Original Article
Identification and validation of novel metastasis-related signatures of clear cell renal cell carcinoma using gene expression databases

Chun-Guang Ma1,2, Wen-Hao Xu1,2*, Yue Xu3*, Jun Wang4*, Wang-Rui Liu5, Da-Long Cao1,2, Hong-Kai Wang1,2, Guo-Hai Shi1,2, Yi-Ping Zhu1,2, Yuan-Yuan Qu1,2, Hai-Liang Zhang1,2, Ding-Wei Ye1,2

1Department of Urology, Fudan University Shanghai Cancer Center, Shanghai 200032, P. R. China; 2Department of Oncology, Shanghai Medical College, Fudan University, Shanghai 200032, P. R. China; 3Department of Ophthalmology, The First Affiliated Hospital of Soochow University, Suzhou 215000, P. R. China; 4Department of Urology, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou 510060, Guangdong, P. R. China; 5Department of Neurosurgery, Affiliated Hospital of Youjiang Medical College for Nationalities, Guangxi, P. R. China. *Equal contributors.

Received February 4, 2020; Accepted July 4, 2020; Epub August 15, 2020; Published August 30, 2020

Abstract: Patients with clear cell renal cell carcinoma (ccRCC) typically face aggressive disease progression when metastasis occurs. Here, we screened and identified differentially expressed genes in three microarray datasets from the Gene Expression Omnibus database. We identified 112 differentially expressed genes with functional enrichment as candidate prognostic biomarkers. Lasso Cox regression suggested 10 significant oncogenic hub genes involved in earlier recurrence and poor prognosis of ccRCC. Receiver operating characteristic curves validated the specificity and sensitivity of the Cox regression penalty used to predict prognosis. The area under the curve indexes of the integrated genes scores were 0.758 and 0.772 for overall and disease-free survival, respectively. The prognostic values of ADAMTS9, C1S, DPYSL3, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2, and TRIB3 were validated through an analysis of 10 hub genes in 380 patients with ccRCC from a real-world cohort. The expression levels of were of high prognostic value for predicting metastatic potential. These findings will likely significantly contribute to our understanding of the underlying mechanisms of ccRCC, which will enhance efforts to optimize therapy.

Keywords: Clear cell renal cell carcinoma, metastasis, signature, bioinformatics, prognosis

Introduction
Renal cell carcinoma (RCC) is one of the most common malignancies of the urinary system, accounting for approximately 2% of cancer deaths [1]. The worldwide morbidity and mortality rates of RCC are increasing by approximately 2%-3% each decade [2]. Until recently, the rates stabilized or declined in many economically advanced countries [3]. As the most predominant histological subtype, clear cell renal cell carcinoma (ccRCC) represents approximately 70% of RCC cases. Although numerous factors contribute to the pathogenesis of RCC, limited information is available that can be applied to explain the aggressive pathogenesis and progression of ccRCC. The aggressiveness of ccRCC is directly associated with metastatic potential, which is typically not efficiently targeted despite great advances in therapeutic strategies [4]. Metastasis is a significant hallmark of tumor progression and thus the major cause of poor overall survival of patients with ccRCC [5]. Therefore, the identification of new prognostic biomarkers and molecular alterations is urgently required to develop more effective treatments of ccRCC.

Accumulating evidence has demonstrated that gene expression panels and related hallmarks participate in the carcinogenesis and aggressive progression of ccRCC [6-8]. For example, the expression levels of a panel of seven genes associated with the extracellular matrix significantly correlate with metastasis and prognosis of ccRCC [9]. Moreover, multigene panels help
predict signatures that detect ccRCC in biopsy specimens [10]. Nomograms employing integrated clinical and gene expression profiles predict pathological nuclear grades and help clinicians to manage personalized regimens [8]. Despite these encouraging advances in strategies to treat metastatic ccRCC, many patients do not achieve longer survival [11]. Therefore, we urgently require methods that identify the underlying biochemical mechanisms associated with metastasis and predict the prognosis of patients with ccRCC. Such methods will facilitate the development of optimal therapeutic strategies.

During the latest decade, high-throughput nucleotide sequencing technologies, combined with bioinformatics analysis, sensitively and specifically detect mRNA expression levels, which identify differentially expressed genes (DEGs) that represent hallmarks associated with the pathogenesis and progression of ccRCC [8, 9, 12]. However, the diversity of genomic alterations and molecular interactions associated with metastasis hinders the identification of patients at high-risk for metastasis who may benefit from available as well as potential therapies. To address these challenges, here we analyzed three transcriptional microarray datasets acquired from the Gene Expression Omnibus (GEO) database to identify DEGs between cancer tissues and adjacent tissues that will serve as biomarkers of ccRCC. We conducted functional pathway enrichment analyses to better understand the associated molecular mechanisms. Moreover, we employed protein-protein interaction (PPI) network analysis to better define the importance of these interactions as they relate to biological processes, molecular functions, and signal transduction [13-15].

Our findings led us to hypothesize that the oncogenicity of significant hub genes correlates with metastasis and that these genes may serve as potential prognostic factors that will facilitate the identification of therapeutic targets for managing ccRCC.

Materials and methods

Raw biological microarray data

Raw transcriptional microarray data from GEO (http://www.ncbi.nlm.nih.gov/geo) [16] were screened for metastatic or primary ccRCC. Corresponding genes were converted into probes and assigned symbols associated with annotation information. We analyzed the datasets GSE22541 (24 primary and 44 metastatic tumors), GSE47352 (4 primary and 4 metastatic tumors), and GSE85258 (14 primary and 14 metastatic tumors) (Affymetrix Human Genome U133 Plus 2.0 Array).

Data normalization and identification of DEGs

Preprocessing and normalization of raw biological data were the first steps in processing the DNA microarrays. This process removes biased microarray data to ensure its uniformity and integrity. Subsequently, background correction, propensity analysis, normalization, and visualization of probe data were performed using the robust multiarray average analysis algorithm 17 in the affy package of R.

DEGs between primary and metastatic ccRCC samples were screened and identified across experimental conditions. Delineating parameters such as adjusted $P$ values (adj. $P$), the Benjamini and Hochberg false-discovery rate (FDR), and fold-change values were used for filtering DEGs and balancing between discovery of significant genes and limitations of false-positives. Probe sets without corresponding gene symbols, or genes with more than one overlapping probe-set, were removed or averaged. Log$_2$FC (fold-change) >1 and adj. $P$ value <0.01 were considered to indicate a significant difference.

Protein-protein interaction (PPI) network and functional annotations

We used Search Tool for the Retrieval of Interacting Genes (http://string-db.org) (version 10.0) to predict a PPI network of DEGs and to analyze the degree of interactions between proteins [17]. A significant edge score was considered as an interaction combined score >0.4. Cytoscape (version 3.5) was used to visualize interactive network data [18].

To identify the role of DEGs in ccRCC, we used Gene Ontology (GO) enrichment analysis to extract functional attributes including biological processes (BP), molecular functions (MF), and cellular component (CC) [19]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used for this purpose as well [20]. The Database for Annotation, Visualization
and Integrated Discovery (DAVID) (http://david.ncifcrf.gov; Version 6.8) was analyzed to explore the role of development-related signaling pathways in ccRCC [21]. P<0.05 was considered to indicate a significant.

ClueGO is a Cytoscape plug-in that visualizes nonredundant biological terms associated with large clusters of genes in a functionally grouped network [22]. The results of KEGG pathway analysis of selected hub genes enriched in the GO term BP were visualized using ClueGO (version 2.5.3) and the Cytoscape plugin CluePedia (version 1.5.3). Potential associations of 24 coordinately expressed hub genes and their potential prognostic values are shown using a heat map. The hub nodes of networks with connectivity degrees >10 were identified. A network comprising 24 genes and their coordinately expressed neighboring genes was constructed using cBioPortal (http://www.cbioportal.org) [23].

Statistical analysis of TCGA data cohort

A lasso Cox regression model was used to identify independent prognostic factors. Functional annotations and enrichment of signaling pathways were predicted and illustrated using DAVID. Phenotype and transcriptomics data of selected hub genes of 515 patients with ccRCC in TCGA data are displayed. Expression profiles were respectively identified as binary variables (high vs low), referring to the median expression level of each hub gene in TCGA cohort data. The primary and secondary endpoints for patients were disease-free survival (DFS) and overall survival (OS), respectively. The duration of follow-up was estimated using the Kaplan-Meier method with 95% confidence intervals (95% CI) and the log-rank test. X-tile software was used to determine the cutoff value [24]. All hypothetical tests were two-sided, and P values <0.05 were considered to indicate a significant difference for all tests. Hierarchical partitioning was performed using transcriptional expression profiles of selected oncogenes in a heat map. Color gradients indicate a high (red) or a low (blue) expression level.

The human protein atlas

The Human Protein Atlas project (https://www.proteinatlas.org) contains immunohistochemistry (IHC) data acquired using a tissue microarray analysis of 17 paired, major cancer types [25]. Data include IHC staining intensity, quantity, location, and patients' information. Here we used the Human Protein Atlas (https://www.proteinatlas.org) to identify representative proteins differentially expressed between ccRCC and corresponding normal tissues.

ccRCC patients from a validated cohort

We analyzed samples acquired from 380 ccRCC patients who underwent nephrectomy at the Department of Urology of Fudan University Shanghai Cancer Center (FUSCC; Shanghai, China) from June 2009 to September 2017. Tissue samples were obtained during surgery, and available from the FUSCC tissue bank.

Quantitative real-time PCR (RT-qPCR) analysis

Total cellular RNA, which was isolated using Trizol (Invitrogen, Carlsbad, CA) in accordance with the source's protocols, was reversed-transcribed using a PrimeScript RT reagent kit (Thermo Fisher, USA). Primers were diluted and mixed in RNase-Free ddH$_2$O, and RT-qPCR was performed using the SYBR Green qPCR method (Takara Biotechnology Co.). The levels of GAPDH mRNA were measured to serve as a standard. Specific amplification conditions were performed using the SYBR Green qPCR Master Mix (Applied Biosystems) according to the manufacturer's protocols. Primes sequences were shown in Table 2. The relative expression level of the target mRNA was calculated using the $2^{-ΔΔCt}$ method.

Statistical analysis of FUSCC-cohort data

We evaluated the significance of the associations of DFS and OS with distinct mRNA expression groups of 10 hub genes expressed by patients with ccRCC. The duration of follow-up was estimated using the Kaplan-Meier method (95% CI) and the log-rank test. Integrated scores represent the sums of each significant oncogenic hub-gene weight. Receiver operating characteristic (ROC) curves were generated to validate the specificity and sensitivity of a diagnosis according to high or low integrated scores of significant hub-gene levels. Area under the curve (AUC) analysis was performed to determine diagnostic ability.

Gene set enrichment analysis (GSEA) and related gene networks

TCGA data were subjected to GSEA using Bioconductor category version 2.10.1 package.
# Table 1. Functional roles of 10 prognostic hub genes

<table>
<thead>
<tr>
<th>No.</th>
<th>Gene symbol</th>
<th>Full name</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ADAMTS9</td>
<td>A Disintegrin And Metalloproteinase With Thrombospondin Motifs 9</td>
<td>Pathway: Metabolism of proteins and O-linked glycosylation. GO: Metalloendopeptidase activity and endopeptidase activity.</td>
</tr>
<tr>
<td>2</td>
<td>C1S</td>
<td>Complement C1s</td>
<td>Pathway: Immune response Lectin induced complement pathway and Creation of C4 and C2 activators. GO: Calcium ion binding and serine-type endopeptidase activity.</td>
</tr>
<tr>
<td>3</td>
<td>DYSL3</td>
<td>Dihydropyrimidinase Like 3</td>
<td>Pathway: Semaphorin interactions and Developmental Biology. GO: Hydrolase activity and phosphoprotein binding.</td>
</tr>
<tr>
<td>4</td>
<td>H2AFX</td>
<td>H2A Histone Family Member X</td>
<td>Pathway: Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3 and ATM Pathway. GO: Sequence-specific DNA binding and enzyme binding.</td>
</tr>
<tr>
<td>5</td>
<td>MINA</td>
<td>MYC-Induced Nuclear Antigen</td>
<td>Pathway: Validated targets of C-MYC transcriptional activation and Chromatin organization.</td>
</tr>
<tr>
<td>6</td>
<td>PLOD2</td>
<td>Procollagen-Lysine, 2-Oxoglutarate 5-Dioxygenase 2</td>
<td>Pathway: Collagen chain trimerization and Lysine degradation. GO: Oxidoreductase activity and oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen.</td>
</tr>
<tr>
<td>7</td>
<td>RUNX1</td>
<td>Runt Related Transcription Factor 1</td>
<td>Pathway: Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds and Embryonic and Induced Pluripotent Stem Cell Differentiation Pathways and Lineage-specific Markers. GO: DNA-binding transcription factor activity and protein homodimerization activity.</td>
</tr>
<tr>
<td>8</td>
<td>SLC19A1</td>
<td>Solute Carrier Family 19 Member 1</td>
<td>Pathway: Metabolism of water-soluble vitamins and cofactors and Antifolate resistance. GO: Oxidoreductase activity and folic acid transmembrane transporter activity.</td>
</tr>
<tr>
<td>9</td>
<td>TPX2</td>
<td>Targeting Protein For Xklp2</td>
<td>Pathway: Gene Expression and Cell Cycle, Mitotic. GO: GTP binding and protein kinase binding.</td>
</tr>
<tr>
<td>10</td>
<td>TRIB3</td>
<td>Tribbles Pseudokinase 3</td>
<td>Pathway: Class I MHC mediated antigen processing and presentation and RET signaling. GO: Transferase activity, transferring phosphorus-containing groups and protein kinase binding.</td>
</tr>
</tbody>
</table>
Metastasis-related signatures in ccRCC

For each analysis, the Student t test was performed, and the mean levels of differentially expressed genes were calculated. A permutation test (1000 times) was used to identify significantly changed pathways. The default adj. P values determined using the BH-FDR method were applied to correct for false-positives [26]. Significantly related genes were defined as those with adj. P < 0.01 and FDR > 0.25. Statistical analyses and graphical presentation were conducted using R software (Version 3.3.2). A detailed PPI network associated with a default set of 10 hub genes was constructed using GeneMANIA (http://genemania.org/).

Results

We first assessed DEGs using three datasets hosted on the GEO platform to identify coordinately-regulated hub genes. Second, we evaluated the significance of the associations between the expression levels of hub genes and prognosis according to TCGA data and then validated their prognostic value using a real-world cohort. Third, we predicted potential functional annotations and hallmark pathways.

Identification of DEGs in ccRCC

After normalization and identification of mRNA microarray data, the DEGs (12,817; 1,817; and 1,766 probe samples in GSE22541, GSE85258, and GSE47352, respectively) were determined as significant according to the analytical and statistical parameters of the data processing steps. The overlap among the three datasets included 112 significant genes (Figure 1A).

Table 2. Primer sequences (5’-3’) of 10 prognostic hub genes for qRT-PCR

<table>
<thead>
<tr>
<th>No.</th>
<th>Gene symbol</th>
<th>Forward</th>
<th>Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ADAMTS9</td>
<td>CAGAAAGGGCTTGGTGGG</td>
<td>TCCTTCATTACCACTTTGTA</td>
</tr>
<tr>
<td>2</td>
<td>C1S</td>
<td>GTTGTACATGACGTTGAGA</td>
<td>GGCTAAATCACCCTGGAAG</td>
</tr>
<tr>
<td>3</td>
<td>DPYS1</td>
<td>GGATCAAGGTGAGCCGCTT</td>
<td>TGTCATGATCAGGGCGCATG</td>
</tr>
<tr>
<td>4</td>
<td>H2AFX</td>
<td>CGCGACTGGTGAAGGCCTCA</td>
<td>AGCTCCCTGTCGGCAGGATG</td>
</tr>
<tr>
<td>5</td>
<td>MINA</td>
<td>CCAAGAAGCTGTTCCTCAGAC</td>
<td>CTGACTGTCAGGCTCCGGATT</td>
</tr>
<tr>
<td>6</td>
<td>PLD2</td>
<td>CAGAGCAGAGGATAATGCTG</td>
<td>AGGGTTGTGTGATCATAAAAA</td>
</tr>
<tr>
<td>7</td>
<td>RUNX1</td>
<td>GCGAGGAAACGGCCCTACCT</td>
<td>GCTATGCTGCAGCGAGGAAAG</td>
</tr>
<tr>
<td>8</td>
<td>SLC19A1</td>
<td>TTGCCCAAGCTATTCAGTCGA</td>
<td>CAGAGACGCCGCCAGGCAAT</td>
</tr>
<tr>
<td>9</td>
<td>TPX2</td>
<td>ACCTGCCGCTAAGATT</td>
<td>AATGTGTCAGGTGAC</td>
</tr>
<tr>
<td>10</td>
<td>TRIB3</td>
<td>ATGCCGCTGGATTTATC</td>
<td>TGCCCTGAAAAAGCCTCC</td>
</tr>
<tr>
<td>11</td>
<td>GAPDH</td>
<td>GTCTCTGACCAGTGGAAGG</td>
<td>CATGCCGCTGAGCTCCCGTGTA</td>
</tr>
</tbody>
</table>

Construction of a PPI network and functional analysis of modules

The PPI network comprising the DEGs is shown in Figure 1B, and the network of the DEGs and their coordinately expressed neighboring genes are shown in Figure 1C. The enrichment profiles determined using DAVID-GO functional analyses suggested that the 112 DEGs in this module were primarily enriched in GO terms such as double-strand break, positive response to stimulus, sarcomere, external side of plasma membrane, metal ion binding, cation binding, ion binding, and cytoskeletal protein binding (Figure 1D).

To further identify underlying signaling pathways, we analyzed KEGG pathways together with GO functional annotations (Figure 2). Detailed functional notes and classification pie charts are shown in Supplementary Figure 1. The frequencies of GO terms determined using ClueGO analysis were as follows: 25.0%, negative regulation of protein complex disassembly; 12.5%, negative regulation of response to oxidative stress; 9.38%, cell-cell junction assembly; 6.25%, maintenance of protein localization in cell; 6.25%, photoreceptor cell differentiation; 6.25%, inorganic cation import across plasma membrane; and 6.25%, lysosome organization.

Selection of significant hub genes

Lasso Cox regression analysis suggested that C1S, PLOD2, ADAMTS9, AK1, CHD2, DCAF8, DNAJC10, DPP3, DPYS1, GSTM3, H2AFX, HBA1, RUNX1, MAMDC4, MFAP3L, MINA, OSBPL1A, PIAS2, RBBP6, RCOR3, RHOBTB3,
Metastasis-related signatures in ccRCC

A

B

GSE22541
10825
GSE47362
879

GSE85258
476

1115
112
63

1115
112
63

Figure 1. Screening and selection of DEGs using multiple GEO database. A. After normalization and identification of mRNA microarray data, the DEGs (12,817; 1,817; and 1,766 probe samples in GSE22541, GSE85258, and GSE47352, respectively) were determined as significant according to the analytical and statistical parameters of the data processing steps. The overlap among the three datasets included 112 significant genes. B. The PPI network of the DEGs was constructed using Cytoscape. MCODE, a plug-in of Cytoscape, predicts the most significant gene panel, marked in light green. C. DEGs and their co-regulated network were analyzed using cBioPortal. Nodes with bold black outline represent hub genes. Nodes with thin black outline represent the co-expression genes. D. The enrichment profiles from DAVID GO functional analyses of the 112 hub genes suggested that the hub genes in this module were primarily enriched in double-strand break, positive response to stimulus, sarcomere, external side of plasma membrane, metal ion biding, cation binding, ion binding, cytoskeletal protein binding and so one.
SLC19A1, SORBS2, SPTAN1, SPTBN1, SYT13, TPM4, TPX2, TRIB3, UGGT1 were significant weighted prognostic factors. Functional annotation of these genes revealed they were primarily enriched in the terms as follows: regulation of actin filament capping, non-membrane-bounded organelle, sarcomere, myofibril, organelle lumen, nucleolus, actin cytoskeleton, and nuclear lumen (Figure 3).

Kaplan-Meier analysis indicated that differentially elevated levels of the hub genes ADAMTS9, C1S, DPYSL3, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2, and TRIB3 significantly correlated with poor prognosis of 513 patients with ccRCC (Figure 4A-T). Functional annotations of each prognostic hub gene are listed in Table 1. Hierarchical partitioning was performed using the transcriptional expression profiles of selected 10 oncogenes (Figure 4U).

External validation of hub DEGs in 380 ccRCC patients from FUSCC cohort

DEGs corresponding to ADAMTS9, C1S, DPYSL3, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2, and TRIB3 mRNAs were identified using the TCGA database and GSE11151 [27] (Figure...
Metastasis-related signatures in ccRCC

5). Significantly increased levels of these mRNAs were expressed by both cohorts. In contrast, MINA mRNA was differentially expressed at lower levels. The levels of the cognate proteins expressed by these hub genes in tumor and adjacent tissues detected using IHC and reported by the Human Protein Atlas are shown in Supplementary Figure 2.

A total of 380 ccRCC patients with available clinical and pathological data were enrolled in this study (Table 3). After integration of qRT-PCR and clinical follow-up data for 380 patients with ccRCC, we confirmed that ADAMTS9, C1S, DPYSL3, H2AFX, PLOD2, RUNX1, SLC19A1, TPX2, and TRIB3 mRNAs were differentially expressed in the FUSCC cohort. Differentially elevated mRNA levels were significantly associated with shorter DFS and OS (P<0.05) (Figure 6).

An integrated gene panel was constructed using 10 hub gene (ADAMTS9, C1S, DPYSL3, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2, TRIB3, UGGT1) are significant weighted prognostic factors. C. GO function annotations of 31 selected genes. Functional annotation of these genes revealed they were primarily enriched in the terms as follows: regulation of actin filament capping, non-membrane-bounded organelle, sarcomere, myofibril, organelle lumen, nucleolus, actin cytoskeleton, and nuclear lumen.

Figure 3. Further identification of hub genes using LASSO regression analysis. A, B. To further select significant prognostic metastasis-related biomarkers, LASSO Cox regression was performed to further search for hub genes associated with metastasis and to narrow the scope of the panel, suggested that a total of 31 genes, including C1S, PLOD2, ADAMTS9, AK1, CHD2, DCAF8, DNAJC10, DPP3, DPYSL3, GSTM3, H2AFX, HBA1, RUNX1, MAMDC4, MFAP3L, MINA, OSBPL1A, PIAS2, RBBP6, RCOR3, RHOB1T3, SLC19A1, SORBS2, SPTAN1, SPTBN1, SYT13, TPM4, TPX2, TRIB3, UGGT1 are significant weighted prognostic factors. C. GO function annotations of 31 selected genes.
Metastasis-related signatures in ccRCC
Figure 4. Prognostic value of significant hub genes from TCGA cohort. A-T. Kaplan-Meier analysis indicated that differentially elevated levels of the hub genes ADAMTS9, C1S, DPYSL3, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2, and TRIB3 significantly correlated with poor prognosis of 513 patients with ccRCC. U. Hierarchical partitioning was performed using transcriptional expression profiles of 10 selected oncogenes.
Metastasis-related signatures in ccRCC

**TRIB3**, which may serve as an independent panel to predict the progression of ccRCC. Kaplan-Meier analysis revealed significant associations of mRNA levels with DFS ($P<0.0001$) and OS ($P<0.0001$) (Figure 7A, 7B). ROC curve analysis indicated the ability of the gene model to predict metastasis (AUCs of the integrated model were 0.758 for OS and 0.772 for DFS) (Figure 7C).

Significantly associated genes and their signaling pathways

GSEA identified 100 DEGs with positive or negative correlations. Among them, we used the hallmarks **ADAMTS9, C15, DPYS13, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2, TRIB3** to perform GSEA. Figure 8 shows the three most relevant hallmark pathways of each
hub gene. The most significant genes associated with 10 hub genes, and their relationships are shown in Figure 9A, and a detailed PPI network related to this gene set shows additional interrelated nodes at the protein level (Figure 9B).

Discussion

Genetic and epigenetic alterations contribute to the development and progression of RCC [28]. ccRCC, which is the most aggressive histological subtype of ccRCC, is associated with elevated mortality owing to its high metastatic potential [29]. Unfortunately, the underlying mechanisms of oncogenesis and metastasis of ccRCC are unknown. We reasoned that microarray technology represents a powerful tool to fill this gap in our knowledge, because it enables comprehensive mRNA expression profiling of ccRCC through its ability to identify and characterize new biomarkers involved in tumorigenesis and progression [30, 31]. Here we conducted an integrated systematic analysis microarray data of well-characterized primary and metastatic ccRCCs to identify unique gene expression profiles characteristic of tumor aggressiveness. We identified 112 DEGs and 10 prognostic hub genes associated with functional annotations in different expression levels. GSEA was used to visualize significantly enriched gene-set hallmarks of 10 selected hub genes. These findings provide the basis for further screening and identification of promising biomarkers of tumor aggressiveness.

Several scoring systems combine clinical and pathological features to determine prognosis and to accurately predict survival outcomes of patients with primary and metastatic ccRCC [32-34]. These predictive models improve the specificity and accuracy of predicting survival outcomes of patients ccRCC, and the incorporated clinicopathological parameters are surrogate measures of major fundamental mechanisms that determine the aggressiveness of malignant tumors. Moreover, proteogenomic characterization of cancers that aims to develop therapeutic strategies will benefit from the acquisition of knowledge of the details of the biological mechanisms that contribute to the oncogenesis and progression of ccRCC.

At present, there have been multiple clinical research models based on multi-center or retrospective studies to predict prognosis, which helps to predict clinical prognosis from the perspective of multiple prognostic factors. These models have the potential to change our clinical practice and guide clinicians to implement individualized research strategies, which has strong clinical guidance value. In 2018, RCCInc4 classifier has been demonstrated to have precise prognostic significance in early ccRCC using four LncRNAs [35]. A retrospective analysis and multicentre validation study also constructed a six-SNP-based classifier for predicting recurrence in localised renal cell carcinoma [36]. Interestingly, there are few researches to analyze the genomic difference between of metastatic and primary ccRCC. The present study represents the first attempt to establish a gene regulatory network incorporating metas...
Figure 6. External validation for prognostic value of 10 hub genes and prediction model in 380 ccRCC patients from FUSCC cohort. After integration of qRT-PCR and clinical follow-up data for 380 patients with ccRCC, we confirmed that ADAMTS9, C1S, DPYS3, H2AFX, PLOD2, RUNX1, SLC19A1, TPX2, and TRIB3 mRNAs were differentially expressed in the FUSCC cohort. Differentially elevated mRNA levels were significantly associated with shorter DFS and OS (P<0.05).
Metastasis-related signatures in ccRCC

The limitations of this study are as follows: First, the microarray data were unbalanced with respect to the numbers of ccRCC and control tissues, which were restricted in quantity and acquired from the GEO databases. Second, these microarray data comprised relatively few ccRCC samples. Furthermore, only 513 patients were enrolled from the TCGA cohort with corresponding transcriptome data and relatively complete phenotypic data. Third, a prospective cohort was not analyzed, and identification of the underlying mechanisms was not addressed. Fourth, only the mRNA levels of hub genes are shown. Thus, further functional analyses of validated cohorts are required to verify these findings.

In conclusion, the present study identifies DEGs and hub genes that may be involved in earlier recurrence and poor prognosis of ccRCC. The expression levels of ADAMTS9, C1S, DYSPLS3, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2, and TRIB3 have high prognostic value and may help us better understand the underlying mechanisms of oncogenesis and progression of ccRCC.

Acknowledgements

We would like to present our sincere gratitude to ccRCC patients from the FUSCC cohort. We also would like to thank Cui-Zhu Zhang, Wei-Yi Yang, and Ying Shen for helping with follow-up of patients. We thank Liwen Bianji, Edanz Editing China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript. This work is supported by Grants from the National Key Research and Development Project (No. 2019YFC1316000) and the National...
Metastasis-related signatures in ccRCC

Figure 8. GSEA was used to perform hallmark analysis for ADAMTS9, C1S, DPYSL3, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2 and TRIB3. GSEA identified 100 DEGs with positive or negative correlations. Among them, we used the hallmarks ADAMTS9, C1S, DPYSL3, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2, and TRIB3 to perform GSEA. Results suggested that the most involved three significant pathways of each oncogene included apical junction, epithelial mesenchymal transition, allograft rejection, G2M checkpoint, mitotic spindle, Kras signaling pathway up, inflammatory response, TNF-alpha signaling via NF-κB, etc.

Figure 9. PPI network of nodes from integrated prediction model. A. The most significant genes associated with ten hub genes and their relationship, including protein-protein interaction, activation, inhibition and activation expression, were visualized in a circos plot. B. A detailed PPI network related to this gene sets were constructed to show more interrelated nodes at protein level.

Natural Science Foundation of China (No. 81802525 and 81772706).

Disclosure of conflict of interest

None.

Abbreviations

ccRCC, clear cell renal cell carcinoma; DEGs, differentially expressed genes; PPI, Protein-protein interaction; TCGA, the Cancer Genome Atlas; DFS, disease-free survival; OS, overall survival; HR, hazard ratio; CI, confidence interval; ROC, receiver operating characteristic curve; AUC, Area under curve; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSEA, Gene set enrichment analysis.

Address correspondence to: Drs. Ding-Wei Ye, Hai-Liang Zhang, Yuan-Yuan Qu and Yi-Ping Zhu, Department of Urology, Fudan University Shanghai Cancer Center, Shanghai 200032, P. R. China. Tel: +86-21-64175590-2805; Fax: +86-21-64434556; E-mail: dwyelie@163.com (Dwy); zhanghl918@163.com (HLZ); quyy1987@163.com (YYQ); qdzhuyping@aliyun.com (YPZ)

References

Metastasis-related signatures in ccRCC

W and Reis S. Expression of micro-RNAs and genes related to angiogenesis in ccRCC and associations with tumor characteristics. BMC Urol 2017; 17: 113.


Metastasis-related signatures in ccRCC


Supplementary Figure 1. The detailed functional annotations and classification pie charts using CluePedia from Cytoscape are provided. Of the ClueGO analysis, 25.0% of terms belonged to the negative regulation of protein complex disassembly, 12.5% to negative regulation of response to oxidative stress, 9.38% to cell-cell junction assembly, 6.25% to maintenance of protein location in cell, 6.25% to photoreceptor cell differentiation, 6.25% inorganic cation import across plasma membrane, and 6.25% to lysosome organization.
Supplementary Figure 2. IHC staining was used to describe differential proteomic expression of ADAMTS9, C1S, DPYSL3, H2AFX, MINA, PLOD2, RUNX1, SLC19A1, TPX2, TRIB3. Antibody staining density described staining status between normal and tumor samples.