Received: 19 July 2021 | Revised: 6 November 2021 | Accepted: 4 December 2021

IET Systems Biology

DOI: 10.1049/syb2.12041

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Identification and validation of a seven-gene prognostic marker in colon cancer based on single-cell transcriptome analysis

Yang Zhou[1] | **Yang Guo**[2] | **Yuanhe Wang**[1]

[1]Medical Oncology Department of Gastrointestinal Cancer, Liaoning Cancer Hospital & Institute, Cancer Hospital of China Medical University, Liaoning Province, China

[2]Shenyang Tenth People's Hospital (Shenyang Chest Hospital), Shenyang, Liaoning, P. R. China

Correspondence

Yuanhe Wang, Medical Oncology Department of Gastrointestinal Cancer, Liaoning Cancer Hospital & Institute, Cancer Hospital of China Medical University, No.44 Xiaoheyan Road, Dadong District, Shenyang 110042, Liaoning Province, China.
Email: wangyuanhe@sina.com

Funding information

2020 Natural Science Foundation of Liaoning Province, Grant/Award Number: 2020-MS-064

Abstract

Colon cancer (CC) is one of the most commonly diagnosed tumours worldwide. Single-cell RNA sequencing (scRNA-seq) can accurately reflect the heterogeneity within and between tumour cells and identify important genes associated with cancer development and growth. In this study, scRNA-seq was used to identify reliable prognostic biomarkers in CC. ScRNA-seq data of CC before and after 5-fluorouracil treatment were first downloaded from the Gene Expression Omnibus database. The data were pre-processed, and dimensionality reduction was performed using principal component analysis and t-distributed stochastic neighbour embedding algorithms. Additionally, the transcriptome data, somatic variant data, and clinical reports of patients with CC were obtained from The Cancer Genome Atlas database. Seven key genes were identified using Cox regression analysis and the least absolute shrinkage and selection operator method to establish signatures associated with CC prognoses. The identified signatures were validated on independent datasets, and somatic mutations and potential oncogenic pathways were further explored. Based on these features, gene signatures, and other clinical variables, a more effective predictive model nomogram for patients with CC was constructed, and a decision curve analysis was performed to assess the utility of the nomogram. A prognostic signature consisting of seven prognostic-related genes, including *CAV2*, *EREG*, *NGFRAP1*, *WBSCR22*, *SPINT2*, *CCDC28A*, and *BCL10*, was constructed and validated. The proficiency and credibility of the signature were verified in both internal and external datasets, and the results showed that the seven-gene signature could effectively predict the prognosis of patients with CC under various clinical conditions. A nomogram was then constructed based on features such as the RiskScore, patients' age, neoplasm stage, and tumor (T), nodes (N), and metastases (M) classification, and the nomogram had good clinical utility. Higher RiskScores were associated with a higher tumour mutational burden, which was confirmed to be a prognostic risk factor. Gene set enrichment analysis showed that high-score groups were enriched in 'cytoplasmic DNA sensing', 'Extracellular matrix receptor interactions', and 'focal adhesion', and low-score groups were enriched in 'natural killer cell-mediated cytotoxicity', and 'T-cell receptor signalling pathways', among other pathways. A robust seven-gene marker for CC was identified based on scRNA-seq data and was validated in multiple independent cohort studies. These findings provide a new potential marker to predict the prognosis of patients with CC.

KEYWORDS

colon cancer (CC), metastasis-associated genes, progression, single-cell RNA sequencing (scRNA-seq), tumour mutational burden (TMB), the cancer genome atlas (TCGA)

The Yang Zhou and Yang Guo are contributed equally.

1 | INTRODUCTION

Colon cancer (CC) is the fourth most frequently diagnosed cancer and the third leading cause of cancer deaths worldwide [1]. In 2018, an estimated 97,220 new cases of CC were diagnosed, accounting for 6.1% of the annual global cancer cases [2, 3]. Although many new treatment methods have been proposed for CC, the long-term survival rate of patients is unsatisfactory due to tumour recurrence and metastasis. Moreover, the 5-year survival rate of patients with CC is only approximately 50% [4]. Previously, systemic chemotherapy based on fluorouracil was the main treatment modality for patients with metastatic CC. 5-fluorouracil (5-Fu), an antimetabolite analogue of pyrimidine, inhibits nucleoside metabolism and DNA synthesis, leading to apoptosis [5]. Several studies have focussed on the development of effective vectors to improve the targeted delivery of 5-Fu; attempts have been made to increase the bioavailability of 5-Fu and decrease its toxicity in vivo by administering lower doses and ensuring the optimum accumulation of the drug at the affected site [6–9]. Despite these advances, drug resistance remains a major limitation to the clinical use of 5-Fu. Novel drug-related combined chemotherapy, which targets both epidermal growth factor receptor (EGFR) signalling and anti-angiogenic pathways that inhibit vascular endothelial growth factor (EGF), is now considered the standard first-line treatment of CC [10–14]. Clinical trials on immunotherapy, represented by immune checkpoint inhibitors, are currently underway in several countries [15–18]. Despite these advances, treatment of CC is limited by a poor progression-free survival (PFS); a minority of cancer cells will continue to proliferate after chemotherapy, leading to treatment failure. There is an urgent need to investigate the molecular mechanisms of drug resistance and discover advanced predictive biomarkers in CC. Numerous biomarkers associated with CC prognosis have been identified, including Kirsten rat sarcoma virus, NRAS or BRAF mutations, Her2 amplification, microsatellite instability, defective mismatch repair, Neurotrophic tyrosine receptor kinase fusion proteins, and PIK3CA mutations [19]. These biomarkers are predictors of efficacy that are used to guide the methods of targeted therapies and immunotherapies.

Single-cell RNA sequencing (ScRNA-seq) is a powerful new technology that allows high-throughput sequencing analysis of the genome, transcriptome, and epigenome at the single-cell level for the detection of clinically important tumour subpopulations. It is an essential tool for studying tumour progression and understanding tumour heterogeneity [20, 21]. Since the first scRNA-seq analysis was conducted in 2009 [22], this approach has begun to address key questions in various tumour types, including glioblastoma, hepatocellular carcinoma, metastatic renal cell carcinoma, and breast and lung adenocarcinoma [23–27]. ScRNA-seq analysis uncovers intercellular heterogeneity by revealing the gene structure and gene expression status of single cells. It overcomes the limitations of traditional high-throughput sequencing and identifies important genes with true tumour cell characteristics [28, 29]. CC is a highly heterogeneous tumour; patients with CC sometimes show significantly different clinical outcomes despite identical aetiologies and treatments. Results from several transcriptomic analyses have shown that stromal cell characteristics are associated with the risk of CC recurrence. These characteristics can predict patient survival, highlighting the importance of multiple cell populations in CC [30, 31]. Dai et al. conducted scRNA-seq to generate comprehensive single-cell expression profiles of cancer tissues from patients with CRC. Their analyses aimed to facilitate the understanding of how activated and deactivated aberrant cell subpopulations contribute to the onset, maintenance, and progression of CC [32]. In the coming years, scRNA-seq is expected to greatly improve our understanding of invasion, metastasis, and therapeutic resistance of cancer cells.

In this study, the scRNA-seq data of CC was first downloaded from the Gene Expression Omnibus (GEO) database to identify all known genomic features and marker genes both before and after 5-Fu treatment. The transcriptomic data, somatic variant data, and clinical data of patients with CC were obtained from The Cancer Genome Atlas (TCGA), and a seven-gene signature associated with CC prognosis was developed using Cox regression analysis and the least absolute shrinkage and selection operator (LASSO) algorithm. Based on these gene signatures and other clinical variables, an effective predictive model nomogram for patients with CC was constructed. The findings of this study suggest that the genes of the prognostic signature are associated with drug resistance and that they play a critical role in CC development. Furthermore, these genes may be employed as potential targets for the treatment of patients with CC.

2 | METHODOLOGY

2.1 | Data source and pre-processing

We had searched for single-cell sequencing data related to CC in the GEO database by keywords 'scRNA-seq' and 'Colon cancer' and obtained the GSE149224 dataset. Illumina HiSeq 4000 was used as the sequencing platform and Drop-seq was used as the sequencing method. This single-cell transcriptome sequencing data included data on 23,768 genes and 11,126 cells. Additionally, GMT files of gene symbols from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway were downloaded from the Molecular Signatures Database (MSigDB).

Data pre-processing, dimensionality reduction, clustering, and visualization of scRNA-seq data of the GSE149224 dataset were performed using the 'Seurat' package in R. Data were pre-processed using the following criteria: each cell showed the expression of at least 200 genes; each gene was expressed in at least three cells; the mitochondrial gene content was less than 5%; the number of QC genes was between 200 and 2500. Dimensionality reduction was performed using principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) algorithms, and single-cell datasets were dimensionally reduced using a plot-based approach and were visualized using the t-SNE method.

Finally, the expression profile dataset containing information on 18,860 genes and 2699 cells was obtained.

2.2 | Construction of the prediction model

Using the feature gene set obtained above, TCGA-COAD data were divided into training and test sets at a 1:1 ratio. The marker genes were obtained by analysing the differences in single-cell sequencing data before and after drug treatment. Univariate and multivariate Cox analyses were performed through the *survival* package [33], and LASSO analysis was conducted through the cv.glmnet function of *glmnet* package [34]. The above operations were all performed in the R software (version 3.6.2). Specifically, we first performed univariate Cox analysis on marker genes before and after drug treatment, and then further screened genes through the LASSO algorithm. Finally, we constructed a multi-gene prediction model through multivariate Cox analysis.

2.3 | Analysis of the entire TCGA-COAD cohort

To investigate the prognostic relationship between these feature genes and samples from the entire TCGA-COAD cohort, univariate and multivariate Cox regression analyses were performed on the entire TCGA-COAD dataset based on the expression pattern of the aforementioned feature genes. RiskScores were analysed for variations in different stages of clinical features (tumour stage, T stage, M stage, and N stage). Receiver operating characteristic (ROC) analysis was performed to verify the accuracy of the model in predicting the 1-, 3-, and 5-year survival rates. Furthermore, Kaplan–Meier survival analysis was performed to investigate the prognostic differences between high- and low-risk groups.

2.4 | Construction of a nomogram

The RiskScores of clinical features (tumour stage, age, T stage, M stage, and N stage) were integrated to construct nomograms for predicting the prognosis of patients with CC. First, the clinical features associated with PFS were screened by univariate ($p < 0.05$) and multivariate Cox regression analyses, followed by construction of the nomograms. The prediction efficacy of the nomograms was evaluated by calibration curves, ROC analysis, and survival analysis.

2.5 | Analysis of the tumour mutational burden

The tumour mutational burden (TMB) is defined as the number of non-synonymous somatic mutations per Mb region. We used the *TCGAbiolinks* package [35] in R to download the MAF file of CC somatic mutations from the TCGA database and used Perl programming to perform calculations. The RiskScore of each patient was obtained based on the clinical features, and the patients were divided into high- and low-risk groups using the median RiskScore as the cut-off value. The mutation data of the high and low-risk groups were visualized using the 'maftools' package in R.

2.6 | Gene set enrichment analysis

Gene set enrichment analysis (GSEA) analysis was implemented using the *clusterProfiler* package [36] in the R software (version 3.6.2). It was performed on the entire TCGA-COAD dataset (reference dataset: KEGG Pathway of MSigDB) using the value of RiskScores as a classification criterion to investigate the differences in significantly enriched pathways between high- and low-risk groups.

3 | RESULTS

3.1 | Single-cell data analysis

The distribution of genes, cells, and mitochondrial genes for each sample is shown in Figure 1a. The distribution was based on the screening criteria mentioned in the methodology. The differences in the distribution of genes, cells, and mitochondrial genes for each sample were more obvious when the number of PCs was 12 (Figure 1b). Therefore, the top 12 PCs were selected for t-SNE dimensionality reduction, and the final clustering yielded eight subpopulations. Additionally, the top 10 marker genes were selected from each subpopulation for visualization (Figure 1c). The distribution of subpopulations, patients, and treatments is shown in Figure 1d–f, from which it can be seen that the difference between medication and non-medication groups was significant. The differentially expressed genes (DEGs) associated with drug resistance were obtained using treatment as the grouping criteria.

3.2 | Construction of the prediction model

Based on the DEGs, a univariate Cox regression analysis was performed in the TCGA-COAD training set, while LASSO regression was performed using the 'glmnet' package in R (Figure 2a,b). The trajectory of the independent variables showed that as the lambda gradually decreased, the number of independent variable coefficients tending to zero gradually increased. The model was constructed using 10-fold cross-validation, and the confidence interval under each lambda was analysed as shown in the figure. The results showed that the model reached the optimum value when log(lambda) = −5.6. For this reason, 12 genes at log(lambda) = −6.3 were selected as target genes, and a multivariate Cox analysis of these target genes was performed to screen out seven genes associated with OS. The RiskScore was calculated using the following formula:

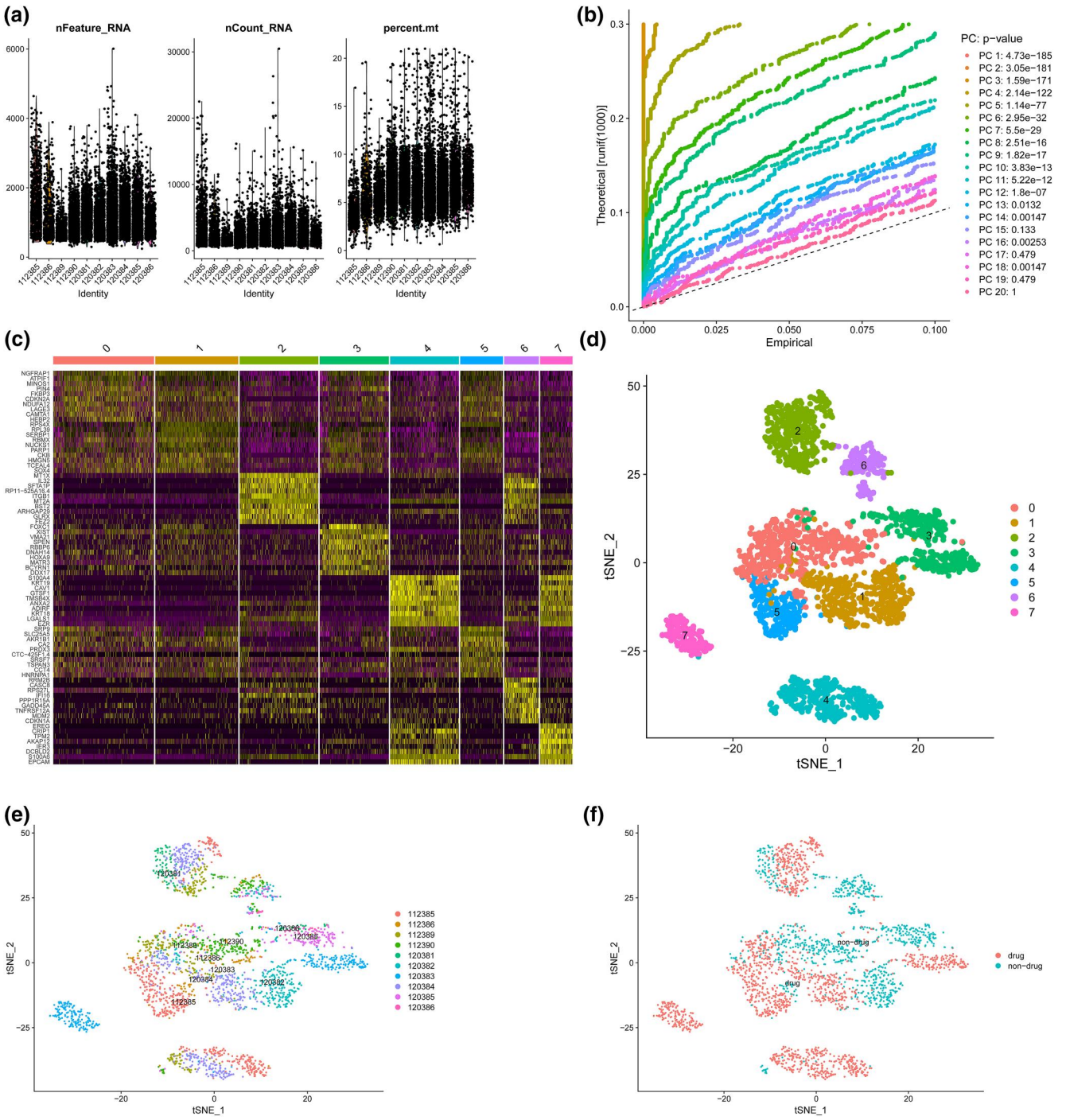

FIGURE 1 (a) Single-cell RNA sequencing data were subjected to quality control, where low quality cells and lowly expressed genes were removed; (b) The number of principal components (PCs) for principal component analysis based on the *p*-value; (c) Heat map of clustered feature genes for each subpopulation; (d) Clustering map of t-distributed stochastic neighbour embedding (t-SNE) dimensionality reduction, where all colon cancer single-cell data were clustered into eight categories; (e) t-SNE distribution of patients with colon cancer; (f) Distribution of patients with or without treatment

RiskScore = 0.557*exp (*CAV2*)−0.3*exp (*EREG*) + 0.239*exp (*NGFRAP1*) + 0.818*exp (*WBSCR22*) + 0.766*exp (*SPINT2*) + 0.7824*exp (*CCDC28A*)−0.710* (*BCL10*). (Note: The coefficient has been mentioned before the exponent, and the gene name has been mentioned within brackets).

These genes were found to be differentially expressed in separate subpopulations of single cells (Figure 2c).

3.3 | Validation of the prediction model

To validate the prediction efficacy of the seven-gene model, ROC and survival analyses were performed in the training set (Figure 3a,b), the entire dataset (Figure 3d,e), and the GSE17536 dataset (Figure 3g,h). The results revealed that the accuracy of the model was better in predicting the 1-, 3-, and 5-

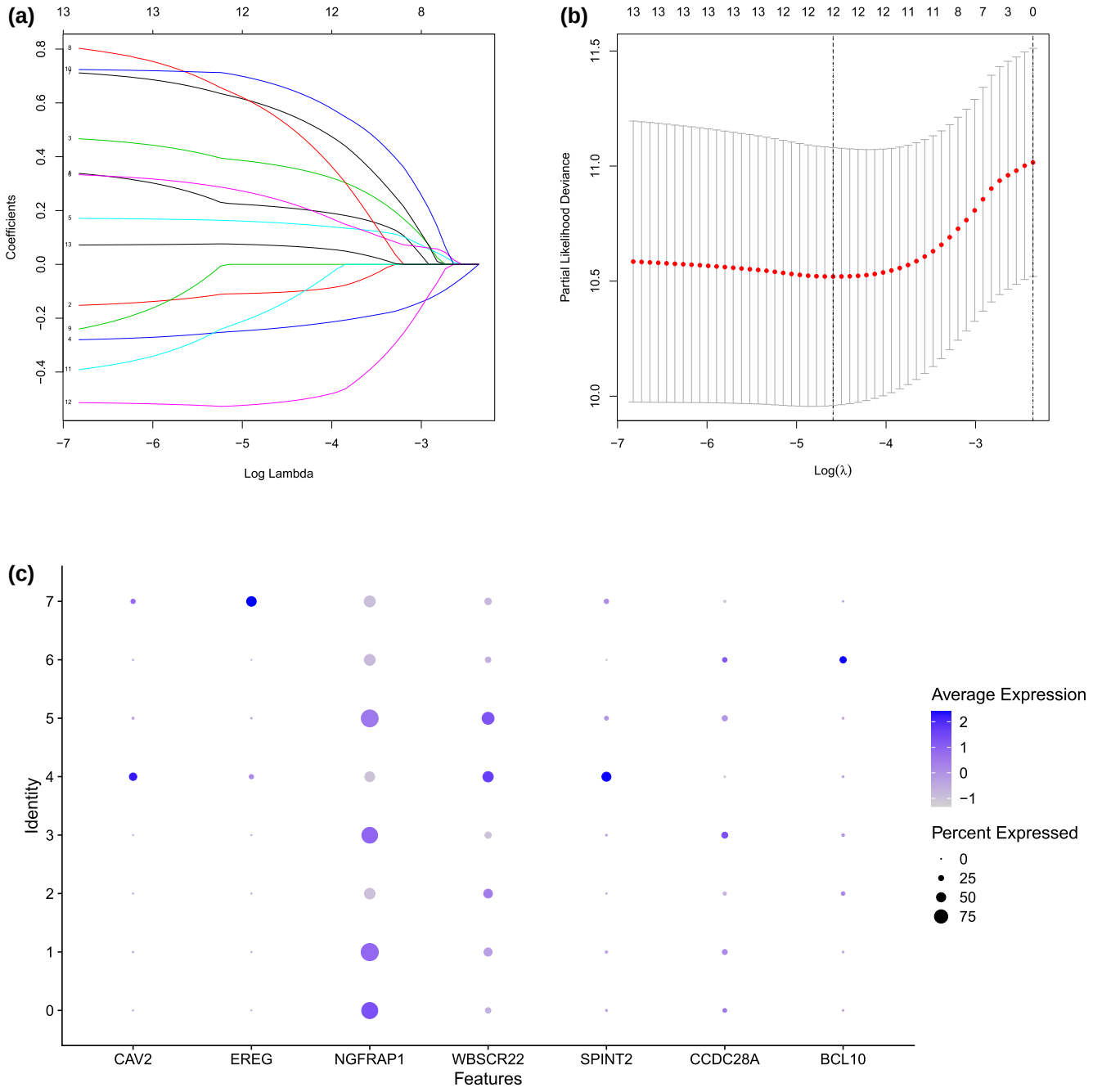

FIGURE 2 (a and b) Twelve prognostic genes were identified in the The Cancer Genome Atlas training cohort based on the least absolute shrinkage and selection operator approach using the 'glmnet' package in R (best cut-off value, −4.6); (c) The expression of these seven genes was analysed by multivariate Cox analysis in each subpopulation

year survival rates in both the training set and the entire TCGA-COAD cohort, as all values of the area under the curve (AUC) were greater than 0.6. Results of Kaplan–Meier survival analysis showed significant differences in OS between the high- and low-risk groups. The OS was higher when the RiskScore was greater than the median OS and vice versa. The risk curves and distribution plots of patients based on RiskScores revealed that mortality was higher in patients with high RiskScores in the training set (Figure 3c), entire TCGA-COAD cohort (Figure 3f), and GEO cohort (Figure 3i).

3.4 | Clinical correlation analysis

In the entire TCGA-COAD dataset, the differences in RiskScores between various clinical features were analysed (Figure 4a–d). This study indicated that the differences in RiskScores between tumour stage, T stage, and N stage were significant and that the RiskScores increased with advanced tumour stage, T stage, and N stage. Therefore, these results further confirmed the accuracy of the risk model. Next, the ROC curves were used to analyse the efficacy of the prediction

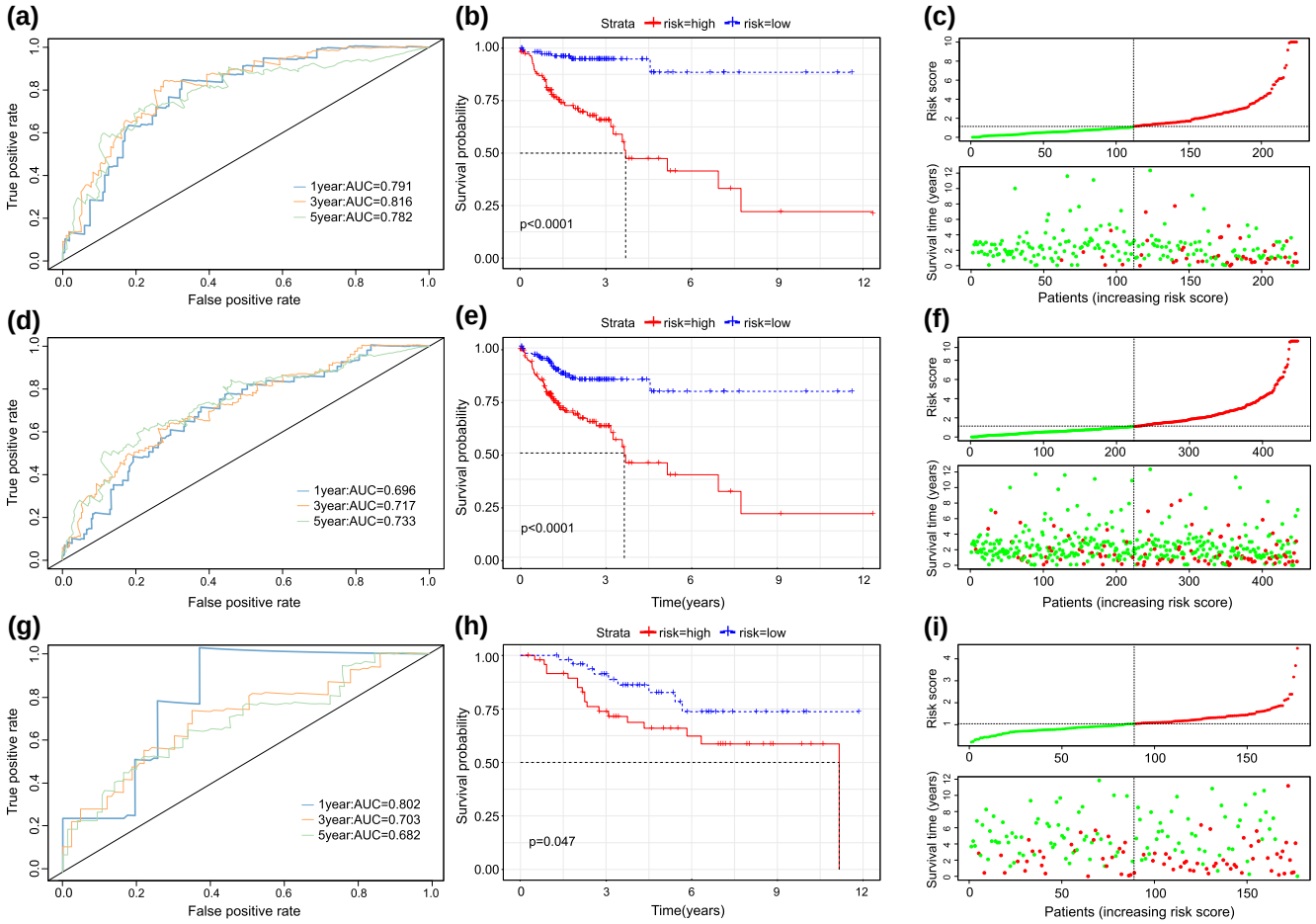

FIGURE 3 (a) Receiver operating characteristic (ROC) analysis of the risk model in the training set; (b) Survival analysis of the risk model in the training set; (c) Distribution of RiskScore and survival status in the training set; (d) ROC analysis of the risk model in the entire The Cancer Genome Atlas (TCGA) cohort; (e) Survival analysis of the risk model in the entire TCGA cohort; (f) Distribution of RiskScore and survival status in the entire TCGA cohort; (g) ROC analysis of the risk model in the GSE17536 validation set; (h) Survival analysis of the risk model in the GSE17536 validation set; (i) Distribution of RiskScore and survival status in the GSE17536 validation set

model in predicting patient prognosis at 1, 3, and 5 years. The results showed that the risk model could accurately predict the prognosis of patients, as all values of AUCs were greater than 0.6. Results of survival analysis also revealed significant differences in PFS between high- and low-risk groups ($p < 0.001$).

3.5 | Nomogram construction

Clinical features such as the RiskScore, age, tumour stage, T stage, M stage, and N stage were used to construct the nomogram. The clinical features associated with prognosis, that is, tumour stage and RiskScore (high- and low-risk groups), were screened by univariate and multivariate Cox analyses ($p < 0.05$; Figure 5a,b). Moreover, nomograms (Figure 5c) were constructed based on the tumour stage and RiskScore, by which the prognosis of patients at 1, 3, and 5 years could be predicted. The calibration curve (Figure 5d), ROC curve (Figure 5e), and survival analysis (Figure 5f) all showed that the nomogram had a high prediction efficacy.

3.6 | TMB analysis

RiskScores for each patient in the entire TCGA-COAD cohort were obtained based on genes associated with drug resistance. Subjects were divided into high- and low-risk groups using the median risk score as the cut-off value, and differences in the mutation frequency among the top 20 genes were identified between the high-risk (Figure 6a) and low-risk groups (Figure 6b). In the high- and low-risk groups, there were significant differences in the mutation frequency of genes, such as *APC* (66% vs. 80%), *TTN* (59% vs. 47%), *TP53* (50% vs. 58%), and *FAT4* (29% vs. 23%).

GSEA of the top 20 genes was performed on the entire TCGA-COAD cohort using the RiskScore as a classification criterion (Figure 6c). The results showed that genes in the high-risk group were enriched in the pathways of 'cytosolic DNA sensing', 'Extracellular matrix receptor interaction', and 'focal adhesion'. However, genes in the low-risk group were enriched in the pathways of 'natural killer cell-mediated cytotoxicity' and 'T cell receptor signalling'.

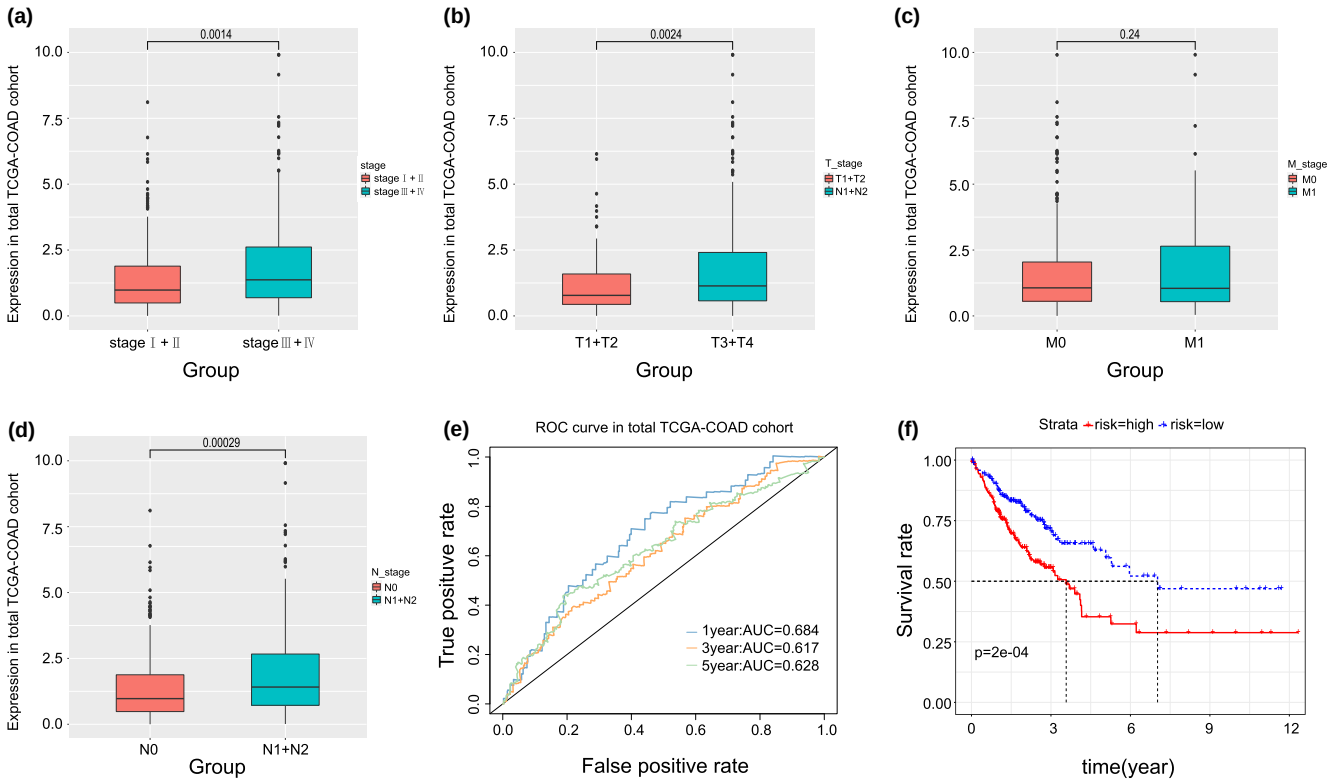

FIGURE 4 (a) Box plots showing the distribution of RiskScores in the entire TCGA-COAD cohort according to different tumour stages; (b) Box plots showing the distribution of RiskScores in the entire TCGA-COAD cohort according to different T stages; (c) Box plots showing the distribution of RiskScores in the entire TCGA-COAD cohort according to different M stages; (d) Box plots showing the distribution of RiskScores in the full set of TCGA-COAD according to different N stages; (e) Receiver operating characteristic (ROC) curves for 1-, 3-, and 5-year survival predicted by the risk model; (f) Survival analysis between high and low-risk groups. TCGA, The Cancer Genome Atlas

4 | DISCUSSION

CC is a frequently diagnosed cancer worldwide. In recent decades, further optimization of 5-Fu-based chemotherapy has improved the survival prospects of patients; however, some patients show a poor response to treatment. Therefore, it is important to elucidate the underlying mechanisms associated with the inefficacy of CC treatment and to identify biomarkers that can predict treatment efficacy in patients with CC. Conventional tissue sequencing uses a mixture of million or more cells, and the results represent information on the average transcriptome expression or the expression of dominant transcripts in a group of cells. In contrast, scRNA-seq can accurately reflect the heterogeneity within and between tumour cells and identify the important genes that truly characterize cancer cells [37]. The genetic markers based on the scRNA-seq data of CC cells can be used as reliable biomarkers for predicting the prognosis of CC.

In this study, data pre-processing, dimensionality reduction, clustering, and visualization of scRNA-seq data of CC were performed by bio-signalling methods to characterize the genomes of CC before and after 5-Fu treatment. PCA was then implemented, and linear dimensionality reduction was performed while maintaining as many data characteristics as possible. Following this, nonlinear dimensionality reduction was performed using the t-SNE algorithm. DEGs were initially screened by univariate Cox regression models, and then variables were further optimized and selected using the multivariate Cox regression method. Finally, key genes associated with CC prognosis were identified. The signature was validated by internal and independent researchers, and the seven-gene signature was found to effectively predict the prognosis of patients with CC under various clinical conditions. Additionally, a nomogram was constructed based on clinical features such as the RiskScore, age, tumour stage, T stage, M stage, and N stage, and it was found to have good clinical utility. A higher RiskScore was associated with a higher TMB, which was shown to be a prognostic risk factor. However, whether these prognostic markers can predict the efficacy of treatment remains to be further studied.

The following seven genes associated with CC prognosis were identified: *CAV2*, *EREG*, *NGFRAP1*, *WBSCR22*, *SPINT2*, *CCDC28A*, and *BCL10*. Some of these genes play important roles in the progression of other human cancer types. Caveolin-2 (CAV2) is a member of the caveolae family [38]. High expression of CAV2 is associated with the progression of different types of cancers, including lung, breast, prostate, pancreatic, breast, and kidney cancers [39–44]. Nevertheless, the exact role of CAV2 in CC remains to be explored. Epiregulin (EREG) belongs to the EGF family, whose members bind to the EGFR or ErbB4 to generate signals for proliferation, migration, differentiation, cytokine

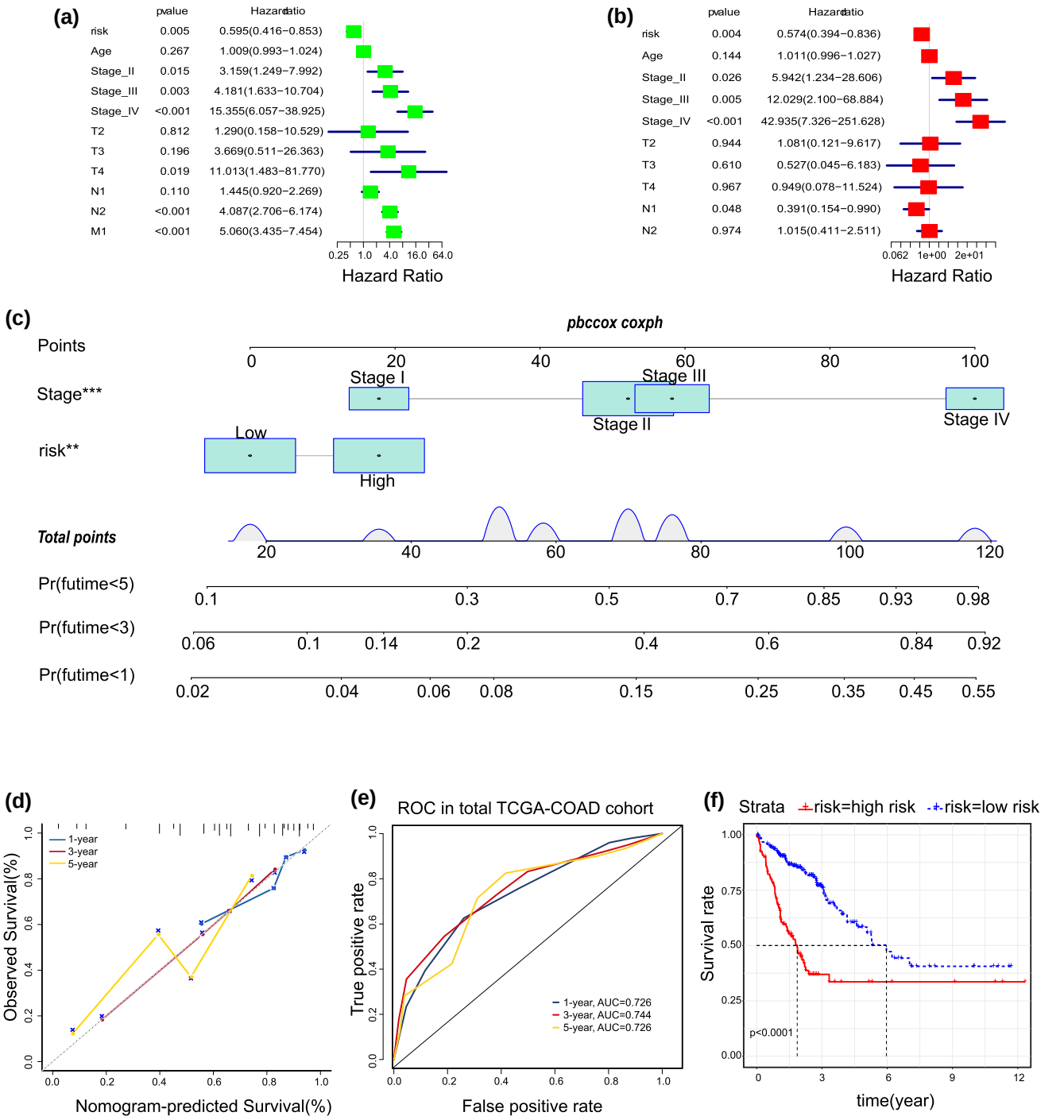

FIGURE 5 (a) Forest plot of univariate Cox analysis; (b) Forest plot of multivariate Cox analysis; (c) Nomogram of the prediction model; (d) 1-, 3-, and 5-year calibration curves of the nomogram; (e) Receiver operating characteristic (ROC) curves for 1-, 3-, and 5-year survival predicted by the risk model; (f) Survival analysis between high- and low-risk groups predicted by the nomogram

secretion, and innate immunity [45]. Elevated expression of EREG is associated with a variety of human cancers; upregulated EREG promotes tumour progression and metastasis, thus reducing the survival time of patients [46–51]. High EREG expression is associated with a better prognosis in patients with CC receiving neoadjuvant concurrent radiotherapy [52]. Therefore, EREG can be used as a potential predictive marker and therapeutic target for patients with CC who received neoadjuvant concurrent radiotherapy. Results from a clinical trial have confirmed that *EREG* gene expression can be used as a predictor of OS in patients with mCRC treated with oxaliplatin/fluoropyrimidine in combination with bevacizumab. Moreover, *EREG* gene expression is considered a predictor of EGFR antibody efficacy [53]. In pre-treated K-Ras wild-type status CC, patients with high *EREG* gene expression appeared to benefit more from cetuximab

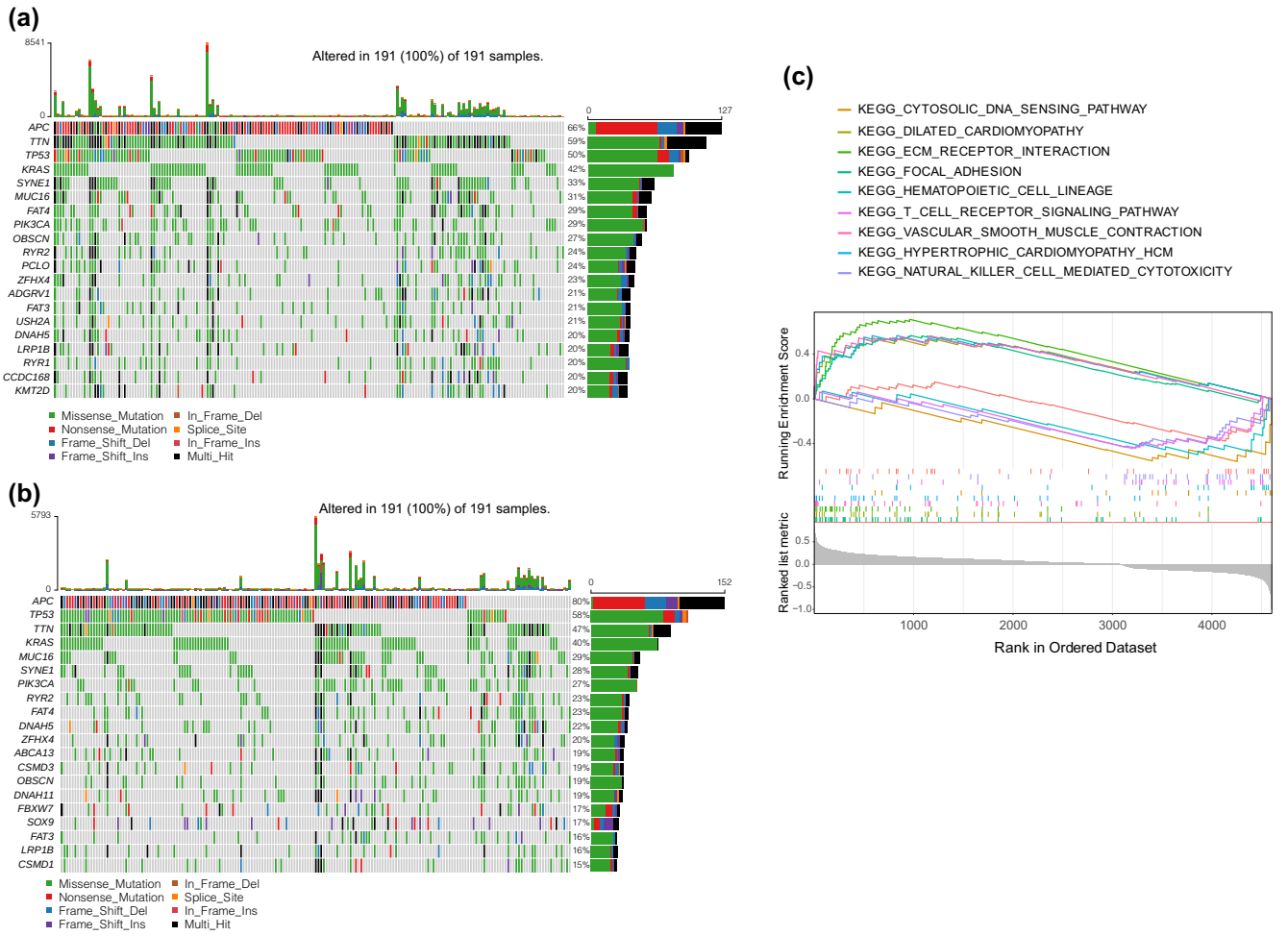

FIGURE 6 (a) The tumour mutational burden (TMB) in the high-risk group predicted by the risk model; (b) The TMB in the low-risk group predicted by the risk model; (c) Results of gene set enrichment analysis

treatment [54]. Many studies have validated our results that EREG can be used as a predictive biomarker for assessing the efficacy of CC treatment.

Nerve growth factor receptor-associated protein 1 (*NGFRAP1*), also known as brain-expressed X-linked protein 3, is an apoptosis-associated gene whose expression is downregulated in certain solid organ malignancies and chronic lymphocytic leukaemia [55, 56] The Williams-Beuren syndrome chromosome region 22 (*WBSCR22*) gene [57] is involved in the proliferation, invasion, and metastasis of cancer cells [58]. In CC, the *WBSCR22* gene is involved in oxaliplatin resistance, suggesting that *WBSCR22* may represent a novel oxaliplatin resistance biomarker as well as a potential target for CRC therapy [59, 60]. This finding is in agreement with the findings of our study. Serine peptidase inhibitor Kunitz type 2 (SPINT2) is a proteinase inhibitor of hepatocyte growth factor activator (HGFA) [61], which plays an important role in deactivating the HGFA-MET pathway and promoting the progression of multiple malignancies [62–66].

The biological function of coiled-coil domain containing protein 28A (CCDC28A) has not been determined; however, some studies have confirmed that it serves a recurrent chromosomal translocation partner of nucleoporin 98 in acute leukaemia [67]. B-cell lymphoma/leukaemia 10 (BCL10) positively regulates the intracellular signalling protein κB in lymphocyte proliferation by coupling antigen receptor-induced signalling in B and T cells to the activation of the transcription factor NF-κ [68]. Additionally, constitutive activation of the NF-κB signalling pathway plays a key role in the pathogenesis of activated B-cell-like diffuse large B-cell lymphoma (ABC-DLBCL), which is the most aggressive and chemo-resistant form of DLBCL [69].

Previous studies have reported that the seven identified signature genes play a crucial role in the progression of several malignancies and have highlighted the importance of these genes in the drug resistance of certain tumours. Significant differences in the expression of these specific genes have been observed in different subpopulations of single cells. Therefore, it is reasonable to believe that all seven genes have great potential as prognostic biomarkers associated with drug resistance in CC.

The GSEA was performed to further study the pathways involved in the seven-gene signature. We divided the samples into high- and low-risk groups based on the value of the Riskscore. The results showed that the high-risk group was

enriched in the pathways of 'cytosolic DNA sensing', 'ECM receptor interaction', and 'focal adhesion' and that the low-risk group was enriched in the pathways of 'natural killer cell-mediated cytotoxicity', and 'T-cell receptor signalling'. Previous studies have shown that 'cytosolic DNA sensing' was closely associated with the secretion of cytokines supporting innate and adaptive anti-tumour immunity [70, 71]. The significance of 'ECM receptor interaction pathways' suggests that tumour cell-environment interactions are dynamic [72]. Related studies have elucidated that ECM is upregulated in prostate cancer tissues and is involved in both tumour invasion and metastasis in gastric, glioblastoma, and breast cancers [73–76]. Moreover, ECM promotes the progression of epithelial-mesenchymal transition in CC cells [77].

The pathways significantly enriched in the low-risk score group are mainly associated with multiple tumour immune mechanisms, among which 'natural killer cell-mediated cellular cytotoxicity' affects the proliferation and migration of tumour cells by altering their immune microenvironment [78–80]. The 'T-cell receptor signalling' balances the differentiation, maintenance, and function of regulatory T (Treg) cells and affects the gene expression, metabolism, cell adhesion, and migration of these cells [81].

In summary, the seven-gene signature is associated with important signalling pathways in tumours. Based on the results of scRNA-seq and subsuequent validation, this signature was found to perform well on both internal and external datasets. Moreover, the seven-gene signature was associated with somatic mutation profiles in patients with CC. This is a simple model with good performance features that can be applied in the clinic. Nevertheless, there are still some limitations to this study. The sample size in this study was limited, and the cohort was not large enough, which may have affected the statistical validity and accuracy of our results. Moreover, this study is based only on bioinformatic analysis. Therefore, complementary and basic experiments are still needed to reveal the specific mechanisms of action of the signature gene markers in promoting tumour progression. Further studies will be necessary to explore the underlying molecular mechanisms of action of these genes, to demonstrate their applicability in clinical applications.

In conclusion, this study explored CC cell heterogeneity and genomic features based on scRNA-seq before and after 5-Fu treatment. It identified a reliable seven-gene drug resistance-associated signature, established a prediction model for CC progression, and provided new clinical guidance for drug sensitivity and prognosis prediction to develop personalized treatment regimens.

ACKNOWLEDGMENTS

Our research was funded by the 2020 Natural Science Foundation of Liaoning Province (2020-MS-064).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

All data, or code generated or used during the study are available from the corresponding author by request.

ORCID

Yuanhe Wang https://orcid.org/0000-0002-0012-1756

REFERENCES

1. Siegel, R.L., et al.: Cancer statistics, 2021. CA Cancer J. Clin. 71(1), 7–33 (2021). https://doi.org/10.3322/caac.21654
2. Sung, H., et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 71(3), 209–249 (2021)
3. Siegel, R.L., et al.: Colorectal cancer statistics, 2020. CA Cancer J. Clin. 70(3), 145–164 (2020)
4. Cheng, L., et al.: Trends in colorectal cancer incidence by anatomic site and disease stage in the United States from 1976 to 2005. Am. J. Clin. Oncol. 34(6), 573–580 (2011)
5. Longley, D.B., Harkin, D.P., Johnston, P.G.: 5-fluorouracil: mechanisms of action and clinical strategies. Nat. Rev. Cancer. 3(5), 330–338 (2003)
6. Astolfi, P., et al.: Lyotropic liquid-crystalline nanosystems as drug delivery agents for 5-fluorouracil: structure and cytotoxicity. Langmuir. 33(43), 12369–12378 (2017)
7. Handali, S., et al.: A novel 5-fluorouracil targeted delivery to colon cancer using folic acid conjugated liposomes. Biomed. Pharmacother. 108, 1259–1273 (2018)
8. Sharma, A., et al.: Stealth recombinant human serum albumin nanoparticles conjugating 5-fluorouracil augmented drug delivery and cytotoxicity in human colon cancer, HT-29 cells. Colloids Surf. B Biointerfaces. 155, 200–208 (2017)
9. Wu, P., et al.: Enhanced antitumor efficacy in colon cancer using EGF functionalized PLGA nanoparticles loaded with 5-Fluorouracil and perfluorocarbon. BMC Cancer. 20(1), 354 (2020)
10. Cremolini, C., et al.: FOLFOXIRI plus bevacizumab versus FOLFIRI plus bevacizumab as first-line treatment of patients with metastatic colorectal cancer: updated overall survival and molecular subgroup analyses of the open-label, phase 3 TRIBE study. Lancet Oncol. 16(13), 1306–1315 (2015)
11. Saltz, L.B., et al.: Bevacizumab in combination with oxaliplatin-based chemotherapy as first-line therapy in metastatic colorectal cancer: a randomized phase III study. J. Clin. Oncol. 26(12), 2013–2019 (2008)
12. Benson, A.B., et al.: Colon cancer, version 2.2021, NCCN clinical practice guidelines in oncology. J. Natl. Compr. Cancer Netw. 19(3), 329–359 (2021)
13. Pietrantonio, F., et al.: First-line anti-EGFR monoclonal antibodies in panRAS wild-type metastatic colorectal cancer: a systematic review and meta-analysis. Crit. Rev. Oncol. Hematol. 96(1), 156–166 (2015)
14. Carrato, A., et al.: First-line panitumumab plus FOLFOX4 or FOLFIRI in colorectal cancer with multiple or unresectable liver metastases: a randomised, phase II trial (PLANET-TTD). Eur. J. Cancer. 81, 191–202 (2017)
15. Overman, M.J., et al.: Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. Lancet Oncol. 18(9), 1182–1191 (2017)
16. Overman, M.J., et al.: Durable clinical benefit with nivolumab plus ipilimumab in DNA mismatch repair-deficient/microsatellite instability-high metastatic colorectal cancer. J. Clin. Oncol. 36(8), 773–779 (2018)
17. Morse, M.A., et al.: Safety of nivolumab plus low-dose ipilimumab in previously treated microsatellite instability-high/mismatch repair-deficient metastatic colorectal cancer. Oncol. 24(11), 1453–1461 (2019)
18. André, T., et al.: Pembrolizumab in microsatellite-instability-high advanced colorectal cancer. N. Engl. J. Med. 383(23), 2207–2218 (2020)
19. Lee, M.K.C., Loree, J.M.: Current and emerging biomarkers in metastatic colorectal cancer. Curr. Oncol. 26(Suppl 1), S7–S15 (2019)

20. Haque, A., et al.: A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 9(1), 75 (2017)
21. Navin, N.E.: The first five years of single-cell cancer genomics and beyond. Genome Res. 25(10), 1499–1507 (2015)
22. Tang, F., et al.: mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods. 6(5), 377–382 (2009)
23. Patel, A.P., et al.: Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 344(6190), 1396–1401 (2014)
24. D'Avola, D., et al.: High-density single cell mRNA sequencing to characterize circulating tumor cells in hepatocellular carcinoma. Sci. Rep. 8(1), 11570 (2018)
25. Kim, K.T., et al.: Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. Genome Biol. 17, 80 (2016)
26. Chung, W., et al.: Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat. Commun. 8, 15081 (2017)
27. Min, J.W., et al.: Identification of distinct tumor subpopulations in lung adenocarcinoma via single-cell RNA-seq. PLoS One. 10(8), e0135817 (2015)
28. Kulkarni, A., et al.: Beyond bulk: a review of single cell transcriptomics methodologies and applications. Curr. Opin. Biotechnol. 58, 129–136 (2019)
29. Nguyen, Q.H., et al.: Experimental considerations for single-cell RNA sequencing approaches. Front. Cell Dev. Biol. 6, 108 (2018)
30. Isella, C., et al.: Stromal contribution to the colorectal cancer transcriptome. Nat. Genet. 47(4), 312–319 (2015)
31. Calon, A., et al.: Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nat. Genet. 47(4), 320–329 (2015)
32. Dai, W., et al.: Single-cell transcriptional profiling reveals the heterogenicity in colorectal cancer. Medicine (Baltim). 98(34), e16916 (2019)
33. Therneau, T.: A package for survival analysis in R. R package version 3.2-13 (2021). https://CRAN.R-project.org/package=survival
34. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Software. 33(1), 1–22 (2010)
35. Colaprico, A., et al.: TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 44(8), e71 (2015). https://doi.org/10.1093/nar/gkv1507
36. Yu, G., et al.: clusterProfiler: an R package for comparing. Biological themes among gene clusters. OMICS A J. Integr. Biol. 16(5), 284–287 (2012). https://doi.org/10.1089/omi.2011.0118
37. Ellsworth, D.L., et al.: Single-cell sequencing and tumorigenesis: improved understanding of tumor evolution and metastasis. Clin. Transl. Med. 6(1), 15 (2017)
38. Hnasko, R., Lisanti, M.P.: The biology of caveolae: lessons from caveolin knockout mice and implications for human disease. Mol. Interv. 3(8), 445–464 (2003)
39. Zhou, X., et al.: CACNA1B (Ca(v)2.2) overexpression and its association with clinicopathologic characteristics and unfavorable prognosis in non-small cell lung cancer. Dis. Markers. 2017, 6136401 (2017)
40. Elsheikh, S.E., et al.: Caveolin 1 and Caveolin 2 are associated with breast cancer basal-like and triple-negative immunophenotype. Br. J. Cancer. 99(2), 327–334 (2008)
41. Sugie, S., et al.: Significant association of Caveolin-1 and Caveolin-2 with prostate cancer progression. Cancer Genom. Protem. 12(6), 391–6 (2015)
42. Jiao, F., et al.: Caveolin-2 is regulated by BRD4 and contributes to cell growth in pancreatic cancer. Cancer Cell Int. 20, 55 (2020)
43. Liu, F., Shangli, Z., Hu, Z.: CAV2 promotes the growth of renal cell carcinoma through the EGFR/PI3K/Akt pathway. OncoTargets Ther. 11, 6209–6216 (2018)
44. Savage, K., et al.: Distribution and significance of caveolin 2 expression in normal breast and invasive breast cancer: an immunofluorescence and immunohistochemical analysis. Breast Cancer Res. Treat. 110(2), 245–256 (2008)
45. Riese, D.J., 2nd, Cullum, R.L.: Epiregulin: roles in normal physiology and cancer. Semin. Cell Dev. Biol. 28, 49–56 (2014)
46. Xia, Q., et al.: Elevated epiregulin expression predicts poor prognosis in gastric cancer. Pathol. Res. Pract. 215(5), 873–879 (2019)
47. Wang, Y., et al.: Epiregulin reprograms cancer-associated fibroblasts and facilitates oral squamous cell carcinoma invasion via JAK2-STAT3 pathway. J. Exp. Clin. Cancer Res. 38(1), 274 (2019)
48. Zong, S., et al.: miR-451a and miR-106b-5p are associated with the cervical cancer development. Arch. Gynecol. Obstet. 299(4), 1089–1098 (2019)
49. Farooqui, M., et al.: Epiregulin contributes to breast tumorigenesis through regulating matrix metalloproteinase 1 and promoting cell survival. Mol. Cancer. 14, 138 (2015)
50. Kohsaka, S., et al.: Epiregulin enhances tumorigenicity by activating the ERK/MAPK pathway in glioblastoma. Neuro Oncol. 16(7), 960–970 (2014)
51. Liu, S., et al.: EREG-driven oncogenesis of head and neck squamous cell carcinoma exhibits higher sensitivity to erlotinib therapy. Theranostics. 10(23), 10589–10605 (2020)
52. Lin, C.Y., et al.: High EREG expression is predictive of better outcomes in rectal cancer patients receiving neoadjuvant concurrent chemoradiotherapy. Oncology. 98(8), 549–557 (2020)
53. Stintzing, S., et al.: Amphiregulin (AREG) and epiregulin (EREG) gene expression as predictor for overall survival (OS) in oxaliplatin/fluoropyrimidine plus bevacizumab treated mCRC patients-analysis of the phase III AIO KRK-0207 Trial. Front. Oncol. 8, 474 (2018)
54. Jonker, D.J., et al.: Epiregulin gene expression as a biomarker of benefit from cetuximab in the treatment of advanced colorectal cancer. Br. J. Cancer. 110(3), 648–655 (2014)
55. Kazi, J.U., Kabir, N.N., Rönnstrand, L.: Brain-expressed X-linked (BEX) proteins in human cancers. Biochim. Biophys. Acta. 1856(2), 226–233 (2015)
56. Gao, W., et al.: BEX3 contributes to cisplatin chemoresistance in nasopharyngeal carcinoma. Cancer Med. 6(2), 439–451 (2017)
57. Pober, B.R.: Williams-Beuren syndrome. N. Engl. J. Med. 362(3), 239–252 (2010)
58. Chi, Y., et al.: WBSCR22 confers cell survival and predicts poor prognosis in glioma. Brain Res. Bull. 161, 1–12 (2020)
59. Yan, D., et al.: WBSCR22 confers oxaliplatin resistance in human colorectal cancer. Sci. Rep. 7(1), 15443 (2017)
60. Zhao, H., et al.: Natural killer cells inhibit oxaliplatin-resistant colorectal cancer by repressing WBSCR22 via upregulating microRNA-146b-5p. Am. J. Cancer Res. 8(5), 824–834 (2018)
61. Roversi, F.M., Olalla Saad, S.T., Machado-Neto, J.A.: Serine peptidase inhibitor Kunitz type 2 (SPINT2) in cancer development and progression. Biomed. Pharmacother. 101, 278–286 (2018)
62. Hwang, S., et al.: Epigenetic silencing of SPINT2 promotes cancer cell motility via HGF-MET pathway activation in melanoma. J. Invest. Dermatol. 135(9), 2283–2291 (2015)
63. Yamamoto, K., et al.: Hepatocyte growth factor activator inhibitor type-2 (HAI-2)/SPINT2 contributes to invasive growth of oral squamous cell carcinoma cells. Oncotarget. 9(14), 11691–11706 (2018)
64. Yue, D., et al.: Epigenetic inactivation of SPINT2 is associated with tumor suppressive function in esophageal squamous cell carcinoma. Exp. Cell Res. 322(1), 149–158 (2014)
65. Pereira, M.S., et al.: Loss of SPINT2 expression frequently occurs in glioma, leading to increased growth and invasion via MMP2. Cell. Oncol. 43(1), 107–121 (2020)
66. Wang, N., et al.: Study on the methylation status of SPINT2 gene and its expression in cervical carcinoma. Cancer Biomark. 22(3), 435–442 (2018)
67. Petit, A., et al.: Functional analysis of the NUP98-CCDC28A fusion protein. Haematologica. 97(3), 379–387 (2012)
68. Thome, M., Tschopp, J.: Bcl10. Curr. Biol. 12(2), R45 (2002)
69. Bao, W., et al.: Targeting BCL10 by small peptides for the treatment of B cell lymphoma. Theranostics. 10(25), 11622–11636 (2020)
70. Kwon, J., Bakhoum, S.F.: The cytosolic DNA-sensing cGAS-STING pathway in cancer. Cancer Discov. 10(1), 26–39 (2020)
71. Vanpouille-Box, C., et al.: Cytosolic DNA sensing in organismal tumor control. Cancer Cell. 34(3), 361–378 (2018)

72. Kai, F., Drain, A.P., Weaver, V.M.: The extracellular matrix modulates the metastatic journey. Dev. Cell. 49(3), 332–346 (2019)
73. Bao, Y., et al.: Transcriptome profiling revealed multiple genes and ECM-receptor interaction pathways that may be associated with breast cancer. Cell. Mol. Biol. Lett. 24, 38 (2019)
74. Andersen, M.K., et al.: Integrative metabolic and transcriptomic profiling of prostate cancer tissue containing reactive stroma. Sci. Rep. 8(1), 14269 (2018)
75. Cui, X., et al.: Hacking macrophage-associated immunosuppression for regulating glioblastoma angiogenesis. Biomaterials. 161, 164–178 (2018)
76. Yan, P., et al.: In silico analyses for potential key genes associated with gastric cancer. PeerJ. 6, e6092 (2018)
77. Rahbari, N.N., et al.: Anti-VEGF therapy induces ECM remodeling and mechanical barriers to therapy in colorectal cancer liver metastases. Sci. Transl. Med. 8(360), 360ra135 (2016)
78. Malmberg, K.J., et al.: Natural killer cell-mediated immunosurveillance of human cancer. Semin. Immunol. 31, 20–29 (2017)
79. Myers, J.A., Miller, J.S.: Exploring the NK cell platform for cancer immunotherapy. Nat. Rev. Clin. Oncol. 18(2), 85–100 (2021)
80. Prager, I., Watzl, C.: Mechanisms of natural killer cell-mediated cellular cytotoxicity. J. Leukoc. Biol. 105(6), 1319–1329 (2019)
81. Li, M.O., Rudensky, A.Y.: T cell receptor signalling in the control of regulatory T cell differentiation and function. Nat. Rev. Immunol. 16(4), 220–233 (2016)

How to cite this article: Zhou, Y., Guo, Y., Wang, Y.: Identification and validation of a seven-gene prognostic marker in colon cancer based on single-cell transcriptome analysis. IET Syst. Biol. 16(2), 72–83 (2022). https://doi.org/10.1049/syb2.12041