Construction and Validation of a Prognostic Nomogram for Clear Cell Renal Cell Carcinoma Based on DNA Methylation-driven Differentially Expressed Genes

Zheng Wang
Shanxi Medical University
Yanlong Zhang
Shanxi Medical University
Shuaishuai Fan
Shanxi Medical University
Yuan Ji
Shanxi Medical University
Jianchao Ren
Shanxi Medical University
Ke Yang
Shanxi Medical University
JINGQI WANG (drwangjq@126.com)
Department of Urology, First Hospital of Shanxi Medical University, Taiyuan, China
https://orcid.org/0000-0003-2820-1026

Research

Keywords: nomogram, risk score, clear cell renal cell carcinoma, DNA methylation, prognosis

Posted Date: September 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-75475/v1

Abstract

Background: Clear cell renal cell carcinoma (ccRCC) is the most frequent type of kidney cancer. This study aimed to establish a nomogram to predict ccRCC prognosis.

Methods: By integrating DNA methylation (DNAm) data and gene expression profiles of ccRCC obtained from The Cancer Genome Atlas (TCGA), DNAm-driven genes were identified by differential and correlation analyses. Next, risk genes were selected by multiple algorithms (univariate Cox and Kaplan-Meier survival analyses) and various databases (TCGA, Clinical Proteomic Tumor Analysis Consortium (CPTAC), and The Human Protein Atlas (HPA)). A risk score model was established by multivariate Cox analyses. ConsensusPathDB and Gene Set Enrichment Analysis (GSEA) were used to identify the biological functions of the selected genes. After comprehensively evaluating the clinical data, we established and assessed a dynamic nomogram available on a webserver.

Results: In total, 220 differentially expressed DNAm-driven genes were identified, and five-gene signature (EPB41L4B, HHLA2, IFI16, CMTM3, and XAF1) was related to overall survival (OS). Next, we integrated the DNAm-driven genes into the prognostic risk score model and found that age, histologic grade, pathological stage, and risk level were correlated with OS in ccRCC patients. Based on these variables, a dynamic nomogram was established to predict the ccRCC prognosis. Finally, Functional enrichment analysis showed that the functions of these genes were relevant to immune reactions.

Conclusions: We identified a 5 DNAm-driven gene signature whose altered status was highly correlated with ccRCC patient OS. We constructed a dynamic nomogram to provide individualized survival predictions for ccRCC patients.

Introduction

Kidney cancer is one of the most common malignancies of the urinary system and ranks fifth in males and eighth in females. (Siegel et al. 2019) Clear cell renal cell carcinoma (ccRCC) is the most frequent type of kidney cancer, accounting for approximately 70–80%. (Rini et al. 2009) Although considerable progress has been made in surgical treatment for early-stage disease, the systematic management of targeted therapy and immunotherapy for advanced stage disease and the overall survival (OS) and recurrence rates remain unsatisfactory. (Greef and Eisen 2016) Therefore, it is important to explore effective biomarkers to improve the early prediction of ccRCC.

To identify factors that can affect the prognosis of ccRCC, scholars have conducted many studies, such as those focusing on the SLC family. Weiting Kang et al. found that the expression levels of SLC22A6, SLC22A7, and other SLC family genes were significantly reduced in ccRCC tissues, and these genes might be potential targets. (Kang et al. 2020) Because of the lack of good sensitivity and specificity, the current use of biological factors for the early prediction of ccRCC is still very limited.

DNA methylation (DNAm) has been shown to be associated with the occurrence and development of tumors. (Lo and Huang 2002) (Guerra et al. 2019)Many studies have reported that DNAm- and DNAm-driven genes can be used as biomarkers to evaluate prognosis. For example, Bai et al. reported 6 DNAm-driven genes as new biomarkers for gastric cancer patients. (Bai et al. 2020) Sailer et al. reported that PITX2 DNAm could act as a novel prognostic biomarker for patients with head and neck squamous cell carcinoma. (Sailer et al. 2017) It is worth noting that DNAm is an inherently reversible change and may provide new drug targets. (Church et al. 2014) (Nielsen et al. 2017) Therefore, DNAm events may provide a new perspective to evaluate the prognosis of cancer patients.

Although the relationship between abnormal DNAm and ccRCC occurrence has been studied(Wang et al. 2020) (Yin et al. 2020), there have been few reports on prognostic models based on DNAm genes. Combining DNAm data with mRNA expression profile data, differentially expressed genes (DEGs) that were related to changes in DNAm status and associated with prognosis were identified. Then, we constructed a risk score model after performing univariate Cox regression, Kaplan-Meier (K-M) survival, least absolute shrinkage and selection operator (LASSO), and multivariate Cox regression analyses. By integrating the values of the risk score in the model with clinically relevant risk factors, we eventually constructed a nomogram for ccRCC patients. Then, functional enrichment analysis indicated that the functions of the selected genes were primarily related to immune reactions and immune infiltration. Finally, we translated our nomogram into a dynamic nomogram and placed it on an online webserver to make the process of predicting prognosis visual and convenient.

Materials And Methods

Data Collection

Data from The Cancer Genome Atlas (TCGA) were downloaded from the National Institutes of Health (NIH) Genomic Data Commons (GDC) (https://portal.gdc.cancer.gov/). We obtained level-3 molecular data from TCGA and the associated clinical data from GDC (Table 1). A value of β (0-1) was used to score the methylation degree of the gene. Zero means unmethylated, and one means totally methylated. The ccRCC and normal kidney tissue protein abundance data were downloaded from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (https://proteomics.cancer.gov/programs/cptac). The immunohistochemistry (IHC) data of renal adenocarcinoma and normal kidney tissues are download from The Human Protein Atlas The Human Protein Atlas (HPA) database (https://www.proteinatlas.org/).

Distinguishing the DEGs Between ccRCC and Nonneoplastic Tissues

The DEGs were distinguished between 539 ccRCC tissues and 72 nearby nonneoplastic renal tissues in the TCGA dataset (HTSeq-Counts of TCGA-Kidney Renal Clear Cell Carcinoma (KIRC) transcriptome data with corresponding prognostic information and a diagnosis of ccRCC) with the DESeq package. (Anders and Huber 2010) The selection criteria were |log fold change (FC)| > 1 and $p < 0.05$. Using Prism 8.0 software (GraphPad, San Diego, CA, USA), we analyzed the expression patterns of 8 DNAm-driven genes in ccRCC and nonneoplastic kidney tissues.

Identifying DNAm-driven Genes

Under the uniform TCGA nomenclature, we combined the transcriptional information of the relevant barcode with the DNAm information. By using linear regression analysis, we identified the DNAm-driven genes because of their higher DNAm degree and lower mRNA expression. Moreover, the diverse DNAm degrees between ccRCC tissues and nearby nonneoplastic renal tissues were compared with the Wilcoxon rank-sum test. (Gevaert 2015)

Functional Enrichment Analysis

We used 220 DNAm-driven genes for Gene Ontology (GO) analysis and pathway enrichment analysis by using the R package clusterProfiler and ConsensusPathDB (http://cpdb.molgen.mpg.de/).

Feature Selection and Construction of the Predictive Signature

After removing invalid sample data, 525 tumor samples were randomly divided into two groups: 350 samples in the training set and 175 samples in the verification set (Table 2). Then, we performed univariate Cox regression analysis and K-M analysis on the data of the training group, and we reduced the number of candidate genes by evaluating the effect of the candidate genes on the survival time of ccRCC patients. Afterward, we screened the candidate genes by comparing the genes with CPTAC data and performing the Wilcoxon signed-rank test. Then, we used LASSO analysis and multivariate Cox regression analysis to further filter the number of screened genes. The coefficient of the gene from multivariate analysis was multiplied by the mRNA expression level to calculate the risk value of each of the five genes.

Risk Score Modeling

Using X-tile (Camp et al. 2004) software to select appropriate cut-off values in the training set, the samples were divided into high-risk and low-risk groups. These two sets of data were then used to draw K-M survival curves for the ccRCC patients.

Screening of Prognostic Factors

Univariate Cox regression analysis was used to evaluate the value of the risk score model and other clinical factors for patient OS. Then, multivariate Cox regression analysis was used to eliminate confounding factors. The significance level was 0.05 (two-sided).

Construction and Evaluation of the Nomogram based on the Training Dataset

Through multivariate regression analysis, we identified factors significantly associated with prognosis and established a predictive model. A nomogram was constructed by the R package survival 3.1.8, and a dynamic nomogram was built by the R package DynNom. The nomogram performance was evaluated after using the Hosmer-Lemeshow test to assess the OS probabilities calibrated by the OS of ccRCC patients in the training set in different years. The nomogram performance was assessed by measuring

Harrell's concordance index (C-index). To obtain a robust C-index, 1000 bootstrap resamples were used for verification. The distribution of the C-index is between 0.5 and 1.0. The closer the score is to 0.5, the weaker the model's ability to distinguish the results. Conversely, the closer the score is to 1, the stronger the model's ability to distinguish the results. In addition, time-dependent receiver operating characteristic (ROC) analysis (Kamarudin et al. 2017) was used to test the nomogram's predictive power. Finally, decision curve analysis (DCA) (Vickers and Elkin 2006) was also used to measure the clinical effect on the decision, considering the clinical results.

Copy Number Variation (CNV) and Gene Set Enrichment Analysis (GSEA) of the Five DNAm-driven Genes

Through cBioPortal (http://www.cbioportal.org/), we obtained the TCGA data set containing all ccRCC tissues with CNV and mutation spectra data for the five genes. GSEA 4.0.3 software was used to annotate the functions of the five genes according to the methods described in the user guide (http://software.broadinstitute.org/gsea/index.jsp).

Statistical Analysis

R software (version 4.0.0) was used to perform all statistical analyses. The statistical tests were two-sided, and *P* values less than 0.05 were considered statistically significant.

Results

Acquisition of DEGs in ccRCC

Figure 1 shows our specific operation process. By comparing the mRNA expression data of ccRCC tissues (n=539) with the data of nontumor samples (n=72), we identified 5825 DEGs (|logFC| > 1, adjusted *P*-value < 0.05), among which the number of upregulated DEGs was 3370 and the number of downregulated DEGs was 2455 (Table S1).

Identification of DNAm-driven Genes in ccRCC

We conducted an analysis of methylation by combining the clinical samples downloaded from TCGA and the identified DNAm-driven genes in ccRCC. The corrected P value was less than 0.05, and the correlation between the DNAm degree and gene expression level was less than -0.3 as the filtering condition. We selected 220 DNAm-driven genes (96 hypermethylated and 124 hypomethylated) and showed the expression degree of these genes through the heat map (Figure 2A Table S2). GO analysis (P < 0.05) showed that the functions of these DNAm-driven genes could be described by eight GO terms (Figure 2B). Through this analysis, the DNAm-driven genes were clearly enriched in the following GO categories: basolateral plasma membrane, apical plasma membrane, brush border membrane, apical part of cell, apical junction complex, cell-cell junction, tight junction, bicellular tight junction, and embryonic skeletal system development (*P* < 0.001). Additionally, according to pathway analysis on ConsensusPathDB, these genes were mainly enriched in the T-cell antigen receptor (TCR) pathway during Staphylococcus aureus infection, inflammatory response pathway, primary immunodeficiency, TCR signaling pathway, T-

cell receptor signaling pathway, rheumatoid arthritis, tight junction, and cell adhesion molecules (CAMs) ($P < 0.05$; Figure 2C).

Selection of Prognostic Genes to Include in the Risk Score Model

Then, we intersected the DEGs and the DNAm-driven genes. Next, we eliminated invalid sample data and randomly divided the 525 samples into two sets: 350 samples for training and 175 for validation. First, univariate Cox analysis was used to assess the effect of the candidate genes on prognosis. Setting the filter condition as P < 0.05, 36 of 103 candidate genes were selected. Then, K-M analysis was used to analyze the influence of 36 DNAm DEGs on the OS of ccRCC patients, and 32 passed the log-rank test and met the criteria for statistical significance ($P < 0.05$).

Validation of the Abnormal Expression of the Prognostic Biomarkers at the Protein Level

To confirm whether there was a differential abundance of the protein associated with the selected genes between the normal kidney tissues and tumor tissues of ccRCC, we downloaded the protein abundance data of 84 normal kidney tissues and 110 ccRCC tissues from the CPTAC database, and 17 of 32 prognostic genes were found in the CPTAC database. Then, we proceeded with the differential analysis using the Wilcoxon signed-rank test. The results showed that 15 of these proteins were abnormally regulated in tumor tissues (P<0.05) (Figure 2D), consistent with the results of the differential gene expression analysis and further confirming the importance of these genes in the occurrence and development of ccRCC.

Establishing a Prognostic Risk Score Model for ccRCC

Subsequently, the 15 prognostic genes were analyzed by LASSO; the filter condition was that the selected genes must be present each time after repeating the process 1000 times (Figure 3A). After LASSO analysis, we narrowed the 15 genes down to 7. Finally, we selected five DNAm-driven DEGs (HHLA2, EPB41L4B, IFI16, CMTM3, and XAF1) by multivariate Cox analysis (with forward selection and backward selection). The products of the respective mRNA expression levels of the five genes and the correlation coefficients were taken as the risk score, and a prediction model was established as follows: risk score = (-0.11958 * *HHLA2* mRNA expression) + (0.2000049**EPB41L4B* mRNA expression) + (0.504096 * *IFI16*) + (0.201696 *CMTM3 mRNA expression) + (0.242373**XAF1* mRNA expression). The higher the expression levels of EPB41L4B, IFI16, XAF1, and CMTM3 were, the shorter the expected survival time of the patients. In contrast, the expression level of HHLA2 was positively correlated with the pre-OS survival time. After we calculated the risk values based on these five genes and determined the most compatible cut-off value (10.75) using X-tile, we divided the 350 patients in the training set with complete clinical data into high-risk (106) and low-risk (244) groups. The K-M results of the two groups clearly show that the OS is greatly shortened in the high-risk group (P <0.0001; Figure 3B). The heat maps in Figure 3C show the gene expression levels and associated risk scores of all patients.

Construction and Assessment of the Nomogram for Predicting OS in ccRCC

To make our nomogram applicable to all ccRCC patients, we changed the risk score to the risk level by the cut-off value. After age, histologic grade, pathological stage, and risk level (divided by the risk score) were identified as important prognostic factors, we conducted univariate regression analyses and then multivariate regression analyses (Figure 4A). By synthesizing the important predictors of interest, we obtained a comprehensive nomogram (Figure 4B). In addition, to assess the reasonableness of these related factors in the model, we plotted Schoenfeld model residuals for age, histologic grade, pathological stage, and risk level (divided by the risk score) in turn. From the Schoenfeld residuals results, it is easy to see that the model we established meets the equally proportional risk hypothesis (Figure 4C). The prediction accuracy of the model was verified by drawing calibration curves for the 3-year and 5-year OS probabilities (Figure 4D). Then, we confirmed the predictive ability of the five-gene signature by using the area under the curve (AUC) of the time-dependent ROC curve. In the ROC diagram, the AUC of nomo-point was higher than that of age, histologic grade, pathological stage, and risk level (Figure 4E), indicating that the evaluation level of the nomogram was significantly higher than that of age, histologic grade, pathological stage, and risk level (divided by the risk score). Compared with the age, histologic grade, pathological stage, and risk level (divided by the risk score) models, the combined model showed more predictive power, so the nomogram was more accurate in evaluating patient prognosis in clinical practice. (Figure 4F). Based on the good performance of this model in both short-term and long-term OS predictions, we believe that this nomogram can help doctors make decisions to some extent when devising follow-up plans for ccRCC patients.

Cross-validation of the Nomogram

We validated the established nomogram using 175 samples from the validation set (Figure 5 A-D). Then, the nomogram was used in the validation set to obtain 1-, 3-, and 5-year OS curves, which were in clear agreement with the predicted results (Figure 5C). Similar results were found in the training set, and we obtained 1- (0.869), 3- (0.824), and 5-year AUCs (0.719) in the validation set (Figure 5D). By comparing the prognostic model containing one variable (Shigeyasu et al. 2015) (Deng et al. 2014) with our risk scoring model from year 1 to year 10 via the C-index, we found that the nomo-point scoring model was the best-performing indicator in the training set and validation set (Figure 5E).

Methylation Levels and Gene Expression Levels of the Five DNAm-driven DEGs

In the five-gene signature, EPB41L4B was hypermethylated, while HHLA2, IFI16, CMTM3, and XAF1 were hypomethylated (Figure 6A and C) based on the TCGA ccRCC cohort. In Figure 6B, we can see that the methylation level is significantly negatively correlated with the mRNA level ($P < 0.05$ and $|R|>0.3$). In addition, the mRNA expression levels of the hypermethylated DEGs in ccRCC tissues were significantly lower than those in nearby nonneoplastic renal tissues ($P < 0.05$, Figure 6D).

CNV, Mutation Characteristics and Kyoto Encyclopedia of Genes and Genomes (KEGG) Enrichment

The five signature genes (HHLA2, EPB41L4B, IFI16, CMTM3, and XAF1) selected were affected by deletions, amplifications, and mutations in addition to methylation. Using the GDC TCGA renal carcinoma

(KIRC) database, the results showed that the percentage of changes in these five genes ranged from 4% to 6%, and the mRNA expression was therefore little affected (Figure S2A). By using regression analysis, we found no significant correlation between the CNV and mRNA expression of each gene (Figure S2C). It can be concluded from the above finding that DNAm plays a more important role in ccRCC. By GSEA of the high- and low-risk groups, we explored some potential signaling pathways that affect the risk scoring model. The cut-off criteria were P<0.05 and |enrichment score (ES)|>0.45. In the high-risk group (risk score >=10.77), the following five signaling pathways ranked the highest: cytokine receptor interaction, glycosaminoglycan biosynthesis-chondroitin sulfate, intestinal immune network for IgA production, primary immunodeficiency, and systemic lupus erythematosus (Figure S2B). In the low-risk group (risk score <10.77), the following five signaling pathways ranked the highest: beta alanine metabolism, fatty acid metabolism, histidine metabolism, peroxisome, and tryptophan metabolism (Figure S2B). Some of these signaling pathways have been reported to be related to tumor immunity, providing more support for further exploring the influence of immune factors on ccRCC.

The IHC of Five DNAm-driven DEGs in HPA

Based on the protein abundance data from the HPA, we compared the five DNAm-driven DEGs in renal adenocarcinoma, and normal kidney tissues and the abundance of these proteins could be divided into four categories: high, medium, low, and not detected by the scoring system and the intensity of staining (strong, moderate, weak, or negative). The EPB41L4B staining in normal kidney tissues was "high", and in renal cancer was "medium". In contrast, CMTM3, HHLA2, XAF1, and IFI16 staining in normal kidney tissues were lower than in renal adenocarcinoma (Figure 7). The result was consistent with the differential analysis based on TCGA and CPTAC, and further confirmed that these genes displayed a crucial role in the occurrence and development of ccRCC.

Generation of an Online Nomogram for Easy Access

To make our nomogram more convenient, we designed a dynamic nomogram and updated it to a website (https://yanlongzhangmodel.shinyapps.io/ccRCCmethylationDynNomapp/) to help clinicians. The OS time of ccRCC patients can be readily estimated by submitting variables (including age, histologic grade, pathological stage, and risk level), and figures and tables created by the dynamic nomogram are then outputted.

Discussion

Finding biomarkers with high sensitivity and specificity to predict prognosis is of great significance for the management of ccRCC patients. In recent years, some prognostic models of ccRCC patients have been reported, including immune-related genes, iron metabolism-related genes, SLC family genes, and autophagy-related genes. Mou et al. reported that TSC1, TSC2, and other genes that affect the prognosis of ccRCC are metabolic pathway operators mediated by iron, oxygen, nutrition, or energy stimulation. (Mou et al. 2020) Some studies have investigated the prognostic value of DNA methylation and indicated that DNA methylation is a crucial factor impacting OS in ccRCC. However, these studies focused on a

single gene or only used DNAm data. (Wang et al. 2020) We combined DNAm data with transcriptome data through various statistical algorithms and selected the five genes that were most relevant to prognosis. On this basis, we first developed and validated a prognostic risk model based on the DNAm signature, and an OS nomogram was constructed to predict the prognosis of ccRCC patients by integrating age, histologic grade, pathological stage, and risk level (divided by the risk score) in the model. After verification, we proved that the accuracy of the model was more satisfactory than that of traditional clinical variables (age, histologic grade, and pathological stage). To make the nomogram more convenient and accurate, we uploaded our dynamic nomogram to the webserver so clinicians can use it quickly for the visualization of prognosis predictions.

Abnormal DNAm often occurs in tumor cells. When the level of DNAm increases, the expression of this gene often decreases, and vice versa. Aberrant DNAm might change the physiological state and accelerate the occurrence and metastasis of cancer. Thanks to the in-depth development of methylation sequencing technology and the improvement of sequencing depth and accuracy, we can identify abnormal DNAm-driven genes with the help of a model-based instrument (MethylMix) and conduct a correlation analysis of these genes with DEGs. These studies could provide a new perspective on the management and treatment of tumor patients. In this study, the enrichment analysis results showed that the 220 preliminarily screened DNAm-driven genes had many immune-related functional annotations, such as rheumatoid arthritis, TCR pathway during Staphylococcus aureus infection, TCR signaling pathway, CAMs, and primary immunodeficiency. Our research findings indicated that four genes (EPB41L4B, IFI16, CMTM3, and XAF1) were highly expressed in tumor samples. The lower the expression of these genes was, the better the prognosis. From this finding, we can infer that the hypermethylation of these genes is beneficial to patients with ccRCC. Conversely, the lower the DNAm level of HHLA2 is, the better the prognosis, suggesting that it may be a tumor suppressor gene. To understand how these genes function in ccRCC, we used GSEA to identify the associated KEGG pathways. Of the five enriched pathways of these genes, three were also associated with tumor immunity, including intestinal immune network for IgA production, primary immunodeficiency, and systemic lupus erythematosus. Moreover, the results of the pathway analysis indicated that these genes' functions were associated with tumor metabolism, for instance, beta alanine metabolism, fatty acid metabolism, histidine metabolism, and tryptophan metabolism. Considering the results of the ConsensusPathDB enrichment analysis mentioned above, it is reasonable to speculate that there is a strong correlation between the occurrence of ccRCC and immune reaction and infiltration. Next, we tried to combine methylation and tumor immunity data for further exploration. Of course, there are some reports exploring the function of these signatures in other cancers. For example, HHLA2 is a new immune checkpoint in ccRCC, (Zhang et al. 2020) EPB41L4B (Ehm2) has been proven to promote the proliferation and invasion of lung cancer cells, (Li et al. 2019) IFI16 can impact the immune reaction, (Sui et al. 2019) XAF1 has a function related to tumor apoptosis, (Jeong et al. 2018) and CMTM3 has been reported to be a promoter in gastric cancer. (Yuan et al. 2017)

Compared with other studies, our study searched for biomarkers in multiple databases (TCGA, CPTAC, and HPA) and at multiple levels (DNAm, mRNA expression, and protein abundance). Then, our nomogram was developed with the mRNA expression data of DNAm driver genes rather than DNAm data and

validated by multiple algorithms. Finally, the dynamic nomogram available on the webserver can conveniently help clinicians predict the OS of ccRCC patients and visualize this prediction. Nevertheless, this paper has some shortcomings. Since both the training set and the validation set were from the same database, the external validation potential of the model is questionable. The development of the prognostic model from the perspective of methylation provides a new perspective for clinical work. However, the model may be inevitably too one-sided, like other prognostic models, and cannot take other factors into account and lacks certain persuasion.

In summary, we developed and verified a prognostic risk score prognostic model composed of five DNAm-driven DEGs through multiple algorithms and verified the accuracy of this model. Because of the integration of other clinical features and the availability of this service on the webserver, the dynamic nomogram provides more precise and convenient prognostic prediction for ccRCC patients than traditional clinical variables. At the same time, we found that these genes regulated by the DNAm level had a close connection with the immune reaction and immune infiltration in ccRCC. Importantly, with only five genes, this model provides a relatively accurate prediction at a lower cost for patients with ccRCC and improves the sensitivity and accuracy of the individualized prediction of OS for patients with ccRCC.

Conclusions

Our study was the first to develop a nomogram that combined the DNAm signature, age, histologic grade, and pathological stage and had high performance at a low cost in the clinical setting, advancing the individualized prediction of OS in ccRCC patients, with high sensitivity and specificity.

Declarations

Funding

Not applicable

Conflicts of interest/Competing interests

Conflict of interest The authors have no conflicts of interest.

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent for publication

Written informed consent for publication was obtained from all participants.

Availability of data and material

The data analyzed in this study can be downloaded from the TCGA, CPTAC and HPA.

Code availability

Not applicable

Authors' contributions

Zheng Wang and Yanlong Zhang designed the study and analyzed the data; Zheng Wang, Yanlong Zhang, Shuaishuai Fan, Yuan Ji, Jianchao Ren, and Ke Yang revised the images;Zheng Wang, Yanlong Zhang, Shuaishuai Fan, Yuan Ji, Jianchao Ren, and Ke Yang performed the literature search and collected data for the manuscript; Jingqi Wang revised the manuscript.

References

1. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106. https://doi.org/10.1186/gb-2010-11-10-r106
2. Bai Y, Wei C, Zhong Y, et al (2020) Development and Validation of a Prognostic Nomogram for Gastric Cancer Based on DNA Methylation-Driven Differentially Expressed Genes. Int J Biol Sci 16:1153–1165. https://doi.org/10.7150/ijbs.41587
3. Camp RL, Dolled-Filhart M, Rimm DL (2004) X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. Clin Cancer Res 10:7252–7259. https://doi.org/10.1158/1078-0432.CCR-04-0713
4. Church TR, Wandell M, Lofton-Day C, et al (2014) Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. Gut 63:317–325. https://doi.org/10.1136/gutjnl-2012-304149
5. Deng J, Liang H, Ying G, et al (2014) Methylation of ras association domain protein 10 (RASSF10) promoter negative association with the survival of gastric cancer. Am J Cancer Res 4:916–923
6. Gevaert O (2015) MethylMix: an R package for identifying DNA methylation-driven genes. Bioinformatics 31:1839–1841. https://doi.org/10.1093/bioinformatics/btv020
7. Greef B, Eisen T (2016) Medical treatment of renal cancer: new horizons. Br J Cancer 115:505–516. https://doi.org/10.1038/bjc.2016.230
8. Guerra JV da S, Pereira BM de S, Cruz JGV da, et al (2019) Genes Controlled by DNA Methylation Are Involved in Wilms Tumor Progression. Cells 8:. https://doi.org/10.3390/cells8080921
9. Jeong S-I, Kim J-W, Ko K-P, et al (2018) XAF1 forms a positive feedback loop with IRF-1 to drive apoptotic stress response and suppress tumorigenesis. Cell Death Dis 9:806. https://doi.org/10.1038/s41419-018-0867-4

10. Kamarudin AN, Cox T, Kolamunnage-Dona R (2017) Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol 17:53. https://doi.org/10.1186/s12874-017-0332-6
11. Kang W, Zhang M, Wang Q, et al (2020) The SLC Family Are Candidate Diagnostic and Prognostic Biomarkers in Clear Cell Renal Cell Carcinoma. Biomed Res Int 2020:1932948. https://doi.org/10.1155/2020/1932948
12. Li S, Ma J, Si Y, et al (2019) Differential expression and functions of Ehm2 transcript variants in lung adenocarcinoma. Int J Oncol 54:1747–1758. https://doi.org/10.3892/ijo.2019.4732
13. Lo K-W, Huang DP (2002) Genetic and epigenetic changes in nasopharyngeal carcinoma. Semin Cancer Biol 12:451–462. https://doi.org/10.1016/s1044579x02000883
14. Mou Y, Zhang Y, Wu J, et al (2020) The Landscape of Iron Metabolism-Related and Methylated Genes in the Prognosis Prediction of Clear Cell Renal Cell Carcinoma. Front Oncol 10:788. https://doi.org/10.3389/fonc.2020.00788
15. Nielsen SN, Grell K, Nersting J, et al (2017) DNA-thioguanine nucleotide concentration and relapse-free survival during maintenance therapy of childhood acute lymphoblastic leukaemia (NOPHO ALL2008): a prospective substudy of a phase 3 trial. Lancet Oncol 18:515–524. https://doi.org/10.1016/S1470-2045(17)30154-7
16. Rini BI, Campbell SC, Escudier B (2009) Renal cell carcinoma. Lancet 373:1119–1132. https://doi.org/10.1016/S0140-6736(09)60229-4
17. Sailer V, Gevensleben H, Dietrich J, et al (2017) Clinical performance validation of PITX2 DNA methylation as prognostic biomarker in patients with head and neck squamous cell carcinoma. PLoS ONE 12:e0179412. https://doi.org/10.1371/journal.pone.0179412
18. Shigeyasu K, Nagasaka T, Mori Y, et al (2015) Clinical Significance of MLH1 Methylation and CpG Island Methylator Phenotype as Prognostic Markers in Patients with Gastric Cancer. PLoS ONE 10:e0130409. https://doi.org/10.1371/journal.pone.0130409
19. Siegel RL, Miller KD, Jemal A (2019) Cancer statistics, 2019. CA Cancer J Clin 69:7–34. https://doi.org/10.3322/caac.21551
20. Sui H, Yang J, Hu X, et al (2019) siRNA containing a unique 5-nucleotide motif acts as a quencher of IFI16-mediated innate immune response. Mol Immunol 114:330–340. https://doi.org/10.1016/j.molimm.2019.08.007
21. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26:565–574. https://doi.org/10.1177/0272989X06295361
22. Wang J, Zhang Q, Zhu Q, et al (2020) Identification of methylation-driven genes related to prognosis in clear-cell renal cell carcinoma. J Cell Physiol 235:1296–1308. https://doi.org/10.1002/jcp.29046
23. Yin H, Zhang H, Wang X, Xu Q (2020) Four methylation-driven genes may be prognostic biomarkers in clear cell renal carcinoma. Clin Transl Med. https://doi.org/10.1002/ctm2.45
24. Yuan W, Liu B, Wang X, et al (2017) CMTM3 decreases EGFR expression and EGF-mediated tumorigenicity by promoting Rab5 activity in gastric cancer. Cancer Lett 386:77–86.

https://doi.org/10.1016/j.canlet.2016.11.015
25. Zhang Z, Liu J, Zhang C, et al (2020) Over-Expression and Prognostic Significance of HHLA2, a New Immune Checkpoint Molecule, in Human Clear Cell Renal Cell Carcinoma. Front Cell Dev Biol 8:280. https://doi.org/10.3389/fcell.2020.00280

Tables

Table 1. Clinical information from the 525 ccRCC patients of TCGA.

Clinical parameters	Variables	n(total=525)	Percentages (%)
Age	<=60	262	0.49904762
	>60	263	0.50095238
Gender	male	342	0.65142857
	female	183	0.34857143
Histologic-grade	G1	13	0.0247619
	G2	225	0.42857143
	G3	204	0.38857143
	G4	78	0.14857143
	Gx	5	0.00952381
	not applicable	3	0.00571429
Pathological-stage	Stage I	262	0.49904762
	Stage II	57	0.10857143
	Stage III	121	0.23047619
	Stage IV	82	0.15619048
	not applicable	3	0.00571429
Tumor	T1	268	0.51047619
	T2	69	0.13142857
	T3	177	0.33714286
	T4	11	0.02095238
Metastasis	M0	417	0.79428571
	M1	78	0.14857143
	Mx	28	0.05333333
	not applicable	2	0.00380952
Lymph Node	N0	236	0.44952381
	N1	16	0.03047619
	Nx	273	0.52

Table 2. Grouping of the ccRCC patients.

Clinical parameter	Variables	Training dataset	Validation dataset
Dead or Alive	Live	228(65%)	127(73%)
	Dead	122(35%)	48(27%)

Figures

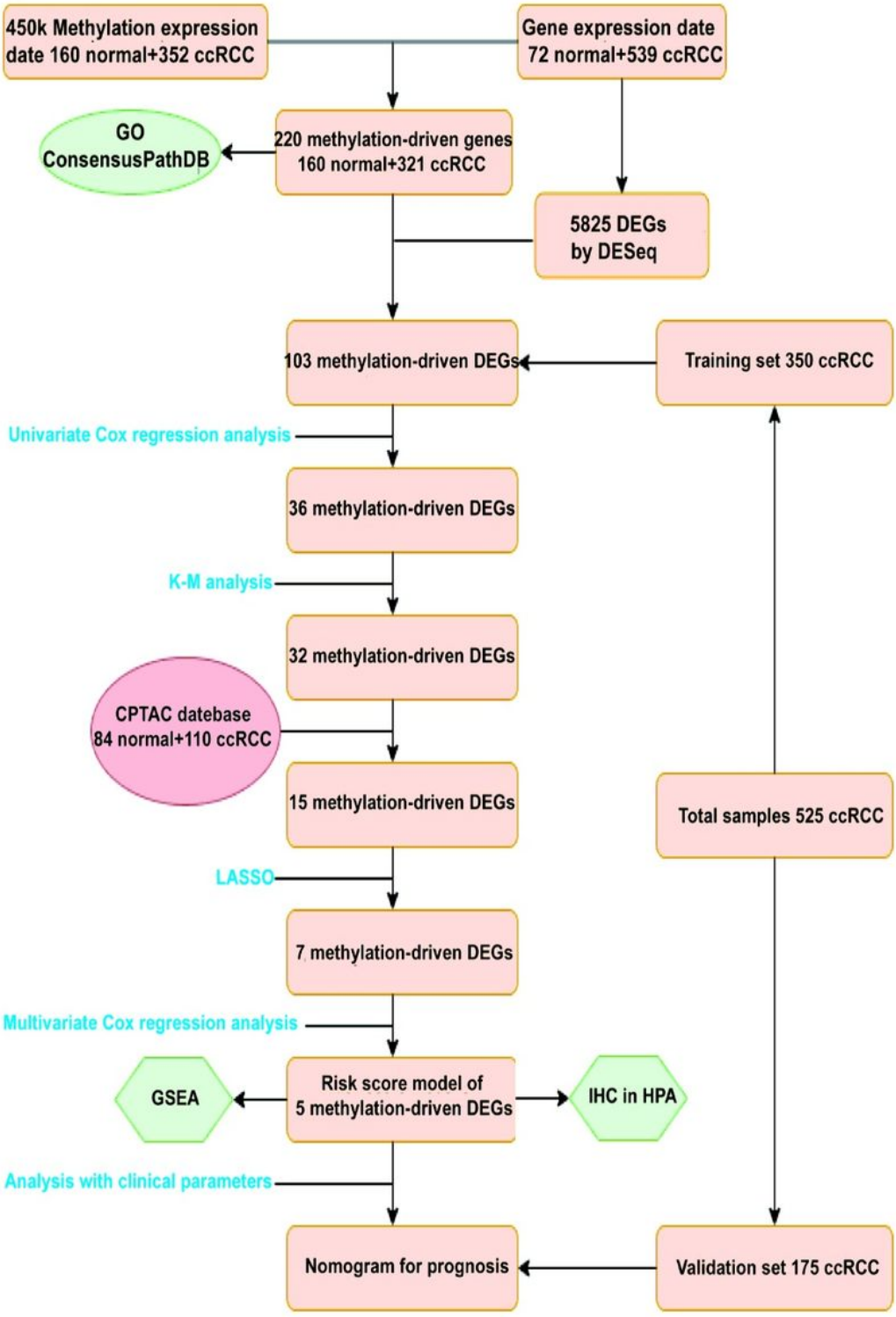

Figure 1

Flowchart depicting how prognostic genes were identified.

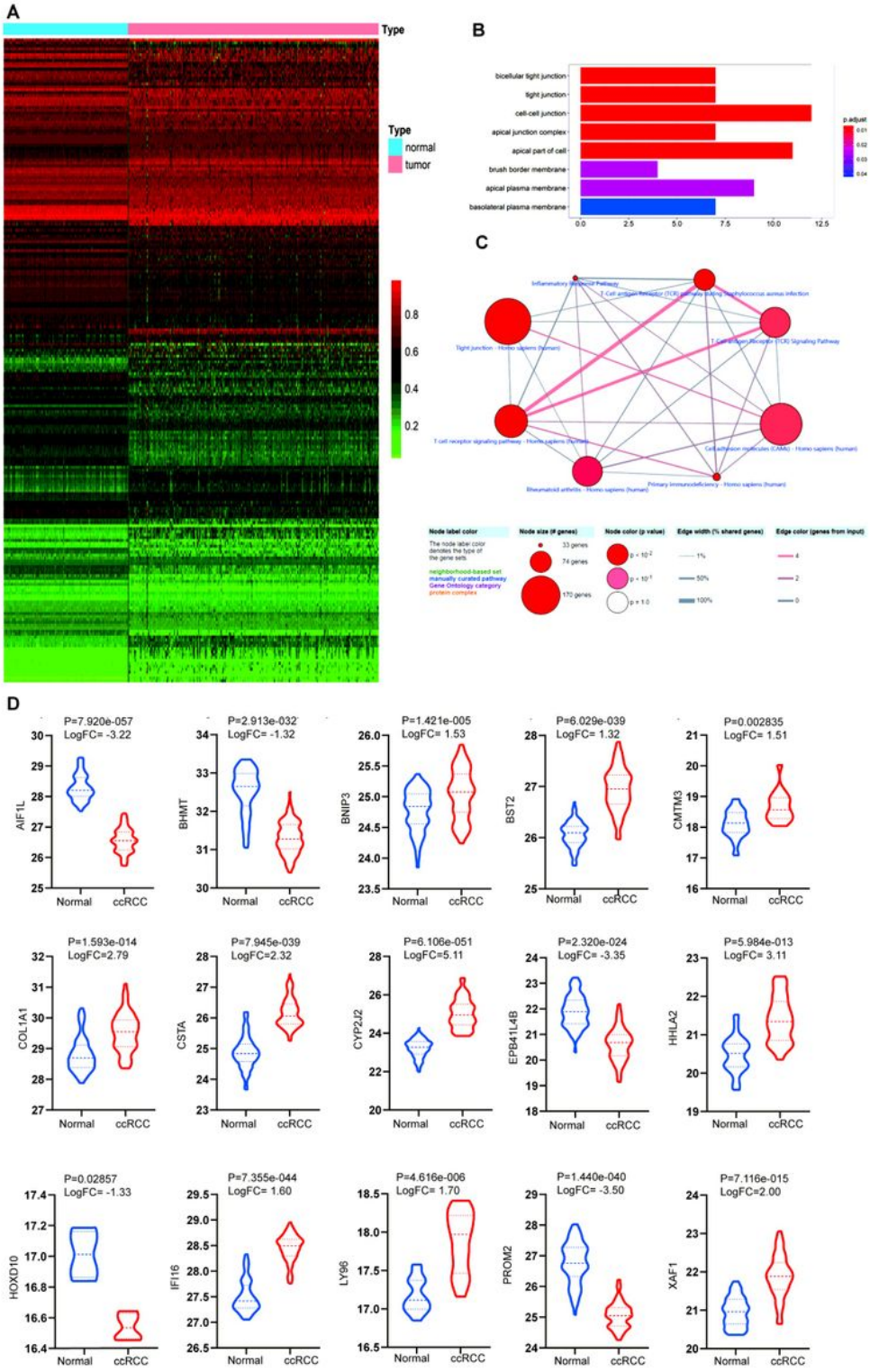

Figure 2

Candidate DNAm-driven genes screened by the Wilcoxon test. (A) Heatmap of the candidate DNAm-driven genes (n=220) in ccRCC and nontumorous renal tissues. (B) GO analysis of 220 DNAm-driven genes. (C) Pathway analysis based on multiple databases. (D) Validation of prognostic biomarkers by CPTAC in protein level.

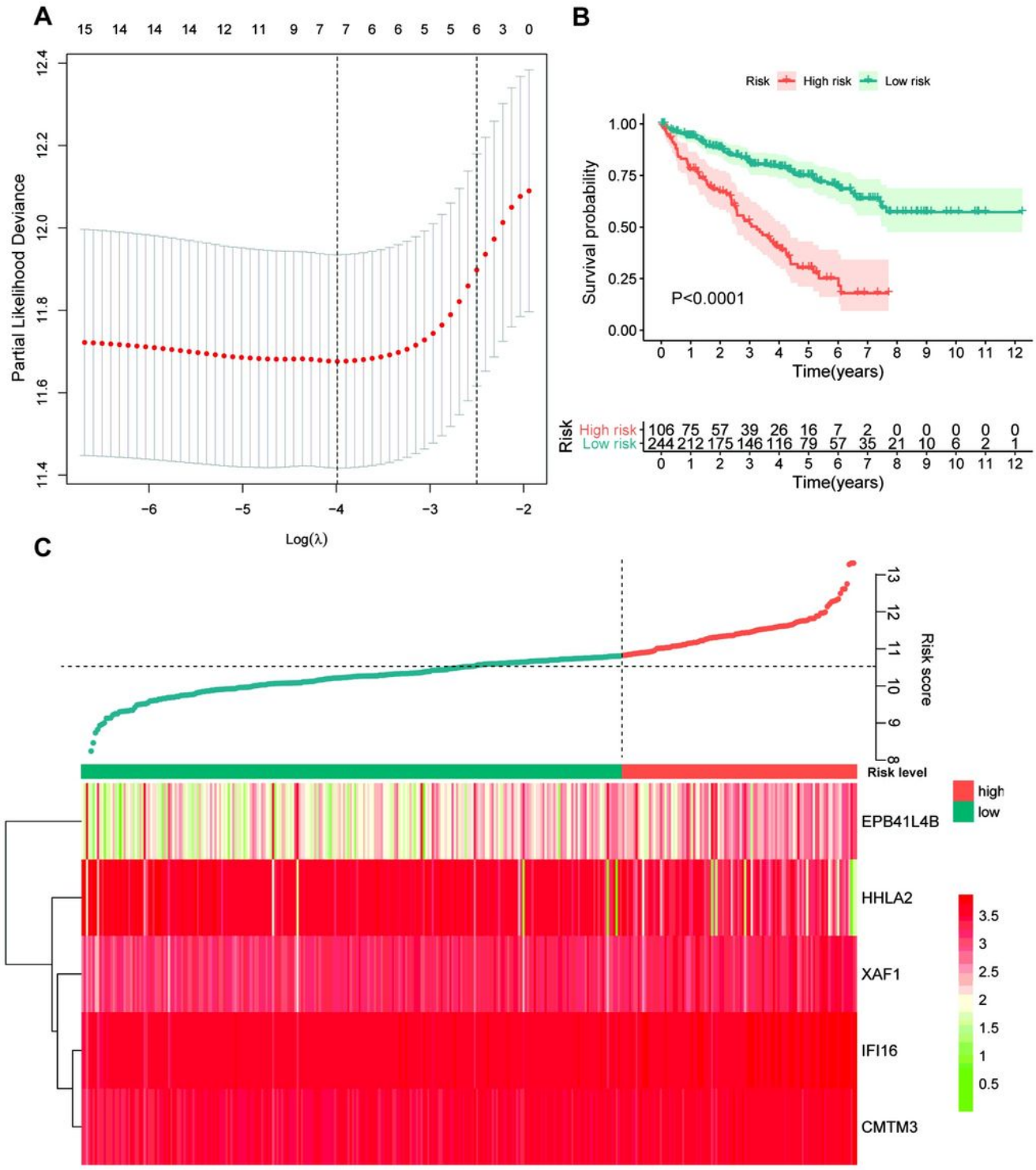

Figure 3

Texture feature selection and five-gene risk score model construction in the TCGA cohort. (A) Tuning parameter (λ) selection in the LASSO model used ten-fold cross-validation via the maximum criteria. The dotted vertical lines were drawn at the optimal values using the maximum criteria and the one standard error of the maximum criteria (the 1-SE criteria). (B) Comparison of OS between the high-risk score and

low-risk score groups. (C) Heatmap of the five-gene expression profiles and distribution of corresponding risk scores in the high-risk and low-risk subgroups in the TCGA database.

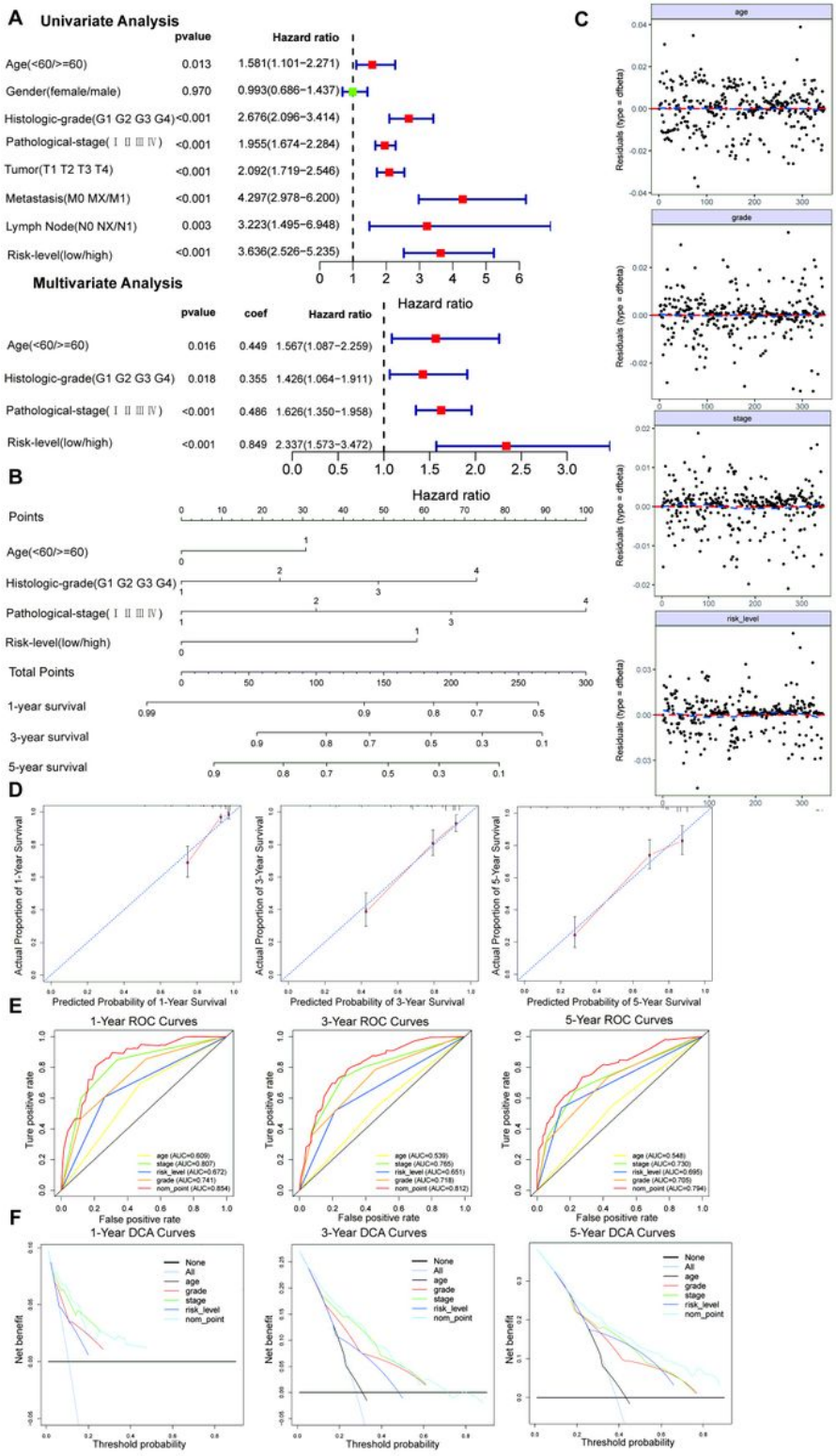

Figure 4

Nomogram to predict 1-, 3- and 5-year OS. The OS nomogram was developed in the TCGA cohort, with age, histologic-grade, pathological-stage, and risk level (DNAm signature) incorporated. (A) Univariate and multivariate analyses of the risk level, clinical factors, and pathological characteristics with OS. The

statistical significance level is indicated by different colors; red indicates statistical significance, and green indicates no significance. (B) Nomogram to predict the 1-, 3- and 5-year OS of GC patients. (C) The Schoenfeld residual suggested that this model met the equally proportional risk hypothesis. Schoenfeld model residuals vs age, histologic-grade, pathological-stage, and risk level were plotted to obtain a preliminary assessment of which of these predictive factors should be incorporated into the model. (D) Calibration curves of 1-, 3-, and 5-year OS. Blue dotted lines represent the ideal predictive model, and the red solid line represents the observed model. (E) Time-dependent ROC analysis was used to evaluate the accuracy of the OS nomograms. The red, yellow, orange, green, and blue solid lines represent the combined model, age, histologic-grade, pathological-stage, and risk level, respectively. (F) DCA curves evaluate OS nomograms from the perspective of clinical benefit and scope of clinical benefits. The y-axis represents the net benefit. The x-axis represents the predicted OS probability. The black dotted line represents the condition that all patients survive in 5 years, while the gray solid line represents the condition that none of the patients survive for more than one year. In the current study, the decision curve showed more benefit with a threshold probability > 0.0% using the OS nomogram.

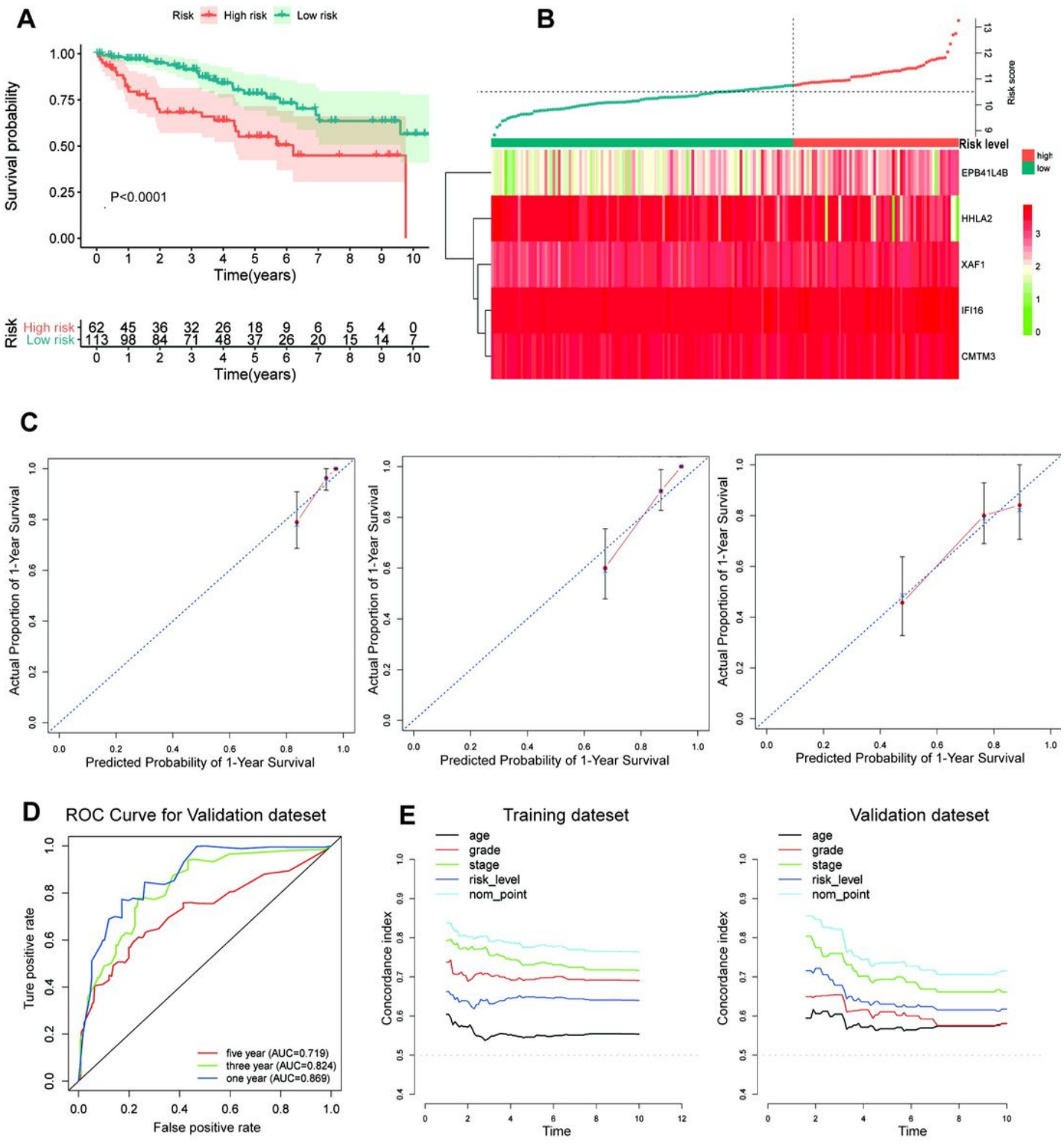

Figure 5

Validation of the prediction model. (A) OS was significantly lower in the high-risk score group than in the low-risk score group. (B) Heatmap and distribution of the five gene expression profiles in the high-risk and low-risk subgroups in the validation set. (C) Calibration curve for the risk score model in the validation cohort. The blue dotted line represents the ideal predictive model, and the red solid line represents the observed model. (D) ROC of the survival prediction model with the combined model, age, histological

grade, pathological stage, and risk level in the validation set. (E) Concordance index of the indicated prognostic model in the training and validation datasets.

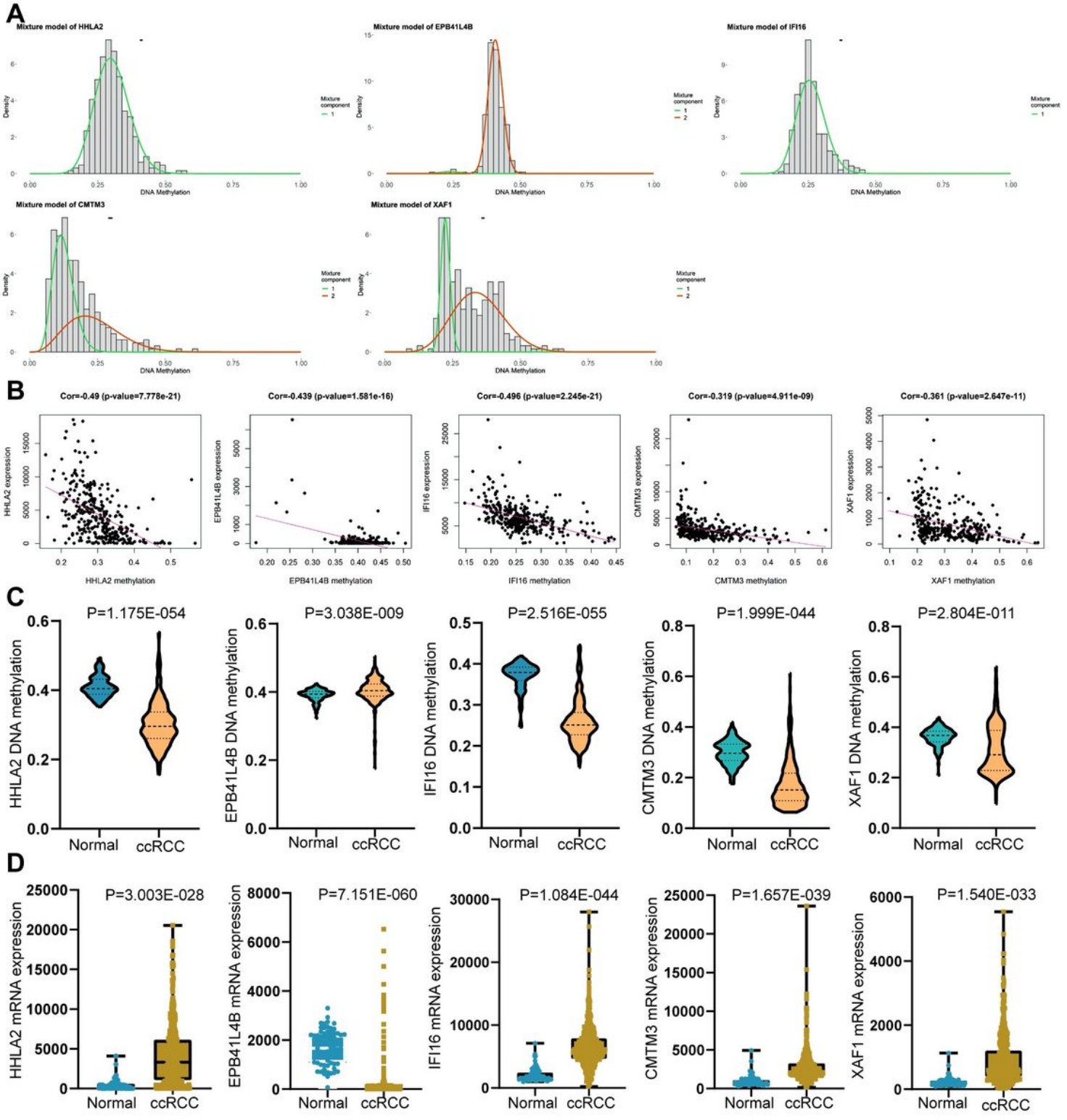

Figure 6

DNAm-driven genes. (A) Differential methylation statuses. The histogram demonstrates the distribution of HHLA2, EPB41L4B, IFI16, CMTM3, and XAF1 methylation in ccRCC samples. Beta values represent the methylation level (range from 0 to 1), and the horizontal black bar indicates the distribution of

methylation values in the nontumorous renal samples. (B) Regression analysis between the mRNA level and DNAm level of the five DNAm-driven DEGs. The vertical axis and the horizontal axis denote the mRNA level and DNAm level, respectively. (C) DNA methylation of the five DNAm-driven DEGs. (D) mRNA expression of the five DNAm-driven DEGs.

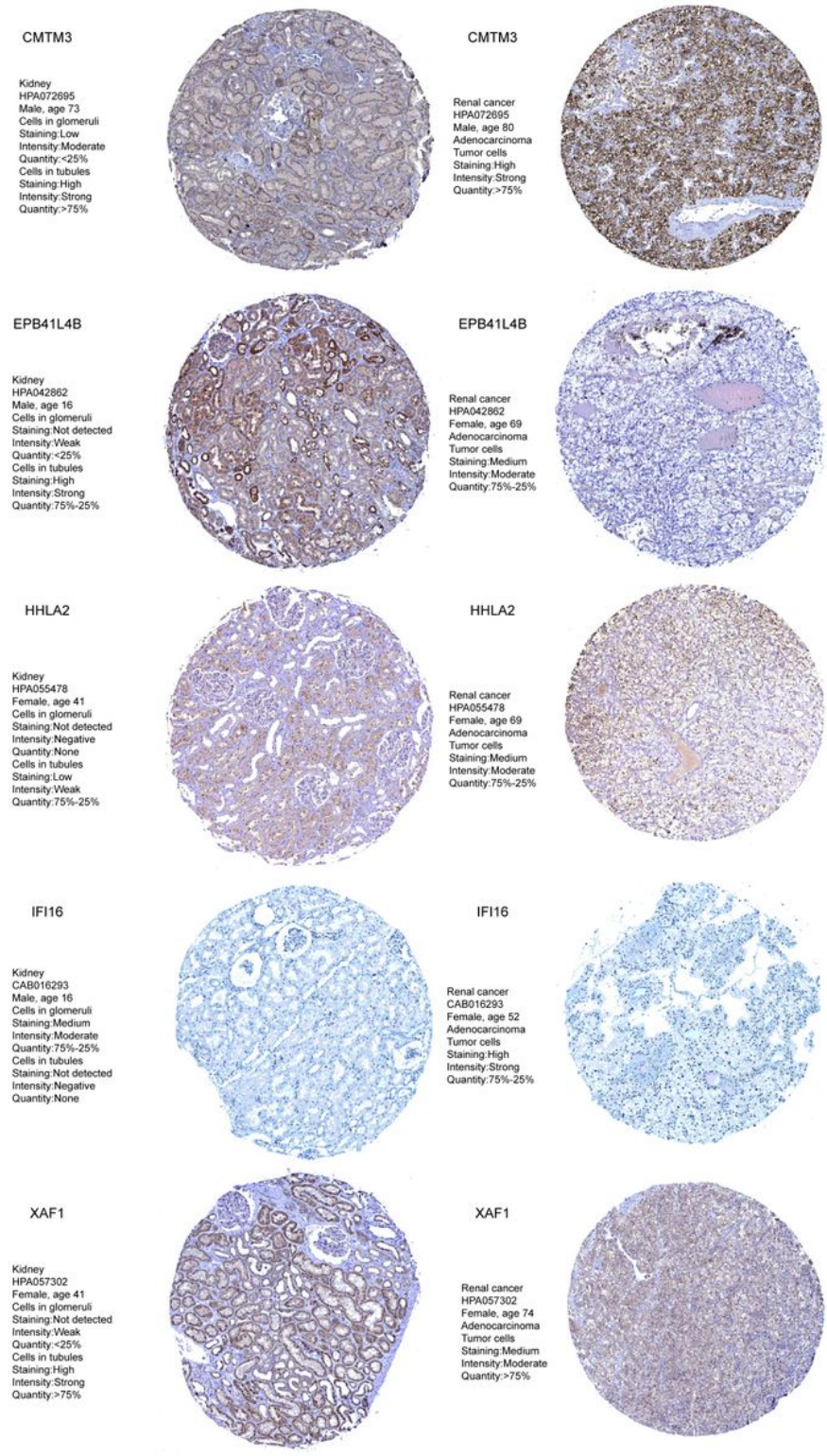

Figure 7

The protein expression of CMTM3, EPB41L4B, HHLA2, IFI16, and XAF1 were detected in normal kidney tissues and kidney adenocarcinoma from the Human Protein Atlas (HPA) database.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xls
- TableS2.xls
- FigS1.tif
- FigS2.tif