Integrative Proteo-Genomic Analysis for Recurrent Survival Prognosis in Colon Adenocarcinoma

FeiYan Ai[1,2], Wenhao Wang[3], Shaojun Liu[1], Decai Zhang[1,2], Zhenyu Yang[1] and Fen Liu[1,2*]

[1] Department of Gastroenterology, The Third Xiangya Hospital of Central South University, Changsha, China, [2] Hunan Key Laboratory of Non-Resolving Inflammation and Cancer The Third Xiangya Hospital, Central South University, Changsha, China, [3] Department of Breast and Thyroid Surgery, The First Affiliated Hospital of University of South China, Hengyang, China

Background: The survival prognosis is the hallmark of cancer progression. Here, we aimed to develop a recurrence-related gene signature to predict the prognosis of colon adenocarcinoma (COAD).

Methods: The proteomic data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and genomic data from the cancer genomic maps [The Cancer Genome Atlas (TCGA)] dataset were analyzed to identify co-differentially expressed genes (cDEGs) between recurrence samples and non-recurrence samples in COAD using limma package. Functional enrichment analysis, including Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway was conducted. Univariate and multivariate Cox regressions were applied to identify the independent prognostic feature cDEGs and establish the signature whose performance was evaluated by Kaplan–Meier curve, receiver operating characteristic (ROC), Harrell's concordance index (C-index), and calibration curve. The area under the receiver operating characteristic (ROC) curve (AUROC) and a nomogram were calculated to assess the predictive accuracy. GSE17538 and GSE39582 were used for external validation. Quantitative real-time PCR and Western blot analysis were carried out to validate our findings.

Results: We identified 86 cDEGs in recurrence samples compared with non-recurrence samples. These genes were primarily enriched in the regulation of carbon metabolic process, fructose and mannose metabolism, and extracellular exosome. Then, an eight-gene-based signature (CA12, HBB, NCF1, KBTBD11, MMAA, DMBT1, AHNAK2, and FBLN2) was developed to separate patients into high- and low-risk groups. Patients in the low-risk group had significantly better prognosis than those in the high-risk group. Four prognostic clinical features, including pathological M, N, T, and RS model status, were screened for building the nomogram survival model. The PCR and Western blot analysis results suggested that CA12 and AHNAK2 were significantly upregulated, while MMAA

and DMBT1 were downregulated in the tumor sample compared with adjacent tissues, and in non-recurrent samples compared with non-recurrent samples in COAD.

Conclusion: These identified recurrence-related gene signatures might provide an effective prognostic predictor and promising therapeutic targets for COAD patients.

Keywords: colon adenocarcinoma, recurrence, prognosis, proteo-genomics, survival (MeSH)

INTRODUCTION

Colon adenocarcinoma (COAD) as the primary pathological type of colon cancer is one of the most commonly diagnosed malignancies and also a major cause of cancerous death throughout the world (1, 2). It was reported that there was an increasing approximately 1.1 million and 0.5 million deaths in COAD per year all over the world (3, 4). So far, continuous advancement of surgical technology has improved the survival rate of patients with COAD (5). However, the prognosis of patients, especially at advanced stage, remains poor due to high recurrent rate (6–8). Consequently, it is necessary to identify more characteristic and valuable biomarkers for the early prediction and treatment of COAD recurrence.

The evidence to date suggests the critical roles for abnormal expression of messenger RNAs (mRNAs) in COAD tumorigenesis by regulating a variety of biological processes, including cell proliferation, migration, and invasion. For example, FOXA1 expression was significantly higher in COAD tissues and associated with worse prognosis, which promoted cell proliferation, migration, and invasion in COAD cells (9). TBX19 mRNA expression was significantly increased in tumorous tissues compared to that in non-tumorous tissues, and increased TBX19 mRNA expression was associated with positive lymph node metastasis on colorectal carcinogenesis (10). Similarly, TMPRSS13 silencing in colorectal cancer cell lines increased apoptosis and impaired invasive potential (11). Moreover, bioinformatics analysis based on the public databases are believed to provide valuable information in disease prediction with the rapid development of gene sequencing technology (12, 13). The Cancer Genome Atlas (TCGA) database (14) and Gene Expression Omnibus (GEO) (15) have publicized various mRNA sequence data, which may provide novel information to improve the understanding of the molecular mechanisms of action in the tumorigenesis and progression of COAD. On the other hand, The Clinical Proteomic Tumor Analysis Consortium (CPTAC) consortium is aimed at characterizing the protein inventory in tumors by leveraging the latest developments in mass spectrometry-based discovery proteomics (16). Nowadays, integrating mRNA data and protein data thus helps to more exactly establish a genetic marker and prognostic model that can predict the recurrent survival prognosis of COAD patients.

The current study aimed to integrate the COAD genome and proteome expression level data to screen the factors that were significantly related to the recurrence of COAD at both mRNA and protein expression levels. We further screened the genes that were significantly related to the recurrent prognosis of COAD by

combining the clinical prognosis information of the sample and constructed an efficient prediction model of recurrence prognosis by combining clinical factors.

MATERIALS AND METHODS

Data Acquisition

The RNAseq data consisted of 372 COAD samples with a complete set of clinical information, including 78 recurrent and 294 non-recurrent samples downloaded from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) uploaded up to the August 15, 2020. Meanwhile, the protein expression profiles consisted of 34 COAD samples with corresponding clinical information, including 6 recurrent and 28 non-recurrent samples obtained from the Clinical Proteomic Tumor Analysis Consortium (CPTAC, https://cptac-data-portal.georgetown.edu/) database. In addition, we downloaded the two gene expression datasets (GSE17538 and GSE39582) of COAD from GEO database based on GPL570 Affymetrix Human Genome U133 Plus 2.0 Array according to the following criteria: (a) colon adenocarcinoma, (b) the organism is *Homo sapiens*, (c) samples containing recurrent information, and (d) COAD sample size exceeding 150 samples.

Identification of Co-Differentially Expression Genes

The limma package of R3.6.1 Version 3.34.7 (17) was used to identify the differentially expression genes (DEGs) between recurrent and non-recurrent samples in TCGA COAD samples. Similarly, the differentially expressed proteins (DEPs) were identified between recurrent and non-recurrent samples in CPTAC database using the limma package. A false discovery rate (FDR) <0.05 and $|\log_2$ fold change (FC)$|$ >0.263 were set as the criteria value for DEGs and DEPs. Visual hierarchical cluster analysis was conducted to display the volcano plot of DEGs and DEPs. The intersection of the DEGs and DEPs was visualized *via* Venn diagrams, which was considered the set of significant co-differentially expressed genes (cDEGs) for further analysis.

Function Enrichment Analysis

To reveal the functions of cDEGs, the Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.ncifcrf.gov/) version 6.8 (18, 19) was used to perform Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis. A $p < 0.05$ was considered to be statistically significant.

Screening of Independent Prognostic cDEGs

First, cDEGs significantly associated with prognosis for overall survival (OS) were screened by performing univariate Cox regression analysis with the survival R3.6.1 package (version R2.41.1; http://bioconductor.org/packages/survivalr/) (20). Afterwards, the multivariate cox regression analysis was carried out to extract independent prognostic cDEGs for OS using survival package in R 3.4.1. For univariate and multivariate Cox regression analyses, the log-rank $p < 0.05$ was thought of as the cutoff for the significant correlation.

Establishing the Prognostic Model and Performance Evaluation

We then established a prognostic score model of independent prognostic cDEGs for OS according to the following formula: prognostic score (PS) = $\Sigma\beta_{cDEGs} \times Exp_{cDEGs}$. Here, the β_{cDEGs} represented the estimated contribution coefficient of independent prognostic cDEGs in the multivariate Cox regression analysis, and Exp_{cDEGs} denoted the level of independent prognostic cDEGs. According to this formula, the PS of each sample was calculated in training dataset and two validation datasets (GSE17538 and GSE39582). Next, all patients in each dataset were divided into high- and low-risk groups with the median of the PS as the cutoff criterion. The Kaplan–Meier survival analysis was used to analyze the survival difference between two risk groups. The corresponding p-value was calculated with the R3.6.1 survival package (20). The predictive performance of the gene panel for OS was estimated using a time-dependent ROC curve by the "survivalROC" package in R software (21).

Screening of Independent Prognostic Parameters

The independent clinical prognostic parameters were identified by performing univariate and multivariate Cox regression analyses using the survival package in R 3.4.1 with log-rank $p < 0.05$ as the threshold for significance (20). Following this, Kaplan–Meier curves were used to further explore the relationships between independent prognostic factors and survival prognosis.

Building and Validating a Predictive Nomogram

Based on all independent prognostic parameters, we built a composite nomogram in the R package "rms" (https://cran.r-project.org/web/packages/rms/index.html) (22, 23) to predict the probability of 3- and 5-year OS for COAD. The calibration curve of the nomogram was plotted by calibrating function of R software to compare predicted OS against observed OS. The predicted and observed outcomes of the nomogram was assessed by using the concordance index (C-index) with a bootstrap method. The area under the ROC curve (AUC) was calculated to make a comparison for discriminatory ability of above prognostic parameters. Subsequently, we compared the nomogram including all with those including only one independent prognostic factor using the area under the ROC

curve (AUROC), C-index, and log-rank p-value. The same methods were used in the validation dataset 1 to validate the results.

Tissue Sample Preparation

Fresh tumor tissues and adjacent non-tumor colon tissues (at least 5 cm away from tumor edge) were collected from 60 cases of COAD patients who underwent surgery at the Third Xiangya Hospital of Central South University for mRNA detection. The patients were enrolled with signed written informed consent according to the following inclusion criteria (1): no neoadjuvant chemotherapy or radiotherapy and (2) no other history of surgery. The study was performed in accordance with the ethical guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the Third Xiangya Hospital of Central South University.

Quantitative Real-Time PCR Analysis

Total RNA was extracted from tissue samples using TRIzol reagent (Invitrogen, Carlsbad, CA, USA), and reverse transcription was performed using PrimeScript™ RT reagent kit (TaKaRa Bio, Inc., Otsu, Japan). Next, quantitative real-time PCR was carried out using Premix Ex Taq™ II (TaKaRa Bio, Inc.) with the following primer sequences, namely, ADH4 (forward, 5′-GGCTGCAATCAACAATGCCA-3′; reverse, 5′-CCAGGGCTTTAGCCTTCACA-3′), AHNAK2 (forward, 5′-CGCGATGTGCGACTGC-3′; reverse, 5′-TCCGTGAGTCCCCTGAATCT-3′), MMAA (forward, 5′-CTTCCGTGGCTTCGGGC-3′; reverse, 5′-AAGCCATCTGACAGCAGCAT-3′), DMBT1 (forward, 5′-TGTCAGCACAGTGAAGACGC-3′; reverse, 5′-TACTGTCGATGCAGGCAAGG-3′), and GAPDH (forward, 5′-GGTGAAGGTCGGAGTCAACG-3′; reverse, 5′-GCATCGCCCCACTTGATTTT-3′) according to the reaction conditions: 95°C for 1 min, then 40 cycles of 95°C for 10 s, 60°C for 30 s, and 75°C for 30 s. Relative mRNA expression levels were calculated with the $2^{-\Delta\Delta Cq}$ method.

Western Blot Analysis

Total protein sample was extracted from tissue specimens using radioimmunoprecipitation assay (RIPA) buffer in the presence of protease and phosphatase inhibitors (Thermo Fisher Scientific, Waltham, MA, USA Inc.). After quantified with a bicinchoninic acid (BCA) protein assay, 30 µm of total protein were separated on 12% sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) and then transferred onto polyvinylidene fluoride (PVDF) membranes (Bio-Rad, Hercules, CA, USA). The membranes were blocked with 5% skim milk dissolved in phosphate-buffered saline (PBS) and Tween 20 (PBST) for 2 h at room temperature, followed by incubated with primary antibodies against CA12 (15180-1-AP, Proteintech Group, Inc., Chicago, Il, USA), AHNAK2 (ab164994, Abcam), MMAA (ab264418, Abcam), DMBT1 (ab276422), and GAPDH (10491-1-AP, Proteintech Group, Inc., Chicago, Il, USA) at 4°C overnight. After incubating with horseradish peroxidase-conjugated secondary antibodies at room temperature for 2 h, protein signals were detected by

SuperSignal West Pico Chemiluminescent Substrate (Thermo Fisher Scientific, Waltham, MA, USA Inc.) with GAPDH as the internal control.

Statistical Analysis

Statistical analyses and graphical presentations were conducted by GraphPad Prism Software 6.0. Quantitative variables were analyzed using a t-test for paired samples or a non-parametric Wilcoxon rank-sum test for unpaired samples as appropriate. Data were expressed as mean ± SD of three independent repeats. The statistical significance was established at $p < 0.05$.
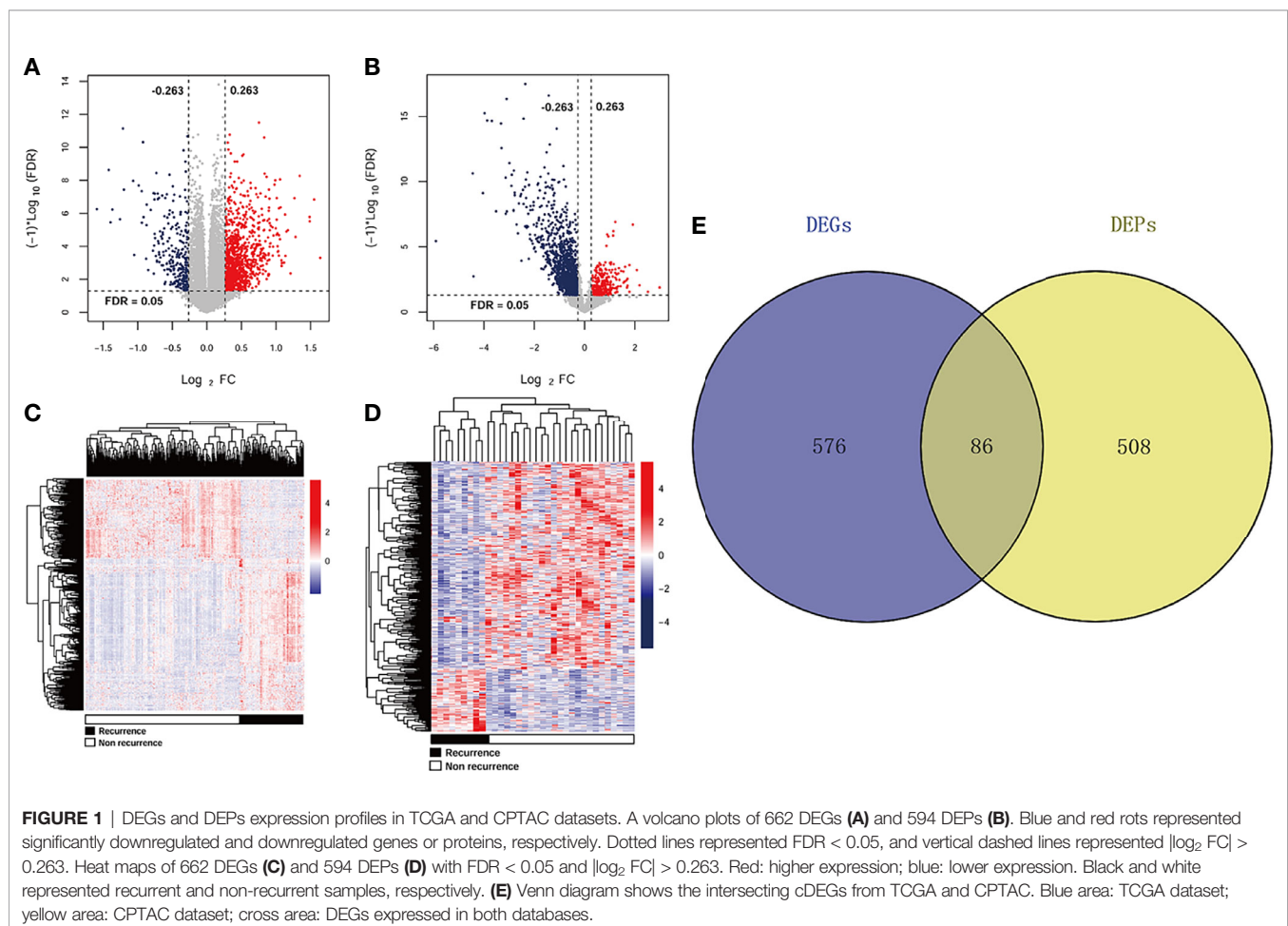
RESULTS

Differential Expression Analysis

A total of 662 DEGs (163 downregulated and 499 upregulated) and 594 DEPs (509 downregulated and 85 upregulated) were identified between recurrent and non-recurrent samples based on the threshold of FDR < 0.05 and $|\log_2 FC| > 0.263$. The expression changes of 662 DEGs (**Figure 1A**) and 594 DEPs (**Figure 1B**) were displayed by volcanic maps. Subsequently, the

expression values of 662 DEGs (**Figure 1C**) and 594 DEPs (**Figure 1D**) were hierarchically clustered, and the results were presented in the form of heat maps. These maps clearly distinguished the recurrent samples from non-recurrent samples. The Venn package was used to screen the cDEGs from both databases and generate the Venn map (**Figure 1E**). Finally, 86 recurrent-related cDEGs with high reliability were obtained (**Supplementary Table S1**).

Functional Annotation for cDEGs

To explore the potential function of cDEGs, online software DAVID was used for the GO and KEGG pathway analysis. As shown in **Figure 2A**, GO annotations of cDEGs were divided into biological process (BP), cell composition (CC), and molecular function (MF). A total of 35 terms (10 gene counts per term), including 9 BP, 17 CC, and 9 MF were arranged in ascending order according to p-value. After screening, it was found that these cDEGs were mainly enriched in oxidation–reduction process and carbon metabolism. In addition, these cDEGs were enriched in 11 KEGG pathways, including metabolic pathway, biosynthesis of antibiotics, and PI3K–Akt signaling pathway (**Figure 2B**).
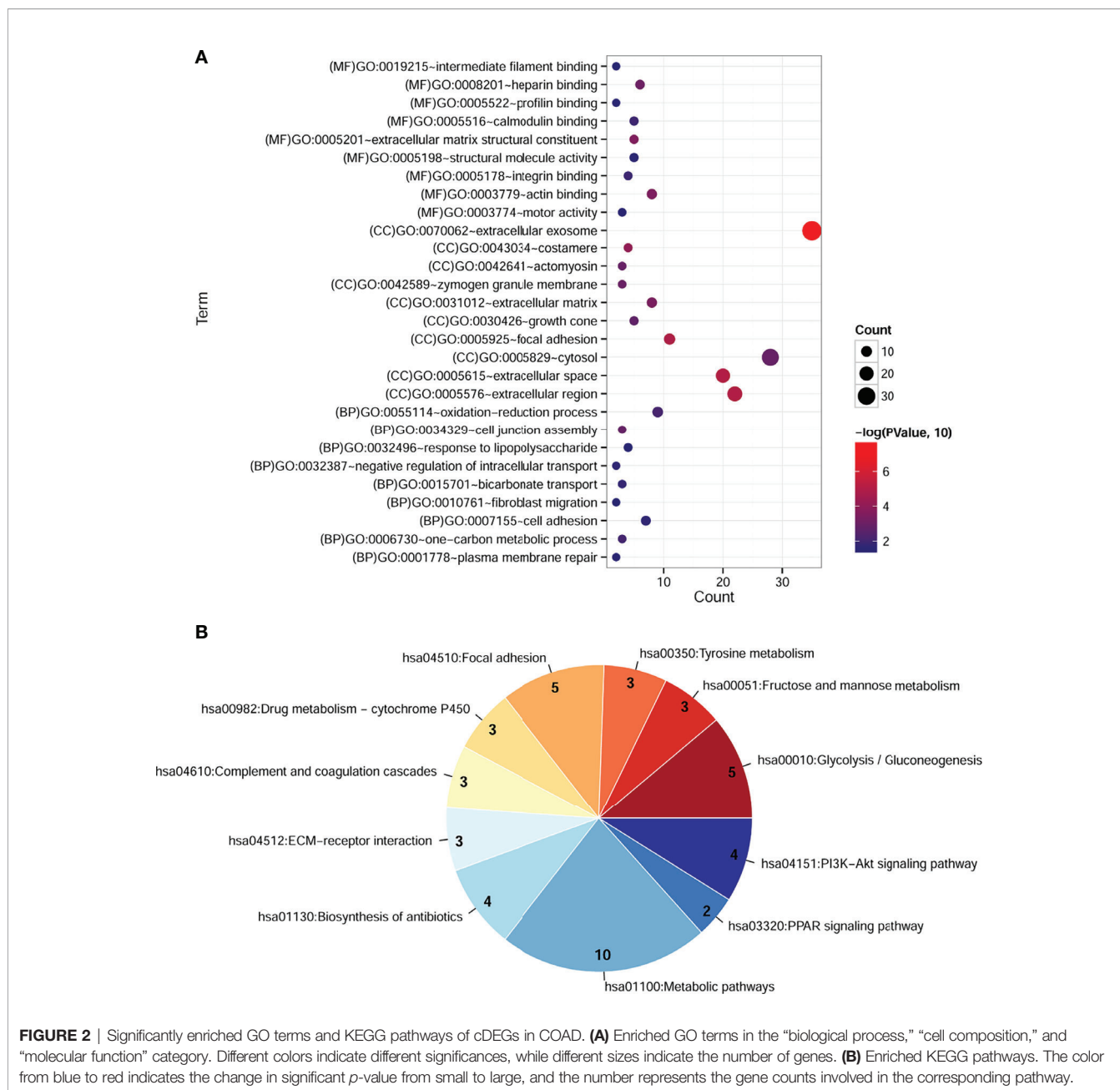

FIGURE 1 | DEGs and DEPs expression profiles in TCGA and CPTAC datasets. A volcano plots of 662 DEGs **(A)** and 594 DEPs **(B)**. Blue and red rots represented significantly downregulated and downregulated genes or proteins, respectively. Dotted lines represented FDR < 0.05, and vertical dashed lines represented $|\log_2 FC| > 0.263$. Heat maps of 662 DEGs **(C)** and 594 DEPs **(D)** with FDR < 0.05 and $|\log_2 FC| > 0.263$. Red: higher expression; blue: lower expression. Black and white represented recurrent and non-recurrent samples, respectively. **(E)** Venn diagram shows the intersecting cDEGs from TCGA and CPTAC. Blue area: TCGA dataset; yellow area: CPTAC dataset; cross area: DEGs expressed in both databases.

FIGURE 2 | Significantly enriched GO terms and KEGG pathways of cDEGs in COAD. **(A)** Enriched GO terms in the "biological process," "cell composition," and "molecular function" category. Different colors indicate different significances, while different sizes indicate the number of genes. **(B)** Enriched KEGG pathways. The color from blue to red indicates the change in significant *p*-value from small to large, and the number represents the gene counts involved in the corresponding pathway.

Screening of Independent Prognostic Feature cDEGs

The univariate Cox regression analysis was performed to identify prognostic cDEGs, and the results revealed that 42 cDEGs were significantly correlated with recurrent survival prognosis. After further screening, the results from multivariate cox regression analysis (**Table 1**) showed that a total of eight cDEGs were found to be independently related to recurrent survival prognosis, including carbonic anhydrase 12 (CA12), hemoglobin subunit beta (HBB), neutrophil cytosolic factor 1 (NCF1), kelch repeat and BTB domain containing 11 (KBTBD11), metabolism of cobalamin associated A (MMAA), deleted in malignant brain

TABLE 1 | The list of independent prognostic feature cDEGs.

Symbol	Coef	P value	Hazard ratio	95%CI
CA12	0.44816	4.360E-04	1.565	1.220-2.010
HBB	0.21914	1.451E-02	1.245	1.044-1.484
NCF1	1.15846	1.722E-02	3.185	1.228-8.262
KBTBD11	-0.55455	1.933E-02	0.574	0.361-0.914
MMAA	-1.04285	3.953E-02	0.352	0.131-0.951
DMBT1	-0.13052	4.151E-02	0.878	0.774-0.995
AHNAK2	0.40594	4.541E-02	1.501	1.300-2.268
FBLN2	-0.37102	4.927E-02	0.690	0.448-0.963

Coef, Coefficient derived from multivariate cox regression analysis; 95%CI, 95% confidence interval.

tumors 1 (DMBT1), AHNAK nucleoprotein 2 (AHNAK2), and fibulin 2 (FBLN2).

Establishing the Prognostic Model and Performance Evaluation

Subsequently, the expression levels of these eight independent prognostic cDEGs in training dataset and two validation datasets were calculated, and prognostic model was constructed as follows:

$$PS = (0.44816) \times Exp_{CA12} + (0.21914) \times Exp_{HBB} + (1.15846) \times Exp_{NCF1} +$$
$$(-0.55455) \times Exp_{KBTBD11} + (-1.04285) \times Exp_{MMAA} + (-0.13052) \times Exp_{DMBT1} +$$
$$(0.40594) \times Exp_{AHNAK2} + (-0.37102) \times ExpFBLN2$$

The PS for each sample in each dataset was then calculated accordingly. Based on the median value of PS, all samples in each dataset were classified into high-risk group and low-risk group.

The survival analysis revealed that there was a significant correlation between two different risk groups and survival outcomes in training dataset (**Figure 3A**; HR = 2.830; 95% CI, 1.746–4.589; log-rank $p = 9.862e{-}06$), validation dataset 1 (**Figure 3B**; HR = 1.460; 95% CI, 1.053–2.024; log-rank $p = 2.245e{-}02$), and validation dataset 2 (**Figure 3C**; HR = 2.306; 95% CI, 1.144–4.648; log-rank $p = 1.562e{-}02$). ROC curve showed that the AUC for OS was 0.870 in the training dataset, 0.771 in the validation dataset 1, and 0.799 in the validation dataset 2, indicating a better predictive performance in established prognostic model (**Figure 3D**).

Screening of Independent Prognostic Parameters

A total of 372 patients from TCGA-COAD database with complete information including gender and clinical stage were
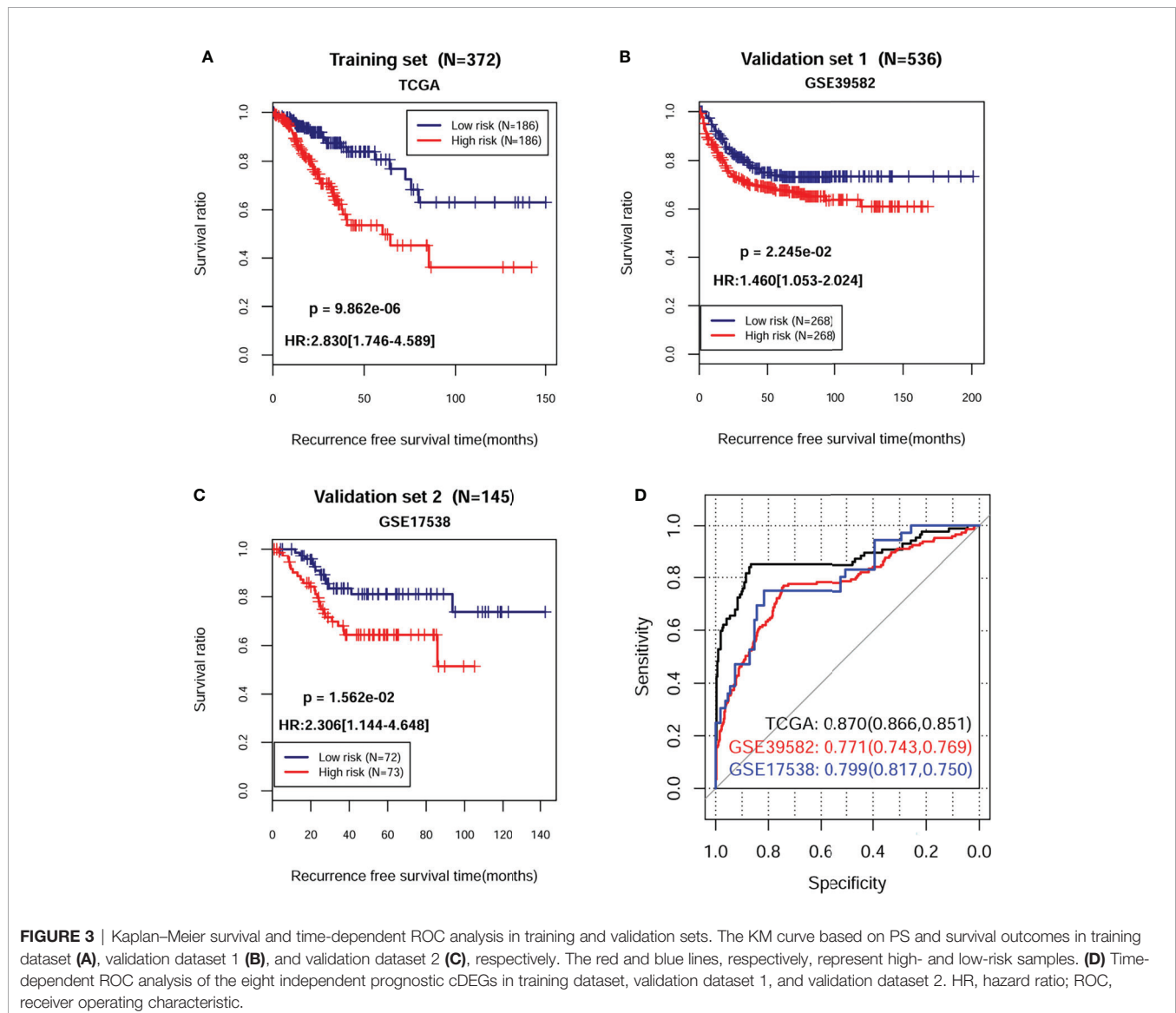

FIGURE 3 | Kaplan–Meier survival and time-dependent ROC analysis in training and validation sets. The KM curve based on PS and survival outcomes in training dataset **(A)**, validation dataset 1 **(B)**, and validation dataset 2 **(C)**, respectively. The red and blue lines, respectively, represent high- and low-risk samples. **(D)** Time-dependent ROC analysis of the eight independent prognostic cDEGs in training dataset, validation dataset 1, and validation dataset 2. HR, hazard ratio; ROC, receiver operating characteristic.

TABLE 2 | The univariable and multi-variable cox regression of clinical parameters and survival outcomes of patients with COAD.

Clinical characteristics	TCGA (N=372)	Uni-variable cox		Multi-variable cox	
		HR (95% CI)	P value	HR (95% CI)	P value
Age (years, mean ± SD)	65.95±12.70	0.989 [0.972-1.007]	2.30E-01	–	–
Gender (Male/Female)	201/171	1.594 [1.007-2.523]	4.37E-02	1.225 [0.744-2.019]	4.25E-01
Pathologic M (M0/M1/-)	273/47/52	3.562 [2.056-6.172]	4.57E-05	3.313 [1.155-9.503]	2.59E-02
Pathologic N (N0/N1/N2)	217/91/64	1.992 [1.521-2.609]	5.44E-07	1.634 [1.160-2.781]	7.03E-03
Pathologic T (T1/T2/T3/T4)	10/64/255/43	2.227 [1.431-3.467]	2.62E-04	1.730 [1.524-3.141]	7.19E-03
Pathologic stage (I/II/III/IV/-)	61/145/109/47/10	1.811 [1.394-2.353]	6.23E-06	1.052 [0.309-1.371]	2.59E-01
KRAS mutation (Yes/No/-)	18/25/329	0.657 [0.238-1.812]	4.11E-01	–	–
Microsatellite instability (Yes/No/-)	10/73/289	0.586 [0.135-2.543]	4.44E-01	–	–
Colon polyps present (Yes/No/-)	75/133/164	0.939 [0.528-1.672]	8.31E-01	–	–
Colon polyps history (Yes/No/-)	99/209/64	0.885 [0.503-1.559]	6.71E-01	–	–
Radiation therapy (Yes/No/-)	10/321/41	0.741 [0.669-6.888]	1.88E-01	–	–
PS model status (High/Low)	186/186	2.830 [1.746-4.589]	9.86E-06	2.508 [1.472-4.276]	7.27E-04
Recurrence (Yes/No)	78/294	–	–	–	–
Recurrence free survival time (months, mean ± SD)	29.61±25.63	–	–	–	–

SD, Standard deviation; HR, Hazard ratio; 95%CI, 95% confidence interval; TCGA, The Cancer Genome Atlas.

included for further analysis. Univariate and multivariate Cox regression analyses indicated that pathological M, N, T, and PS model status calculated from the eight-gene signature were independent prognostic parameters for OS (**Table 2**). Moreover, survival analysis showed that pathological M (**Figure 4A**; HR = 3.562; 95% CI, 2.056–6.172; log-rank p = 4.565e−05), N (**Figure 4B**; HR = 1.992; 95% CI, 1.521–2.609; log-rank p = 5.442e−07), and T (**Figure 4C**; HR = 1.742; 95% CI, 1.228–2.470; log-rank p = 5.442e−04) were markedly related to overall survival in COAD patients from TCGA database. Furthermore, we divided the samples according to different pathological M, N, and T and analyzed the association between high- and low-risk group in the samples of each stage. The results indicated that the recurrent-free survival prognosis of the low-risk group was significantly better than that of the high-risk group in patients from pathological M (**Figure 5A**), N (**Figure 5B**), and T (**Figure 5C**) regardless of their stage.

Building and Validating a Predictive Nomogram

Next, we built a nomogram to predict 3- and 5-year OS in the 372 COAD patients using four independent prognostic parameters, including pathological M, N, T, and PS. As shown in **Figure 6A**, the score of every indicator was presented by points scale located at the top of nomogram. After summing the points of each indicator, we could estimate the survival probability at 3 and 5 years. Moreover, we plotted a calibration curve to evaluate the performance of the nomogram model. As described in **Figure 6B**, the C-index was 0.714 and 0.688 for predicted probability of OS at the 3 and 5 years, respectively, suggesting that the constructed nomogram prediction for survival rates showed good agreement with the actual observation for OS patients. Furthermore, compared with nomogram including only the pathological M, N, T, clinical, and PS model, the combined model (**Table 3**) showed the largest
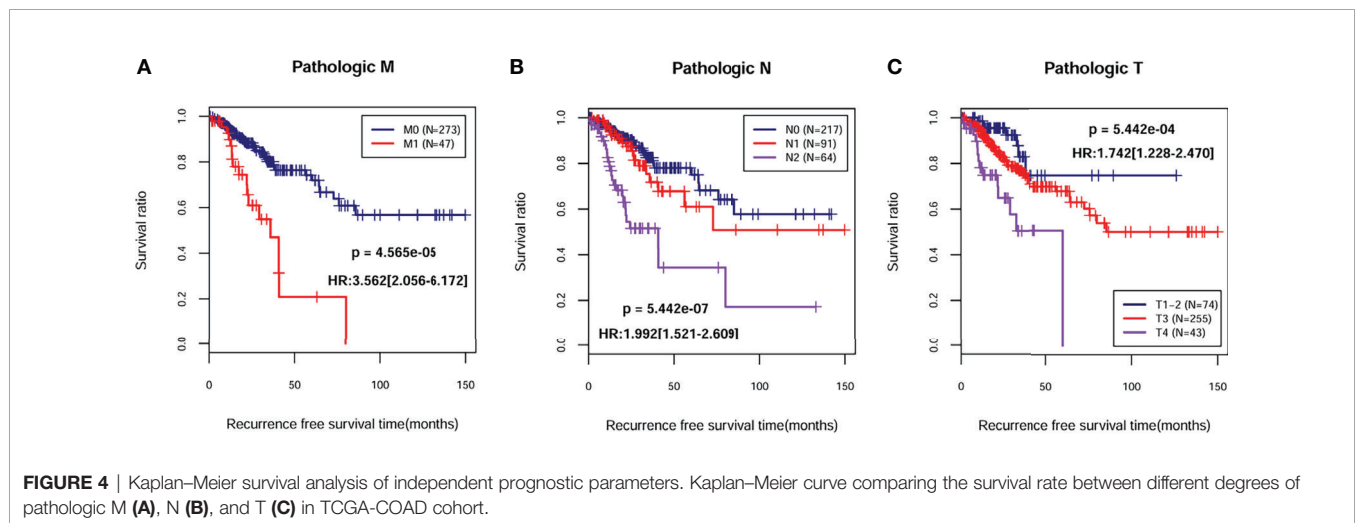

FIGURE 4 | Kaplan–Meier survival analysis of independent prognostic parameters. Kaplan–Meier curve comparing the survival rate between different degrees of pathologic M **(A)**, N **(B)**, and T **(C)** in TCGA-COAD cohort.
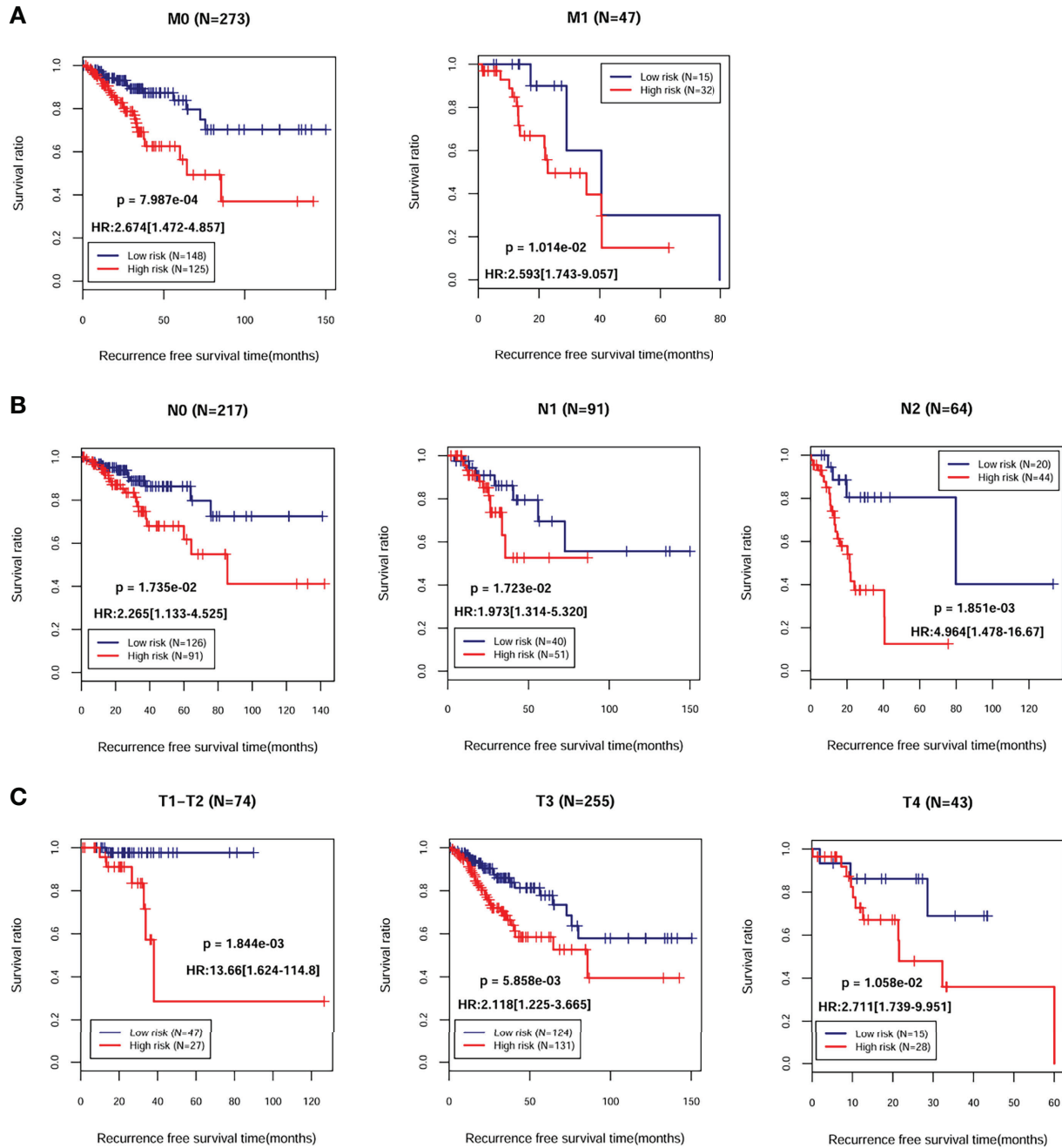
FIGURE 5 | Kaplan–Meier survival analysis of the eight-gene signature in independent prognostic parameters. Kaplan–Meier curve comparing the survival rate between high- and low-risk groups in patients in different pathologic M **(A)**, N **(B)**, and T **(C)**.

AUROC, C-index, and smallest *p*-value for 3-year survival in both training dataset (**Supplementary Figure 1SA**) and validation dataset 1 (**Supplementary Figure 1SB**). Taking together, these results indicated that the nomogram built with the combined model might be the best nomogram for predicting short-term survival, in comparison with nomograms built with a single prognostic factor.

Validation of mRNA by Quantitative Real-Time PCR

According to the results analyzed by bioinformatics, the mRNA expression levels of eight independent prognostic feature cDEGs selected was then validated in clinical tissues originating from COAD patients. Consistent with the results obtained with bioinformatics, it was revealed that CA12 and AHNAK2 mRNA
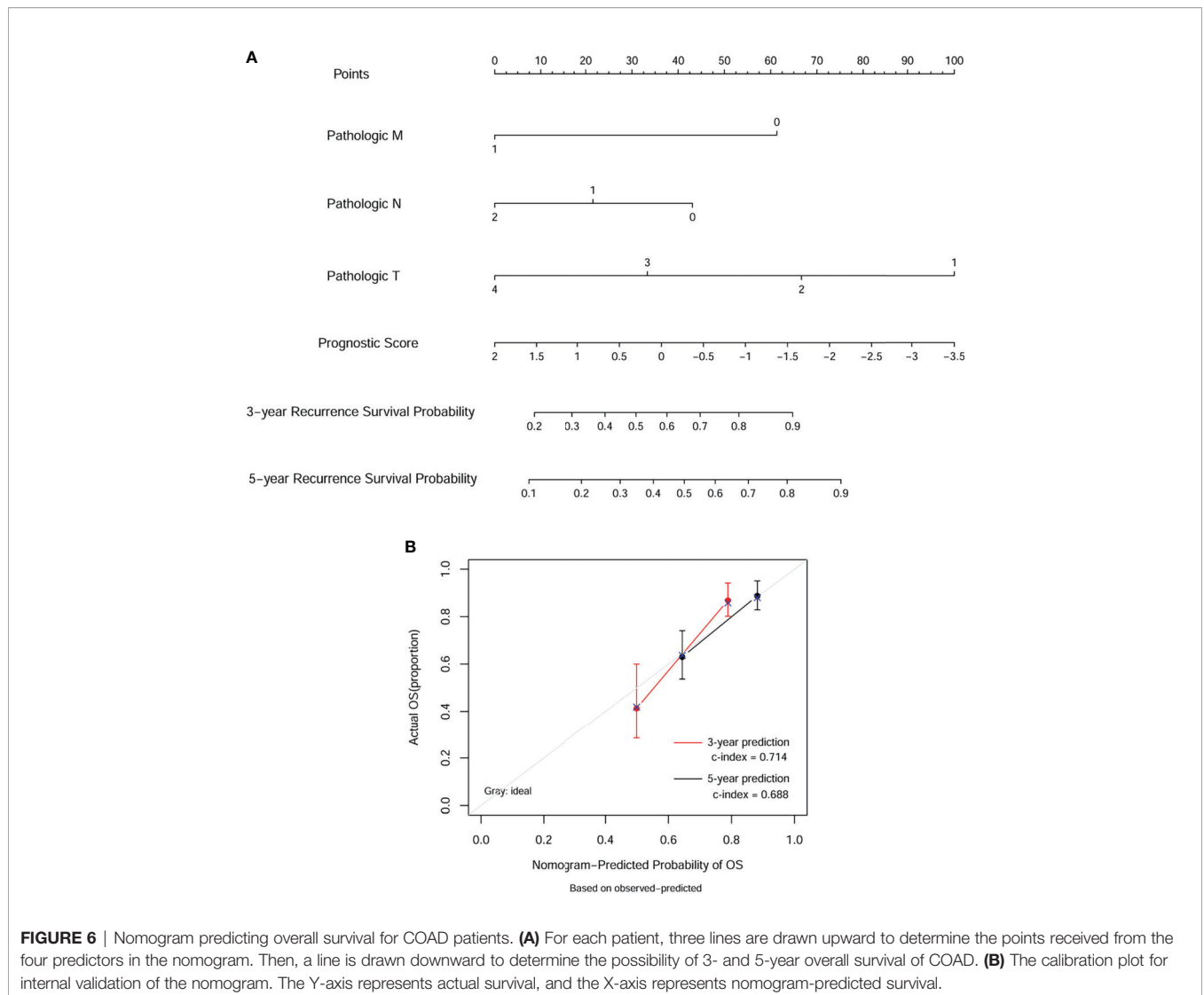
FIGURE 6 | Nomogram predicting overall survival for COAD patients. **(A)** For each patient, three lines are drawn upward to determine the points received from the four predictors in the nomogram. Then, a line is drawn downward to determine the possibility of 3- and 5-year overall survival of COAD. **(B)** The calibration plot for internal validation of the nomogram. The Y-axis represents actual survival, and the X-axis represents nomogram-predicted survival.

TABLE 3 | Comparison of the nomogram with different models.

Type	AUROC (95% CI)		C-index		P Value	
	TCGA	**GSE39582**	**TCGA**	**GSE39582**	**TCGA**	**GSE39582**
Pathologic M	0.556 (0.908,0.402)	0.505 (0.847,0.463)	0.580	0.516	5.592E-03	1.770E-01
Pathologic N	0.593 (0.622,0.564)	0.584 (0.861,0.506)	0.617	0.604	2.937E-03	3.336E-02
Pathologic T	0.558 (0.425,0.869)	0.515 (0.825,0.523)	0.587	0.560	5.357E-03	1.439E-02
Clinical model	0.744 (0.730,0.585)	0.579 (0.769,0.583)	0.645	0.610	1.762E-04	1.388E-02
Prognostic score model	0.870 (0.878,0.833)	0.771 (0.746,0.769)	0.683	0.611	2.85E-06	3.14E-02
Combine model	0.951 (0.929,0.939)	0.777 (0.797,0.715)	0.714	0.666	3.71E-08	3.45E-04

AUROC, area under the ROC curve; CI, confidence interval; C-index, concordance index.

levels were significantly upregulated in tumor tissues compared to adjacent tissues, whereas MMAA and DMBT1 mRNA levels were downregulated analogously (**Figure 7A**). Moreover, we divided tissue samples into recurrent and non-recurrent samples and further analyzed the mRNA expression levels of the above four genes. As expected, CA12 and AHNAK2 mRNA levels were significantly upregulated in recurrent tissues compared to non-recurrent tissues, whereas MMAA and DMBT1 mRNA levels were downregulated analogously (**Figure 7B**). Consistently, Western blot analysis obtained the same protein trends between tumor tissues and adjacent tissues (**Figure 7C**) and between recurrent and non-recurrent samples (**Figure 7D**).
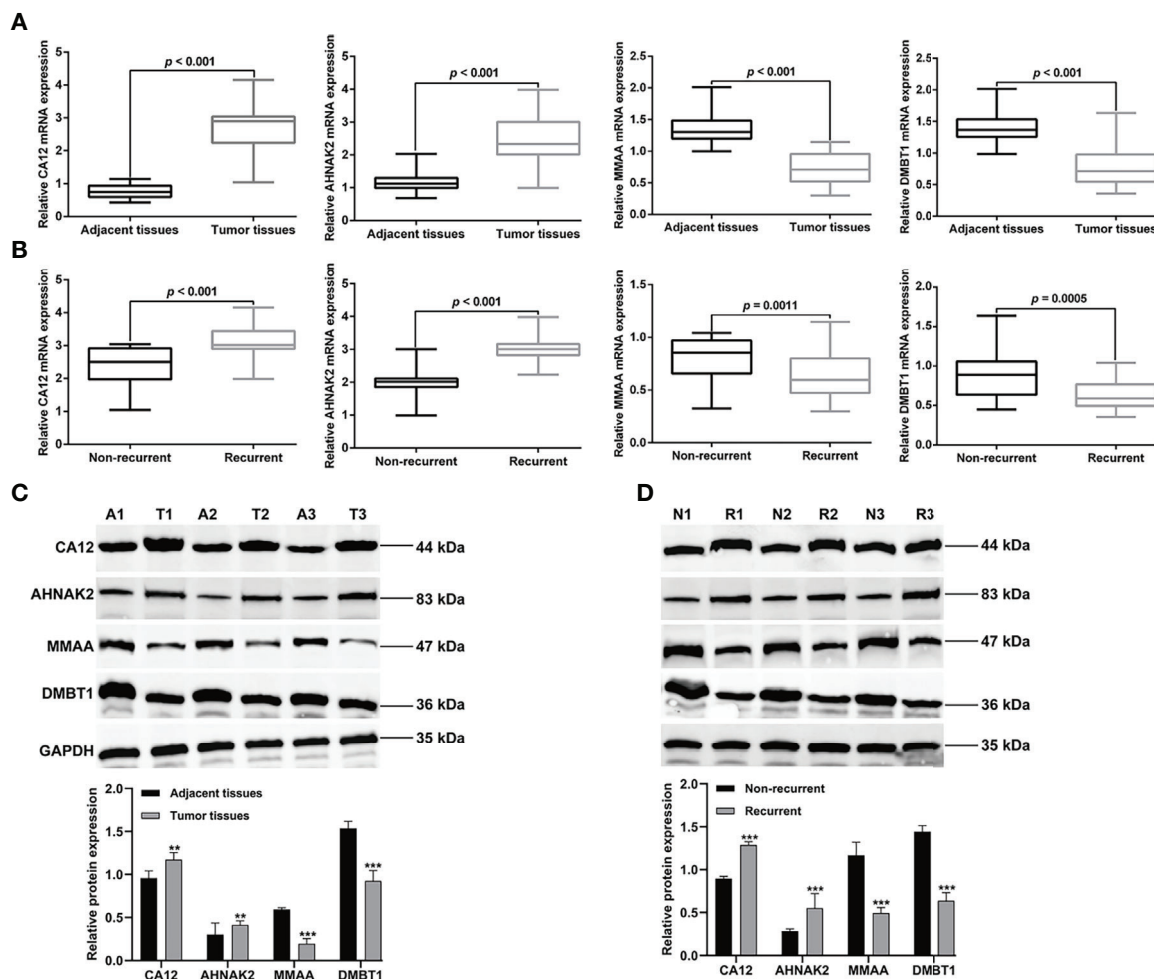
FIGURE 7 | Validation of eight independent prognostic cDEGs. Quantitative real-time PCR analysis was performed to determine the mRNA expression levels of ADH4, AHNAK2, MMAA, and DMBT1 in tumor tissue and adjacent tissues **(A)**, and in recurrent tissues and non-recurrent tissues **(B)**. Western blot analysis was utilized to measure the protein levels of ADH4, AHNAK2, MMAA, and DMBT1 in tumor tissue and adjacent tissues **(C)**, and in recurrent tissues and non-recurrent tissues **(D)**. Data were expressed as mean ± SD of three independent repeats. **$p < 0.01$, ***$p < 0.001$, compared with adjacent or non-recurrent tissues. A, adjacent tissues; T, tumor tissues; N, non-recurrent; R, recurrent.

DISCUSSION

COAD remains one of the most life-threatening malignancies in the world due to the complex molecular mechanisms. With the development of gene sequencing technology, some potential gene markers, including INHBA (24), COL8A1 (25), and GABRD (26), correlated with poor prognosis for COAD patients have been identified. Nevertheless, the number of such biomarkers in COAD tumorigenesis is still limited. Therefore, it is urgently needed to screen out more biomarkers with higher prediction accuracy for improving the prognosis and outlining an individualized treatment plan for COAD patients.

In this study, by combined TCGA and CPTAC database, a total of 86 cDEGs between recurrent and non-recurrent samples were identified. After univariate and multivariate Cox regression analysis, we finally obtained eight-related risk signatures (CA12,

HBB, NCF1, KBTBD11, MMAA, DMBT1, AHNAK2, and FBLN2) that were related to recurrent OS. KM survival analysis and AUC further illustrated that our model had a great predictive performance. Additionally, the prognostic scores of the recurrence-free survival model could be considered as independent predictive indicators. Finally, the nomogram and calibration chart showed that the risk signatures could accurately assess the survival of COAD patients. Carbonic anhydrase 12 (CA12) is a transmembrane protein, which was found to be differentially expressed in colon cancer (27) and a useful clinical tool in discriminating the prognosis of cervical cancer (28). Hemoglobin subunit beta (HBB) is a member of the globin family and a structurally conserved group of proteins often containing the heme group, which has been identified as a potential serum biomarker for the diagnosis and prognosis of ovarian cancer (29, 30). Neutrophil cytoplasmic factor 1 (NCF1)

belongs to the NADPH oxidase complex and is a cytoplasmic component, which plays a role in tumor growth and metastasis (31). The study by Kelkka et al. (32) demonstrated that mice lacking NCF1 exhibited reduced growth of implanted melanoma and carcinoma tumors. Kelch repeat and BTB domain containing 11 (KBTBD11) is a member of the KBTBD subfamily of proteins that possess a BTB domain and Kelch repeats, which has been identified as a negative regulator of osteoclast differentiation (33) and reported to promote adipocyte homeostasis (34). In addition, Li et al. (35) manifested the association of high fibulin 2 gene (FBLN2) expression with worse disease-specific survival and metastasis-free survival in urothelial carcinoma.

Next, we selected four genes (CA12, MMAA, DMBT1, and AHNAK2) from the above eight risk biomarkers for quantitative PCR (qPCR) and Western blot analysis verification according to literature retrieval. The experimental results confirmed that ADH4 and AHNAK2 were significantly upregulated, while MMAA and DMBT1 were downregulated analogously in tumor tissues or recurrent tissues compared with controls. Consistent with our data, CA12 is highly expressed in glial tumors compared with normal tissue and predicts for poor clinical course of tumor patients (36). However, Watson et al. (37) put forward an opposite conclusion; they believed that CA12 expression was associated with a better prognosis in an unselected series of invasive breast carcinoma patients. Metabolism of cobalamin-associated A (MMAA) gene is essential for the proper functioning of a cofactor of the methylmalonyl-CoA mutase, whose mutation was associated with the clinical treatment effects of methylmalonic aciduria (38). Malignant brain tumor 1 (DMBT1) was decreased in cervical squamous cell carcinoma (CSCC), whereas its overexpression cannot only inhibit the proliferation, migration, and invasion but also induce the apoptosis of human CSCC cells (39), which strongly supported our results. AHNAK nucleoprotein 2 (AHNAK2), also known as C14orf78, is a member of the AHNAK family, which also includes its homologous gene AHNAK (40). Upregulated ANHAK2 activates the PI3K/AKT signaling pathway and promotes melanoma cell metastasis (41). In agreement with our data, AHNAK2 was upregulated in tumor samples and correlated with poor prognosis in lung adenocarcinoma patients (42). Recent studies also revealed that ANHAK2 could act as a biomarker for various tumors, including thyroid cancer (43), pancreatic ductal cancer (44), bladder cancer (45), and gastric cancer (46). However, the unrecognized roles of CA12, MMAA, DMBT1, and AHNAK2 in COAD are worth further investigating to identify their biological functions and underlying mechanisms in the development and progression of the disease.

Subsequently, KM curve, ROC curve, and risk plot analyses verified that the eight-based risk signature performs well in stratifying the risk groups of recurrent COAD in TCGA. Our data also indicated that the nomogram built with the combined model might be the best nomogram for predicting short-term survival, in comparison with nomograms built with a single prognostic factor.

CONCLUSIONS

In summary, our results suggest that the eight-gene-based recurrent prognosis model is a reliable tool in risk stratification and serves as an independent prognostic factor of the OS in recurrent COAD patients. These model genes are expected to be diagnostic and treatment indicators, which will face the challenge on how to apply various genes signature reasonably in a particular stage of COAD.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The study was performed in accordance with the ethical guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the Third Xiangya Hospital of Central South University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FL made substantial contributions to conception and design of the research. FA and WW carried out data collection and analysis. SL and DZ were involved in drawing figures and tables. DZ and ZY wrote the paper. FA, SL, and ZY edited the manuscript and provided critical comments. All authors contributed to the article and approved the submitted version.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.871568/full#supplementary-material

Supplementary Figure 1 | Performance of gene-based nomogram in predicting survival probability. Time-dependent ROC curves of the pathologic M, N, T, clinical model and combined model in OS prediction in the training dataset **(A)** and validation dataset 1 **(B)**. OS, overall survival; ROC, receiver operating characteristic.

REFERENCES

1. Fleming M, Ravula S, Tatishchev SF, Wang HL. Colorectal Carcinoma: Pathologic Aspects. *J gastrointest Oncol* (2012) 3(3):153–73. doi: 10.3978/j.issn.2078-6891.2012.030

2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: Cancer J Clin* (2018) 68 (6):394–424. doi: 10.3322/caac.21492

3. Obaro AE, Burling DN, Plumb AA. Colon Cancer Screening With CT Colonography: Logistics, Cost-Effectiveness, Efficiency and Progress. *Br J radiol* (2018) 91(1090):20180307. doi: 10.1259/bjr.20180307

4. Zeng JH, Liang L, He RQ, Tang RX, Cai XY, Chen JQ, et al. Comprehensive Investigation of a Novel Differentially Expressed lncRNA Expression Profile Signature to Assess the Survival of Patients With Colorectal Adenocarcinoma. *Oncotarget.* (2017) 8(10):16811–28. doi: 10.18632/oncotarget.15161

5. Sugarbaker PH. Improving Oncologic Outcomes for Colorectal Cancer at High Risk for Local-Regional Recurrence With Novel Surgical Techniques. *Expert Rev Gastroenterol hepatol* (2016) 10(2):205–13. doi: 10.1586/17474124.2016.1110019

6. Bertelsen CA, Larsen HM, Neuenschwander AU, Laurberg S, Kristensen B, Emmertsen KJ. Long-Term Functional Outcome After Right-Sided Complete Mesocolic Excision Compared With Conventional Colon Cancer Surgery: A Population-Based Questionnaire Study. *Dis colon rectum.* (2018) 61(9):1063–72. doi: 10.1097/DCR.0000000000001154

7. Tarazona N, Gimeno-Valiente F, Gambardella V, Huerta M, Rosello S, Zuniga S, et al. Detection of Postoperative Plasma Circulating Tumour DNA and Lack of CDX2 Expression as Markers of Recurrence in Patients With Localised Colon Cancer. *ESMO Open* (2020) 5(5):e000847. doi: 10.1136/esmoopen-2020-000847

8. Jung EE, Heinemann FS, Egelston CA, Wang J, Pollock RE, Lee PP, et al. Synchronous Recurrence of Concurrent Colon Adenocarcinoma and Dedifferentiated Liposarcoma. *BMJ Case Rep* (2019) 12(5):e228868. doi: 10.1136/bcr-2018-228868

9. Pan J, Xu Z, Xu M, Lin X, Lin B, Lin M. Knockdown of Forkhead Box A1 Suppresses the Tumorigenesis and Progression of Human Colon Cancer Cells Through Regulating the Phosphatase and Tensin Homolog/Akt Pathway. *J Int Med Res* (2020) 48(12):300060520971453. doi: 10.1177/0300060520971453

10. Ando J, Saito M, Imai JI, Ito E, Yanagisawa Y, Honma R, et al. TBX19 is Overexpressed in Colorectal Cancer and Associated With Lymph Node Metastasis. *Fukushima J Med* (2017) 63(3):141–51. doi: 10.5387/fms.2017-08

11. Varela FA, Foust VL, Hyland TE, Sala-Hamrick KE, Mackinder JR, Martin CE, et al. TMPRSS13 Promotes Cell Survival, Invasion, and Resistance to Drug-Induced Apoptosis in Colorectal Cancer. *Sci Rep* (2020) 10(1):13896. doi: 10.1038/s41598-020-70636-4

12. Ji F, Sadreyev RI. RNA-Seq: Basic Bioinformatics Analysis. *Curr Protoc Mol Biol* (2018) 124(1):e68. doi: 10.1002/cpmb.68

13. Tao Z, Shi A, Li R, Wang Y, Wang X, Zhao J. Microarray Bioinformatics in Cancer- a Review. *J BUON.* (2017) 22(4):838–43.

14. Rau A, Flister M, Rui H, Auer PL. Exploring Drivers of Gene Expression in the Cancer Genome Atlas. *Bioinformatics.* (2019) 35(1):62–8. doi: 10.1093/bioinformatics/bty551

15. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for Functional Genomics Data Sets–Update. *Nucleic Acids Res* (2013) 41(Database issue):D991–5. doi: 10.1093/nar/gks1193

16. Xie H, Wang W, Sun F, Deng K, Lu X, Liu H, et al. Proteomics Analysis to Reveal Biological Pathways and Predictive Proteins in the Survival of High-Grade Serous Ovarian Cancer. *Sci Rep* (2017) 7(1):9896. doi: 10.1038/s41598-017-10559-9

17. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res* (2015) 43(7):e47. doi: 10.1093/nar/gkv007

18. Huang da W, Sherman BT, Lempicki RA. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat Protoc* (2009) 4(1):44–57. doi: 10.1038/nprot.2008.211

19. Huang da W, Sherman BT, Lempicki RA. Bioinformatics Enrichment Tools: Paths Toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Res* (2009) 37(1):1–13. doi: 10.1093/nar/gkn923

20. Wang P, Wang Y, Hang B, Zou X, Mao JH. A Novel Gene Expression-Based Prognostic Scoring System to Predict Survival in Gastric Cancer. *Oncotarget.* (2016) 7(34):55343–51. doi: 10.18632/oncotarget.10533

21. Heagerty PJ, Lumley T, Pepe MS. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics.* (2000) 56(2):337–44. doi: 10.1111/j.0006-341X.2000.00337.x

22. Anderson WI, Schlafer DH, Vesely KR. Thyroid Follicular Carcinoma With Pulmonary Metastases in a Beaver (Castor Canadensis). *J wildlife diseases.* (1989) 25(4):599–600. doi: 10.7589/0090-3558-25.4.599

23. Eng KH, Schiller E, Morrell K. On Representing the Prognostic Value of Continuous Gene Expression Biomarkers With the Restricted Mean Survival Curve. *Oncotarget.* (2015) 6(34):36308–18. doi: 10.18632/oncotarget.6121

24. Li X, Yu W, Liang C, Xu Y, Zhang M, Ding X, et al. INHBA is a Prognostic Predictor for Patients With Colon Adenocarcinoma. *BMC cancer.* (2020) 20 (1):305. doi: 10.1186/s12885-020-06743-2

25. Zhang L, Jiang X, Li Y, Fan Q, Li H, Jin L, et al. Clinical Correlation of Wnt2 and COL8A1 With Colon Adenocarcinoma Prognosis. *Front Oncol* (2020) 10:1504. doi: 10.3389/fonc.2020.01504

26. Wu M, Kim KY, Park WC, Ryu HS, Choi SC, Kim MS, et al. Enhanced Expression of GABRD Predicts Poor Prognosis in Patients With Colon Adenocarcinoma. *Trans Oncol* (2020) 13(12):100861. doi: 10.1016/j.tranon.2020.100861

27. Viikila P, Kivela AJ, Mustonen H, Koskensalo S, Waheed A, Sly WS, et al. Carbonic Anhydrase Enzymes II, VII, IX and XII in Colorectal Carcinomas. *World J gastroenterol* (2016) 22(36):8168–77. doi: 10.3748/wjg.v22.i36.8168

28. Yoo CW, Nam BH, Kim JY, Shin HJ, Lim H, Lee S, et al. Carbonic Anhydrase XII Expression is Associated With Histologic Grade of Cervical Cancer and Superior Radiotherapy Outcome. *Radiat Oncol* (2010) 5(1):101–. doi: 10.1186/1748-717X-5-101

29. Bonaventura C, Henkens R, Alayash AI, Banerjee S, Crumbliss AL. Molecular Controls of the Oxygenation and Redox Reactions of Hemoglobin. *Antioxid Redox Signaling* (2013) 18(17):2298–313. doi: 10.1089/ars.2012.4947

30. Woong-Shick A, Sung-Pil P, Su-Mi B, Joon-Mo L, Sung-Eun N, Gye-Hyun N, et al. Identification of Hemoglobin-Alpha and -Beta Subunits as Potential Serum Biomarkers for the Diagnosis and Prognosis of Ovarian Cancer. *Cancer sci* (2005) 96(3):197–201. doi: 10.1111/j.1349-7006.2005.00029.x

31. Chen Y, He F, Wang R, Yao M, Li Y, Guo D, et al. NCF1/2/4 Are Prognostic Biomarkers Related to the Immune Infiltration of Kidney Renal Clear Cell Carcinoma. *BioMed Res Int* (2021) 2021:5954036. doi: 10.1155/2021/5954036

32. Kelkka T, Pizzolla A, Laurila JP, Friman T, Gustafsson R, Kallberg E, et al. Mice Lacking NCF1 Exhibit Reduced Growth of Implanted Melanoma and Carcinoma Tumors. *PloS One* (2013) 8(12):e84148. doi: 10.1371/journal.pone.0084148

33. Narahara S, Sakai E, Kadowaki T, Yamaguchi Y, Narahara H, Okamoto K, et al. KBTBD11, a Novel BTB-Kelch Protein, is a Negative Regulator of Osteoclastogenesis Through Controlling Cullin3-Mediated Ubiquitination of Nfatc1. *Sci Rep* (2019) 9(1):3523. doi: 10.1038/s41598-019-40240-2

34. Watanabe K, Yokota K, Yoshida K, Matsumoto A, Iwamoto S. Kbtbd11 Contributes to Adipocyte Homeostasis Through the Activation of Upstream Stimulatory Factor 1. *Heliyon.* (2019) 5(11):e02777. doi: 10.1016/j.heliyon.2019.e02777

35. Li WM, Chan TC, Huang SK, Wu WJ, Ke HL, Liang PI, et al. Prognostic Utility of FBLN2 Expression in Patients With Urothelial Carcinoma. *Front Oncol* (2020) 10:570340. doi: 10.3389/fonc.2020.570340

36. Li G, Chen TW, Nickel AC, Muhammad S, Steiger HJ, Tzaridis T, et al. Carbonic Anhydrase XII is a Clinically Significant, Molecular Tumor-Subtype Specific Therapeutic Target in Glioma With the Potential to Combat Invasion of Brain Tumor Cells. *OncoTargets Ther* (2021) 14:1707–18. doi: 10.2147/OTT.S300623

37. Watson PH, Chia SK, Wykoff CC, Han C, Leek RD, Sly WS, et al. Carbonic Anhydrase XII is a Marker of Good Prognosis in Invasive Breast Carcinoma. *Br J cancer.* (2003) 88(7):1065–70. doi: 10.1038/sj.bjc.6600796

38. Wesol-Kucharska D, Kaczor M, Pajdowska M, Ehmke VE-SE, Bogdanska A, Kozlowski D, et al. Clinical Picture and Treatment Effects in 5 Patients With

Methylmalonic Aciduria Related to MMAA Mutations. *Mol Genet Metab Rep* (2020) 22:100559. doi: 10.1016/j.ymgmr.2019.100559

39. Zhang CX. The Protective Role of DMBT1 in Cervical Squamous Cell Carcinoma. *Kaohsiung J Med Sci* (2019) 35(12):739–49. doi: 10.1002/kjm2.12117

40. Marg A, Haase H, Neumann T, Kouno M, Morano I. AHNAK1 and AHNAK2 are Costameric Proteins: AHNAK1 Affects Transverse Skeletal Muscle Fiber Stiffness. *Biochem Biophys Res Commun* (2010) 401(1):143–8. doi: 10.1016/j.bbrc.2010.09.030

41. Li M, Liu Y, Meng Y, Zhu Y. AHNAK Nucleoprotein 2 Performs a Promoting Role in the Proliferation and Migration of Uveal Melanoma Cells. *Cancer biotherapy radiopharm.* (2019) 34(10):626–33. doi: 10.1089/cbr.2019.2778

42. Zhang S, Lu Y, Qi L, Wang H, Wang Z, Cai Z. AHNAK2 Is Associated With Poor Prognosis and Cell Migration in Lung Adenocarcinoma. *BioMed Res Int* (2020) 2020:8571932. doi: 10.1155/2020/8571932

43. Kim MJ, Sun HJ, Song YS, Yoo SK, Kim YA, Seo JS, et al. CXCL16 Positively Correlated With M2-Macrophage Infiltration, Enhanced Angiogenesis, and Poor Prognosis in Thyroid Cancer. *Sci Rep* (2019) 9(1):13288. doi: 10.1038/s41598-019-49613-z

44. Klett H, Fuellgraf H, Levit-Zerdoun E, Hussung S, Kowar S, Kusters S, et al. Identification and Validation of a Diagnostic and Prognostic Multi-Gene Biomarker Panel for Pancreatic Ductal Adenocarcinoma. *Front Genet* (2018) 9:108. doi: 10.3389/fgene.2018.00108

45. Witzke KE, Grosserueschkamp F, Jutte H, Horn M, Roghmann F, von Landenberg N, et al. Integrated Fourier Transform Infrared Imaging and Proteomics for Identification of a Candidate Histochemical Biomarker in Bladder Cancer. *Am J pathol* (2019) 189(3):619–31. doi: 10.1016/j.ajpath.2018.11.018

46. Zhou YY, Kang YT, Chen C, Xu FF, Wang HN, Jin R. Combination of TNM Staging and Pathway Based Risk Score Models in Patients With Gastric Cancer. *J Cell Biochem* (2018) 119(4):3608–17. doi: 10.1002/jcb.26563