcinoma
RNA-seq
RNA sequencing
EPIC
Estimate the Proportion of Immune and Cancer Cells
MCP-counter
microenvironment cell populations-counter
ESTIMATE
Estimation of Stromal and Immune cells in Malignant Tumor tissues using Expression data
TIDE
Tumor Immune Dysfunction and Exclusion
GS
Gene significance
MM
Module membership
MCC
Maximal Clique Centrality
GO
Gene Ontology
KEGG
Kyoto Encyclopedia of Gene
LASSO
Least absolute shrinkage and selection operator
TAL
tumor-associated Leukocyte
CIBERSORT
Cell type Identification By Estimating Relative Subsets Of known RNA Transcripts
scRNA-seq
Single-cell RNA-sequencing
TISCH
Tumor Immune Single Cell Hub
TIME
Tumor immune microenvironment
OS
Overall survival
AUC
Area under the cover
ROC
Receiver operating characteristic curve
HCC
Hepatocellular carcinoma
HIF1α
Hypoxia-inducible factor 1 alpha
ZEB1
Zinc finger enhancer-binding protein 1
TME
Tumor mutational burden
TP53
tumor protein p53
TKIs
tyrosine kinase inhibitors
LUAD
Lung adenocarcinoma
MMPs
Matrix metalloproteinases

**Declarations**

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Funding**

This is a part research accomplishment of the project “National Natural Science Foundation of China Youth Program (32200977)” and National Natural Science Foundation of China (81874214), which is supported by National Natural Science Foundation.
Availability of data and materials

All the datasets used in our analysis are publicly available and all web links are described in the “Methods” section.

References


Figures

A

Figure 1

The kaplan-Meier analysis of (A) TCGA cohort and (B) GEO cohort, including CAF_EPIC, Fibroblasts_MCPcounter, Fibroblasts_xCell, TIDE score and StromalScore_estimate.

Scale indenpendence and mean connectivity in (C) TCGA cohort and (G) GEO cohort.

Module trait relationship of (D) TCGA cohort and (H) GEO cohort. Topographical overlap matrix (TOM) of the co-expression network of 1000 randomly chose genes in (E) TCGA cohort and (I) GEO cohort. Scatterplot of MS and GS of CAFs_EPIC in (F) TCGA cohort and (J) GEO cohort.
Figure 2

(A) Venn of the CAF-associated genes.

(B) The PPI network of 348 genes.

(C) GO (D) KEGG enrichment analysis.
(E) Top 10 hub genes calculated by Cytoscape.

(F) LASSO regression analysis.

(G) Kaplan-Meier analysis of cohorts.

(H) Composite nomogram to predict survival for patients from TCGA - LUSC cohort.
Figure 3

Correlation analysis of CAF scores, stromal score and risk score in (A) TCGA - LUSC, (B) GEO cohort. Heatmap of expression of CAF markers in different risk groups of (C) TCGA - LUSC, (D) GEO cohort. Correlation analysis of 5 genes involved in risk score and CAF markers of (E) TCGA - LUSC, (F) GEO cohort.

Figure 4
Histogram of immunotherapy response in (A) TCGA - LUSC cohort and (D) GEO cohort.

Violin chart of immunotherapy response in (B) TCGA - LUSC cohort and (E) GEO cohort.

ROC of immunotherapy response in (C) TCGA - LUSC cohort and (F) GEO cohort.

Abundance of TAL subsets between low- and high-risk group in (G) TCGA - LUSC cohort and (H) GEO cohort.

Figure 5

GSEA plot in the high-risk group of (A) TCGA - LUSC and (B) GEO cohort based on hallmark gene set.

GSEA plot in the high-risk group of (C) TCGA - LUSC and (D) GEO cohort based on KEGG gene set.
Correlation analysis between risk score and (a) angiogenesis, (b) apical junction, (c) EMT, (d) oxidative phosphorylation, (e) TGF-β signaling in (E) TCGA cohort based on hallmark gene set, correlation analysis between risk score and (a) angiogenesis, (b) apical junction, (c) EMT, (d) TGF-β signaling, (f) k-ras signaling up in (F) GEO cohort based on hallmark gene set.

Correlation analysis between risk score and (a) ECM receptor interaction, (b) focal adhesion, (c) pathways in cancer in (G) TCGA cohort based on KEGG gene set, correlation analysis between risk score and (a) ECM receptor interaction, (b) focal adhesion in (H) GEO cohort based on KEGG gene set.

Figure 6

Gene mutations in (A) high-risk and (B) low-risk groups of TCGA cohort.
(C) Spearman analysis between risk score and TBM.

(D) Correlation analysis between CAF scores and TBM.

(E) IC50 values between different risk groups, including 5-Fluorouracil, Afatinib, Afuresertib, Cisplatin, Erlotinib, Gefitinib, Ibrutinib, Ipatasertib, Nilotinib, Osimertinib, Savolitinib, Staurosporine, Tamoxifen, Temozolomide, Uprosertib, Entospletinib.

Figure 7

(A) Expression difference between fibroblast and lung based on CCLE.
(B) Heatmap of genes expression between fibroblast and lung.

(C) Cell type in GSE 127465.

(D) Different distribution of gene expression in different type of cells.

(E) GSEA of genes in different cell types.
Figure 8

Expression of genes involved in risk score and results downloaded from HPA database.