A 44-gene set constructed for predicting the prognosis of clear cell renal cell carcinoma

YONGGANG WANG, YAO WANG and FENG LIU

Department of Urology, China-Japan Union Hospital of Jilin University, Changchun, Jilin 130033, P.R. China

Received May 1, 2018; Accepted September 7, 2018

DOI: 10.3892/ijmm.2018.3899

Abstract. Clear cell renal cell carcinoma (ccRCC) is the most frequent type of renal cell carcinoma (RCC). The present study aimed to examine prognostic markers and construct a prognostic prediction system for ccRCC. The mRNA sequencing data of ccRCC was downloaded from The Cancer Genome Atlas (TCGA) database, and the GSE40435 dataset was obtained from the Gene Expression Omnibus database. Using the Limma package, the differentially expressed genes (DEGs) in the TCGA dataset and GSE40435 dataset were obtained, respectively, and the overlapped DEGs were selected. Subsequently, Cox regression analysis was applied for screening prognosis-associated genes. Following visualization of the co-expression network using Cytoscape software, the network modules were examined using the GraphWeb tool. Functional annotation for genes in the network was performed using the clusterProfiler package. Finally, a prognostic prediction system was constructed through Bayes discriminant analysis and confirmed with the GSE29609 validation dataset. The results revealed a total of 263 overlapped DEGs and 161 prognosis-associated genes. Following construction of the co-expression network, 16 functional terms and three pathways were obtained for genes in the network. In addition, red, yellow (Involving chemokine ligand 10 (CXCL10), CD27 molecule (CD27) and runt-related transcription factor 3 (RUNX3)), green (Involving angiopoietin-like 4 (ANGPTL4), stanniocalcin 2 (STC2), and sperm associated antigen 4 (SPAG4)), and cyan modules were extracted from the co-expression network. Additionally, the prognostic prediction system involving 44 signature genes, including ANGPTL4, STC2, CXCL10, SPAG4, CD27, MMP9 and RUNX3, was identified and confirmed. In conclusion, the 44-gene prognostic prediction system involving ANGPTL4, STC2, CXCL10, SPAG4, CD27, MMP9 and RUNX3 may be utilized for predicting the prognosis of patients with ccRCC.

Introduction

Renal cell carcinoma (RCC), which is derived from the lining of the proximal convoluted tubule, represents 90-95% of cases of kidney cancer (1). The early symptoms of RCC are usually undetected, which leads to advanced disease stages in patients newly diagnosed with RCC (2). With the progression of RCC, tumor cells may metastasize to other organs, including the liver, lymph nodes, lungs, brain, adrenal glands and bones (3). RCC has a relatively higher incidence in men than women, particularly in those >65 years old (4). Clear cell RCC (ccRCC), characterized by the clear cytoplasm in cells, is the most frequent type of RCC (5). Therefore, examining the pathogenesis of ccRCC is necessary and of significance.

There are several studies reporting the molecular mechanisms of ccRCC. For example, metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) is upregulated in patients with ccRCC, which may be utilized to predict overall survival (OS) and serve as a promising therapeutic target for the disease (6,7). Hypoxia-inducible factor-1α, which is involved in tumoral adaptation to hypoxic conditions, can be a critical prognostic factor in metastatic ccRCC (8). The over-expression of cannabinoid receptor 2 is important in the cell cycle, cellular proliferation and migration of RCC cells, and predicts poor outcomes for patients with RCC (9). Previous studies have demonstrated that X-linked inhibitor of apoptosis protein is a potential prognostic marker in RCC, and that its downregulation may weaken immune resistance (10,11). Carbonic anhydrase 9 belongs to the carbonic anhydrase family, and its low expression is correlated with the poor prognosis of patients with ccRCC (12,13). However, these studies are insufficient and the prognostic mechanisms of ccRCC remain to be fully elucidated. In addition, the majority of the aforementioned results obtained from these microarray data were not validated by other datasets.

Bioinformatics analysis has been increasingly applied for revealing the genetic changes in the high-throughput data of tumors (14,15). In the present study, comprehensive bioinformatics analyses for gene expression data were performed to construct a prognostic prediction system for ccRCC with specific signature genes. Additionally, these predicted genes were validated by another dataset. The aim of the present study

Correspondence to: Dr Feng Liu, Department of Urology, China-Japan Union Hospital of Jilin University, 126 Xiantai Street, Erdao, Changchun, Jilin 130033, P.R. China

E-mail: lfdctch@sina.com

Key words: clear cell renal cell carcinoma, co-expression network, module analysis, signature genes, prognosis
was to provide additional, and reliable, prognostic markers for patients with ccRCC, and shed light on the molecular mechanisms of disease progression.

Materials and methods

Data source and data preprocessing. From The Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov/) database, the mRNA sequencing data of ccRCC, which was sequenced on the Illumina HiSeq 2000 RNA Sequencing platform, were downloaded with relevant clinical information on December 18th, 2016. There were a total of 606 samples, and 605 of these had information on sample source (normal or tumor tissues), including 533 ccRCC and 72 normal samples. The expression profile dataset GSE40435 (GPL10558 platform) in the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) database was also obtained, which included 101 ccRCC and 101 normal samples. According to the annotation platform, probes in the raw data of GSE40435 were converted into gene symbols. If multiple probes matched one gene, the average expression value of the gene was obtained. Subsequently, the expression data underwent logarithmic transformation using the R package Limma (http://www.bioconductor.org/packages/release/bioc/html/limma.html) (16), to reach an approximately normal distribution from the skewed distribution. Thereafter, the median normalization method (17) was utilized to normalize the data.

Analysis of differentially expressed genes (DEGs). The TCGA data was combined with the 19,004 protein coding gene annotation information in the HUGO Gene Nomenclature Committee (HGNc) database (18), and mRNAs in the TCGA dataset were identified. The differential expression analysis of genes or mRNAs was then performed between ccRCC and normal samples in the TCGA dataset and GSE40435 dataset separately, using the R package Limma (16). Subsequently, the false discovery rate (FDR) values were calculated by the R package multtest (http://www.bioconductor.org/packages/release/bioc/html/multtest.html) (19). An FDR <0.05 and log2 fold change (FC)>0.585 were the thresholds for selection of the DEGs. In addition, the overlapped DEGs in the TCGA and GSE40435 datasets were identified for the following analyses.

Screening of prognosis-associated genes. From the TCGA dataset, 596 samples (525 ccRCC samples and 71 normal samples) that matched with survival information (survival time and survival status) were identified. Subsequently, the 525 ccRCC samples were used for screening prognosis-associated genes. Cox regression analysis in the survival package (20) was applied for selecting prognosis-associated genes, and the log-rank test (21) was then utilized to calculate significant p-values. The top six genes were selected according to -logRank (P-value), following which Kaplan-Meier (KM) survival curves (22) were produced for the six genes.

Co-expression network analysis for significant prognostic genes and functional annotation. The expression values were extracted from TCGA database for the prognosis-associated genes, following which the correlation coefficients between the expression values of two genes were calculated using the COR function (23) in R. P<0.05 and a correlation coefficient |r|≥0.6 were used for selecting gene co-expression interactions. Subsequently, the co-expression network was visualized using Cytoscape software 3.4.0 (http://www.cytoscape.org/) (24).

Using Fisher’s exact test method in the R package clusterProfiler (http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) (25), the significant Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were enriched for the gene nodes in the co-expression network. In addition, the transcription factors (TFs) significantly associated with the nodes in the co-expression network were searched using The Database for Annotation, Visualization and Integrated Discovery (DAVID; https://david.ncifcrf.gov/) (26). Additionally, sub modules in the co-expression network were examined using the GraphWeb tool (http://biit.cs.ut.ee/graphweb/) (27).

Construction of the prognostic prediction system. The 525 ccRCC samples with survival information in the TCGA dataset were considered as the training dataset for the prognostic prediction system. First, the above samples were classified into two groups based on their survival status: Alive and deceased groups. Subsequently, combining survival status with survival times, the samples were further divided into good prognosis (alive and survival time ≥15 months) and bad prognosis (succumbed to mortality and survival time <15 months) groups. As the median OS time was ~15 months for all the ccRCC samples, the cut-off value for grouping was set as 15 months. The prior probability based on the Bayesian approach was then determined. The nodes of the co-expression network were ranked (-logRank P-values from largest to smallest). Using the discriminant Bayes function in the e1071 package of R (https://cran.r-project.org/web/packages/e1071/index.html) (28), Bayes discriminant analysis was performed for the network nodes through adding genes one by one and removing the genes that influenced prediction accuracy.

The prognostic score was defined as the discriminant coefficient of each sample when the prediction accuracy was the highest, and the gene set was considered as signature genes. The prediction system under the highest prediction accuracy was defined as the prognostic prediction system.

Validation of the prognostic prediction system. To examine the predictive effect of the constructed prediction system, KM survival analysis (22) was performed for the TCGA dataset to assess the association between the classification results of the prognostic prediction system and the real survival time and survival status. In addition, GSE29609 in the GEO database was obtained as the validation dataset, which included 596 samples (525 ccRCC samples and 71 normal samples) that matched with survival information (survival time and survival status). In addition, GSE29609 in the GEO database was obtained as the validation dataset, which included 596 samples (525 ccRCC samples and 71 normal samples). The expression profile dataset GSE40435 was used for screening prognosis-associated genes. Subsequently, the co-expression network was visualized using Cytoscape software 3.4.0 (http://www.cytoscape.org/) (24).

Using Fisher’s exact test method in the R package clusterProfiler (http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) (25), the significant Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were enriched for the gene nodes in the co-expression network. In addition, the transcription factors (TFs) significantly associated with the nodes in the co-expression network were searched using The Database for Annotation, Visualization and Integrated Discovery (DAVID; https://david.ncifcrf.gov/) (26). Additionally, sub modules in the co-expression network were examined using the GraphWeb tool (http://biit.cs.ut.ee/graphweb/) (27).

Construction of the prognostic prediction system. The 525 ccRCC samples with survival information in the TCGA dataset were considered as the training dataset for the prognostic prediction system. First, the above samples were classified into two groups based on their survival status: Alive and deceased groups. Subsequently, combining survival status with survival times, the samples were further divided into good prognosis (alive and survival time ≥15 months) and bad prognosis (succumbed to mortality and survival time <15 months) groups. As the median OS time was ~15 months for all the ccRCC samples, the cut-off value for grouping was set as 15 months. The prior probability based on the Bayesian approach was then determined. The nodes of the co-expression network were ranked (-logRank P-values from largest to smallest). Using the discriminant Bayes function in the e1071 package of R (https://cran.r-project.org/web/packages/e1071/index.html) (28), Bayes discriminant analysis was performed for the network nodes through adding genes one by one and removing the genes that influenced prediction accuracy.

The prognostic score was defined as the discriminant coefficient of each sample when the prediction accuracy was the highest, and the gene set was considered as signature genes. The prediction system under the highest prediction accuracy was defined as the prognostic prediction system.

Validation of the prognostic prediction system. To examine the predictive effect of the constructed prediction system, KM survival analysis (22) was performed for the TCGA dataset to assess the association between the classification results of the prognostic prediction system and the real survival time and survival status. In addition, GSE29609 in the GEO database was obtained as the validation dataset, which included 596 samples (525 ccRCC samples and 71 normal samples). The expression profile dataset GSE40435 was used for screening prognosis-associated genes. Subsequently, the co-expression network was visualized using Cytoscape software 3.4.0 (http://www.cytoscape.org/) (24).

Using Fisher’s exact test method in the R package clusterProfiler (http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) (25), the significant Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were enriched for the gene nodes in the co-expression network. In addition, the transcription factors (TFs) significantly associated with the nodes in the co-expression network were searched using The Database for Annotation, Visualization and Integrated Discovery (DAVID; https://david.ncifcrf.gov/) (26). Additionally, sub modules in the co-expression network were examined using the GraphWeb tool (http://biit.cs.ut.ee/graphweb/) (27).

Construction of the prognostic prediction system. The 525 ccRCC samples with survival information in the TCGA dataset were considered as the training dataset for the prognostic prediction system. First, the above samples were classified into two groups based on their survival status: Alive and deceased groups. Subsequently, combining survival status with survival times, the samples were further divided into good prognosis (alive and survival time ≥15 months) and bad prognosis (succumbed to mortality and survival time <15 months) groups. As the median OS time was ~15 months for all the ccRCC samples, the cut-off value for grouping was set as 15 months. The prior probability based on the Bayesian approach was then determined. The nodes of the co-expression network were ranked (-logRank P-values from largest to smallest). Using the discriminant Bayes function in the e1071 package of R (https://cran.r-project.org/web/packages/e1071/index.html) (28), Bayes discriminant analysis was performed for the network nodes through adding genes one by one and removing the genes that influenced prediction accuracy.

The prognostic score was defined as the discriminant coefficient of each sample when the prediction accuracy was the highest, and the gene set was considered as signature genes. The prediction system under the highest prediction accuracy was defined as the prognostic prediction system.
Table I. Information of the 44 signature genes in The Cancer Genome Atlas dataset.

<table>
<thead>
<tr>
<th>Gene</th>
<th>logFC</th>
<th>P-value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C10orf99</td>
<td>-2.344</td>
<td>6.88x10^{-44}</td>
<td>3.23x10^{-41}</td>
</tr>
<tr>
<td>ADAMDEC1</td>
<td>-1.896</td>
<td>1.81x10^{-55}</td>
<td>6.03x10^{-53}</td>
</tr>
<tr>
<td>H3SST2</td>
<td>-1.641</td>
<td>2.44x10^{-48}</td>
<td>5.72x10^{-46}</td>
</tr>
<tr>
<td>IL20RB</td>
<td>-1.514</td>
<td>3.55x10^{-37}</td>
<td>7.49x10^{-35}</td>
</tr>
<tr>
<td>SLC17A2</td>
<td>-1.182</td>
<td>7.10x10^{-28}</td>
<td>5.23x10^{-26}</td>
</tr>
<tr>
<td>LAG3</td>
<td>-0.863</td>
<td>6.59x10^{-22}</td>
<td>3.17x10^{-20}</td>
</tr>
<tr>
<td>SPAG4</td>
<td>-0.856</td>
<td>1.76x10^{-29}</td>
<td>1.43x10^{-27}</td>
</tr>
<tr>
<td>CD27</td>
<td>-0.820</td>
<td>2.44x10^{-22}</td>
<td>1.24x10^{-20}</td>
</tr>
<tr>
<td>SH2D1A</td>
<td>-0.804</td>
<td>1.55x10^{-18}</td>
<td>5.34x10^{-17}</td>
</tr>
<tr>
<td>ST2</td>
<td>-0.728</td>
<td>7.73x10^{-32}</td>
<td>7.47x10^{-30}</td>
</tr>
<tr>
<td>MMP9</td>
<td>-0.797</td>
<td>5.24x10^{-22}</td>
<td>2.54x10^{-20}</td>
</tr>
<tr>
<td>ANGPTL4</td>
<td>-0.786</td>
<td>5.94x10^{-37}</td>
<td>7.60x10^{-35}</td>
</tr>
<tr>
<td>OSCAR</td>
<td>-0.753</td>
<td>2.63x10^{-17}</td>
<td>7.89x10^{-16}</td>
</tr>
<tr>
<td>HHLA2</td>
<td>-0.742</td>
<td>1.70x10^{-24}</td>
<td>9.79x10^{-23}</td>
</tr>
<tr>
<td>RASD2</td>
<td>-0.734</td>
<td>6.34x10^{-20}</td>
<td>2.43x10^{-18}</td>
</tr>
<tr>
<td>NKG7</td>
<td>-0.728</td>
<td>2.15x10^{-21}</td>
<td>9.80x10^{-20}</td>
</tr>
<tr>
<td>INHBB</td>
<td>-0.723</td>
<td>1.73x10^{-24}</td>
<td>9.92x10^{-23}</td>
</tr>
<tr>
<td>CD96</td>
<td>-0.718</td>
<td>1.58x10^{-18}</td>
<td>5.42x10^{-17}</td>
</tr>
<tr>
<td>NOD2</td>
<td>-0.687</td>
<td>2.47x10^{-14}</td>
<td>5.39x10^{-13}</td>
</tr>
<tr>
<td>P2RX7</td>
<td>-0.674</td>
<td>1.82x10^{-16}</td>
<td>5.02x10^{-15}</td>
</tr>
<tr>
<td>PGF</td>
<td>-0.655</td>
<td>7.44x10^{-22}</td>
<td>3.54x10^{-20}</td>
</tr>
<tr>
<td>RASAL3</td>
<td>-0.627</td>
<td>5.19x10^{-16}</td>
<td>1.36x10^{-14}</td>
</tr>
<tr>
<td>ST8SIA4</td>
<td>-0.627</td>
<td>2.64x10^{-19}</td>
<td>9.66x10^{-18}</td>
</tr>
<tr>
<td>CDC7</td>
<td>-0.624</td>
<td>1.16x10^{-11}</td>
<td>1.96x10^{-10}</td>
</tr>
<tr>
<td>CHSY3</td>
<td>-0.624</td>
<td>3.61x10^{-13}</td>
<td>7.12x10^{-12}</td>
</tr>
<tr>
<td>CXCL10</td>
<td>-0.622</td>
<td>4.34x10^{-17}</td>
<td>1.27x10^{-15}</td>
</tr>
<tr>
<td>CD6</td>
<td>-0.622</td>
<td>1.38x10^{-14}</td>
<td>3.09x10^{-13}</td>
</tr>
<tr>
<td>RUNX3</td>
<td>-0.615</td>
<td>1.35x10^{-16}</td>
<td>3.79x10^{-15}</td>
</tr>
<tr>
<td>SLC1A3</td>
<td>-0.609</td>
<td>6.02x10^{-16}</td>
<td>1.56x10^{-14}</td>
</tr>
<tr>
<td>CXCL9</td>
<td>-0.609</td>
<td>2.13x10^{-18}</td>
<td>7.18x10^{-17}</td>
</tr>
<tr>
<td>TRIM9</td>
<td>-0.605</td>
<td>5.80x10^{-15}</td>
<td>1.36x10^{-13}</td>
</tr>
<tr>
<td>SEMA5B</td>
<td>-0.599</td>
<td>3.69x10^{-20}</td>
<td>1.47x10^{-18}</td>
</tr>
<tr>
<td>PDGFRB</td>
<td>0.605</td>
<td>9.85x10^{-21}</td>
<td>4.11x10^{-19}</td>
</tr>
<tr>
<td>ADH1B</td>
<td>0.667</td>
<td>8.24x10^{-26}</td>
<td>5.35x10^{-24}</td>
</tr>
<tr>
<td>G6PC</td>
<td>0.693</td>
<td>3.07x10^{-21}</td>
<td>1.37x10^{-19}</td>
</tr>
<tr>
<td>KRIT7</td>
<td>0.718</td>
<td>3.93x10^{-29}</td>
<td>3.09x10^{-27}</td>
</tr>
<tr>
<td>TEMEM30B</td>
<td>0.754</td>
<td>1.11x10^{-29}</td>
<td>9.13x10^{-28}</td>
</tr>
<tr>
<td>FAM167A</td>
<td>0.911</td>
<td>8.99x10^{-37}</td>
<td>1.13x10^{-34}</td>
</tr>
<tr>
<td>TFFAP2A</td>
<td>0.948</td>
<td>4.13x10^{-36}</td>
<td>5.13x10^{-34}</td>
</tr>
<tr>
<td>SLC13A3</td>
<td>1.009</td>
<td>9.21x10^{-39}</td>
<td>2.29x10^{-36}</td>
</tr>
<tr>
<td>TEMEM45B</td>
<td>1.036</td>
<td>1.43x10^{-37}</td>
<td>3.12x10^{-35}</td>
</tr>
<tr>
<td>C1orf116</td>
<td>1.062</td>
<td>5.48x10^{-47}</td>
<td>1.14x10^{-44}</td>
</tr>
<tr>
<td>SCNN1B</td>
<td>1.404</td>
<td>2.16x10^{-45}</td>
<td>2.74x10^{-42}</td>
</tr>
<tr>
<td>ATP6V1B1</td>
<td>1.456</td>
<td>1.36x10^{-92}</td>
<td>1.91x10^{-89}</td>
</tr>
</tbody>
</table>

FC, fold change; FDR, false discovery rate.

the classification results of the prognostic prediction system and the actual survival time and survival status were detected by KM survival analysis (22).

Figure 1. Expression density distributions of the mRNAs prior to (solid line) and following (dotted line) filtering.

**Results**

**DEG analysis.** Based on the HGNC database, the expression values of 18,531 protein-coding mRNAs were obtained from the TCGA dataset. A total of 12,669 mRNAs remained following removal of the mRNAs with low-abundance expression. There were numerous mRNAs expressed at low levels prior to filtering, but the peak of expression density was markedly elevated following filtering (Fig. 1).

Through differential expression analysis, a total of 621 and 2,764 DEGs were identified in the TCGA dataset and GSE40435 dataset, respectively. Among them, 263 overlapping DEGs in both the TCGA dataset and GSE40435 dataset were selected.

According to the logFC values, the top 50 DEGs (top 25 upregulated and top 25 downregulated) in the TCGA dataset (Fig. 2A) and GSE40435 dataset (Fig. 2B), respectively, were selected and subjected to bidirectional hierarchical clustering analysis. The heatmap showed that it was possible to divide the samples into two groups by the above DEGs.

**Screening of prognosis-associated genes.** A total of 161 prognosis-associated genes were identified using Cox regression analysis from the TCGA dataset. According to -logRank (P-value), the top six genes (transcription factor AP-2a (TFAP2A); family with sequence similarity 167, member A (FAM167A); interleukin 20 receptor b (IL20RB); leucine-rich repeat LGI family, member 4 (LGI4); sodium channel, nonvoltage-gated 1b (SCNN1B); and solute carrier family 17, member 2 (SLC17A2)] were selected. The KM survival curves were generated (Fig. 3). It was shown that low expression of these genes was linked to high survival time, indicating that these six genes may be used as predictors for prognosis.

**Co-expression network analysis and functional annotation.** Following the selection of gene co-expression interactions for the prognosis-associated genes, the co-expression network (involving 141 nodes and 1,937 edges) was constructed (Fig. 4).
In addition, four significant network modules were identified, including a red module, a yellow module, involving chemokine ligand 10 (CXCL10), CD27 molecule (CD27), and runt-related transcription factor 3 (RUNX3), a green module, involving angiopoietin-like 4 (ANGPTL4), stanniocalcin 2 (STC2), and sperm associated antigen 4 (SPAG4), and a cyan module. Nervous system development, immune system process, regulation of secretion, and NAD metabolic process were the most significantly enriched terms for genes in the red, yellow, green and cyan modules, respectively (Fig. 4).

By performing functional and pathway enrichment analysis, a total of 16 GO terms, including immune response, three KEGG pathways, including cytokine-cytokine receptor interaction, and 11 TFs, including nuclear factor-kB, were obtained for genes in the network nodes (Fig. 5A and B).

Construction and validation of prognostic prediction system. The 525 ccRCC samples in the TCGA dataset were divided into good prognosis (202 samples) and bad prognosis (323 samples) groups. Through a series of processes (Fig. 6A), the prognostic prediction system containing 44 signature genes, including ANGPTL4, STC2, CXCL10, SPAG4, CD27, matrix metallo-

peptidase 9 (MMP9), and RUNX3, were constructed (Table I). The 44-gene prediction system had the highest prognostic accuracy for the patients with ccRCC.

The prognostic scores of the samples varied between -1.5 and 1.5 (good prognosis group between -1.5 and 0; bad prognosis group between 0 and 1.5; Fig. 6B). The discriminant scoring system of the prognostic prediction system was as follows:

Prognostic score = \( \alpha_i \) (Bayes discriminant) = [0-1.5, bad; -1.5-0, good], where \( \alpha \) represents the prognostic score, and \( i \) means gene.

To assess the effect of the prediction system, KM survival analysis was performed in the TCGA dataset first. The result showed that the survival ratio of the good prognosis group was significantly higher than that of the bad prognosis group (P=7.008x10^{-15}; Fig. 6C).

In addition, in the GSE29609 validation dataset, the survival ratio of the good prognosis group was also significantly higher that of the bad prognosis group (P=8.46x10^{-8}; Fig. 6D). These results suggested that the prognostic prediction system was able to accurately and practically classify ccRCC samples according to their prognosis.
Discussion

Previously, several studies have utilized microarray analysis to identify prognostic genes for ccRCC. By genome-wide expression analyses of the expression profiles of patients with primary ccRCC with different disease-free survival rates, platelet and endothelial cell adhesion molecule 1, endothelin receptor type B and tetraspanin 7 have been considered as prognostic markers potentially involved in tumor metastases (29). Based on array-comparative genomic hybridization,
the genetic clustering of ccRCC was considered as a potential prognostic indicator in patients with RCC that is closely associated with DNA methylation alteration (30). However, these studies are insufficient for prognostic marker identification in ccRCC. In the present study, there were a total of 263 DEGs overlapped in the TCGA dataset and GSE40435 dataset, and 161 of these were associated with prognosis. Enrichment analysis showed that they were correlated with three KEGG pathways, including the cytokine-cytokine receptor interaction pathway. In addition, four significant network modules (red, yellow, green, and cyan modules) were identified from the co-expression network. Notably, ANGPTL4, STC2 and SPAG4 were present in the green module; whereas CXCL10, CD27 and RUNX3 were present in the yellow module. Through a series of bioinformatics methods, a prognostic prediction system was established comprising 44 signature genes, including ANGPTL4, STC2, CXCL10, SPAG4, CD27, MMP9, and RUNX3, and its prediction accuracy was confirmed. ANGPTL4, which is a member of the angiopoietin/ANGPTL family, has a high expression and can be used as a diagnostic marker in primary ccRCC (31,32). It is suggested that an increased serum level of ANGPTL4 may function as a promising diagnostic and prognostic marker for patients with RCC (33). The overexpression of STC2 is involved in the metastasis of RCC and may be an indicator for the shorter OS of patients with RCC (34,35). In addition, multilevel whole-genome analysis has revealed that STC2 is one of the genes hypomethylated in copy number gains in ccRCC (36). SPAG4 is upregulated in human RCC and has influences on the growth and invasion capability of tumor cells (37,38). The mRNA expression of SPAG4 is negatively correlated with tumor stage, grade and size, suggesting that SPAG4 can act as a marker for the diagnosis and prognosis of RCC (39). SPAG4 contributes to the survival of cancer cells via suppressing hypoxia-induced tetraploid formation, and thus SPAG4 can independently predict poor prognosis in RCC (40). In the present study, ANGPTL4, STC2 and SPAG4 were all involved in the green module. Therefore, ANGPTL4, STC2 and SPAG4 may function in ccRCC through their co-expression, making them the prognostic factors for ccRCC.

CXCL10 inhibits tumor growth in RCC via restraining angiogenesis and decreasing the expression levels of vascular endothelial growth factor, platelet derived growth factor, fibroblast growth factor, and MMP9 (41). In addition, the interferon-inducible CXCR3 ligands score based on expression levels of CXCL9, CXCL10 and CXCL11, is suggested to be linked with different risk subgroups of recurrence and mortality in patients with ccRCC (42). CD27+ lymphocyte infiltration and the overexpression of CD70 are correlated with poor prognosis in ccRCC, and an elevated serum level of CD27 may be used for anti-CD70 therapy by predicting CD70-expressing ccRCC (43,44). RUNX3 is closely associated with RCC progression, and its high expression can significantly suppress the migration, invasion and angiogenesis in RCC (45-47). RUNX3 inhibits RCC migration and invasion through mediating the microRNA-6780a-5p/E-cadherin/epithelial-mesenchymal transition signaling pathway, therefore, RUNX3 serves as a potential prognostic factor of RCC (48). In terms of the association of this gene with ccRCC, RUNX3 is decreased in ccRCC tissues, and it functions as an inhibitor of ccRCC cell growth and metastasis via regulating cyclins and tissue inhibitors of matrix metal-
Figure 5. Functional and pathway enrichment analysis. (A) Functional terms and pathways enriched for the network nodes. (B) Transcription factors targeting the network nodes.
loproteinase 1 (TIMP1) (49). In the present study, CXCL10, CD27 and RUNX3 were all involved in the yellow module, indicating that these co-expressed genes may also be associated with the prognosis of patients with ccRCC.

MMPs and their inhibitors (TIMPs) have an important function in the maintenance of extracellular matrix homeostasis. In RCC, the mRNA or protein expression of MMP2, MMP3, MMP9, TIMP1 and TIMP2 are relevant to the clinicopathological parameters (50). MMP9 correlates with high metastasis and poor survival rates in RCC, indicating that MMP9 may be utilized for predicting the disease-free survival rates of patients with RCC (51,52). The Notch ligand d-like 4 (DLL4) is tied up with tumor invasion and metastasis. It has been found that DLL4 facilitates RCC cell migration and invasion via upregulating MMP2 and MMP9 (53). It is known that, in the majority of ccRCC cases, inaction of the von Hippel-Lindau (VHL) tumor suppressor gene is an important hallmark, and the protein isoform of VHL coordinately regulates the metastasis-associated genes CXCR4/CXCL12 and MMP2/MMP9 (54). MMP9 has been selected as one of the 10 important genes in the protein-protein interaction network that associates with the progression of ccRCC (55). In the present study, this gene was one of the 44 signature genes predicting ccRCC prognosis, suggesting it may be used as a prognostic factor for patients with ccRCC.

Previously, a study reported that cytokine-cytokine receptor interaction was the most significant pathway for DEGs in RCC tissue (56). In the present study, this pathway was also significantly enriched for DEGs identified in ccRCC samples, suggesting that these crucial DEGs may function through the regulation of this pathway.

The predictive accuracy of the 44-signature gene-prognostic prediction system was confirmed by the validation dataset (GSE29609), indicating this system may be applied for the prognosis of patients with ccRCC. Although comprehensive bioinformatics analysis was performed, and hundreds of
samples were used in the present study, a limitation remained that the validation dataset had a relatively small sample size, and thus the results require experimental validation, particularly the co-expression of genes identified in the same module.

In conclusion, the 44-gene prognostic prediction system, involving ANGPTL4, STC2, CXCL10, SPAG4, CD27, MMP9, and RUNX3, may be important in predicting the prognosis of patients with ccRCC. However, these key genes and the 44-gene prognostic prediction system require further validation by experimental investigations.

Acknowledgements

Not applicable.

Funding

No funding was received.

Availability of data and materials

The datasets used during the current study are available from the corresponding author on reasonable request.

Authors' contributions

YoW and YaW were involved in the design of this study and performed the statistical analysis. YaW collected important background information. FL drafted the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References


This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.