Genes & Diseases (2023) 10, 1657–1674

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/en/journals/genes-diseases

FULL LENGTH ARTICLE

Hallmark guided identification and characterization of a novel immune-relevant signature for prognostication of recurrence in stage I–III lung adenocarcinoma

Yongqiang Zhang [a,b,1], Zhao Yang [a,1], Yuqin Tang [c,1], Chengbin Guo [d], Danni Lin [d], Linling Cheng [d], Xun Hu [e,f,**], Kang Zhang [a,d,*], Gen Li [b,***]

[a] *West China School of Medicine, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China*
[b] *Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, Guangdong 510620, China*
[c] *State Key Laboratory of Southwestern Chinese Medicine Resources, School of Basic Medical Sciences, Chengdu University of Traditional Chinese Medicine, Chengdu, Sichuan 611137, China*
[d] *Center for Biomedicine and Innovations, Faculty of Medicine, Macau University of Science and Technology, Macau 999078, China*
[e] *Clinical Research Center, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310003, China*
[f] *Biorepository, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China*

Received 12 March 2022; received in revised form 7 June 2022; accepted 16 July 2022
Available online 5 August 2022

KEYWORDS
Immune infiltration;
Lung adenocarcinoma;

Abstract The high risk of postoperative mortality in lung adenocarcinoma (LUAD) patients is principally driven by cancer recurrence and low response rates to adjuvant treatment. Here, A combined cohort containing 1,026 stage I–III patients was divided into the learning (n = 678) and validation datasets (n = 348). The former was used to establish a 16-mRNA risk signature

* Corresponding author. Center for Biomedicine and Innovations, Faculty of Medicine, Macau University of Science and Technology, Macau 999078, China.
** Corresponding author. Clinical Research Center, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310003, China.
*** Corresponding author. Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, Guangdong 510620, China.
E-mail addresses: hxxhu99@163.com (X. Hu), kang.zhang@gmail.com (K. Zhang), superleegen@hotmail.com (G. Li).
Peer review under responsibility of Chongqing Medical University.
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.gendis.2022.07.005
2352-3042/© 2022 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Nomogram;
Recurrence;
Risk signature

for recurrence prediction with multiple statistical algorithms, which was verified in the validation set. Univariate and multivariate analyses confirmed it as an independent indicator for both recurrence-free survival (RFS) and overall survival (OS). Distinct molecular characteristics between the two groups including genomic alterations, and hallmark pathways were comprehensively analyzed. Remarkably, the classifier was tightly linked to immune infiltrations, highlighting the critical role of immune surveillance in prolonging survival for LUAD. Moreover, the classifier was a valuable predictor for therapeutic responses in patients, and the low-risk group was more likely to yield clinical benefits from immunotherapy. A transcription factor regulatory protein—protein interaction network (TF-PPI-network) was constructed via weighted gene co-expression network analysis (WGCNA) concerning the hub genes of the signature. The constructed multidimensional nomogram dramatically increased the predictive accuracy. Therefore, our signature provides a forceful basis for individualized LUAD management with promising potential implications.
© 2022 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Lung cancer is still one of the most frequently diagnosed malignancies worldwide, with the second-highest incidence of all kinds of cancers (more than 2.2 million new cases in 2020), and it ranks the leading cause of cancer mortality in most countries.[1] Lung adenocarcinoma (LUAD) is the predominant histological subtype of lung cancer, accounting for approximately 40% of all cases.[2] Despite the substantial progress in therapeutic strategies such as radiation, chemotherapy, molecularly targeted agents, and immunotherapy in recent years, curative resection is still the standard and main approach in LUAD treatment. However, due to local recurrence or distant metastasis, the postoperative 5-year survival rate of LUAD remains disappointing. Currently, the American Joint Committee on Cancer (AJCC) tumor node metastasis (TNM) cancer staging system is widely used to assess the recurrence risk of LUAD, and provides valuable information for patient prognosis after surgical resection. Nevertheless, due to the tumor heterogeneity on both histological and molecular levels,[3–5] conflict clinical outcomes are frequently appeared even in patients of same LUAD stage. Thus, new prognostic biomarkers are urgently needed for risk stratification to predict the recurrence of LUAD.

Progressions in molecular oncology have facilitated the investigations of candidate predictors to distinguish different risk groups, especially at the mRNA level. Numbers of gene expression signatures have been previously identified to classify cancer patients into distinct risk groups, including breast cancer,[6–8] liver cancer,[9–13] colon cancer,[14,15] and lung cancer.[16–21] Interestingly, previously proposed gene signatures may also predict therapeutic resistance in cancer. For instance, a *TP53*-associated gene signature in lung squamous cell carcinoma was linked to different sensitivities to chemotherapy and targeted therapy, as well as immune checkpoint blockade (ICB) (targeting *CTLA4* and *PD-1*).[22] Similar conclusions were reached for an m^{6}A-related gene signature in LUAD.[23] However, very few studies have focused on the recurrence outcome rather than overall survival (OS) as the endpoint.[17,18,24] Besides, most of these efforts used differentially expressed genes of tumor samples (in comparison with normal samples) or limited gene sets (such as immune-related genes) as targets in prognostic signatures screening. As a result, significant genes would be neglected if they were beyond these criteria. Gene set enrichment analysis (GSEA) is a powerful computational approach to determine statistically significant gene sets between two different disease statuses[25] and thus may provide novel insight for the development of new prognostic biomarkers. Here we focused on assessing the usefulness of hallmark guided GSEA for the identification of transcriptomic signature predicting cancer recurrence with multiple statistical approaches. At first, a hallmark guided gene prognostic signature (HGGPS) composed of 16 mRNA was developed and validated based on a combined cohort of 1026 LUAD patients, followed by the comprehensive analysis to screen out the differences of genome alterations, hallmark pathways, immune checkpoints, the composition of infiltrating immune cells, and drug responses regarding the high- and low-risk groups. We also identified a transcription factor regulatory protein—protein interaction-network (TF-PPI-network) to better elucidate the potential mechanisms leading to different clinical outcomes. Finally, a predictive nomogram was built for better discrimination of recurrence risk.

Materials and methods

Data source

We reviewed carefully the retrospective mRNA expression datasets deposited in public databases and conducted the strict inclusion/exclusion criteria: large sample size ($n > 50$); complete information for TNM stage, recurrent status, and recurrence-free survival time. Four eligible datasets from Gene Expression Omnibus (GEO) containing 576 stage I—III LUAD patients were selected, including GSE31210[26] ($n = 226$), GSE50081[27] ($n = 125$), GSE41271[28] ($n = 173$), and GSE37745[29] ($n = 52$). Gene expression profilings were obtained from two high-throughput microarray platforms: Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) and Illumina HumanWG-6 v3.0 expression

beadchip (GPL6884). We also achieved the Cancer Genome Atlas (TCGA)-LUAD expression dataset with 450 stage I–III patients from the UCSC XENA project (https://xena.ucsc.edu), as well as the corresponding clinicopathological information. The corresponding clinicopathologic characteristics were shown in Table S1.

All of the 1026 subjects were merged into one single cohort and subsequently randomly divided into the learning dataset ($n = 678$) and the validation dataset ($n = 348$) with the approximate ratio of 2:1. Considering the non-recurrent patients ($n = 659$) were predominant versus recurrent patients ($n = 367$), we verified that both states were well balanced in the learning set (1.94:1) and the validation set (1.96:1). The learning set was used to screen recurrence-free survival (RFS) -related genes and to train the RFS prognostic signature, and the validation set was applied to validate the robustness of the signature. The baseline clinical features in each dataset were described in Table S2.

The mutation annotation format (MAF) data including the somatic mutation (single nucleotide polymorphisms and small insertion-deletion polymorphisms) from 443 of 450 cases were acquired from the TCGA data portal (http://tcga-data.nci.nih.gov/tcga/) and analyzed with R package "maftools". Tumor mutation burden (TMB) was computed and compared between distinct groups. GISTIC 2.0 was used to detect the significant deletions or amplifications with copy number variation (CNV) profiles. Among the 450 TCGA LUAD patients, 399 included the paired DNA methylation data generated by the Infinium HumanMethylation450 beadchip platform, and the methylation value ranging from 0 to 1 for each probe was downloaded from the UCSC XENA.

Gene expression data preprocessing

All gene expression data were preprocessed with R software (version 3.6.0, https://www.r-project.org/), except for the exact matches of probes to gene symbols for microarray expression sets, which was determined by using a Perl script. Briefly, for microarray datasets, the processed format of expression profiles was obtained from the GEO database and then log 2 transformed. Next, all probes for each dataset were annotated with gene symbols according to its platform, using a Perl script. If multiple probes were mapped to one unique gene, the average value of all probes was calculated and retained. For the TCGA LUAD dataset, level three expression profile was derived from the UCSC XENA browser with normalized RSEM format (log 2 transformed). Subsequently, we conducted a z-score transformation to normalize each gene value across all samples, and the batch effects were removed with ComBat function of the sva package.[30]

Establishment of a prognostic signature

Three phases were included to establish and validate the recurrence-related prognostic signature. In the discovery phase, 678 LUAD patients from the learning set were firstly assigned to two groups according to their recurrent status. Gene set enrichment analysis (GSEA) with the command line version was applied to identify significant hallmark gene sets between the two groups and h.all.v7.1.symbols gene set was selected as the reference database.[25,31] Permutations were employed 1000 times for each analysis. A maximum Q-value (FDR) of 0.25 was set as the significant cutoff as recommended. Besides, the candidate genes that appeared in the significant gene sets were further subject to univariate Cox (UniCox) regression to estimate their prognostic values for RFS, allowing the identification of RFS prognostic genes (Wald test, $P < 0.05$).

In the training phase, we performed three steps for feature selection on the predefined RFS prognostic genes. First, we conducted the popular support vector machine (SVM) -based algorithm -SVM recursive feature elimination (SVM-RFE) to narrow down the candidate genes (e1071 package). Second, the least absolute shrinkage and selection operator (LASSO) algorithm was taken 100 times to minimize overfitting risk (glmnet package). Only those genes that repeatedly occurred for more than 80 times were selected for further analysis. For each time, the optimal lambda parameter was calculated by ten-fold cross-validation, and the corresponding lambda, known as lambda.min was used to filter significant genes. At last, to make the model simpler and more practical, a stepwise backward variable selection was conducted by SPSS (version 23) with the default settings.

Based on the expression values of the selected prognostic genes in the learning dataset, a linear combination of risk signature was constructed to evaluate RFS or OS of LUAD patients. The risk score of each patient was calculated with the detailed formula: risk score $= \Sigma$ (regression coefficient $\times$ expression value of each selected mRNA). All patients of the learning dataset were then classified into two groups according to the median risk score. Kaplan–Meier curves combined with a log-rank test were used to assess the statistical significance between the two groups by the survival package. Time-dependent receiver operating characteristic (tROC) curve was plotted by timeROC package to measure the performance of predictive power. Besides, the validation dataset was used to estimate the robustness of the signature.

Additional bioinformatic and statistical analyses

The other detailed methods, including function enrichment analysis and gene set variation analysis (GSVA), differential analysis of checkpoint molecules and tumor infiltrating immune cells, WGCNA and hub genes identification, establishment of a protein–protein interaction (PPI) network with predicted transcription factors, construction of a methylation score system and nomogram, therapeutic response prediction in clinical treatment, and statistical analysis are described in the supporting material (supplementary materials and methods).

Results

Hallmark guided identification of prognostic genes for RFS of LUAD

A series of gene expression datasets were collected from the GEO database and TCGA, containing at least 52 LUAD patients, with complete information of RFS status and RFS time, ranging from TNM stage I to stage III (Table S1). To

establish an RFS prognostic classifier for LUAD patients, and to explore its underlying molecular mechanisms and clinical application, we adopted a novel integrated approach based on the hallmark gene sets using the GSEA algorithm (Fig. 1). We firstly split the merged cohort into the learning and validation datasets with the ratio of 2:1. Figure 2A and Table S2 indicate the balanced distributions of clinical characters for the two groups.

In the discovery phase, GSEA was preliminarily implemented to screen the recurrence-related hallmark gene sets. As a result, with the cutoff of FDR <0.25, GSEA identified seven hallmark gene sets that were significantly altered in the recurrent group, including COMPLEMENT (FDR = 0.144, NES = 2.768), ESTROGEN_RESPONSE_EARLY (FDR = 0.164, NES = 2.520), INTERFERON_ALPHA_RESPONSE (FDR = 0.169, NES = 2.169), UV_RESPONSE_UP (FDR = 0.178, NES = 2.780), INFLAMMATORY_RESPONSE (FDR = 0.190, NES = 2.948), KRAS_SIGNALING_DN (FDR = 0.178, NES = 2.230), and MYC_TARGETS_V1 (FDR = 0.202, NES = 2.237) (Fig. 2B and Table S3), implying that these hallmark pathways may play important roles during the recurrence of LUAD. Next, we performed a univariate Cox (UniCox) procedure to reduce the dimensionality of the 978 genes in the above pathways, and 249 genes were found to be significantly associated with the RFS of LUAD patients (Table S4). KEGG enrichment indicated these 249 genes were mainly involved in cell cycle, DNA replication, central carbon metabolism in cancer and other pathways, and well-known oncogenetic pathways including HIF-1 signaling pathway, PI3K-Akt signaling pathway, and Jak-STAT signaling pathway were also enriched (Fig. 2C).

Construction and validation of the RFS prognostic signature

In the training phase, we performed three approaches to narrow down candidate signature genes. Generally, we adopted a learning vector quantization (LVQ) method and obtained a ranking list of all the 249 prognostic genes by their importance (Table S5), and then the recursive feature

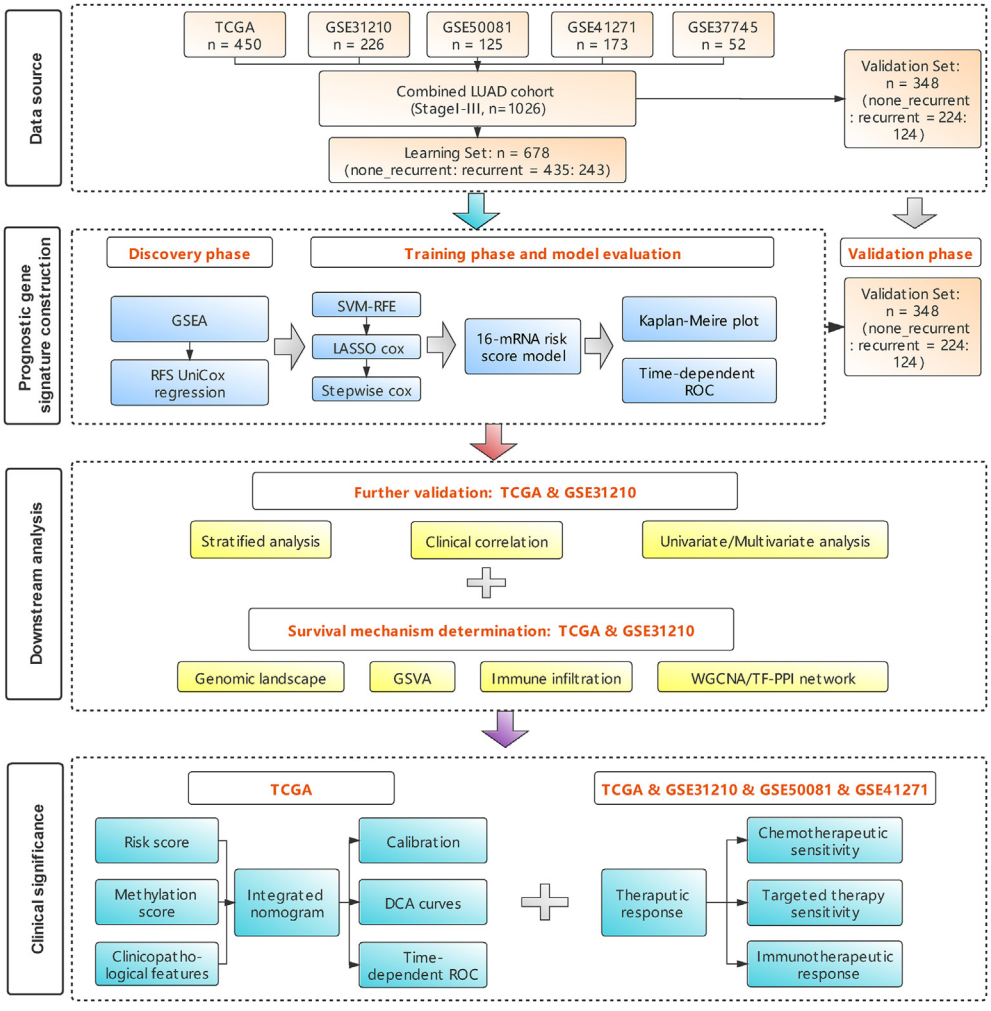

Figure 1 Schematic workflow of the study showing data collection and steps of analysis. DCA, decision curve analysis; GSEA, gene set enrichment analysis; GSVA, gene set variation analysis; LASSO, least absolute shrinkage and selection operator; LUAD, lung adenocarcinoma; RFS, recurrence-free survival; ROC, receiver operating characteristic; SVM-RFE, support vector machine recursive feature elimination; TCGA: The Cancer Genome Atlas; TF-PPI-network, transcription factor regulatory protein—protein interaction network; UniCox: univariate Cox; WGCNA, weighted gene co-expression network analysis.

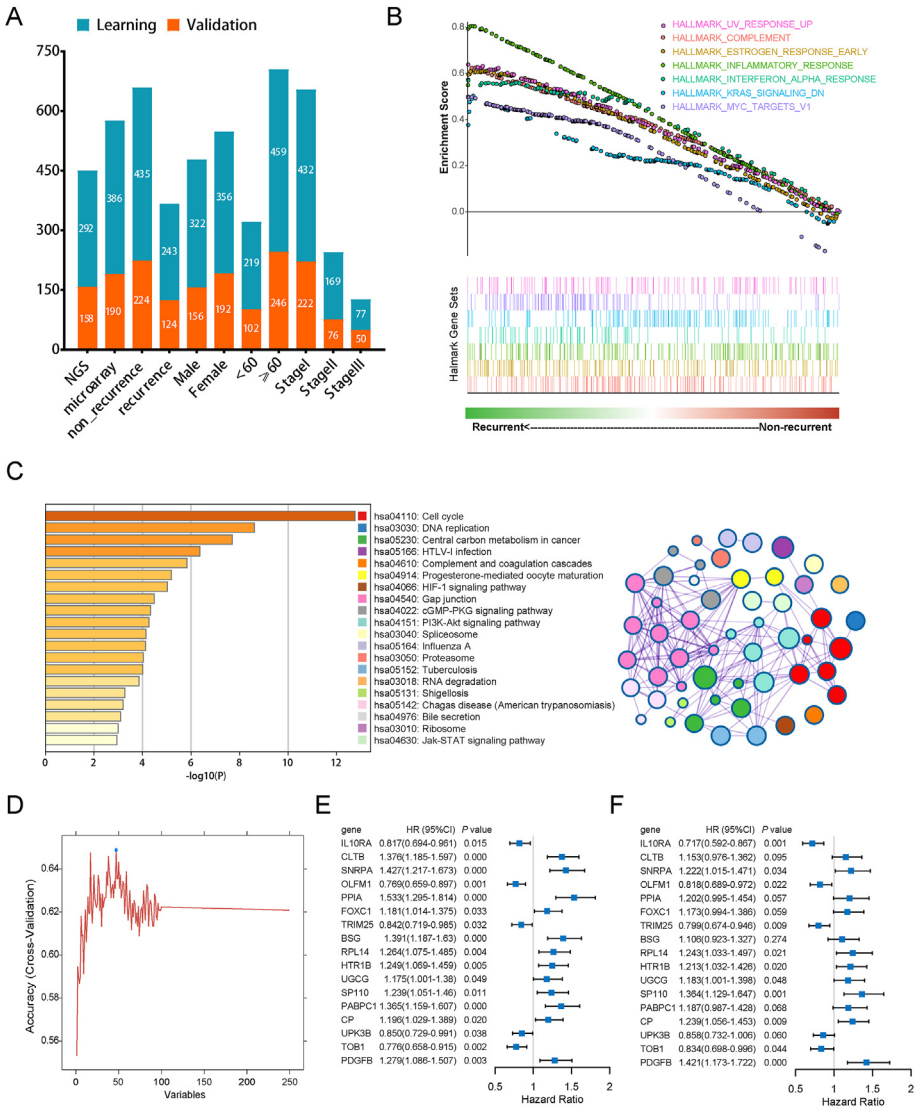

Figure 2 Establishment of a 16-mRNA risk signature for the prognostation of LUAD recurrence. **(A)** Unbiased division of the whole combined cohort into the learning and validation datasets. **(B)** GSEA result showed seven significant gene sets with the cutoff of FDR <0.25. **(C)** Top 20 KEGG pathways enriched by the 249 prognostic genes. The right panel shows the network of enriched pathways clustered by Kappa-statistical similarities. Terms with a similarity score of >0.3 are labeled with the same color and connected by an edge to form a cluster. The size of a node represents the number of genes that fall into the term. **(D)** A total of 47 candidate genes were filtered by SVM-RFE. **(E)** List of the 17 genes retained after LASSO Cox selection with forest plot presenting HR and corresponding *P* values for RFS of each gene by univariate Cox. **(F)** Forest plot showed HR and corresponding *P* values of the 17 genes with multivariate cox model using backward stepwise selection. (Default setting: 0.1). 95% CI, confidence interval; HR, hazard ratio; KEGG, Kyoto Encyclopedia of Genes and Genomes.

elimination procedure was taken with a random forest selection function, which resulted in 47 candidate genes for the following analysis (Fig. 2D and Table S5). Next, the LASSO algorithm was applied to further reduce the features by which 47 potential signature genes were shrunk to 17 (Fig. 2E), and 16 of them showed consistently significant in the stepwise Cox regression analysis: *IL10RA*, *CLTB*, *SNRPA*, *OLFM1*, *PPIA*, *FOXC1*, *TRIM25*, *RPL14*, *HTR1B*, *UGCG*, *SP110*, *PABPC1*, *CP*, *UPK3B*, *TOB1*, and *PDGFB* (Fig. 2F). Subsequently, an RFS prognostic signature (HGGPS) was built according to their expression values and the corresponding coefficients with the following formula: risk score = $(0.2245 \times EXP_{SNRPA}) + (0.3624 \times EXP_{PDGFB}) - (0.33927 \times EXP_{IL10RA}) + (0.2958 \times EXP_{SP110}) + (0.1884 \times EXP_{PPIA}) + (0.16785 \times EXP_{FOXC1}) + (0.1993 \times EXP_{HTR1B}) + (0.2187 \times EXP_{CP}) - (0.2232 \times EXP_{TRIM25}) - (0.2001 \times EXP_{OLFM1}) + (0.1608 \times EXP_{UGCG}) + (0.2210 \times EXP_{RPL14}) + (0.1704 \times EXP_{PABPC1}) - (0.1668 \times EXP_{UPK3B}) + (0.1654 \times EXP_{CLTB}) - (0.1831 \times EXP_{TOB1})$. The risk scores for all patients in the learning dataset were calculated with the formula. Using the median value of 1.018223, patients were designated as high- or low-risk (Fig. 3A). Figure 3B suggested that not only the risk scores were significantly different between the high- and low-risk groups ($P < 0.0001$), but also the

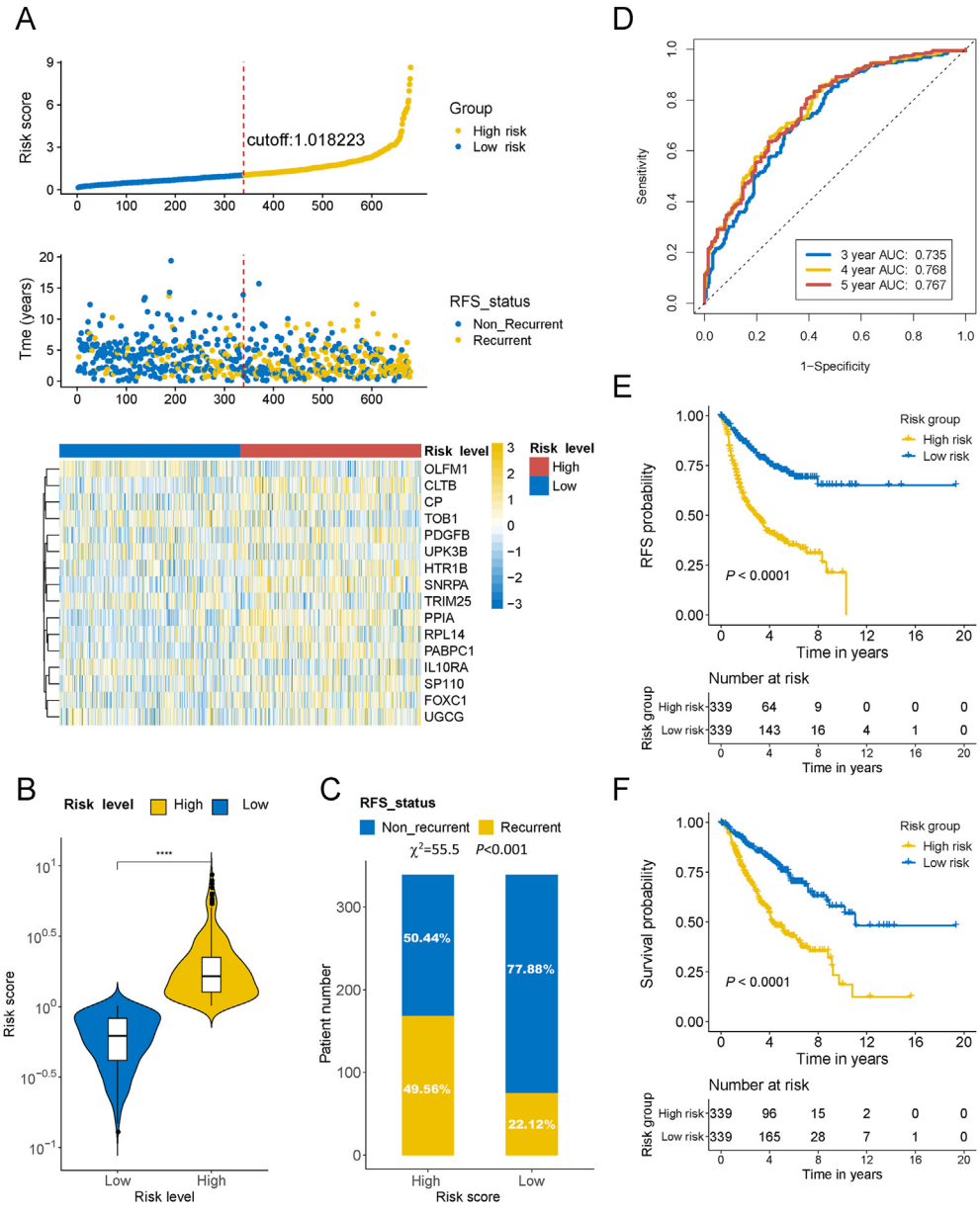

Figure 3 Prognostic value of the generated risk signature in the learning dataset. **(A)** Risk score, recurrence status, and gene expression profile for each patient. **(B)** Violin plot presenting the significant difference between the high- and low-risk groups. **(C)** Recurrence rates distribution between the high- and low-risk groups. **(D)** ROC curves at 1-, 3-, and 5-year of the signature for prediction of recurrence. **(E)** Kaplan—Meier plot of RFS in LUAD patients based on the risk score. **(F)** Kaplan—Meier plot of OS in LUAD patients based on the risk score. **** $P < 0.0001$.

recurrence rate was significantly higher in the high-risk group (Fig. 3C, $P < 0.0001$). The tROC showed good discriminatory capacities for 3-, 4-, and 5-year recurrence rates (Fig. 3D), with the area under the curve (AUC) of 0.735, 0.768, and 0.767, respectively. Kaplan—Meier curves of RFS or OS generated by risk score indicated a significant difference between high- and low-risk groups with the log-rank test (Fig. 3E, F; $P < 0.0001$ for both RFS and OS). Interestingly, although the risk signature was capable of classifying early stage (stage I and stage II) patients into high- or low-risk subgroups based on their difference in RFS or OS ($P < 0.0001$ for both RFS and OS), these trends were not significant in stage III patients (Fig. S1).

In the validation dataset, patients were divided into different risk groups (Fig. 4A) based on the same formula and cutoff value (1.018223). Similarly, different risk scores were observed in the two groups, as well as the recurrence rate (Fig. 4B, C). Kaplan—Meier survival analyses exhibited a significantly better RFS or OS in the low-risk patients than the high-risk patients (Fig. 4D, E; $P = 0.00047$ for RFS and $P = 0.0086$ for OS, respectively). Additionally, while the high-risk group had a significantly higher rate of recurrence (Fig. 4F; $P = 0.00063$) and mortality (Fig. S2; $P = 0.02$) than the low-risk group in the early stage, only a borderline difference for RFS or OS was observed for stage III patients (data not shown).

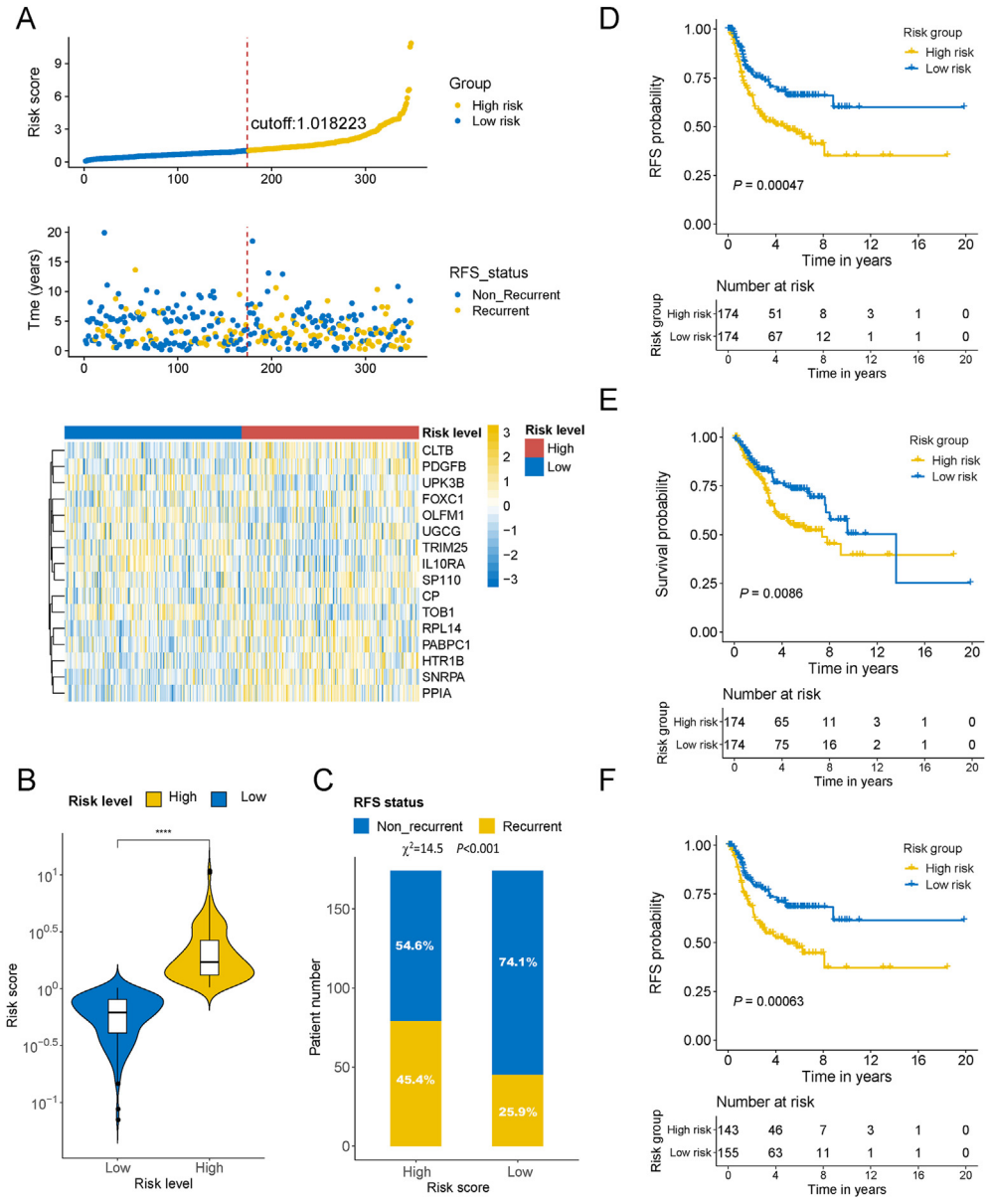

Figure 4 Verifying the robustness of the 16-mRNA risk signature using the validation dataset. **(A)** Risk score, recurrence status, and gene expression profile for each patient. **(B)** Violin plot showing the significant difference between the high- and low-risk groups. **(C)** Recurrence rate distribution between the high- and low-risk groups. **(D)** Kaplan–Meier plot of RFS in LUAD patients based on the risk score. **(E)** Kaplan–Meier plot of OS in LUAD patients based on the risk score. **(F)** Kaplan–Meier plot of RFS in early stage LUAD patients (stage I and stage II) based on the risk score. **** $P < 0.0001$.

HGGPS is an independent predictor for RFS or OS of LUAD

Using the combined cohort of GSE31210 and TCGA-LUAD, we compared the prognostic power of HGGPS with that of other clinicopathological features for RFS. As a result, univariate analysis found five variables including TNM stage ($P < 0.001$), age ($P = 0.002$), smoking status ($P = 0.001$), and tumor burden ($P < 0.001$) besides HGGPS ($P < 0.001$) were significantly associated with RFS. Whereas, in multivariate analysis, only TNM stage ($P = 0.006$), tumor burden ($P < 0.001$), and HGGPS ($P = 0.003$) were still significant for RFS, indicating their independent values in RFS prediction in LUAD patients (Fig. S3A).

We also evaluated the prognostic value of HGGPS for OS prediction. Results showed that after adjusting for TNM stage, age, smoking status, *EGFR* mutation status, and tumor burden, HGGPS remained significantly correlated with the OS outcome ($P < 0.001$), revealing its effectiveness and independence in prognosis (Fig. S3B).

Besides, the correlations between HGGPS and other clinicopathological features were analyzed using the available patients in the combined cohort. As demonstrated in Figure S3C and Table S6, the HGGPS risk group was positively

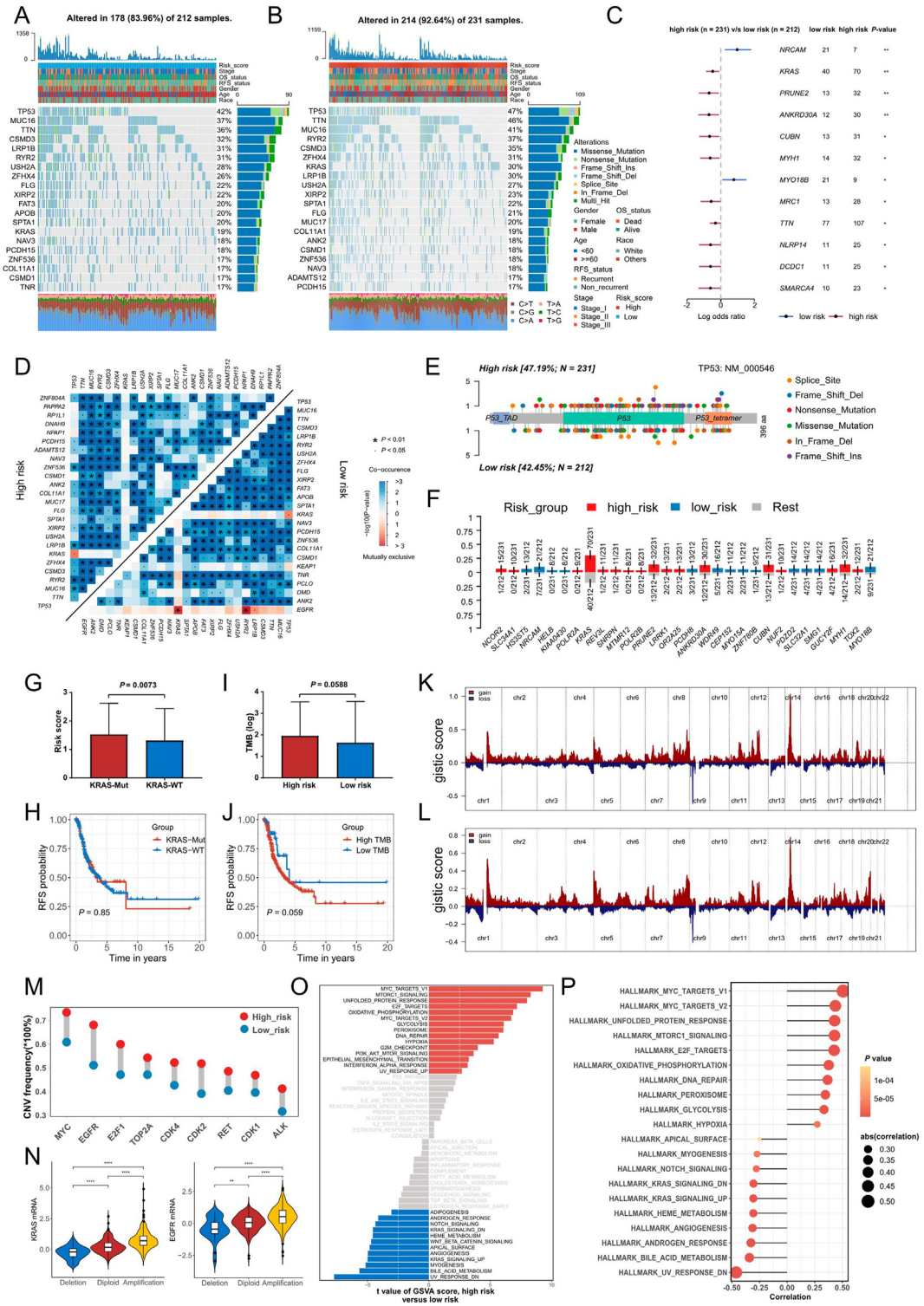

Figure 5 Distinctive mutation patterns and significant hallmark pathways in terms of the risk signature. **(A, B)** Waterfall plot illustrating the mutation profile of the top 20 most frequently mutated genes in each risk group from the TCGA-LUAD dataset. The middle columns indicate the mutation types of the individual patient. The bar plots in the upper and right panels represent the mutation frequency of each sample and that of mutation type of genes. The stacked barplot (bottom part) shows the proportion of transversion or transition in each sample. **(C)** Forest plot displaying the common driver genes mutating significantly differentially in the high- and low-risk groups. **(D)** Co-occurrence and mutual exclusivity of mutations in the most frequently mutated genes of the high- and low-risk groups. **(E)** Lollipop diagram visualizing the differential mutation site for *TP53* between two risk groups. **(F)** Mutation enrichment analysis for risk group. **(G)** Risk score was significantly increased in the *KRAS* mutant group versus the wild-type group. **(H)** Kaplan–Meier analysis to reveal the relevance between *KRAS* mutation and RFS of LUAD. **(I)** The comparison of TMB

correlated with gender ($P = 0.038$), smoking status ($P = 0.011$), TNM stage ($P < 0.001$), and tumor burden ($P < 0.001$), while other clinicopathological variables did not show significant association with our risk signature.

Next, we performed stratification analysis of HGGPS for OS or RFS in subgroups of clinical characteristics, including age ($\geq$60 and $<$60), gender (female and male), smoking status (yes and no), race (white and others), anatomic neoplasm subdivision (left and right), and location in lung parenchyma (central and peripheral). Surprisingly, as the RFS curves illustrated (Fig. S4), after subdivision by the above clinical variables, HGGPS was remained to be a significant prognostic factor in all subgroups. Notably, even though age and smoking status had been proved to be prognostic factors in UniCox analysis (Fig. S3A), HGGPS could still stratify the patients into distinct risk groups. Hence, HGGPS could exert additional prognostic value to existing risk models. For the stratification analysis of OS, similar results were achieved. In most subgroups, a significantly shorter OS time was observed in the high-risk than the low-risk group (Fig. S5), confirming the considerable power of HGGPS for the prognosis prediction in LUAD.

Given that the *EGFR* was the most common mutant gene in LUAD, we also evaluated the prognostic performance in RFS and OS in LUAD patients with different *EGFR* mutation statuses. The results suggested that in both the *EGFR* wild-type (*EGFR*-WT) subgroup and the *EGFR* mutant (*EGFR*-MUT) subgroup, the recurrence rate was significantly higher in the high-risk group than the low-risk group, as well as the OS result (Fig. S6).

Genomic alterations between different HGGPS risk groups

To give a hint of HGGPS-related mechanisms for RFS classification of LUAD from genomic layer, available somatic mutations of the TCGA-LUAD dataset were acquired, and the distribution differences in the high- and low-risk groups were analyzed by the package "maftools". The top 20 driver genes that were most frequently mutant in each risk group were shown in Figure 5A and B. Generally, *TP53*, *TTN*, and *MUC16* occupied the top three most frequently mutated positions in both groups. Nevertheless, the Fisher's exact test by mafCompare function revealed that among common driver genes (at least 20 mutation events in each group), 10 of them (*KRAS*, *PRUNE2*, *ANKRD30A*, *CUBN*, *MYH1*, *MRC1*, *TTN*, *NLRP14*, *DCDC1*, and *SMARCA4*) exhibited higher mutation frequencies in the high-risk group (Fig. 5C). Besides, significant co-occurrence and mutual exclusivity for the 25 most frequently mutated genes were investigated. Despite the pervasive pattern of co-occurring mutations in both groups, mutually exclusive mutations were observed in a unique case of *TP53-KRAS* in both groups and three cases of *EGFR-KRAS*, *EGFR-RYR2*, and *EGFR-LRP1B* in the low-risk group (Fig. 5D), implying their potentially redundant effect in the same pathway, and the selective advantages to retain more copies of the mutations. The observation that most somatic mutations in both the high- and low-risk groups were missense mutations suggested the necessity of the classification of mutation types. Figure 5E illustrates an example of different mutation spots between different risk groups and Figure S7 summarized the mutation profiles (including the variant type, standard normal variate (SNV) class, and so on) of different risk groups. Moreover, the mutation enrichment trends of different groups were manifested (Fig. 5F) and the detailed list of significant genes was shown in Table S7.

Given that *KRAS* mutation was the most enriched position among the common variants in the high-risk group, a possible correlation between risk score and *KRAS* mutation status was proposed, and the mutant group (*KRAS*-Mut) displayed significant higher risk score than that of the wild group (*KRAS*-WT) (Fig. 5G). Nonetheless, the specific distribution of risk score between the two groups does not denote its prognostic value in the whole dataset of TCGA-LUAD (Fig. 5H), but rather the heterogeneity and complexity inherent in subgroups of LUAD patients.[20,32] The TMB quantification results demonstrated an elevated level in the high-risk group, although in a non-significant mode (Fig. 5I), and patients with a lower TMB score presented a better RFS survival (Fig. 5J).

Next, we investigated the CNV features of the different mRNA risk groups. Figure 5K, L exhibited the genomic landscape of CNV segments across all human chromosomes after germline CNVs were eliminated, and typical oncogenes such as *CDK1*, *RET*, *CDK2*, *ALK*, *CDK4*, *KRAS*, *TOP2A*, *E2F2*, *EGFR*, and *MYC* were prevalently amplified in the high-risk group compared with the low-risk group (Fig. 5M). Furthermore, the expressions of *KRAS* and *EGFR* presented remarkably positive correlations with their copy numbers (Fig. 5N).

Identification of related hallmark pathways by GSVA

To investigate the underlying pathways involved in LUAD risk stratification, GSVA enrichment scores of all 50 hallmark gene sets from MSigDB were computed for each sample. Gene expression profiling of GSE31210 was converted to a pathway matrix using the GSVA package. Comparison of the high-risk group versus the low-risk group of LUAD patients revealed 27 significantly changed pathways at a *P*-value cutoff of 0.01 (of which 15 pathways showed significantly activated and 12 were inhibited) (Fig. 5O and Table S8), indicating the difference in pathway activities between the two risk groups. Moreover, we conducted the Spearman correlation test to evaluate the correlation between the risk score and each of all 50 hallmark gene sets. As a result, 20 gene sets showed significant association with risk score by a cutoff of $P < 0.01$ and correlation efficient $>$0.25 (Fig. 5P and Table S9). All of

between different risk groups. **(J)** The association of TMB level with RFS in LUAD patients. **(K, L)** Distribution of CNV segments across the human genome for the high- (K) and low-risk (L) groups. **(M)** Typical oncogenes were widely amplified in the high-risk group. **(N)** The relationship between copy number and mRNA expression for *KRAS* and *EGFR*. **(O)** 27 hallmark pathways exhibited significantly different between two risk groups ($P < 0.01$, $t > 2.5$). Enrichment scores were calculated by GSVA algorithm. **(P)** Top 20 important pathways that significantly correlated with risk score ($P < 0.01$ and correlation efficient $>$0.25). TMB, tumor mutation burden.

the 20 gene sets showed significantly activated or inhibited in the high-risk group compared with the low-risk group (Fig. 5O). Next, we selected the top six gene sets that had the strongest correlations with risk score ($P < 0.01$ and correlation coefficient >0.4) to perform unsupervised hierarchical clustering, as presented in Figure S8. All samples were grouped into two clusters by these pathways. We noticed that cluster 1 was mainly composed of samples from the low-risk group, and cluster 2 was mainly composed of samples of the high-risk group. Besides, considering its negative correlation with the risk score, the expression level of the UV_RESPONSE_DN pathway in cluster 1 was clearly higher than that in cluster 2, while the other five pathways that had positive associations with risk score were highly expressed in cluster 2 compared with that in cluster 1. These findings further confirmed the important roles of the identified pathways during the RFS risk determination by HGGPS. HGGPS was tightly associated with the immune microenvironment.

Early efforts trying to identify candidate therapeutic targets that function as determining factors in multiple activities of various immune cells have revealed their critical and diverse roles during carcinogenesis and cancer progression, and evading immune destruction has been widely considered to be an immerging hallmark of cancer.[3] Besides, increasing lines of evidence have suggested that the immune system may possess great potential for the development of clinical biomarkers.[18,33–36] In the current study, to better clarify the underlying mechanism of the immune response concerning HGGPS, we analyzed the contribution of immune phenotypes to RFS risk classification in GSE31210. As shown in Figure 6A, five immune checkpoint molecules including *LAG3*, *CD86*, *B7-H3*, and *VISTA* exhibited significant dysregulation between the different risk groups with a cutoff of *P*-value < 0.01. We then utilized the ssGSEA algorithm to generate an estimation of immune profiles for 24 types of tumor-infiltrating immune cells. The interactive network of the tumor microenvironment (TME) cells revealing the correlations among the TME cell populations was depicted (Fig. 6B). While the abundance of T helper 2 (Th2) cell presents a markedly negative correlation with the RFS in LUAD patients, mast cells, $CD8^+$ cells, regulatory T (Treg) cells, dendritic cells (DC), and effector memory T (Tem) cells were significantly associated with a better RFS. The comprehensive network of TME cells interactions and prognostic impact may reflect the reciprocal crosstalk among infiltrating immune cells in the formation of distinct risk groups of LUAD. Differential analysis suggested the significantly aberrant infiltration of $CD8^+$ cells, DC, Eosinophils, immature DC (iDC), Mast cells, natural killer (NK) $CD56^{dim}$ cells, central memory T cells (Tcm), Tem, T follicular helper (TFH), and Th2 cells in the high-risk group with a *P*-value cutoff of 0.01 (Fig. 6C).

When conducting the correlation analysis, Th2 cells and NK $CD56^{dim}$ cells showed a positive association with the identified risk score, while Tcm, Mast cells, $CD8^+$ T cells, Eosinophils, Tem, DC, TFH, and T helper cells were negatively related to our risk score (Table S10). These data confirmed the distinct immune infiltration status in the TME between different risk groups, which might partly cause the different clinical outcomes of LUAD patients. Moreover, the top six immune cell types that gave the largest absolute values of the Pearson correlation coefficient (PCC) were selected to perform the hierarchical clustering for all LUAD patients, and all patients were divided into two groups with different immune phenotypes (Fig. 6D). As displayed in Figure 6D, the high-infiltration group ("hot tumor") with active immune cells corresponded to the low-risk score group, while the low-infiltration group ("cold tumor") with inhibited immune cells corresponded to the high-risk score group, which further demonstrated the potential contribution of immune surveillance in the low-risk group of LUAD patients. In agreement with the clustering result, the PCA 3D plot in Figure 6E also confirmed the divergence between these immune groups. When contrasting the risk scores between the two immune groups, a significantly higher risk score level was observed in the low-infiltration group (Fig. 6F), verifying the important mechanism of immune regulation for RFS risk categorization. In addition, Kaplan–Meier survival curves generated by immune groups for RFS or OS also suggested distinct clinical outcomes of LUAD patients (Fig. 6G, H), which might support the explanation that altered immune surveillance is responsible for a poor RFS prognosis in LUAD.

Establishment of the gene co-expression network via WGCNA

In order to further uncover the underlying biological function of HGGPS, we used WGCNA to establish a scale-free gene co-expression network based on the gene expression profile of 226 LUAD samples with 5000 filtered genes. As the first step, GSM773616, GSM773601, GSM773559, and GSM773630 were recognized as outliers and removed by hierarchical clustering dendrogram of samples with average linkage method (Fig. S9A). After carrying out the topological analysis (scale independence and mean connectivity) for the network with thresholding powers ranging from 1 to 20, we picked $\beta = 4$ as the optimal soft-threshold power to obtain a scale-free fitting index (R^2) of >0.85 (Fig. S9B). Next, a hierarchical clustering tree of genes was constructed and eight modules were screened out with a minimum size of 100 genes. Additionally, similar modules were merged using a minimum height of 0.3 and six modules were eventually obtained (Fig. 7A, B) with variable sizes ranging from 310 to 1795 (grey module excluded). To determine the relevance between the identified modules and clinicopathological characteristics, such as age, gender, smoking status, immune group, TNM stage, RFS time, RFS status, risk group, and so on, we computed the Pearson correlation coefficients and corresponding *P* values, and a module-trait heatmap was depicted. As illustrated in Figure 7C, the blue module and the black module were found to be the two most significant modules with the highest absolute values of PCC for risk level. It should be noted that the blue module also showed the highest significantly positive correlation with TNM stage, OS status, and RFS status, while it was most negatively associated with the immune group, OS time, and RFS time. For the black module, however, an opposite trend was observed—it displayed the positive correlations with the immune group, OS time, and RFS time, while revealing strong negative correlations with TNM stage, OS status, and RFS status. Function enrichment suggested these two modules

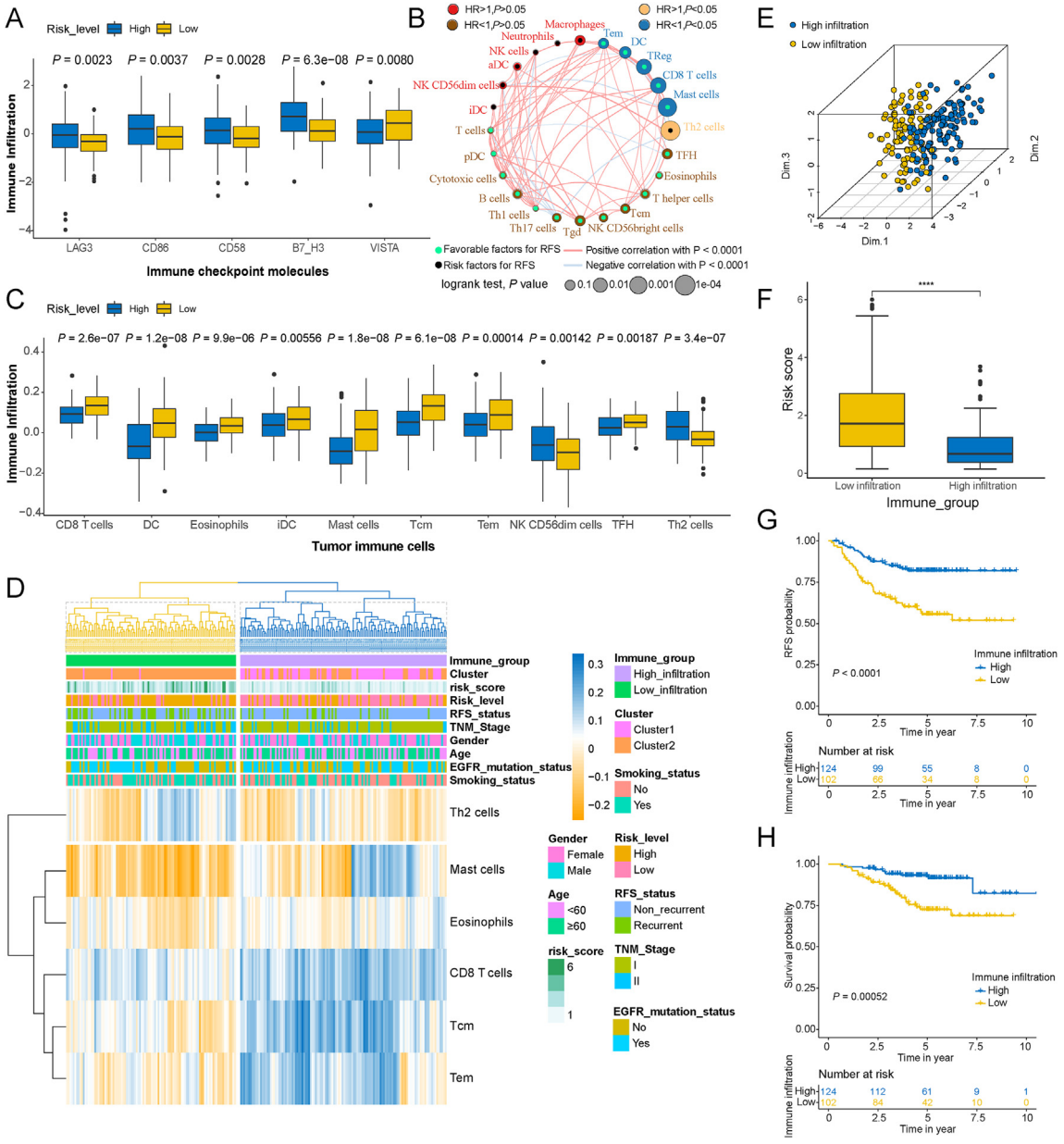

Figure 6 The mRNA risk signature tightly linked to tumor immune infiltration based on GSE31210. **(A)** Relationship of the risk signature to immune checkpoint molecules. **(B)** Interactive network of TME cells with prognostic values. The size of each cell represents the survival influence evaluated by the log-rank test, and the blue (red) lines represent the positive (negative) correlation between cells. **(C)** Differential infiltration levels of estimated immune cell types. **(D)** Hierarchical clustering of all samples with top-six immune cell types that most correlated with a risk score. **(E)** Principal component analysis of the six immune cell types regarding the immune groups. **(F)** Comparison of risk scores between the high- and low-immune infiltration groups. **(G, H)** Kaplan—Meier curves of RFS (G) and OS (H) for two distinct immune groups. TME, tumor microenvironment.

were mostly enriched in those GO terms linked to early events of cancer metastasis, especially relevant to epithelial—mesenchymal transition (EMT), like extracellular matrix organization, extracellular matrix, adherens junction, focal adhesion, and cell adhesion molecule binding (Fig. S10A). For KEGG pathway analysis, it was found that the selected modules were mainly involved in PI3K-Akt signaling pathway, Focal adhesion, Hippo signaling pathway, TNF signaling pathway, ECM-receptor interaction, and so on (Fig. S10B, C).

Key hub gene identification and putative TF regulatory PPI network construction

Following the prespecified cutoff criteria ($|GS| > 0.4$ and $|MM| > 0.8$), a total of 26 genes were extracted and deemed as candidate hub genes from the blue and black modules (Fig. 7D, E), which were subsequently used to generate a highly confident PPI network by STRING database to identify key hub genes. After eliminating the disconnected nodes, a tightly connected network cluster

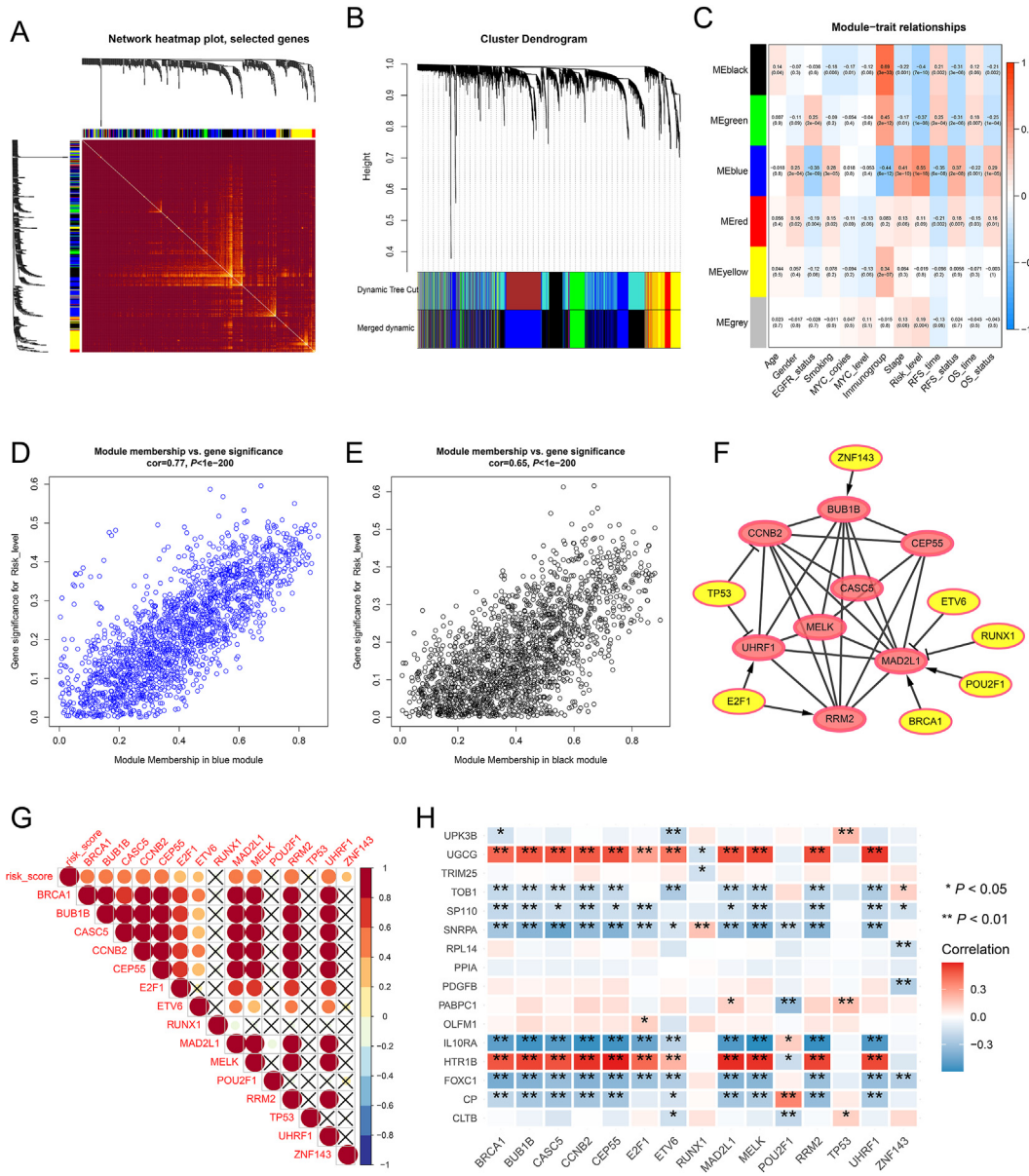

Figure 7 Construction of a TF-PPI—network and correlation analysis using GSE3121. **(A)** Heatmap visualizing the co-expression network with six gene modules, which was generated from the topological overlap matrix among all genes. **(B)** Hierarchical clustering tree of genes using the dissimilarity (1-TOM). Each color signifies an individual gene module. **(C)** Module-trait heatmap showing correlations between module eigengenes and clinicopathological traits. Each row represents a module eigengene and columns indicate clinicopathological traits. Each cell contains PCC and corresponding *P* value. **(D, E)** Scatter plots selected candidate hub genes in the blue and black modules. Dots in the right-up section were candidate hub genes ($|GS| > 0.4$ and $|MM| > 0.8$). **(F)** The established TF-PPI-network contained eight key hub genes and seven upstream transcription factors. **(G)** Pearson correlation analysis between risk score and each node of the TF-PPI-network. The color depth represents the degree of correlation, and " × " indicates no significant difference. **(H)** Correlation matrix of the 16 risk signature genes and each node in the TF-PPI-network. * $P < 0.05$, ** $P < 0.01$. GS, gene significance; MM, module membership; PCC, Pearson correlation coefficient; TOM, topological overlap matrix.

was formed using the remaining eight genes (*CASC5*, *MAD2L1*, *CCNB2*, *RRM2*, *MELK*, *CEP55*, *UHRF1*, and *BUB1B*), which were considered as the key hub genes for further analysis. We preliminary explored the expression levels and the prognostic values of these genes with the GEPIA database (http://gepia.cancer-pku.cn/),[37] an interactive online tool, to analyze RNA sequencing data from the TCGA project. Notably, the expression levels of *MAD2L1*, *CCNB2*, *RRM2*, *MELK*, *CEP55*, *UHRF1*, and *BUB1B* were significantly elevated in tumor samples, compared with adjacent normal lung tissues with a *P*-value cutoff of 0.01 (Fig. S11). *CASC5* expression was also up-regulated in the

tumor, although the change was not statistically significant. In addition, increased expressions of all the identified key hub genes were found to confer unfavorable OS outcomes (Fig. S12A), and five of them (*CASC5*, *CCNB2*, *RRM2*, *MELK*, and *BUB1B*) were demonstrated to be significantly associated with worse RFS of LUAD patients (Fig. S12B). We then predicted the putative upstream transcription factors of these eight key hub genes by the TRRUST database, and seven transcription factors (*BRCA1*, *E2F1*, *ETV6*, *POU2F1*, *RUNX1*, *TP53*, and *ZNF143*) were predicted to activate or repress the expression levels of key hub genes and thus used to construct the TF transcription factor regulatory PPI network (Fig. 7F). When subjected to Pearson correlation analysis, all nodes in the TF regulatory PPI network showed significant correlations ($P < 0.01$) with the risk score derived from HGGPS, except for *RUNX1*, *POU2F1*, and *TP53* (Fig. 7G). Eventually, we evaluate the correlations between the 16 identified risk signature genes and each of the network nodes. As illustrated in Figure 7H, *UGCG* and *HTR1B* were significantly positively correlated with most key hub genes and the corresponding transcription factors, while *TOB1*, *SP110*, *SNRPA*, *IL10RA*, *FOXC1*, and *CP* showed significant negative associations with all the key hub genes. Collectively, these findings disclosed that the key hub genes and the established TF-PPI-network played pivotal roles in determining the RFS risk of LUAD patients.

Building a multi-omic nomogram based on HGGPS

Evidence revealed that DNA methylation patterns could predict prognosis and survival in common cancers including LUAD. Thus, we anticipated that an integrated nomogram with the multi-omic strategy would improve the prediction accuracy tremendously. On the basis of previous work,[38] we retrieved the methylation profiles of 399 LUAD patients in TCGA, and screened the reported 82 CpG markers identified by LASSO and boosting. UniCox analysis with the "multi-split" approach was implemented for dimensionality reduction. We introduced a methylation score for each patient with a linear equation using the four resulting methylation markers: cg05556202, cg00620629, cg23389061, and cg19928450. All patients were subsequently assigned to two risk groups by the median methylation score. The generated Kaplan–Meier curves revealed a longer median survival time for RFS ($P = 0.016$) or OS ($P = 0.041$) in the low-risk group (Fig. 8A, B), and a tSNE plot of the four methylation markers showing all cases clustered by methylation score levels was also depicted (Fig. 8C).

For the purpose of developing a clinically quantitative tool to optimize prediction accuracy for LUAD recurrence, significant prognostic risk characteristics including age, TNM stage, smoking status, tumor burden, HGGPS, and methylation score were combined to construct an integrated nomogram based on the multivariate cox analysis for RFS prediction (Fig. 8D). The calibration plots for the possibility of 3- and 5-year RFS were highly consistent with the observed RFS rates, showing good performance for RFS prediction (Fig. 8E). As expected, the C-index of the integrated nomogram was higher than each of the single variables (Fig. 8F), strengthening its predictive capacity.

Furthermore, decision curve analysis (DCA) curves for 1- and 3-year RFS indicated the integrated nomogram may be more beneficial than clinicopathological characteristics (Fig. 8G, H) for the complete range of threshold probabilities. The results of the Kaplan–Meier graph also demonstrated a considerable value of the derived nomogram for RFS or OS prognosis prediction (Fig. 8I, J). Notably, the tROC analysis indicated that the integrated nomogram was the best predictor for recurrence-free survival with the estimates of the AUCs of 0.802, 0.860, and 0.919 for the 3-, 5-, and 10-year RFS, respectively (Fig. 8K). Taking together, these findings suggested that the established multi-dimensional nomogram may provide a powerful method for RFS prediction in LUAD patients, which was helpful for clinical administration.

HGGPS predicts therapy responses in patients with LUAD

Conventional chemotherapy or targeted therapy are systemic therapies that are prevalently applied after surgery to eradicate micrometastases and decrease the recurrence rate in LUAD, so recognition of which specific subsets are more sensitive to related compounds or drugs is fundamental to prolong the median RFS time. The estimated 50% inhibiting concentration (IC_{50}) values of 138 drugs were inferred by the pRRophetic algorithm to predict the treatment responses in different risk groups derived from the HGGPS classifier. We found the low-risk group was more sensitive to Methotrexate, Nilotinib, Lenalidomide, ATRA, etc., while the high-risk group was more sensitive to Docetaxel, Paclitaxel, RDEA119, Dasatinib, Bortezomib, Elesclomol, etc. (Fig. S13A).

The above findings have established the correlations between HGGPS and the immune system activities, as well as common immune checkpoints, which are emerging as new promising targets in immune therapy. Thus, we speculated that HGGPS may be used to predict the anti-*PD-1/PDL1* immunotherapy. Based on the TCGA-LUAD dataset, a significantly higher immunophenoscore in the low-risk group was observed, which indicated a more immunogenic tumor with higher sensitivity to immunotherapy (Fig. S13B). We further utilized tumor immune dysfunction and exclusion (TIDE) score to evaluate the ICB responses. As shown in Figure S13C, the TIDE scores was significantly decreased in the low-risk group, suggesting more promise in response to ICB therapy.

Discussion

LUAD is a highly heterogeneous malignancy at the histological and molecular level, posing a huge challenge for the accurate prognostication of LUAD patients. We have previously demonstrated the presence of solid subtype was significantly associated with poor overall survival of LUAD.[39] In this study, given the fact that universal postoperative recurrence or metastasis takes responsibility for the majority of lung cancer deaths, and better recognition of LUAD cases likely to relapse can guide standard disease management and maximize clinical benefit, it is crucial that we focus on the development and validation of a useful and applicable gene expression signature to predict cancer recurrence of LUAD patients.

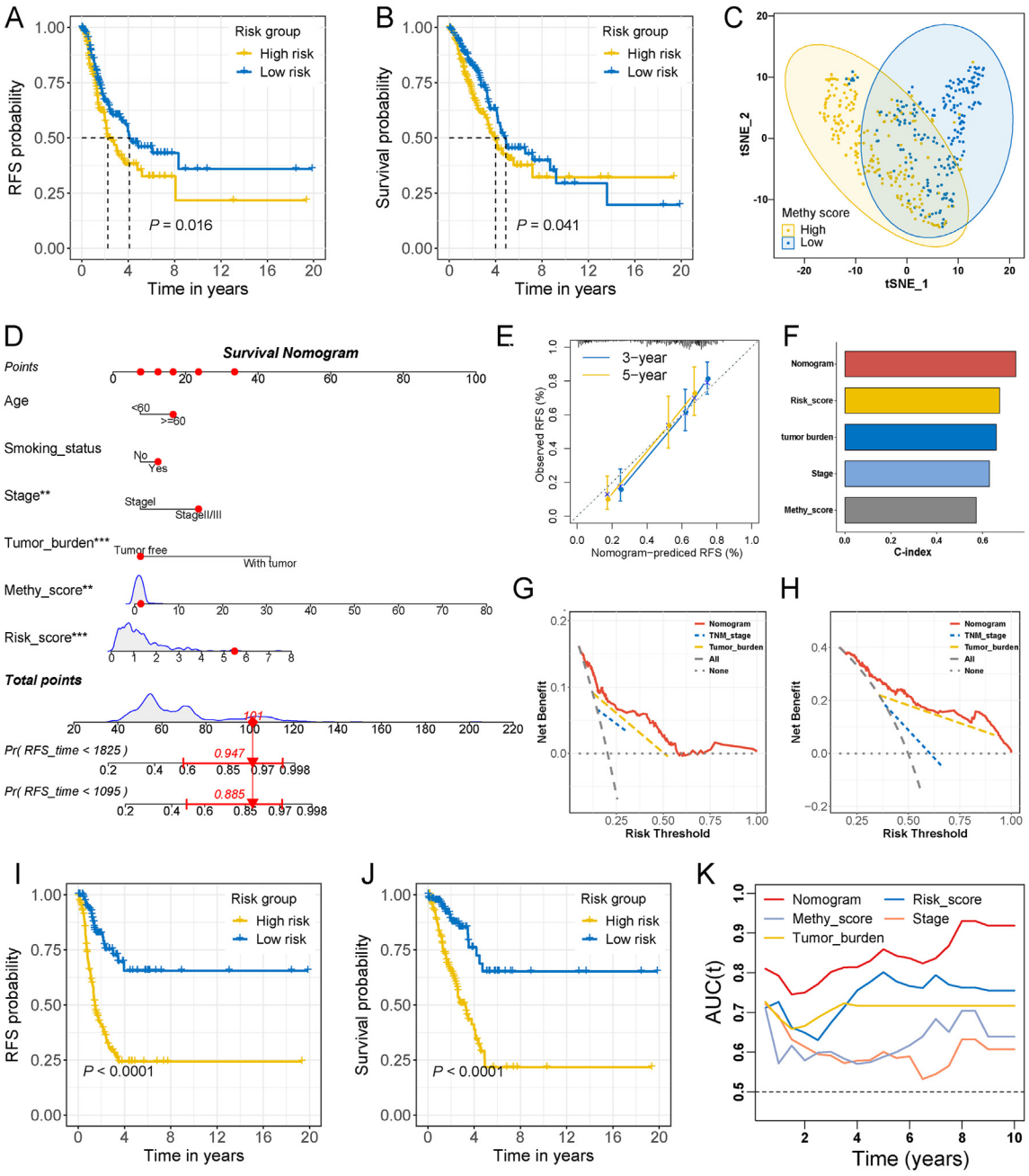

Figure 8 Methylation score system and nomogram construction based on the TCGA-LUAD dataset. **(A, B)** Kaplan—Meier curves of the methylation score system for RFS (A) and OS (B). **(C)** A tSNE plot of four methylation markers with methylation score levels. **(D)** The established nomogram. The red arrow signifies an example to visualize the assessment of risk for 3-year RFS and 5-year RFS. **(E)** Calibration curves for RFS prediction with the integrated nomogram showing the agreement between prediction and observation at 3 and 5 years. **(F)** C-indexes for the generated nomogram and single variables in predicting RFS of LUAD. **(G, H)** DCA plots for 1-year (G) and 3-year (H) RFS indicating a better net benefit of the nomogram than TNM stage or tumor burden. **(I, J)** Kaplan—Meier analysis of the nomogram for RFS (I) and OS (J) in LUAD patients. **(K)** The tROC analysis of the nomogram. AUC, area under the curve; tROC, time-dependent receiver operating characteristic.

In the present study, we presume that GSEA is a powerful tool to screen significant and meaningful pathways for LUAD recurrence, and it is possible to develop a prognostic signature combining GSEA with multiple statistical algorithms. With this goal, we retrospectively enrolled a pooled set of 1026 patients from five individual datasets of gene expression profile and obtained the corresponding clinicopathological characters and survival information, which allowed the following unbiased division of the learning and validation datasets to minimize the risk of potential overfitting. After the multi-step process of feature selection, we fitted an effective 16-mRNA risk signature — HGGPS — for

LUAD RFS prediction with the learning dataset. Analysis with the validation dataset found that it was a robust classifier for RFS risk determination. Moreover, multivariate analysis involving other significant clinicopathological covariates (including TNM stage, age, smoking status, and *EGFR* mutation) showed the signature remained an independent prognostic indicator for both RFS and OS of LUAD patients. Remarkably, stratification analysis based on subsets of several clinical variables suggested that our signature can impose additional prognostic value for RFS and OS prediction, which reinforced its ability for risk classification.

Thanks to the rapid development of biological technology and bioinformatics, a large number of cancer prognostic biomarkers were identified and applied in clinical practice. These biomarkers including protein, mRNA, miRNA and metabolites, and their examination platforms are quantitative PCR, microarray, next generation sequencing or mass spectrometry. The prognostic panel could be single biomarker or a few biomarkers in combination.[40–42] Recently, mounting published studies have revealed that gene expression signature may hold potential clinical utility in various solid cancers,[6–11,14–18,43–50] in line with our results in this article. In comparison with previous efforts seeking to make prognostic models, this study possesses at least four major advances. First, unlike previous research that mainly focused on overall survival which is likely to be influenced by competing risks (such as pre-existing comorbidities), we took cancer recurrence as the primary endpoint with the aim of exploring a gene expression classifier to discriminate individuals that are at risk of relapse and might achieve the most benefit during adjuvant treatment. Second, we applied the "pooled set" strategy to get a large sample size that contained 1026 cases, which exceeds that of many studies and yielded sufficient statistical power. Third, compared with most studies that only used one or two approaches to select variables, we adopted a multi-step process that integrated multiple algorithms for comprehensive feature selection. Last but not least, most previous studies used prefiltered dysregulated genes between tumors and adjacent normal tissues, or only a certain gene set (such as immune-related genes,[18,36,48] Tyrosine kinases,[6] and WEE Family Kinases[49]), our study used a gene set-based strategy to prescreen significant gene sets that hold prognostic potential to decrease the possibility of missing important markers.

Strikingly, most of the 16 signature genes have been linked to lung cancer. As an example, *CLTb* was found to be specifically upregulated in non-small cell lung cancer (NSCLC) and associated with poor prognosis, and upregulation of *CLTb*, together with *Dyn1* can regulate the activity of clathrin-mediated endocytosis (CME) to selectively modulate *EGFR* recycling, resulting in elevated migration capacity of NSCLC cell lines and increased invasion and metastatic efficiency *in vivo*.[51–53] *OLFM1* was an immune-related gene that significantly down-regulated in endometrial cancer, colorectal cancer (CRC), and neuroblastoma,[54–56] and overexpression of *OLFM1* attenuated CRC cells' proliferation and migration *in vitro*.[55] On the other hand, however, *OLFM1* protein showed an up-regulated level in LUAD than squamous cell carcinoma and normal lung tissues, indicating its potential utility as a diagnostic marker.[57] Besides, dysregulated expressions or prognostic values of *FOXC1*,[58] *PPIA*,[59] *PDGFB*,[59,60] *TOB1*,[61] *CP*,[62] *HTR1B*,[63,64] *TRIM25*,[65] and *PABPC1*[66] in lung cancer were also observed previously. Despite these concerns, the oncogenic roles of the remaining signature genes (*UPK3B*, *IL10*RA, *SP110*, *RPL14*, *SNRPA*, and *UGCG*) in LUAD have not been well investigated. For example, *SNRPA* was only reported in gastric cancer[67] to promote tumor growth, and little was known about *SP110* as to its relationship with cancer, providing new insight into cancer progression and candidate target identification, especially for lung adenocarcinoma.

Genomic alterations and CNVs between different risk groups were analyzed, and more aggressive molecular disorders including somatic mutations such as *KRAS*, and the amplification of widely known oncogenes in LUAD (*KRAS*, *ALK*, *EGFR*, etc.) were identified. Through GSVA, we discovered several important hallmark pathways involved in the risk categorization by the defined gene signature. Moreover, to determine the underlying mechanism of immune regulations, we compared the expression levels of common immune checkpoints between distinct risk groups and found significant differences in *LAG3*, *CD86*, *CD58*, *B7-H3*, and *VISTA*. A collection of immune cells was also indicated to play critical roles during risk classification. The high-risk group showed significantly lower proportions of $CD8^+$ cells, DC, Eosinophils, iDC, Mast cells, Tcm, Tem, and TFH, denoting their decreased activities may have negative impacts on LUAD patient's survival. Interestingly, immune groups clustered by the top six immune cells that mostly correlated with the risk score also revealed significantly associated with clinical outcome. These findings revealed a tight linkage between the dynamic immune environment and our gene signature in LUAD.

WGCNA is a popular method to identify cluster modules of highly related genes or hub genes related to external sample traits.[68,69] In this study, we utilized WGCNA to discover two important gene modules that highly correlated with the risk score, which then proved to be mostly enriched in early events of metastasis and well-known pathways that implicated in tumor initiation or progression through GO and KEGG analysis. Based on WGCNA results, we also identified a TF-PPI-network including eight key hub genes and seven closely connected transcription factors, and a majority of nodes were found to be significantly correlated with the risk score and at least one of the 16 signature genes (except for *PPIA*). The results give a hint that these genes may act as essential regulators during tumorigenesis or tumor progression, and future investigations along this line are required to elucidate their oncogenic or anti-tumor roles.

It is worth noting that although some nomograms have been constructed as to the outcome prediction of NSCLC,[70–72] little research has been published regarding nomogram construction to predict LUAD recurrence, especially based on gene expression or DNA methylation. We here built a nomogram integrating significant clinicopathological features including age, smoking status, TNM stage, tumor burden, HGGPS, and methylation score to improve the predictive accuracy. Calibration curves, tROC plots, and decision curve analysis verified its ideal performance. Thus, the nomogram may serve as a simple, reliable, and useful instrument in recurrence prediction for LUAD patients. We also proved the advantage of the incorporation of multi-omic data for clinical model development.

The main cause for the current failure of postsurgical treatment (chemotherapy, targeted therapy, or immunotherapy) lies in drug resistance induced by tumor heterogeneity. So it is critical to elucidate the sensitivity and efficacy of these routine interventions in the patient populations with heterogeneous risk of recurrence. At a significance level of 0.01, we found a total of 78 chemotherapeutic or targeted drugs were responded differently between the high- and low-risk groups. For instance, patients in the high-risk group may be more sensitive to Docetaxel, Paclitaxel, Dasatinib, Bortezomib, and Elesclomol, but resistant to Methotrexate, Nilotinib, Lenalidomide, and ATRA. We also explored the relationship of the risk signature and ICB therapy response for the sake of recognizing the LUAD patients most likely to benefit from immunotherapy. These results raise the possibility that HGGPS might be used as candidate biomarkers to predict therapeutic resistance and provide valuable cues for optimizing regimens in clinical practice to aid personalized medicine.

Two important limitations should be addressed. First, due to the lack of an independent cohort in the current study, the predictive capability of our risk signature and the established nomogram should be assessed by adequate external validation cohorts in the future, which is the indispensable step to move molecular models to ultimate clinical application. Second, since the biological functions of certain candidate markers or key genes in the identified TF-PPI-network are not well characterized, more *in vitro* and *in vivo* experiments are expected to deepen our understanding of their relevance to carcinogenesis and cancer development, which will be our main focus in the future research.

In summary, we developed an effective 16-mRNA risk signature to predict the RFS and therapeutic response of stage I—III LUAD patients, which might be correlated to the distinct genomic alterations and pathways, and dynamic tumor immune phenotype. A hub TF-PPI-network was identified with regard to the risk signature. We also established a nomogram based on the mRNA signature and methylation score to provide an optimal and applicable approach for the quantification of recurrence risk, which would help clinicians to make personalized treatment decisions for LUAD patients. The identified markers or key genes may offer a firm basis for further studies concerning tumorigenesis or progression in LUAD.

Author contributions

Y. Zhang, Z. Yang, and Y. Tang performed the overall study and drafted the manuscript. C. Guo, L. Cheng and D. Lin contributed to data collection and participated in data analysis. G. Li, X. Hu, and K. Zhang conceived and supervised the research, and revised the manuscript for critical review. All authors read and approved the final version of the manuscript.

Conflict of interests

The authors have declared that no competing interest exists.

Acknowledgements

We sincerely thank Gene Expression Omnibus (GEO) database and The Cancer Genome Atlas (TCGA) for data sharing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gendis.2022.07.005.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J Clin*. 2021;71(3):209—249.
2. Bender E. Epidemiology: the dominant malignancy. *Nature*. 2014;513(7517):S2—S3.
3. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646—674.
4. Marusyk A, Janiszewska M, Polyak K. Intratumor heterogeneity: the Rosetta stone of therapy resistance. *Cancer Cell*. 2020;37(4):471—484.
5. Travis WD, Brambilla E, Noguchi M, et al. International association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma. *J Thorac Oncol*. 2011;6(2):244—285.
6. de Nonneville A, Finetti P, Adelaide J, et al. A tyrosine kinase expression signature predicts the post-operative clinical outcome in triple negative breast cancers. *Cancers*. 2019;11(8):1158.
7. Shimizu H, Nakayama KI. A 23 gene-based molecular prognostic score precisely predicts overall survival of breast cancer patients. *EBioMedicine*. 2019;46:150—159.
8. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. 2004;351(27):2817—2826.
9. Long J, Chen P, Lin J, et al. DNA methylation-driven genes for constructing diagnostic, prognostic, and recurrence models for hepatocellular carcinoma. *Theranostics*. 2019;9(24):7251—7267.
10. Hoshida Y, Villanueva A, Kobayashi M, et al. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *N Engl J Med*. 2008;359(19):1995—2004.
11. Lin T, Gu J, Qu K, et al. A new risk score based on twelve hepatocellular carcinoma-specific gene expression can predict the patients' prognosis. *Aging*. 2018;10(9):2480—2497.
12. Zhang Y, Tang Y, Guo C, Li G. Integrative analysis identifies key mRNA biomarkers for diagnosis, prognosis, and therapeutic targets of HCV-associated hepatocellular carcinoma. *Aging*. 2021;13(9):12865—12895.
13. Tang Y, Zhang Y, Hu X. Identification of potential hub genes related to diagnosis and prognosis of hepatitis B virus-related hepatocellular carcinoma via integrated bioinformatics analysis. *BioMed Res Int*. 2020;2020:4251761.
14. Zhou Z, Mo S, Dai W, et al. Prognostic nomograms for predicting cause-specific survival and overall survival of stage I-III colon cancer patients: a large population-based study. *Cancer Cell Int*. 2019;19:355.
15. Jiang H, Du J, Gu J, Jin L, Pu Y, Fei B. A 65-gene signature for prognostic prediction in colon adenocarcinoma. *Int J Mol Med*. 2018;41(4):2021—2027.
16. Okayama H, Schetter AJ, Ishigame T, et al. The expression of four genes as a prognostic classifier for stage I lung

adenocarcinoma in 12 independent cohorts. *Cancer Epidemiol Biomarkers Prev*. 2014;23(12):2884—2894.
17. Larsen JE, Pavey SJ, Passmore LH, Bowman RV, Hayward NK, Fong KM. Gene expression signature predicts recurrence in lung adenocarcinoma. *Clin Cancer Res*. 2007;13(10):2946—2954.
18. Zhang C, Zhang Z, Zhang G, et al. Clinical significance and inflammatory landscapes of a novel recurrence-associated immune signature in early-stage lung adenocarcinoma. *Cancer Lett*. 2020;479:31—41.
19. Wu J, Li L, Zhang H, et al. A risk model developed based on tumor microenvironment predicts overall survival and associates with tumor immunity of patients with lung adenocarcinoma. *Oncogene*. 2021;40(26):4413—4424.
20. Shi R, Bao X, Unger K, et al. Identification and validation of hypoxia-derived gene signatures to predict clinical outcomes and therapeutic responses in stage I lung adenocarcinoma patients. *Theranostics*. 2021;11(10):5061—5076.
21. Sun J, Zhao T, Zhao D, et al. Development and validation of a hypoxia-related gene signature to predict overall survival in early-stage lung adenocarcinoma patients. *Ther Adv Med Oncol*. 2020;12:1758835920937904.
22. Xu F, Lin H, He P, et al. A *TP53*-associated gene signature for prediction of prognosis and therapeutic responses in lung squamous cell carcinoma. *Oncoimmunology*. 2020;9(1):1731943.
23. Li Y, Gu J, Xu F, et al. Molecular characterization, biological function, tumor microenvironment association and clinical significance of m6A regulators in lung adenocarcinoma. *Briefings Bioinf*. 2021;22(4):bbaa225.
24. Cai J, Tong Y, Huang L, et al. Identification and validation of a potent multi-mRNA signature for the prediction of early relapse in hepatocellular carcinoma. *Carcinogenesis*. 2019;40(7):840—852.
25. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545—15550.
26. Okayama H, Kohno T, Ishii Y, et al. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res*. 2012;72(1):100—111.
27. Der SD, Sykes J, Pintilie M, et al. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *J Thorac Oncol*. 2014;9(1):59—64.
28. Sato M, Larsen JE, Lee W, et al. Human lung epithelial cells progressed to malignancy through specific oncogenic manipulations. *Mol Cancer Res*. 2013;11(6):638—650.
29. Botling J, Edlund K, Lohr M, et al. Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin Cancer Res*. 2013;19(1):194—204.
30. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882—883.
31. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1(6):417—425.
32. Zhang Y, Yang M, Ng DM, et al. Multi-omics data analyses construct TME and identify the immune-related prognosis signatures in human LUAD. *Mol Ther Nucleic Acids*. 2020;21:860—873.
33. Liu J, Nie S, Wu Z, et al. Exploration of a novel prognostic risk signatures and immune checkpoint molecules in endometrial carcinoma microenvironment. *Genomics*. 2020;112(5):3117—3134.
34. Wang S, Zhang Q, Yu C, Cao Y, Zuo Y, Yang L. Immune cell infiltration-based signature for prognosis and immunogenomic analysis in breast cancer. *Briefings Bioinf*. 2021;22(2):2020—2031.
35. Brooks JM, Menezes AN, Ibrahim M, et al. Development and validation of a combined hypoxia and immune prognostic classifier for head and neck cancer. *Clin Cancer Res*. 2019;25(17):5315—5328.
36. Wang Z, Zhu J, Liu Y, et al. Development and validation of a novel immune-related prognostic model in hepatocellular carcinoma. *J Transl Med*. 2020;18(1):67.
37. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*. 2017;45(W1):W98—W102.
38. Hao X, Luo H, Krawczyk M, et al. DNA methylation markers for diagnosis and prognosis of common cancers. *Proc Natl Acad Sci U S A*. 2017;114(28):7414—7419.
39. Zhang YQ, Wang WY, Xue JX, et al. microRNA expression profile on solid subtype of invasive lung adenocarcinoma reveals a panel of four miRNAs to Be associated with poor prognosis in Chinese patients. *J Cancer*. 2016;7(12):1610—1620.
40. Zhu CQ, Tsao MS. Prognostic markers in lung cancer: is it ready for prime time? *Transl Lung Cancer Res*. 2014;3(3):149—158.
41. Mathé EA, Patterson AD, Haznadar M, et al. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Res*. 2014;74(12):3259—3270.
42. Zhang C, He Z, Cheng L, Cao J. Investigation of prognostic markers of lung adenocarcinoma based on tumor metabolism-related genes. *Front Genet*. 2021;12:760506.
43. Wang Y, Yang Z. A Gleason score-related outcome model for human prostate cancer: a comprehensive study based on weighted gene co-expression network analysis. *Cancer Cell Int*. 2020;20:159.
44. Millstein J, Budden T, Goode EL, et al. Prognostic gene expression signature for high-grade serous ovarian cancer. *Ann Oncol*. 2020;31(9):1240—1250.
45. Kang H, Chen IM, Wilson CS, et al. Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. *Blood*. 2010;115(7):1394—1405.
46. Liu J, Nie S, Li S, et al. Methylation-driven genes and their prognostic value in cervical squamous cell carcinoma. *Ann Transl Med*. 2020;8(14):868.
47. Wu J, Jin S, Gu W, et al. Construction and validation of a 9-gene signature for predicting prognosis in stage *III* clear cell renal cell carcinoma. *Front Oncol*. 2019;9:152.
48. Xiao H, Wang B, Xiong HX, et al. A novel prognostic index of hepatocellular carcinoma based on immunogenomic landscape analysis. *J Cell Physiol*. 2021;236(4):2572—2591.
49. Liu Y, Qi J, Dou Z, et al. Systematic expression analysis of WEE family kinases reveals the importance of PKMYT1 in breast carcinogenesis. *Cell Prolif*. 2020;53(2):e12741.
50. Tang Y, Guo C, Yang Z, Wang Y, Zhang Y, Wang D. Identification of a tumor immunological phenotype-related gene signature for predicting prognosis, immunotherapy efficacy, and drug candidates in hepatocellular carcinoma. *Front Immunol*. 2022;13:862527.
51. Chen PH, Bendris N, Hsiao YJ, et al. Crosstalk between CLCb/Dyn1-mediated adaptive clathrin-mediated endocytosis and epidermal growth factor receptor signaling increases metastasis. *Dev Cell*. 2017;40(3):278—288. e5.
52. Wilson BJ, Allen JL, Caswell PT. Vesicle trafficking pathways that direct cell migration in 3D matrices and *in vivo*. *Traffic*. 2018;19(12):899—909.
53. Majeed SR, Vasudevan L, Chen CY, et al. Clathrin light chains are required for the gyrating-clathrin recycling pathway and thereby promote cell migration. *Nat Commun*. 2014;5:3891.

54. Wong YF, Cheung TH, Lo KWK, et al. Identification of molecular markers and signaling pathway in endometrial cancer in Hong Kong Chinese women by genome-wide gene expression profiling. *Oncogene*. 2007;26(13):1971—1982.
55. Shi W, Ye Z, Zhuang L, et al. Olfactomedin 1 negatively regulates NF-κB signalling and suppresses the growth and metastasis of colorectal cancer cells. *J Pathol*. 2016;240(3):352—365.
56. Khan J, Wei JS, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*. 2001;7(6):673—679.
57. Wu L, Chang W, Zhao J, et al. Development of autoantibody signatures as novel diagnostic biomarkers of non-small cell lung cancer. *Clin Cancer Res*. 2010;16(14):3760—3768.
58. Gong R, Lin W, Gao A, et al. Forkhead box C1 promotes metastasis and invasion of non-small cell lung cancer by binding directly to the lysyl oxidase promoter. *Cancer Sci*. 2019;110(12):3663—3676.
59. Zhang M, Zhu K, Pu H, et al. An immune-related signature predicts survival in patients with lung adenocarcinoma. *Front Oncol*. 2019;9:1314.
60. Neri S, Miyashita T, Hashimoto H, et al. Fibroblast-led cancer cell invasion is activated by epithelial-mesenchymal transition through platelet-derived growth factor BB secretion of lung adenocarcinoma. *Cancer Lett*. 2017;395:20—30.
61. Jiao Y, Sun KK, Zhao L, Xu JY, Wang LL, Fan SJ. Suppression of human lung cancer cell proliferation and metastasis *in vitro* by the transducer of ErbB-2.1 (TOB1). *Acta Pharmacol Sin*. 2012; 33(2):250—260.
62. Matsuoka R, Shiba-Ishii A, Nakano N, et al. Heterotopic production of ceruloplasmin by lung adenocarcinoma is significantly correlated with prognosis. *Lung Cancer*. 2018;118:97—104.
63. Takai D, Yagi Y, Wakazono K, et al. Silencing of HTR1B and reduced expression of EDN1 in human lung cancers, revealed by methylation-sensitive representational difference analysis. *Oncogene*. 2001;20(51):7505—7513.
64. Zhang Y, Fan Q, Guo Y, Zhu K. Eight-gene signature predicts recurrence in lung adenocarcinoma. *Cancer Biomarkers*. 2020; 28(4):447—457.
65. Han Q, Cheng P, Yang H, Liang H, Lin F. Altered expression of microRNA-365 is related to the occurrence and development of non-small-cell lung cancer by inhibiting TRIM25 expression. *J Cell Physiol*. 2019;234(12):22321—22330.
66. Comtesse N, Keller A, Diesinger I, et al. Frequent overexpression of the genes FXR1, CLAPM1 and EIF4G located on amplicon 3q26-27 in squamous cell carcinoma of the lung. *Int J Cancer*. 2007;120(12):2538—2544.
67. Dou N, Yang D, Yu S, Wu B, Gao Y, Li Y. SNRPA enhances tumour cell growth in gastric cancer through modulating NGF expression. *Cell Prolif*. 2018;51(5):e12484.
68. Yin L, Cai Z, Zhu B, Xu C. Identification of key pathways and genes in the dynamic progression of HCC based on WGCNA. *Genes*. 2018;9(2):92.
69. Guo C, Tang Y, Zhang Y, Li G. Mining TCGA data for key biomarkers related to immune microenvironment in endometrial cancer by immune score and weighted correlation network analysis. *Front Mol Biosci*. 2021;8:645388.
70. She Y, Jin Z, Wu J, et al. Development and validation of a deep learning model for non-small cell lung cancer survival. *JAMA Netw Open*. 2020;3(6):e205842.
71. Liang W, Zhang L, Jiang G, et al. Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. *J Clin Oncol*. 2015;33(8):861—869.
72. Keam B, Kim DW, Park JH, et al. Nomogram predicting clinical outcomes in non-small cell lung cancer patients treated with epidermal growth factor receptor tyrosine kinase inhibitors. *Cancer Treat Res*. 2014;46(4):323—330.