Shang *et al. Journal of Translational Medicine* (2025) 23:263
https://doi.org/10.1186/s12967-025-06316-6

Journal of Translational Medicine

RESEARCH Open Access

Molecular subtyping of stage I lung adenocarcinoma via molecular alterations in pre-invasive lesion progression

Jun Shang[1,2,3,6†], He Jiang[3†], Yue Zhao[1,2,6†], Jingcheng Yang[3], Yicong Lin[4], Naixin Zhang[3], Luyao Ren[3], Qingwang Chen[3], Ying Yu[3], Leming Shi[3,7*], Yuan Li[4,5*], Haiquan Chen[1,2,6*] and Yuanting Zheng[3*]

Abstract

Background Patients with adenocarcinoma in situ (AIS) and minimally invasive (MIA) lung adenocarcinoma (LUAD) are curable by surgery, whereas 20% stage I patients die within five years after surgery. We hypothesize that poor-prognosis stage I patients may exhibit key molecular characteristics deviating from AIS/MIA. Therefore, we tried to reveal molecularly and prognostically distinct subtypes of stage I LUAD by applying key molecular alterations from AIS/MIA to invasive LUAD progression.

Methods The RNA and whole-exome sequencing data of 197 tumor-normal matched samples from patients with AIS, MIA, and invasive LUAD were analyzed. ddPCR quantified 202 samples from 182 patients at the absolute expression level. Immunohistochemical quantified the protein expression levels of ACTA2. RNA-seq data from 954 LUAD patients, including 541 stage I patients, along with 12 published datasets comprising 1,331 stage I LUAD patients, were used to validate our findings.

Results Focal adhesion (FA) was identified as the only pathway significantly perturbed at both genomic and transcriptomic levels by comparing 98 AIS/MIA and 99 LUAD. Then, two FA genes (COL11A1 and THBS2) were found strongly upregulated from AIS/MIA to stage I while steadily expressed from normal to AIS/MIA. Furthermore, unsupervised clustering separated stage I patients into two molecularly and prognostically distinct subtypes (S1 and S2) based on COL11A1 and THBS2 expressions (FA2). Subtype S1 resembled AIS/MIA, whereas S2 exhibited more somatic alterations and activated cancer-associated fibroblast. Immunohistochemistry on 73 samples also observed that CAF was more active in S2 compared to S1 and AIS/MIA. The prognostic value of these two genes identified

†Jun Shang, He Jiang and Yue Zhao contributed equally to this work.

*Correspondence:
Leming Shi
lemingshi@fudan.edu.cn
Yuan Li
lumoxuan2009@163.com
Haiquan Chen
hqchen1@yahoo.com
Yuanting Zheng
zhengyuanting@fudan.edu.cn

Full list of author information is available at the end of the article

from our knowledge-driven process was confirmed by 541 stage I patients in a prospective dataset, ddPCR and 12 published datasets.

Conclusions We successfully revealed two molecularly and prognostically distinct subtypes of stage I LUAD by applying key molecular alterations from AIS/MIA to invasive LUAD progression. Our model may help reliably identify high-risk stage I patients for more intensive post-surgery treatment.

Keywords Lung adenocarcinoma, Pre/minimally invasive, Molecular subtypes, Prognosis, COL11A1, THBS2, Overfitting-resistant, Unsupervised clustering

Introduction

Lung adenocarcinoma (LUAD) is the most common histological subtype of lung cancer with greatly varied five-year survival rate [1, 2]. Adenocarcinoma in situ (AIS) and minimally invasive adenocarcinoma (MIA), defined as pre/minimally invasive stage lesions with no or less than 5 mm of invasion, have excellent five-year survival rate of virtually 100% [2–4]. However, the five-year survival rate of invasive LUAD, even in early pathological stage I, drops to about 80% [1]. Meanwhile, there is controversy over whether patients with stage I LUAD benefit from adjuvant therapy [5, 6], causing uncertainties in clinical treatment of these patients.

There is an urgent need to accurately classify stage I LUAD patients into good-prognosis subgroup like AIS/MIA who can be cured by surgical resection alone and poor-prognosis subgroup with high risk of recurrence or death who may benefit from more aggressive post-surgery treatment such as adjuvant therapy. Over the past 15 years, many predictive models based on gene-expression data from microarray and high-throughput sequencing technologies have been developed for risk stratification of LUAD patients [7]. However, there is still a lack of simple and robust predictive models for molecular subtyping of stage I LUAD suitable for clinical applications [8]. So far, the Oncocyte DetermaRx test (https://oncocyte.com) based on the expression of 14 genes appears to be the only one used in clinic [9].

The small number of genes (features) and the over-fitting-resistant process by which how such genes are selected are two critically important characteristics for the robustness of a predictive model [10]. For most models of molecular classification of stage I LUAD [11, 12], genes are generally selected based on direct comparison of the expression levels between patients with good and bad prognosis. Over-fitting caused by the complex modeling process including the selection of many genes to explicitly fit the prognosis endpoint of the training set, accompanied by inadequate cross-validation and external validation with truly independent datasets, is an important reason why many published models have not been adopted for clinical applications [8, 10]. In addition to risk stratification, understanding the molecular characteristics and tumor microenvironment of subtypes of stage I LUAD may help lay the foundation for precision patient treatment.

The molecular alterations from AIS/MIA to invasive LUAD can provide useful insight beyond the degree of pathological invasion for better understanding the disease but have not been adequately studied. Thus, in a previous study [13], we comprehensively analyzed the genomic and immune profiling of AIS/MIA and invasive LUAD, and identified potential driver mutation events such as TP53 mutation, arm-level copy number variation (CNV), and HLA loss of heterozygosity. However, details about the differentially expressed or mutated genes (DEGs or DMGs) between AIS/MIA and invasive LUAD remain to be fully explored. We hypothesize that such DEGs and DMGs, when combined, may help identify key genes associated with the early progression to stage I LUAD from AIS/MIA, and such key genes may then be used to further stratify stage I LUAD patients into subtypes with divergent molecular characteristics and prognosis, in a completely knowledge-driven, unsupervised, and robust manner. Partition Around Medoids (PAM) [14, 15], which is well-known for its robust performance in identifying the true underlying number of clusters within a dataset by resisting noise and isolated data points, may serve this purpose.

In this study, we successfully revealed two molecularly and prognostically distinct subtypes of stage I LUAD, by applying key molecular alterations from AIS/MIA to invasive LUAD progression. First, the focal adhesion (FA) pathway and associated DEGs were identified after we thoroughly compared the differences in genomics and transcriptomics between AIS/MIA and invasive LUAD, without any training based on patient prognosis information. Secondly, stage I LUAD patients were further clearly separated into two subtypes (S1 and S2) based on a simple unsupervised partition model using the expression levels of only two genes (COL11A1 and THBS2) in the FA pathway. Thirdly, we comprehensively analyzed the molecular characteristics of S1 and S2. S1 was closer to pre/minimally invasive LUAD in genomics, transcriptomics, and tumor microenvironment; whereas subtype S2 demonstrated elevated expression of COL11A1 and THBS2, higher somatic alterations, and more active in cancer-associated fibroblast (CAF), diverging markedly

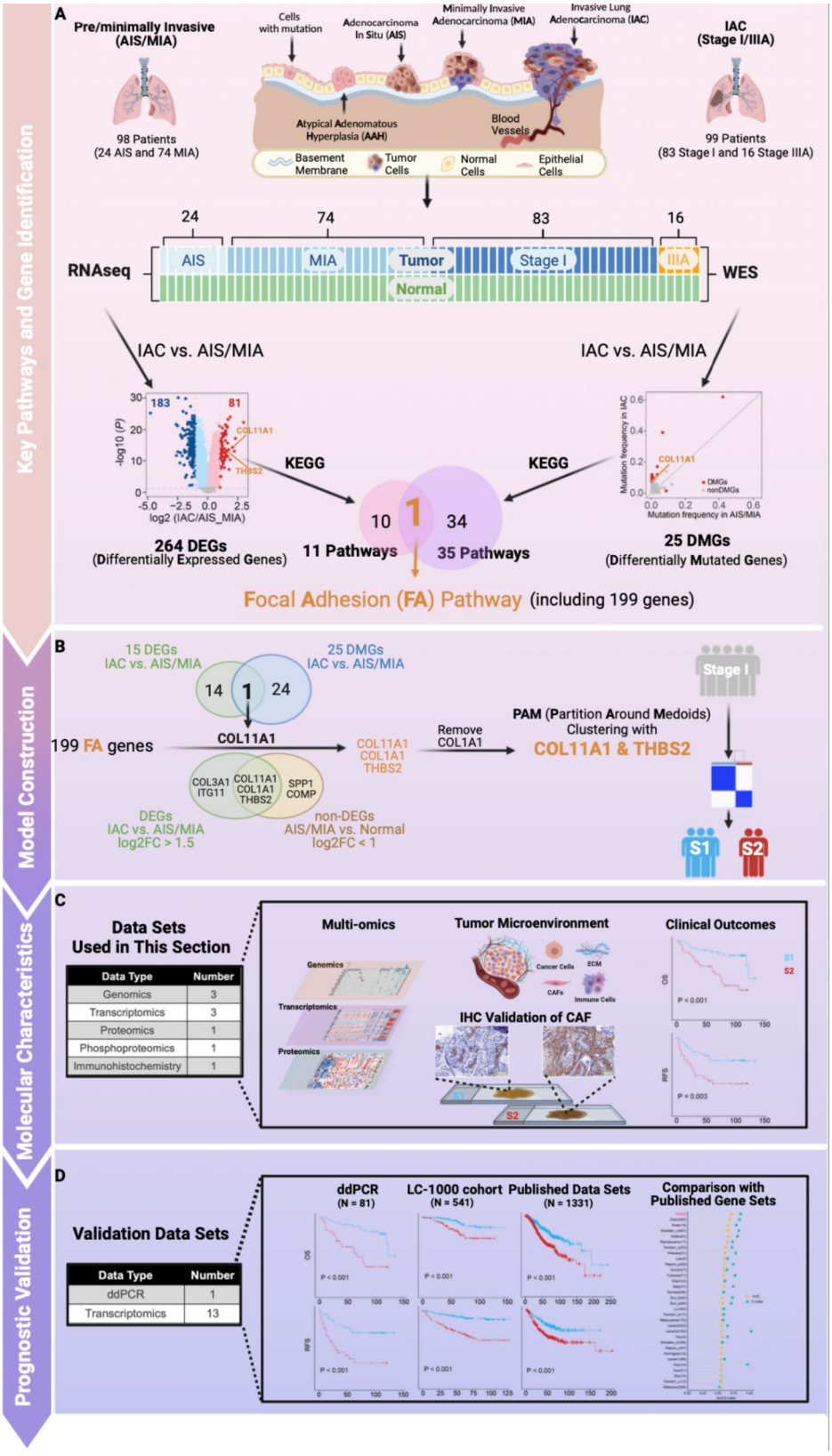

Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Workflow of identification of stage I LUAD subtypes and associated molecular characteristics. (**A**) Patients with AIS/MIA and invasive LUAD were enrolled and tissue samples were collected for sequencing with WES and RNA-seq. DMGs and DEGs between AIS/MIA and invasive LUAD were identified. DMGs and DEGs were both enriched in the focal adhesion (FA) pathway by KEGG enrichment analysis. (**B**) COL11A1 and THBS2 were retained to construct the clustering model, and the PAM consensus clustering using expression of COL11A1 and THBS2 classified stage I LUAD into subtypes S1 and S2. (**C**-**D**) Extensive differences in multi-omics molecular characters, tumor microenvironment (TME), and clinical outcomes between subtypes S1 and S2 were explored in the FUSCC and external datasets

from pre/minimally invasive LUAD. Importantly results from a prospective dataset of 541 stage I LUAD patients and 12 published datasets comprising 1,331 patients, along with validation by ddPCR, confirmed our findings that S2 patients had a significantly worse prognosis compared to S1 patients.

Results

Study design and workflow

The study design and workflow are shown in Fig. 1. Briefly, a total of 197 patients with primary tumor tissues and matched normal tissues were enrolled in this study at Fudan University Shanghai Cancer Center (FUSCC). Among them, 98 were AIS (24) or MIA (74) LUAD, termed pre/minimally invasive, and 99 were stage I (83) or IIIA (16) invasive LUAD, termed invasive, according to the 8th TNM staging (Fig. 1A). RNA sequencing (RNA-seq) and whole-exome sequencing (WES) were carried out for the 197 (98 + 99) tumor-normal pairs of samples as described in our previous study [13] (Fig. 1A). First, we identified 264 DEGs and 25 DMGs between pre/minimally invasive and invasive LUAD corresponding to the transcriptomic and genomic alterations, respectively. Secondly, we identified 11 and 35 pathways that were enriched with the DEGs and DMGs, respectively. Strikingly, focal adhesion (FA), involving a total of 199 genes including 15 DEGs and 1 DMGs (with an overlap of COL11A1), was the only pathway that was significantly perturbed both transcriptomically and genomically (Fig. 1A and B). Thirdly, two genes (COL11A1 and THBS2) that were most significantly differentially expressed between invasive and AIS/MIA, and that at the same time showed no difference in expression between AIS/MIA and normal, were identified to develop a model (FA2) using an unsupervised consensus clustering method called Partition Around Medoids (PAM), clearly separating stage I invasive LUAD patients into two distinct subtypes (S1 and S2) (Fig. 1B and C). Finally, the differences between subtypes S1 and S2 in molecular characteristics including genomics, transcriptomics, proteomics, tumor microenvironment, and clinical outcomes were comprehensively evaluated and confirmed with multiple datasets previously reported in the literature (Fig. 1C and D).

Genomic and transcriptomic alterations from pre/minimally invasive to invasive LUAD identified focal adhesion (FA) pathway and elevated expression of COL11A1 and THBS2 as key changes of LUAD progression

AIS and MIA are similar in genomic and transcriptomic characters. The principal component analysis (PCA) suggested that the expression profiles of AIS and MIA were similar to each other and were closer to that of invasive LUAD than to normal lung tissue (Fig. 2A). Meanwhile, almost no significant DMGs and DEGs between AIS and MIA could be identified (Figures S1A and S1B). Therefore, we combined AIS and MIA into one group (AIS/MIA) in the following analyses.

We set out to determine important and reliable disease progression-associated pathways by first detecting DEGs and DMGs between AIS/MIA and invasive LUAD. We thus identified 264 DEGs (|log2FC| >=1 and $P < 0.05$) and 25 DMGs ($P < 0.05$) (Fig. 2B and C, S1C and S1D). Except for BRAF (AIS/MIA vs. invasive, 8% vs. 1%), the other 24 DMGs, such as TP53 (AIS/MIA vs. invasive, 6% vs. 38%), showed much higher mutation frequency in invasive LUAD than in AIS/MIA (Fig. 2C, S1C and S1D). We performed Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis using the 264 DEGs and 25 DMGs separately. As a result, the focal adhesion (FA) pathway was commonly shared by the 11 DEG-enriched pathways and the 35 DMG-enriched pathways (Fig. 2D, S1E and S1F). It has been reported that the FA complex is a bridge between cells and the extracellular matrix, and plays an important role in cell proliferation, invasion, and migration [16]. We further identified 15 DEGs from the 199 FA pathway genes, which were downloaded from the molecular signature database (MsigDB) (Fig. 2E). COL11A1, which was the only gene shared by the 15 DEGs and the 25 DMGs in the FA pathway, may play an important role in the progression of AIS/MIA to invasive LUAD (Fig. 2F).

We hypothesize that genes whose expression is significantly increased only from AIS/MIA to invasive (corresponding to good and bad prognosis, respectively), but not from normal to AIS/MIA (both with good prognosis), may play a more prominent role in disease progression and prognosis. By setting a more stringent log2FC > 1.5 cutoff, we retained five genes (SPP1, COL11A1, COL1A1, COMP, and THBS2) with significantly increased expression level from AIS/MIA to invasive. However, two (SPP1 and COMP) of them had

already showed significantly higher expression level in AIS/MIA than in normal, and therefore were removed from further consideration (Fig. 2G). This process led to three remaining genes, namely COL11A1, THBS2, and COL1A1 (Fig. 2H). COL11A1 and THBS2 were still selected as the two key genes according to the screening process shown in Fig. 1 without considering the 16 IIIA samples, indicating the robustness of the gene-selection process and the reliability of the two selected genes (Figure S2). Considering that COL1A1 and COL11A1 are from the same gene family with similar functions, we only used COL11A1 and THBS2 for subsequent molecular subtyping analysis of stage I LUAD. Obviously, we can see that the expression levels of COL11A1 and THBS2 both increased significantly from normal/AIS/MIA to stage IA (Fig. 2I), whereas no appreciable change in expression was observed from normal to AIS to MIA. The fold change in COL11A1 and THBS2 expression for AIS&MIA over Normal was 1.21 and 1.16, and for Invasive over AIS&MIA was 4.17 and 2.99, respectively.

Unsupervised consensus clustering classified stage I LUAD patients into AIS/MIA-like subtype S1 and AIS/MIA-diverging subtype S2 based solely on the expression of COL11A1 and THBS2

We assumed that stage I LUAD patients may be further divided into multiple molecular subtypes depending on how they are similar or dissimilar to AIS/MIA in different degrees based on the expression of the two FA genes (FA2). Thus, we used the unsupervised Partition Around Medoids (PAM) consensus clustering method, which is well-known for its robust performance in identifying the true underlying number of clusters within a dataset, to cluster stage I LUAD patients using the expression profiles of COL11A1 and THBS2. The expression levels of COL11A1 and THBS2 were used to calculate the Euclidean distance between each sample and the center point of each cluster. After the number of clusters was evaluated from 2 to 10, we identified two subtypes (clusters) named S1 (low expression of COL11A1 and THBS2) and S2 (high expression of COL11A1 and THBS2). This clustering showed the clearest cut (between-cluster distance) and the highest Average Silhouette Width (ASW) [14, 15], a popular cluster validation index to estimate the number of clusters within a dataset (Figures S3A and S3B).

We extensively explored the differences in molecular characteristics between S1 and S2 subtypes within stage I LUAD, and included AIS/MIA as a control group in subsequent analyses. We identified seven genes with significantly different mutation frequency among AIS/MIA, S1, and S2 using Fisher's exact test (Fig. 3A). Four (EGFR, TP53, TTN, and CSMD3) of the seven genes were among the top 20 most frequently mutated genes in our datasets (Figure S4A). Except for EGFR and MGAM, the mutation frequency of the other five genes (TP53, TTN, CSMD3, DST, and FSCB) significantly increased in S2 over S1 (Fig. 3B). Similarly, tumor mutation burden (TMB) gradually increased from AIS/MIA to S1 to S2 (Fig. 3C). The same trend was also seen in APOBEC-related mutations, although only AIS/MIA vs. S2 was significantly different (Fig. 3D). These results indicated that S1 was genetically closer to AIS/MIA than S2.

Consistent with the trend of genomic characteristics, transcriptomic analysis also indicated that S1 was similar to AIS/MIA. PCA demonstrated that gene expression profiling of S1 was closer to AIS/MIA than S2 (Fig. 3E). We further compared the expression profiles between AIS/MIA, S1 and S2, and identified 83 DEGs between AIS/MIA and S1, 881 DEGs between AIS/MIA and S2, and 383 DEGs between S1 and S2 (Fig. 3F). Several pathways (ECM-receptor interaction, protein digestion and absorption, and focal adhesion) were enriched with most genes upregulated in S2 (Figure S4B). Meanwhile, the expression of all 15 FA pathway DEGs between AIS/MIA and invasive LUAD (Fig. 2E) were also significantly altered between S1 and S2 (Figure S4C). We further explored cancer-associated biological functions of DEGs between AIS/MIA, S1, and S2 using Get Set Variation Analysis (GSVA). A total of 22 hallmarks of cancer were identified using the hallmark gene sets from MSigDB (Fig. 3G). The enrichment scores of these identified hallmarks showed continuous changes from AIS/MIA, S1, S2, to IIIA, indicating that S1 may be an intermediate biological stage during the development of AIS/MIA to S2 or IIIA.

To evaluate the molecular subtypes underlying the whole FUSCC set of 394 samples including normal, AIS, MIA, and invasive LUAD, we performed PAM clustering using the expression profiles of COL11A1 and THBS2. Consistent with clustering of stage I LUAD samples alone, the average silhouette width indicated that the optimal number of clusters was two (Figure S3C). It was interestingly and gratifying to notice that 100% normal, 95.8% AIS, 94.6% MIA, 64.3% IA, 40.7% IB, and 37.5% IIIA were assigned to S1 (Figures S3D and S3E). These results indicated that S1 was closer to AIS/MIA, and that as the disease stage progresses, more and more patients became S2-like.

S2 with COL11A1 upregulation had more activated cancer-associated fibroblasts than S1

We explored the differences in tumor microenvironment (TME) between AIS/MIA, S1, and S2. We analyzed the composition of TME using EPIC [17] and MCP-counter [18], two widely used software packages for such purposes. Associations between CAF and molecular subtypes were observed in that S2 with COL11A1

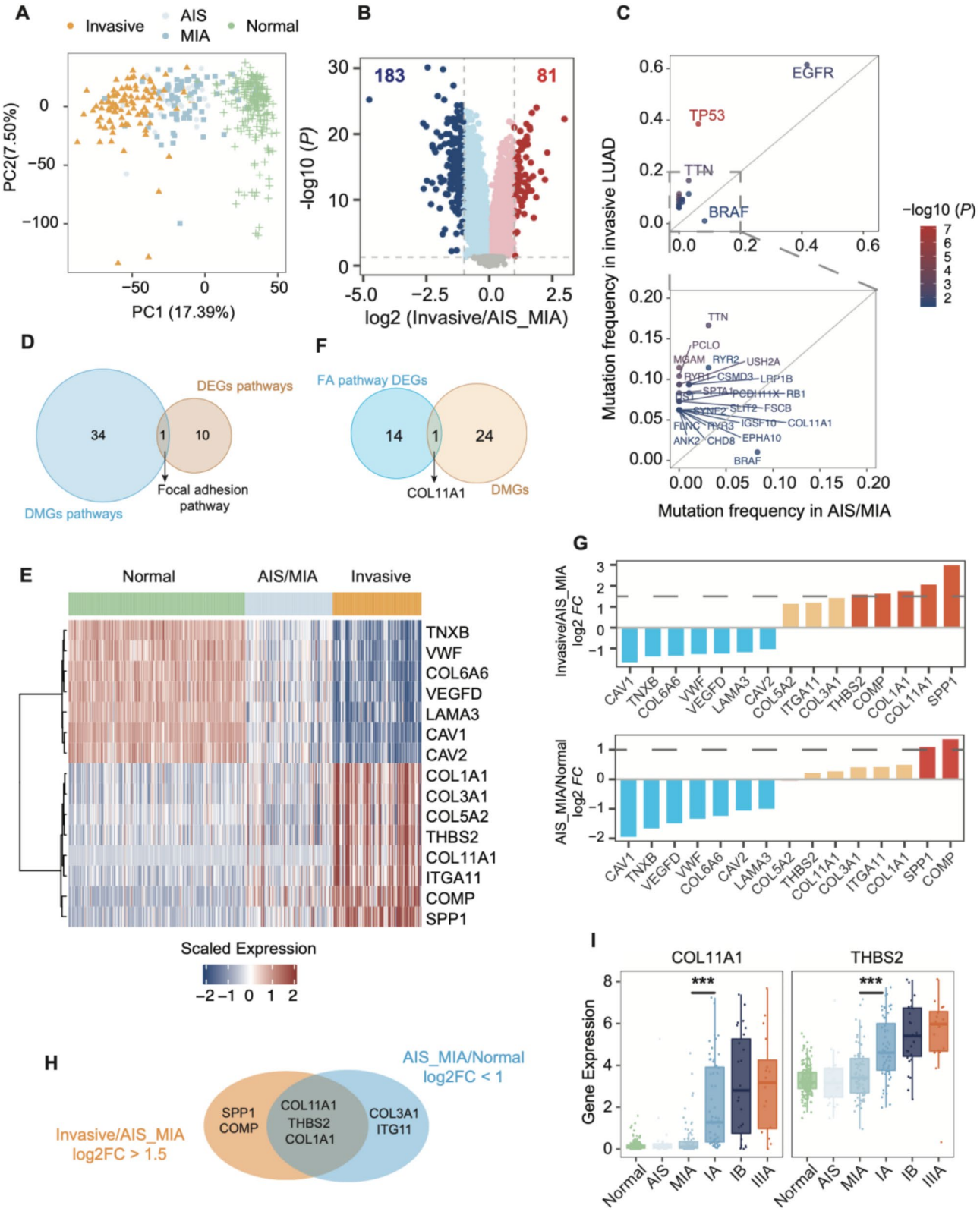

Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 Identification of COL11A1 and THBS2 in the FA pathway as key determinants for invasive LUAD deviating from pre/minimally invasive status. (**A**) PCA of the expression profiles of 39,476 genes in 197 pairs of LUAD samples including 24 pairs of AIS, 74 pairs of MIA, and 99 pairs of invasive LUAD. (**B**) Volcano plot shows differential gene expression between invasive LUAD and AIS/MIA. (**C**) Comparison of gene mutation frequency between AIS/MIA and invasive LUAD. A total of 25 genes show significantly different mutation frequencies between AIS/MIA and invasive LUAD. Color bar shows -log10 (*P*). (**D**) Venn diagram shows the intersection of pathways enriched by DMGs and DEGs. DMGs and DEGs were both enriched in the FA pathway. (**E**) The expression of 15 DEGs in the FA pathway between AIS/MIA and invasive LUAD. (**F**) Venn diagram shows the intersection of 15 DEGs in the FA pathway and 25 DMGs between AIS/MIA and invasive LUAD. (**G**) The log2FC of the 15 DEGs in the FA pathway: invasive vs. AIS/MIA (top) and AIS/MIA vs. normal (bottom). (**H**) Venn diagram shows genes with expression significantly increased from AIS/MIA to invasive LUAD, but no significant difference between AIS/MIA and normal. (**I**) The expression of COL11A1 and THBS2 from normal to stage IIIA. (*** $P < 0.001$)

upregulation had more activated CAF than S1 (Fig. 3H). By performing immunohistochemistry on 78 samples (Normal: 10, AIS: 7, MIA: 9, S1: 29, S2: 14, IIIA: 9) from 68 patients using ACTA2 (α-SMA), a marker protein for CAF, we also observed that CAF was more active in S2 compared to S1 and AIS&MIA (Fig. 4A and B). We found that the expression of ACTA2, a CAF marker protein, was closely linked to poor prognosis (Figures S5A and 5B). Consistent with published research, COL11A1 in CAF was increased compared with normal fibroblasts in non-small cell lung cancer (NSCLC) [19]. CAF, one of the most abundant cell types in tumor tissues, was closely associated with promoting lung cancer development [20–22]. In fact, many clinical studies aimed at inhibiting the interplay between CAF and tumors are ongoing [22]. Meanwhile, several studies suggested that CAF may promote cancer invasion by remodeling the extracellular matrix (ECM) [22]. We also observed that CAF and the ECM-receptor interaction pathways were more active in S2 (Figs. 3H and 4B and S4B), suggesting that CAF and ECM played important roles in LUAD progression. Therefore, patients with S2, which showed more activated CAF than S1 and AIS/MIA, may benefit from treatment aimed at preventing CAF activation.

To better characterize the CAF subclusters linked to lung adenocarcinoma progression, we analyzed single-cell data from 57 lung adenocarcinoma and normal lung tissue samples, focusing on isolating CAF cells for detailed subclusters analysis (Figures S6A-C). We further subtyped CAFs using single-cell data from lung adenocarcinoma and identified 7 subclusters (Fig. 4C). Interestingly, we found that COL11A1 was specifically expressed in subgroup 1 CAFs, which we subsequently defined as COL11A1 + CAFs (Fig. 4D and E). Consistent with RNA-seq results, COL11A11 + CAF was largely absent in Normal and MIA samples, but its proportion increased as lung adenocarcinoma advanced (Fig. 4F and G). This suggests that elevated expression of COL11A1 is closely linked to increased COL11A1 + CAF activity, and the rising levels of COL11A1 + CAF are strongly associated with lung adenocarcinoma progression.

Distinct S1 and S2 subtypes were also observed with proteogenomic data

We subsequently reanalyzed the multi-omics data from the study of Gittelle et al. [23]. to explore differences in proteogenomic characteristics between S1 and S2 subtypes. Gene mutation and expression data were downloaded from Genomic Data Commons (GDC) [24] and proteomics and phosphoproteomics data were downloaded from Clinical Proteomic Tumor Analysis Consortium (CPTAC) [25]. In this unique dataset, PAM consensus clustering was also used to classify stage I LUAD patients based on the expression of COL11A1 and THBS2. Again, the optimal number of clusters was determined to be two (Figure S7A). Therefore, PAM consensus clustering was performed to identify two subtypes based on the expression of COL11A1 and THBS2 of stage I patients. Consistent with what were observed in the FUSCC dataset, S2 showed more mutation events, more death or relapse events than S1 (Fig. 5A and B). The mutation frequency of TP53, RYR2, USH2A, KRAS, and XIRP2 was much higher in S2 than S1 (Fig. 5B). Moreover, the event of copy number variations, such as amplification peaks, was less common in S1 than S2 (Fig. 5C). Consistent with our FUSCC dataset, the genomes of S1 were relatively simpler than S2.

Quantitative omics, including transcriptomics, proteomics, and phosphorylated proteomics analysis confirmed the distinct differences between S1 and S2. We performed differential expression analysis between S1 and S2 and identified 371 DEGs, 64 differentially expressed proteins (DEPs), and 121 differentially expressed phosphoproteins (DEPPs) (Figures S7B and S7C). To further explore biological functions associated with the DEGs and DEPs, we performed KEGG pathway enrichment analysis. We found that the DEGs and DEPs were both enriched in protein digestion and absorption, ECM – receptor interaction, focal adhesion, bladder cancer, and steroid hormone biosynthesis pathways (Fig. 5D). Meanwhile, we also found that S2 showed more activated CAF than S1 (Figure S7D). Consistent with what we identified from our FUSCC dataset, more activated CAFs and ECM – receptor interaction were also identified for S2 in the Gittelle et al. dataset.

We observed a strong correlation between gene and protein expression levels for both COL11A1 and THBS2

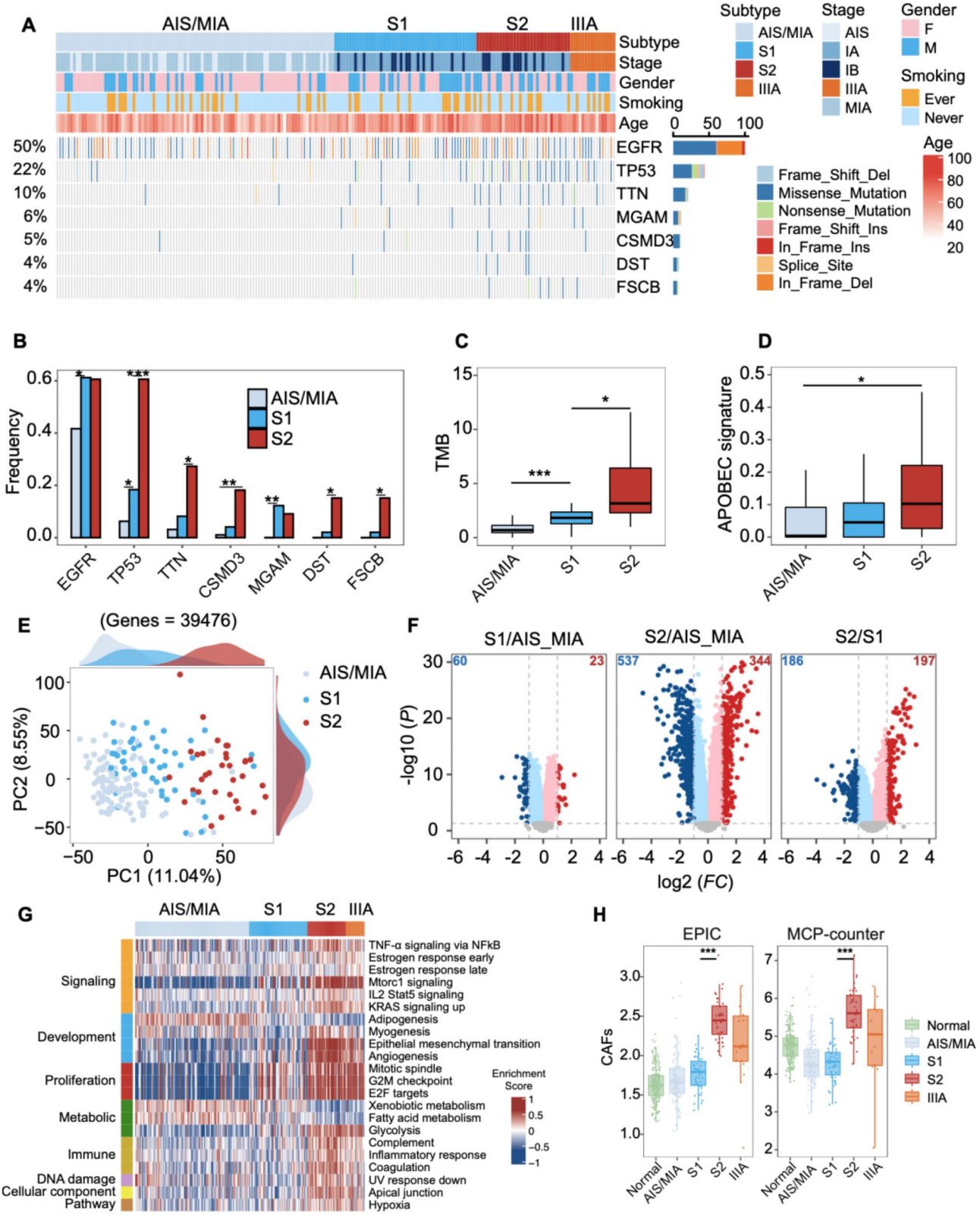

Fig. 3 (See legend on next page.)

(See figure on previous page.)

Fig. 3 Molecular subtypes of stage I LUAD and associated distinct genomic and transcriptomic characteristics. (**A**) Classification of stage I LUAD into S1 and S2 subtypes. DMGs among AIS/MIA, S1, and S2 were shown from AIS/MIA to IIIA LUAD. (**B**) Gene mutation frequency of DMGs for AIS/MIA, S1, and S2. AIS/MIA had lower EGFR and MGAM mutation frequency than S1 and S2. Gene mutation frequency of TP53, TTN, CSMD3, DST, and FSCB increased significantly from AIS/MIA to S2. (**C**) Boxplot shows that S2 had higher TMB than S1 and AIS/MIA. (**D**) Boxplot shows that S2 had higher APOBEC-related mutation than AIS/MIA. (**E**) PCA of the expression profiles of 39,476 genes in AIS/MIA, S1, and S2. (**F**) Volcano plots show between-group differences in gene expression, S1 vs. AIS/MIA, S2 vs. AIS/MIA, and S2 vs. S1. (**G**) Enrichment scores from the get set variation analysis of DEGs between AIS/MIA, S1, and S2. (**H**) Boxplots show CAF in different pathological stages, of which stage I was divided into S1 and S2. S2 had a higher CAF than S1. (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$)

(Fig. 5E, S7E and S7F), indicating that protein expression, like gene expression, may also be used for molecular subtyping of stage I LUAD patients. To explore the relationship between clinical outcomes and molecular subtypes identified by consensus clustering based on the protein expression of COL11A1 and THBS2, we downloaded proteomics data and the corresponding clinical information from the study of Xu et al. [26]. We identified two subtypes closely associated with relapse-free survival (RFS) ($P < 0.001$), although the association with overall survival did not achieve statistical significance ($P = 0.31$, Fig. 5F).

The prognosis model was locked down and allows for its easy clinical applications

We locked down our prognosis model with detailed information and parameter settings, making it suitable for performing molecular subtyping of a single new sample. The centroids of S1 and S2 in the FUSCC RNA-seq dataset are M1 (0.557, 4.141) and M2 (4.756, 6.479), respectively. For a sample from a new patient from other datasets, if its Euclidean distance to M1 is shorter than that to M2, then it is classified as subtype S1; otherwise, it is classified as subtype S2 (Fig. 6A). Similar molecular characteristics between S1 and AIS/MIA indicated that they may exhibit similarly excellent prognosis. In our RNA-seq dataset of LC-197, the OS and RFS of S1 were significantly better than S2 for stage I patients (Fig. 6B and C, $P < 0.05$). We did not find any significant correlations between the clinical demographic data (including age, smoking history, gender, and histology) and molecular subtypes (S1 and S2) (Table S1).

Our model was successfully validated with droplet digital PCR (ddPCR), further confirming its robustness and making its clinical application more easily and reliable than RNA-seq. We quantified 204 samples (Normal: 20, AIS: 57, MIA: 63, IA: 57, IB: 22, IIIA: 16) from 184 patients of the LC-197 dataset at the absolute expression level by ddPCR. Three replicates of each sample were measured and a total of seven batches of 96-well plates were run, with the Human Brain Reference RNA [27, 28] profiled in each batch for positive control and batch monitoring. The inter-plate CVs for COL11A1 and THBS2 were less than 9% among the seven batches, indicating reliability of the ddPCR data (Figure S8). Consistent with RNA-seq, the expression levels of COL11A1 and THBS2 both increased significantly from normal/AIS/MIA to stage IA (Figure S9A). Then, we used the PAM consensus clustering method to cluster stage I LUAD patients using the absolute expression levels of COL11A1 and THBS2, and two subtypes were clearly present (Figure S9B). For the ddPCR data, the centroids of S1 and S2 were determined to be M1 (1.451, 5.812) and M2 (6.398, 8.838), respectively. The subtype of a new patient is decided based on its distances between M1 and M2 (Fig. 6D). In our dataset, the OS and RFS of S1 were significantly better than S2 for stage I patients (Fig. 6E and F, $P < 0.05$).

AIS/MIA-like S1 subtype had better prognosis than AIS/MIA-diverging S2 subtype as validated with a prospective and 12 published datasets

The prognostic performance of the knowledge-driven and overfitting-resistant FA2 model was validated with 541 stage I LUAD patients from a prospective data set. The prospective cohort of 968 LUAD tumors and normal lung tissue RNA-seq was constructed, including 541 patients in stage I. The consistent distribution of COL11A1 and THBS2 expression in the prospective dataset and the LC-197 dataset provides the basis for typing of individual samples based on the centroid of LC-197 dataset (Figure S10). We calculated the S1 or S2 subtypes of each of the 541 stage I patients based on M1 (0.557, 4.141) and M2 (4.756, 6.479) centroids (Fig. 6G). The FA2 model still had superior prognostic predictive performance in this cohort (OS, HR = 3.02, $P < 0.001$; RFS, HR = 4.14, $P < 0.001$) (Fig. 6H and I). Multivariate Cox analysis also indicated that the two-class subtyping was an independent prognostic variable (Figures S11A and S11B).

In addition to our LC-1000 dataset, we downloaded 12 external published datasets containing both gene-expression data and prognosis information of stage I LUAD (Table S2). The prognostic performance of the two FA genes (COL11A1 and THBS2) identified from a knowledge-driven and overfitting-resistant process was validated by 12 external published datasets containing both gene-expression data and prognosis information of stage I LUAD (Table S2 and S3). When each of the 12 data sets was subjected to the PAM analysis using the expression data of COL11A1 and THBS2, the optimal number of clusters for all data sets was found to be two (Figure S13), indicating the true number of underlying molecular

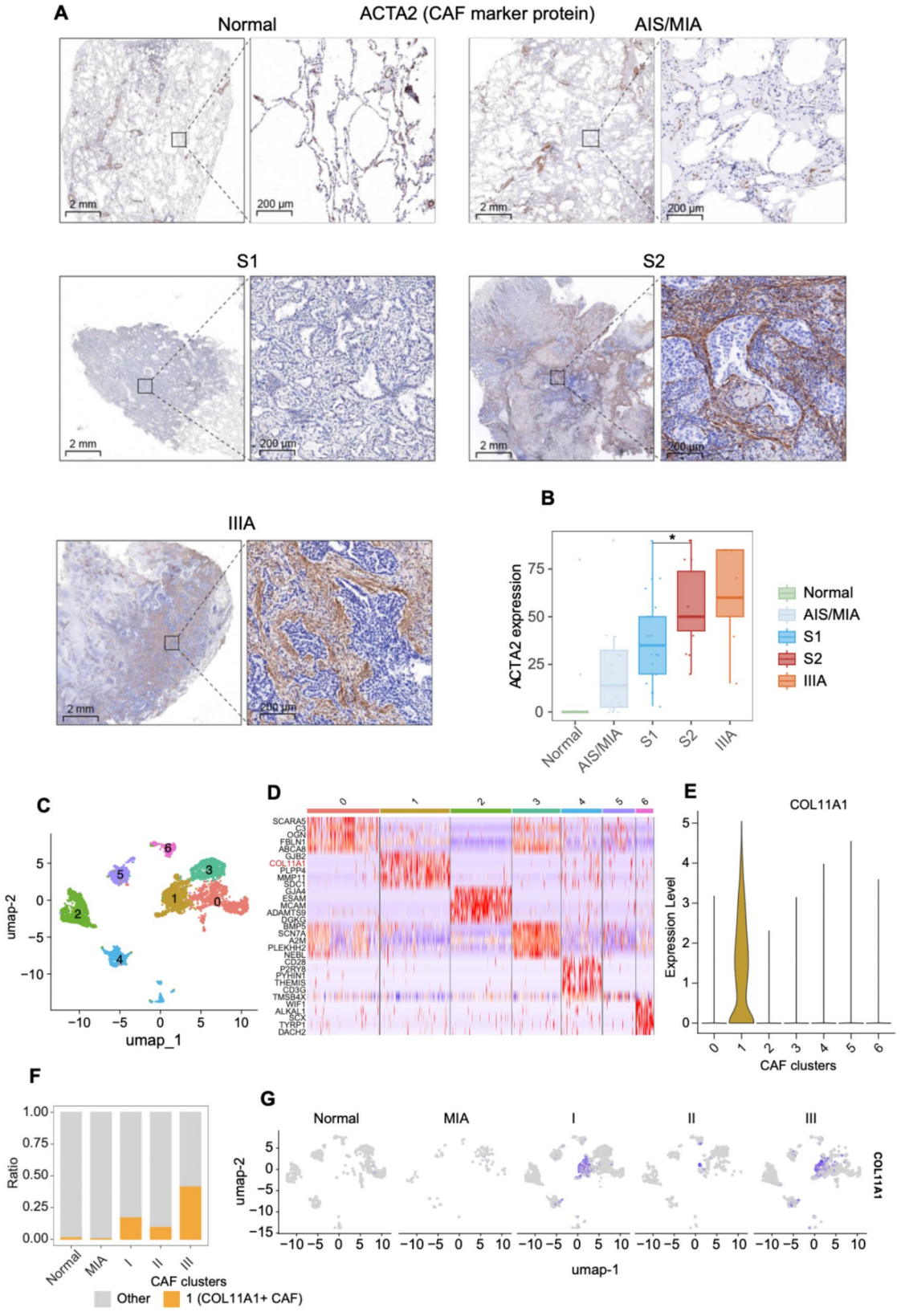

Fig. 4 (See legend on next page.)

(See figure on previous page.)

Fig. 4 Immunohistochemical validation of CAF. (**A**) Representative immunohistochemical stained image of ACTA2 (marker protein of CAF) in LUAD containing normal, AIS/MIA, S1, S2 and IIIA. (**B**) Proportion of ACTA2 expression in different subtypes or stages. (**C**) UMAP visualization of the CAF transcriptome highlights the seven subgroups identified through unsupervised clustering. (**D**)The heatmap displays the top five subgroup-specific genes for each subgroup, with COL11A1 specifically expressed in subgroup 1. (**E**) The violin plot illustrates COL11A1 expression across the seven subgroups, showing a high expression level in subgroup 1. (**F**) The bar plot shows the distribution of subgroup 1, with proportions gradually increasing from normal to pathological stage III. (**G**) The UMAP plot further demonstrates COL11A1 expression across each cell subpopulation. (* $P < 0.05$)

subtypes of stage I LUAD. After subtyping each data set, we merged the 1,331 patients from the 12 data sets together and performed survival analysis (Figure S14). As we expected, survival analysis indicated that subtype S1 had better overall survival (Figure S14A-C, $P < 0.001$) and relapse-free survival (Figure S14D-F, $P < 0.001$) than S2. Multivariate Cox analysis also indicated that the two-class subtyping was an independent prognostic variable (Figures S15A and S15B). Combining IA and S1 would help identify patients who may be truly at low risk (Figures S15C and S15D). These findings also suggested that the FA2 model had better prognostic stratification for early-stage IA.

To comprehensively evaluate and compare the prognostic predictive power of the simple FA2 model involving COL11A1 and THBS2, we identified 42 published prognosis gene signatures of lung cancer through literature research and the review of Tang et al. [7, 9, 29]., with mean and median number of genes of 65 and 42, respectively (Table S4). To facilitate fair comparisons between different signatures (models), the optimal number of clusters was chosen to be two for all models. AUC of the time-dependent receiver-operating characteristics (ROC) curve and concordance index (C-index) were used to evaluate the performance of the 42 literature signatures plus the FA2 signature with the 1,331 patients. The 43 models were ranked by the mean of AUC and C-index. As a result, the FA2 model, with the least number of genes, ranked the second (OS and RFS, Figure S12A) and the first (RFS, Figure S12B) in prognostic predictive power for the prospective dataset. For published datasets, the FA2 model ranked top nine (OS, Figure S16A) and top six (RFS, Figure S16B). These results further suggested that FA2 genes were biologically important and had a good predictive performance for prognosis of stage I patients.

Discussion

Changes in molecular characteristics from pre-invasive AIS/MIA to invasive LUAD may provide us with insights for the accurate classification of stage I LUAD with divergent prognosis. Many studies on constructing models for risk stratification of stage I LUAD based on gene expression have been reported [11, 12, 29]. It is usually a straightforward choice to obtain gene features of "high" prognostic prediction performance through training with prognosis as the endpoint, but this approach is prone to over-fitting [10]. Different from previous studies, we identified two genes (COL11A1 and THBS2) in the FA pathway based mainly on somatic alterations and gene-expression alterations between AIS/MIA and invasive LUAD. In the process of selecting features and conducting molecular subtyping, we did not perform any training with knowledge of a patient's prognosis and thus effectively avoided overfitting, as can be seen from the performance validation results with 11 external datasets.

The two genes played important roles in the progression of lung adenocarcinoma. COL11A1 showed higher mutation frequency in invasive LUAD than AIS/MIA (Fig. 2C). Meanwhile, the expression of COL11A1 was almost undetectable in normal and AIS/MIA, but started to increase dramatically in stage IA LUAD (Fig. 2I). These results indicated that COL11A1 may promote AIS/MIA progression. Many studies have also demonstrated that COL11A1 plays an important role in tumor progression including NSCLC [30–33]. Similarly, there was no significant difference in the expression of THBS2 between normal and AIS/MIA, whereas there was a steady increase from IA to IB and IIIA LUAD (Fig. 2I). The evolving mutation or expression characteristics of COL11A1 and THBS2 indicated their close association with the invasiveness of LUAD. Indeed, FA2 consisting of COL11A1 and THBS2 helped identify two molecular subtypes S1 and S2 with different degrees of invasion in stage I LUAD.

The consistency and robustness of genomic, transcriptomic, and proteomic differences between S1 and S2 demonstrated that the two subtypes of stage I LUAD were biologically and clinically relevant. In our FUSCC dataset, the genomic and transcriptomic characteristics of S1 were similar to AIS/MIA, and the prognosis of S1 was closer to AIS/MIA. Furthermore, the high correlation between gene and protein expression for COL11A1 and THBS2 indicated that protein expression data may also be used for molecular classification. We performed consensus clustering using protein expression of COL11A1 and THBS2 to divide stage I LUAD into S1 and S2 using a publicly available multi-omics dataset. The significant differences in RFS between S1 and S2 further suggested the prospect of clinical applications of COL11A1 and THBS2. Successfully validated at the absolute expression level (ddPCR) further validated the robustness of our model and made it apply to clinical scenarios more easily.

The differences in molecular characteristics between S1 and S2 may have potential therapeutic implications.

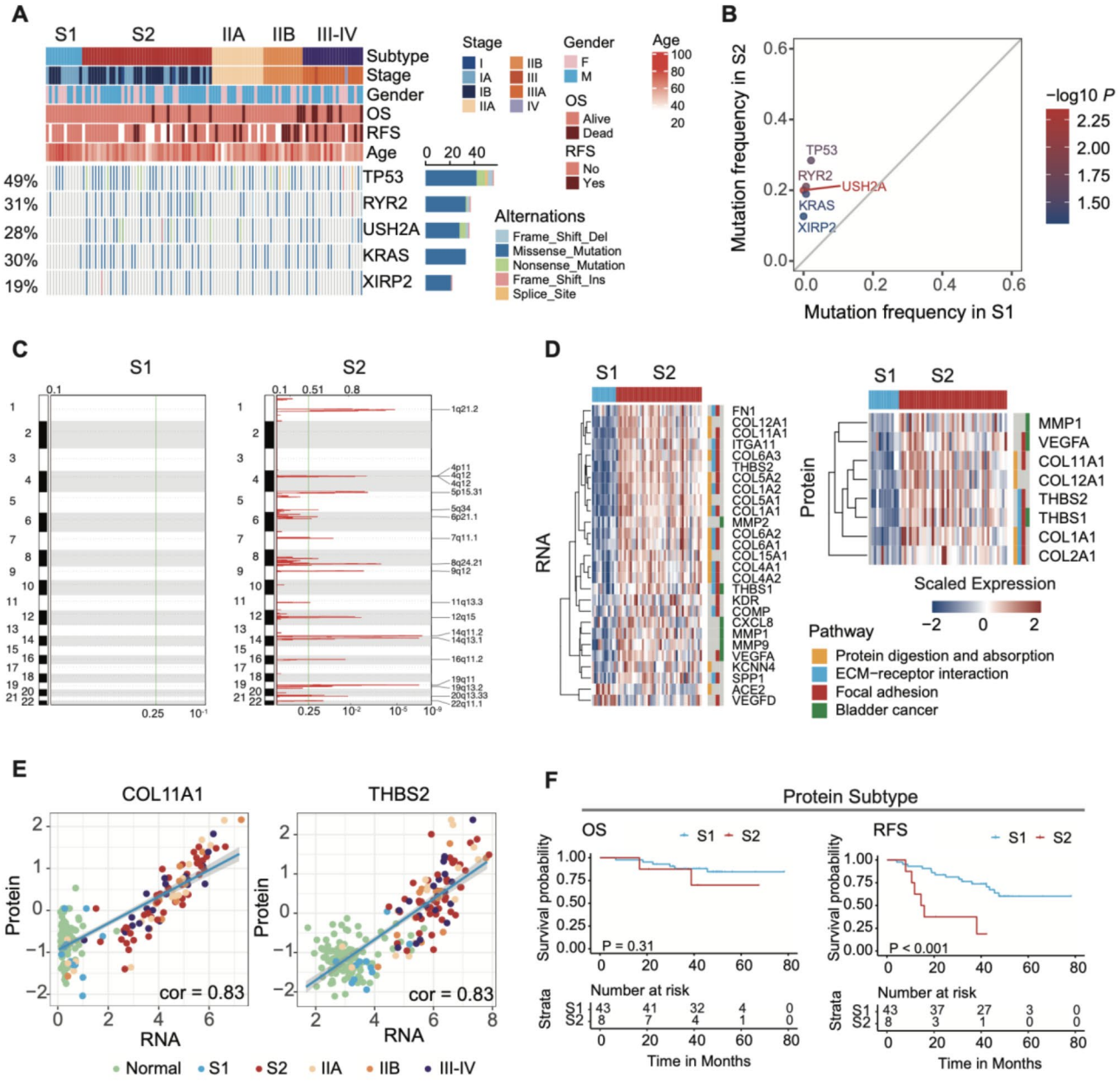

Fig. 5 Proteogenomic relationships between S1 and S2 subtypes. (**A**) Classification of all samples into pathological clusters, of which stage I was divided into S1 and S2 subtypes. Oncoplot shows DMGs between S1 and S2. (**B**) S2 had a higher gene mutation frequency than S1. (**C**) Significant amplification peaks based on copy number profiling of S1 and S2. (**D**) Heatmaps show the DEGs and DEPs and the five pathways which were both enriched with DEGs and DEPs. (**E**) Scatterplots show the correlation between RNA and protein expression for COL11A1 and THBS2. (**F**) Clinical outcomes of S1 and S2 subtypes based on protein expression of COL11A1 and THBS2 in the Xu et al. dataset. S1 had better RFS than S2

The poor clinical outcomes and high TMB of S2 may provide hint for clinical decision-making. At present, the guidance for clinical adjuvant therapy is still based on pathological staging. What is more, there is no obvious evidence that patients with stage I LUAD can benefit from adjuvant therapy, so that most stage I patients do not undergo adjuvant therapy systematically after surgery [5, 6]. According to our analysis, the classification of stage I LUAD into S1 and S2 was better than IA and IB in prognostic predictive performance (Figure S11). Stage I LUAD patients who were classified as S2 had a higher risk of recurrence or death. At the same time, S2 had higher TMB than S1. Many studies have shown that patients with high TMB may benefit from immune checkpoint inhibitors [34]. Thus, our results indicated that S2 patients may benefit from receiving adjuvant therapy, such as immunotherapy. Meanwhile, a high score of CAF in S2 may suggest possible treatment. Many

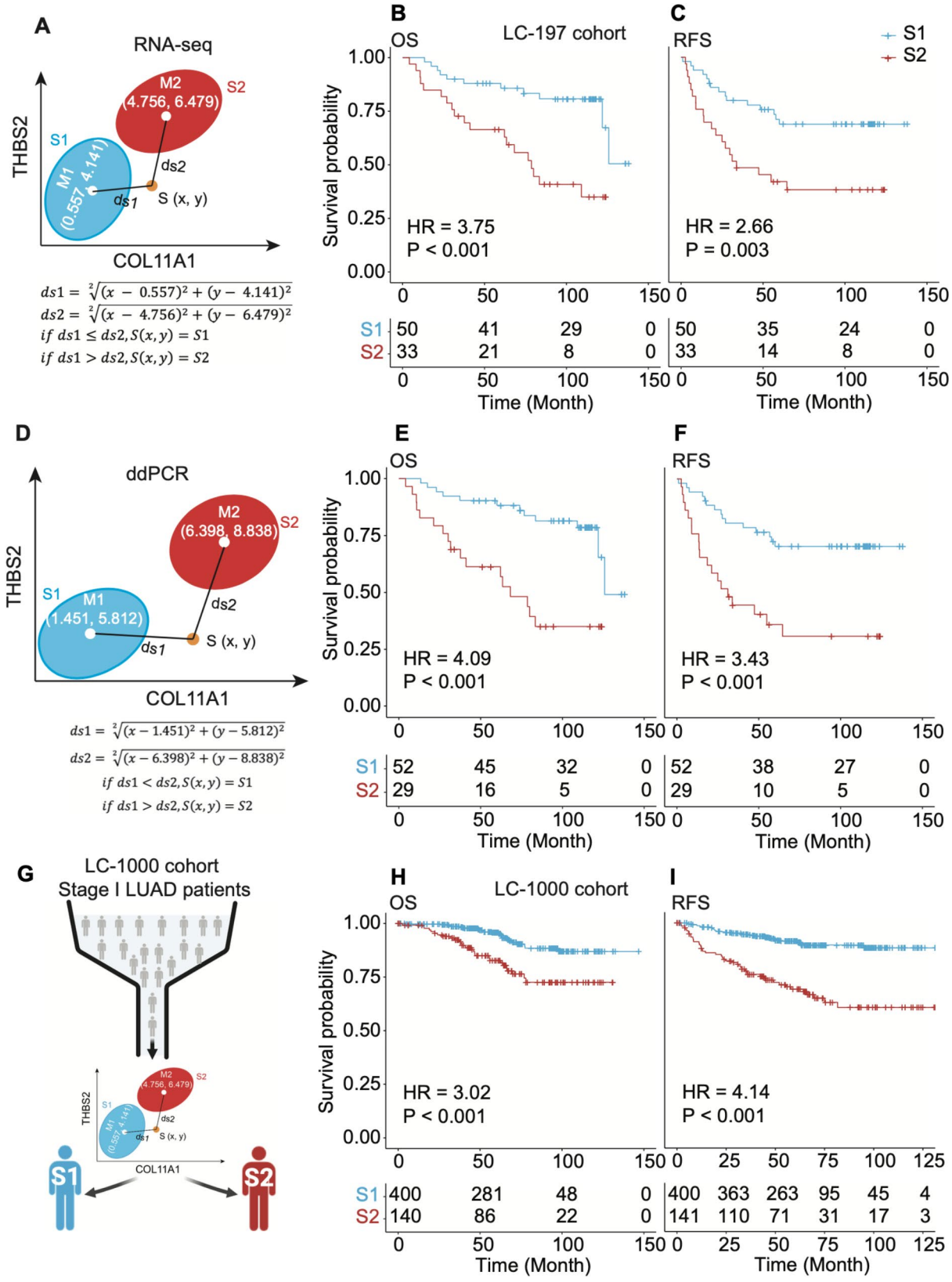

Fig. 6 (See legend on next page.)

(See figure on previous page.)

Fig. 6 Schematic diagram of assigning a single sample into S1 or S2 subgroup. The centroids of S1 and S2 at RNAseq dataset (**A**) and ddPCR (**D**) in the FUSCC dataset are M1 and M2, respectively. Survival analysis of FA2-based subtypes (S1 and S2) for 83 stage I patients' samples quantified by RNA-seq. S1 showed significantly better OS (**B**) and RFS (**C**) than S2 for stages I ($P < 0.05$). Survival analysis of FA2-based subtypes (S1 and S2) for 81 stage I patients' samples quantified by ddPCR. S1 showed significantly better OS (**E**) and RFS (**F**) than S2 for stages I ($P < 0.05$). (**G**) A schematic representation of an individual patient classified as either S1 or S2 subtype based on the expression levels of COL11A1 and THBS2 on a case-by-case basis. Survival analysis of FA2-based subtypes (S1 and S2) for 541 stage I patients from a prospective data set. S1 showed significantly better OS (**H**) and RFS (**I**) than S2 for stages I ($P < 0.001$)

studies found that TME played an important role in the development of tumor invasion [35, 36]. CAF, an important component in the TME, is distributed among tumor cells to provide a beneficial tumor stroma [22]. It was reported that CAF promoted cancer invasion by remodeling ECM [22, 37]. Coincidentally, CAF level was almost not changed among paired normal, AIS/MIA and S1, but there was obvious activation in S2 (Figs. 3H and 4B). Furthermore, the ECM-receptor interaction pathway was more active in S2 than in S1 (Figures S4B and 5D). These results indicated that CAF may promote early LUAD invasion by remodeling ECM. In fact, several clinical trials of targeted therapeutic for the interaction between CAF and tumor have been initiated [22]. Our results suggested that S2 patients may benefit from drugs that attenuate CAF activation.

Compared with published models, the FA2 model proposed in our study consisted of the least number of genes (two), but performed better than most published gene signatures (Figures S12 and S16). Furthermore, the prognostic predictive power of FA2 was better than pathological staging and FA2 was an independent prognosis predictor (Fig. 6, S11 and S15). These results suggested that FA2 had the potential to complement pathological staging. However, there were some limitations in the evaluation of the performance of FA2 compared to published gene signatures. The best evaluation approach would be combining the published gene signatures with the corresponding classification methods, and then evaluating the performance of each published signature-classification method pair. However, it is very difficult if not impossible for us to completely replicate the published method [7]. Therefore, we compared the prognostic predictive performance of published gene signatures using the same widely adopted unsupervised clustering algorithm of PAM for its ability of finding the optimal number of clusters underlying a dataset.

In conclusion, we applied PAM consensus clustering with COL11A1 and THBS2 expression to classify stage I LUAD into S1 (AIS/MIA-like) and S2 (CAF-rich) subtypes, which showed clear differences in multi-omics, tumor microenvironment, and clinical outcomes. The molecular classification of stage I LUAD showed good prognosis predictive performance, which may provide more precise management of these patients in clinical practice. Combining IA and S1 would help to identify patients who may be truly at low risk to avoid overtreatment. S2 (CAF-rich) with higher TMB and CAF may benefit from adjuvant therapies, such as immunotherapy and CAF suppression therapy. Meanwhile, prospective studies and functional or mechanistic experiments need to be completed to further verify our conclusions. Nevertheless, our simple and robust FA2 model may serve as a foundation toward reliable identification of high-risk stage I LUAD patients for more intensive post-surgery treatment.

Methods

Patients

During September 2011 and May 2016, we collected tumor-normal matched samples from a total of 197 patients during surgery, including AIS, MIA, IA, IB, and IIIA. No patient received neoadjuvant therapy before surgery. A decision to take the patient to surgery was reached based on pre-surgical examinations and TNM stage of the disease. For patients with a newly detected lung mass on CT scan, clinical stage was determined first. To determine the clinical stage before surgery, patients received enhanced chest computed tomography (CT) scanning and positron emission tomography-CT (PET-CT) scanning. Those patients who did not PET-CT scanning received an enhanced brain CT or magnetic resonance imaging, whole-body bone scanning, and abdominal CT or ultrasonography. Fiber optic bronchoscopy was routinely performed. For patients with newly detected ground-glass opacity (GGO), surgery would be performed on patients whose lung nodules were highly suspected for invasive lung cancer, otherwise we would recommend regular follow-up. Follow-up period would be based on the size and solid component of the GGO. Surgery will be performed later if the size or solid component of the GGO grows during follow-up. RFS and OS time was recorded according to clinical or telephone follow-up. This study has been approved by the research ethics review committee of Fudan University Shanghai Cancer Center (FUSCC) Institutional Review Board (No. 090977-1).

The LC-1000 cohort was derived from our CTLX study (https://www.researchsquare.com/article/rs-4977481/v1), which collected 1,008 lung adenocarcinoma samples from 954 patients and conducted whole genome and whole transcriptome sequencing. Detailed follow-up information was available for all patients. In this study, we selected patients with pathological stage I lung

adenocarcinoma from the CTLX cohort. Using quantitative RNA-seq results, we applied our FA2 model on a sample-by-sample basis to classify these patients into S1 or S2 subtypes. Ultimately, 541 patients with pathological stage I lung adenocarcinoma were categorized into the S1 and S2 subtypes. The CTLX cohort was generated after our FA2 model was locked-down, making it comparable to a prospective cohort. This enables us to fully validate the robustness of our FA2 model using the CTLX cohort.

RNA-seq data analysis

Hisat2-StringTie pipeline was used to obtain expression profiles from raw FASTA data [38]. Trimmomatic (v0.36) was used to remove adapters in the raw RNA-seq reads [39]. The quality of raw RNA-seq reads was assessed through FastQC (v0.11.5). FastQ Screen (v0.11.0) was used to evaluate whether there was contamination from other species in RNA-seq reads. We used Hisat2 (v2.0.5) to align reads to the human reference genome (GRCh38, release-84), which was downloaded from GDC [38]. The reads aligned to the human reference genome were assembled by StringTie (v1.3.3) and annotated as transcripts or genes by genome annotation file (gencode.v22.annotation.gtf) [38]. Finally, Fragments Per Kilobase of exon model per Million mapped fragments (FPKM) was used to measure gene expression.

Mutation profiling

As described in the previous study [13], the gene mutation data were generated through whole-exome sequencing (WES). In this study, we continued to use the results of the previous analysis. The TMB- and APOBEC-related mutation data also came from the previous study and can be downloaded from the supplementary information (https://doi.org/10.1038/s41467019-13460-3).

Single cell data analysis

Data for single cells were obtained from published study [40] and re-analyzed after we acquired count data for each cell in each sample. Using Seurat (v5.0.0) [41] and Harmony(v1.2.0) [42], we processed the single-cell data as follows:

1. The entire count matrix was normalized using default parameters, and the top 2,000 variable genes were selected for scaling and PCA.
2. Harmony was applied with default settings to remove batch effects between samples.
3. A K-nearest neighbor graph was constructed based on the top 30 Harmony-corrected PCs, with a resolution of 0.1 set for subpopulation analysis of all cells.
4. CAF subpopulations were identified using the CAF marker gene COL1A1.
5. Steps 1–3 were repeated to further identify subpopulations within CAF cells.
6. Genes specifically expressed in each CAF subpopulation were identified using the FindMarkers function, with validation criteria set at log2FC > 1 and $P < 0.05$.

Processing of publicly available datasets

Data collection was conducted from October 2020 to March 2021. All gene mutation data of Gillette et al. [23]. were downloaded from GDC [24]. All gene expression microarray data and corresponding clinical phenotypes were obtained from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/). Gene symbols were used to represent genes from different platforms. If there were multiple probes corresponding to the same gene symbol, the one with the highest signal intensity was used to represent the expression level of the corresponding gene. RNA-seq gene expression datasets (TCGA and Gillette et al. [23]). and corresponding clinical phenotypes were downloaded from GDC. The proteomic data of the two studies used in our analysis were obtained as follows. Normalized protein and phosphorylated protein expression data of Gillette et al. were downloaded from Clinical Proteomic Tumor Analysis Consortium (CPTAC) [25]. Normalized protein expression data of Xu et al. [26]. were downloaded as an attachment table of the article. As shown in Table S2, our study used 14 published datasets including two genomic datasets, 11 microarray gene expression datasets, two RNA-seq transcriptomics datasets, two proteomics datasets, and one phosphoproteomics dataset. Summary of the published datasets was shown in Table S2.

Differential gene expression and mutation analysis

R package limma (v3.42.2) was used to perform differential expression analysis between normal, pre-invasive, and invasive LUAD, or between S1 and S2 subtypes [43]. The commonly used cutoffs ($P < 0.05$,|log2(fold change)| >= 1) were used to identify differentially expressed genes [27]. Fisher's exact ($P < 0.05$) was used to identify differentially mutated genes between pre-invasive and invasive LUAD or between S1 and S2 subtypes.

KEGG and SsGSEA analysis

R package clusterProfiler (v3.14.3) was used to perform KEGG pathway enrichment analysis [44]. Focal adhesion pathway genes and hallmark gene sets were downloaded from MSigDB. Pathways significantly enriched with genes in an input set were identified by adjusted *P*-value ($P < 0.05$). R package GSVA [45](v1.34.0) with default gsva method was used to estimate gene-set enrichment score.

Partition around medoids (PAM)

Unsupervised clustering using the partition around medoids (PAM) cluster algorithm and Euclidean distance was performed through R package ConsensusClusterPlus (v1.50.0) [46]. The two genes (COL11A1 and THBS2) in the FA pathway, which were commonly identified both as DEGs or DMGs between AIS/MIA and invasive LUAD, were used as features for consensus clustering. R package factoextra (v1.0.7) was used to count average silhouette width [15] and choose the optimal number of clusters in a dataset. The consensus matrix with K = 2 was selected for further analysis after the number of clusters was evaluated from 2 to 10.

Assessment of tumor microenvironment

EPIC [17] and MCP-counter [18] were used to identify the composition and density of cells based on gene expression of each sample. R package immunedeconv (v2.0.3) [48] was used to perform EPIC and MCP-counter functions.

ddPCR

For ddPCR, synthesis of cDNA utilized 500 ng of RNA, which was reverse-transcribed using the HiScript III RT SuperMix (+gDNA wiper) kit (Vazyme, no.R323-01) according to the manufacturer's protocol. Droplets for each cDNA sample were generated in triplicate using 25 ng of cDNA, 900 nM of primers, 250 nM of probes in 1× ddPCR Supermix for Probes (No dUTP) (Bio-Rad, no.1863024) and 55 uL Droplet Generation Oil for Probes (Bio-Rad, no.1863005) on a QX200 Droplet Generator (Bio-Rad), followed by PCR amplification on a T100 Thermal Cycler (Bio-Rad). Two positive control (Human Brain Reference RNA) [27, 28] and one negative control (UltraPure Distilled Water, Invitrogen, no.2085372) were added to each plate to monitor the inter-plate assay performance. The sequences of primers and Taqman probes used for the quantitation of COL11A1 and THBS2 transcripts are listed in **Figure S17A.** Primers and Taqman probes were designed by Beacon Designer (v8.14) and validated by Oligo 7 and Primer Premier 5. The high specificity of primers was confirmed by Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) and capillary gel electrophoresis on a Qsep100 Advanced (Bioptic lnc.) (Figures S17B-E). We found that 52℃ was the best extension temperature which can best distinguish the positive and negative droplets (Figure S18). PCR cycling conditions included an initial enzyme activation step for 10 min at 95 °C, followed by 40 cycles of denaturation at 94 °C for 30 s and extension at 52 °C for 1 min with a 1.5 °C per second ramp rate and an enzyme deactivation step at 98 °C for 10 min. Following PCR amplification, droplets containing the transcripts of interest were detected via fluorescence (FAM and HEX) with the QX200 Droplet Reader System (Bio-Rad), and the QuantaSoft software (v1.7.4, Bio-Rad) was used to calculate the concentration (copies/uL) of COL11A1 and THBS2. The log2-transformed concentration of COL11A1 and THBS2 was used for performing unsupervised clustering.

IHC

The LUAD formalin-fixed paraffin-embedded (FFPE) specimens from 2011 to 2016 at FUSCC were collected. FFPE slides (4 μm) were heated for 2 h at 65°C and then submerged twice in xylene for 15 min. Slides were then gradually re-hydrated by submerging for 5 min in each of the following ethanol solutions: 100%, 85%, 75%, and finally in ddH2O for 5 min. Antigen retrieval was performed in citrate buffer (pH 6.0) by the microwave oven. Next, 3% hydrogen peroxide in ddH2O was utilized to quench endogenous peroxidase activity, followed by interdiction for nonspecific binding. Mouse anti-ACTA2 (1:1000, no. Abcam, ab7817) were incubated with the slides overnight in a moist chamber at 4℃. After washing in PBS, slides were treated with a secondary antibody, stained by a 3,3'-diaminobenzidine (DAB) system, and counterstained with hematoxylin (DAKO, no. K5007). The expression status of immunostaining was reviewed and scored independently by two pathologists based on the whole tumor bed (ACTA2).

Statistical analysis

All statistical analysis was performed with R (v3.6.3). Statistical tests included t-test, Fisher's exact test, and Pearson correlation. R package survival (v3.1-8) and survminer (0.4.8) were used to perform survival and Cox analysis. Kaplan-Meier survival analysis combined with log-rank test was used for overall survival (OS) and relapse-free survival (RFS) analysis. R package ComplexHeatmap (v2.2.0) was used to draw heatmaps [48]. Principal component analysis (PCA) was performed using R package stats (v3.6.3). Oncoplot and lollipop plots were performed with maftools (v2.6.05) [49]. Amplification peaks were identified by GISTIC2.0. Boxplots and scatter plots were drawn with R package ggpubr (v0.4.0) and ggplot2 (v3.3.3).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-025-06316-6 .

Supplementary Material 1

Acknowledgements

We thank Biorender.com for creating Fig. 1.

Authors' contributions
YZheng, HC, LS, JS and HJ conceived the study. JS, HJ and JY analyzed the data. HJ and JS performed the ddPCR experiment and analyzed the data. YL, HJ, YCL, and JS performed the IHC experiment and analyzed the data. YZhao collected and interpreted clinical data. NZ, LR, QC, and YY participated in the verification and interpretation of the data. JS and HJ drafted the manuscript. LS, HC and YZheng revised the manuscript and supervised the work. All authors reviewed and approved the manuscript.

Funding
This study was supported in part by the State Sponsored Postdoctoral Fellowship Programme (GZC20230501), National Key R&D Program of China (2018YFE0201603), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01) and National Natural Science Foundation of China (31720103909).

Data availability
WES and RNA-seq raw data of LC-197 have been deployed in the National Omics Data Encyclopedia (NODE) (https://www.biosino.org/node) with the accession number OEP000325. RNA-seq raw data of LC-1000 have been deployed in the NODE with the accession number OEP002580. RNA-seq data of Gillette et al. and TCGA LUAD were downloaded from GDC (https://portal.gdc.cancer.gov/). Normalized protein and phosphorylated protein expression data of Gillette et al. were downloaded from CPTAC (https://cptac-data-portal.georgetown.edu/study-summary/S056). All gene expression microarray data and corresponding clinical phenotypes of Table S2 were obtained from GEO (https://www.ncbi.nlm.nih.gov/geo/). Source data is provided in the supplementary tables of this paper.

Declarations

Ethics approval and consent to participate
This study has been approved by the research ethics review committee of Fudan University Shanghai Cancer Center (FUSCC) Institutional Review Board (No. 090977-1).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing financial interests.

Author details
[1]Departments of Thoracic Surgery and State Key Laboratory of Genetic Engineering, Fudan University Shanghai Cancer Center, Shanghai, China
[2]Institute of Thoracic Oncology, Fudan University, Shanghai, China
[3]State Key Laboratory of Genetic Engineering, School of Life Sciences, Human Phenome Institute and Shanghai Cancer Center, Fudan University, Shanghai, China
[4]Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai, China
[5]Cancer Institute, Shanghai Cancer Center, Fudan University, Shanghai, China
[6]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China
[7]International Human Phenome Institutes (Shanghai), Shanghai, China

Received: 10 December 2024 / Accepted: 23 February 2025
Published online: 04 March 2025

References

1. Goldstraw P, et al. The IASLC lung Cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (Eighth) edition of the TNM classification for lung Cancer. J Thorac Oncology: Official Publication Int Association Study Lung Cancer. 2016;11:39–51.
2. Yotsukura M, et al. Long-Term prognosis of patients with resected adenocarcinoma in situ and minimally invasive adenocarcinoma of the lung. J Thorac Oncol. 2021;16:1312–20.
3. Travis WD, et al. International association for the study of lung Cancer/american thoracic Society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncology: Official Publication Int Association Study Lung Cancer. 2011;6:244–85.
4. Chen T, et al. Should minimally invasive lung adenocarcinoma be transferred from stage IA1 to stage 0 in future updates of the TNM staging system? J Thorac Dis. 2018;10:6247–53.
5. Strauss GM, et al. Adjuvant Paclitaxel plus carboplatin compared with observation in stage IB non-small-cell lung cancer: CALGB 9633 with the Cancer and leukemia group B, radiation therapy oncology group, and North central Cancer treatment group study groups. J Clin Oncology: Official J Am Soc Clin Oncol. 2008;26:5043–51.
6. Pignon JP, et al. Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE collaborative group. J Clin Oncology: Official J Am Soc Clin Oncol. 2008;26:3552–9.
7. Tang H, et al. Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies. Annals Oncology: Official J Eur Soc Med Oncol. 2017;28:733–40.
8. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst. 2010;102:464–74.
9. Kratz JR, et al. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. Lancet (London England). 2012;379:823–32.
10. Shi L, et al. The microarray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010;28:827–38.
11. Kadara H, et al. A five-gene and corresponding protein signature for stage-I lung adenocarcinoma prognosis. Clin cancer Research: Official J Am Association Cancer Res. 2011;17:1490–501.
12. Bianchi F, et al. Survival prediction of stage I lung adenocarcinomas by expression of 10 genes. J Clin Investig. 2007;117:3436–44.
13. Chen H, et al. Genomic and immune profiling of pre-invasive lung adenocarcinoma. Nat Commun. 2019;10:5472.
14. Kaufman L, Rousseeuw P. Clustering by means of medoids. 1987;405–16.
15. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Wiley. 2009.
16. Yam JWP, Tse EYT, Ng IO-L. Role and significance of focal adhesion proteins in hepatocellular carcinoma. J Gastroenterol Hepatol. 2009;24:520–30.
17. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* 2017;6:e26476.
18. Becht E, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17:218.
19. Iwai M, et al. Cancer-associated fibroblast migration in non-small cell lung cancers is modulated by increased integrin A11 expression. Mol Oncol. 2021;15:1507–27.
20. Zhou Z, et al. VCAM-1 secreted from cancer-associated fibroblasts enhances the growth and invasion of lung cancer cells through AKT and MAPK signaling. Cancer Lett. 2020;473:62–73.
21. Lee S, et al. Cancer-associated fibroblasts activated by miR-196a promote the migration and invasion of lung cancer cells. Cancer Lett. 2021;508:92–103.
22. Sahai E, et al. A framework for advancing our Understanding of cancer-associated fibroblasts. Nat Rev Cancer. 2020;20:174–86.
23. Gillette MA, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell. 2020;182:200–e225235.
24. Heath AP, et al. The NCI genomic data commons. Nat Genet. 2021;53:257–62.
25. Edwards NJ, et al. The CPTAC data portal: A resource for Cancer proteomics research. J Proteome Res. 2015;14:2707–13.
26. Xu JY, et al. Integrative proteomic characterization of human lung adenocarcinoma. Cell. 2020;182:245–e261217.
27. Shi L, et al. The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006;24:1151–61.
28. Su Z, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. Nat Biotechnol. 2014;32:903–14.
29. Wistuba II, et al. Validation of a proliferation-based expression signature as prognostic marker in early stage lung adenocarcinoma. Clin cancer Research: Official J Am Association Cancer Res. 2013;19:6261–71.

30. Li A, Li J, Lin J, Zhuo W, Si J. COL11A1 is overexpressed in gastric cancer tissues and regulates proliferation, migration and invasion of HGC-27 gastric cancer cells in vitro. Oncol Rep. 2017;37:333–40.
31. Shen L, et al. COL11A1 is overexpressed in recurrent non-small cell lung cancer and promotes cell proliferation, migration, invasion and drug resistance. Oncol Rep. 2016;36:877–85.
32. Wu YH, Chang TH, Huang YF, Huang HD, Chou CY. COL11A1 promotes tumor progression and predicts poor clinical outcome in ovarian cancer. Oncogene. 2014;33:3432–40.
33. Wu YH, Huang YF, Chang TH, Chou CY. Activation of TWIST1 by COL11A1 promotes chemoresistance and inhibits apoptosis in ovarian cancer cells by modulating NF-κB-mediated IKKβ expression. Int J Cancer. 2017;141:2305–17.
34. Rizvi H, et al. Molecular determinants of response to Anti-Programmed cell death (PD)-1 and Anti-Programmed death-Ligand 1 (PD-L1) Blockade in patients with Non-Small-Cell lung Cancer profiled with targeted Next-Generation sequencing. J Clin Oncology: Official J Am Soc Clin Oncol. 2018;36:633–41.
35. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. Nat Med. 2013;19:1423–37.
36. Chaffer CL, Weinberg RA. A perspective on cancer cell metastasis. Sci (New York N Y). 2011;331:1559–64.
37. Kechagia JZ, Ivaska J, Roca-Cusachs P. Integrins as Biomechanical sensors of the microenvironment. Nat Rev Mol Cell Biol. 2019;20:457–73.
38. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, stringtie and ballgown. Nat Protoc. 2016;11:1650–67.
39. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30:2114–20.
40. Deng Y et al. Multicellular ecotypes shape progression of lung adenocarcinoma from ground-glass opacity toward advanced stages. Cell Rep Med. 2024;101489.
41. Hao Y, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024;42:293–304.
42. Korsunsky I, et al. Fast, sensitive and accurate integration of single-cell data with harmony. Nat Methods. 2019;16:1289–96.
43. Ritchie ME, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
44. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: an R package for comparing biological themes among gene clusters. Omics. 2012;16:284–7.
45. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
46. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics. 2010;26:1572–3.
47. Sturm G, et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. Bioinformatics. 2019;35:i436–45.
48. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32:2847–9.
49. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28:1747–56.

Publisher's note