Construction and Validation of a novel signature for Immune response prediction in clear cell Renal Cell Carcinoma

Nan Wu  
The Second Hospital of ShanXi Medical University

Ziwei Gui  
The Second Clinical Medical College of ShanXi Medical University

Juan Du  
The Second Hospital of ShanXi Medical University

Ningning Shen  
The Second Hospital of ShanXi Medical University

Zhiqing Yang  
The Second Hospital of ShanXi Medical University

Huijun Yang  
The Second Clinical Medical College of ShanXi Medical University

Zixin Zeng  
The Second Clinical Medical College of ShanXi Medical University

Wei Lu  
The Second Clinical Medical College of ShanXi Medical University

Zijia Leng  
The Second Clinical Medical College of ShanXi Medical University

Rong Wei  
The Second Hospital of ShanXi Medical University

Wenxia Ma  
The Second Hospital of ShanXi Medical University

Chen Wang  ( wangchen@sxmu.edu.cn )  
The Second Hospital of ShanXi Medical University

Research Article

Keywords: clear cell renal cell carcinoma (ccRCC), immune response, LASSO analysis, gene signature, prediction biomarker

Posted Date: November 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2301282/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Background
clear cell renal cell carcinoma (ccRCC) is the most common renal malignancy, although newly developing targeted therapy and immunotherapy have been showing promising effects in clinical treatment, the effective biomarkers for immune response prediction are still lacking. The study is to construct a gene signature according to ccRCC immune cells infiltration landscape, thus aiding clinical prediction of patients response to immunotherapy.

Methods
Firstly, ccRCC transcriptome expression profiles from Gene Expression Omnibus (GEO) database as well as immune related genes information from IMMPORT database were combine applied to identify the differently expressed meanwhile immune related candidate genes in ccRCC comparing to normal control samples. Then, based on protein-protein interaction network (PPI) and following module analysis of the candidate genes, a hub gene cluster was further identified for survival analysis. Further, LASSO analysis was applied to construct a signature which was in succession assessed with Kaplan-Meier survival, Cox regression and ROC curve analysis. Moreover, ccRCC patients were divided as high and low-risk groups based on the gene signature followed by the difference estimation of immune treatment response and exploration of related immune cells infiltration by TIDE and Cibersort analysis respectively among the two groups of patients.

Results
Based on GEO and IMMPORT databases, a total of 269 differently expressed meanwhile immune related genes in ccRCC were identified, further PPI network and module analysis of the 269 genes highlighted a 46 genes cluster. Next step, Kaplan-Meier and Cox regression analysis of the 46 genes identified 4 genes that were supported to be independent prognosis indicators, and a gene signature was constructed based on the 4 genes. Furthermore, after assessing its prognosis indicating ability by both Kaplan-Meier and Cox regression analysis, immune relation of the signature was evaluated including its association with environment immune score, Immune checkpoint inhibitors expression as well as immune cells infiltration. Together, immune predicting ability of the signature was preliminary explored.

Conclusions
Based on ccRCC genes expression profiles and multiple bioinformatic analysis, a 4 genes containing signature was constructed and the immune regulation of the signature was preliminary explored. Although more detailed experiments and clinical trials are needed before potential clinical use of the signature, the results shall provide meaningful insight into further ccRCC immune researches.

Background
Renal cell carcinoma has been the most common kidney malignancy which comprises ccRCC and non clear cell Renal Cell Carcinoma (nccRCC), and ccRCC accounts for approximately 70%~75% of all the cases[1, 2]. Attributing to the rapid development of molecular pathology, the genome mechanism behind ccRCC occurrence has been gradual clear, of which short arm of chromosome 3 (3p) genes variations were showing defining characteristic roles involving most importantly VHL gene known by symbolic “double hit”. The aberrant change of VHL including gene mutation and promoter methylation causes the “first hit”, followed by "second hit", namely the 3p chromosome deletion leading to tumor occurrence[3]. Besides VHL, other 3p gene variations were also reported in ccRCC, for instance SETD2[4], BAP1[5] and PBRM1[6], which were reported to be survival related.
Regarding the clinical treatment, over the past two decades, molecular targeted therapies and immune therapies have been showing great potential. Currently, at least 13 drugs in 6 categories have been approved for metastatic ccRCC, including VEGFR, mTORC1, c-Met, FGFR inhibition, cytokines, and most recently anti PD-1/PD-L1 immune checkpoint inhibitors (ICIs) which has been a promising pillar of nowadays clinical treatment[2, 7]. Multi clinical trials most notably KEYNOTE-564 results supported ccRCC as immune sensitive and demonstrated the efficiency of ICIs in advanced patients clinical treatment using independent ICI therapies or ICI + TKI combination therapies[8].

Although evidence-based medicine supported RCC as one of the malignancies that could benefit from neo-adjuvant PD-1/PD-L1 blockade[9–11], the immune response obviously vary among individuals indicating the importance of usable and effective immune biomarkers for selecting the potential patients that were most likely to be able to benefit from the treatment[12]. Currently, PD-L1 expression, MSI and tumor mutation burden (TMB) were three most widely used biomarkers in other types of cancers, however, their use in ccRCC are still in dispute.

As for PD-L1 expression, although it has been an effective biomarker in other cancers for predicting immune response[13–15], it’s function in ccRCC is inconclusive yet. For example, clinical trial Checkmate025 showed that advanced ccRCC patients benefited from immunotherapy regardless of PD-L1 expression[16, 17]. Checkmate214 revealed that the objective remission rate (ORR) in immunotherapy received group was higher than that in sunitinib treatment group no matter PD-L1 expression was higher or less than 1%[18, 19]. Meanwhile, Keynote426 results also revealed that both PD-L1 negative and positive ccRCC patients benefited from pablituzumab and axitinib combination therapy[20, 21].

Meanwhile, as for the use of TMB in renal malignancy, Robert M Samstein.et al reported an pan-cancer MSK-IMPACT genome sequencing analysis based on 1600 cases of cancer samples, of which 151 cases were RCC samples, and the results showed that TMB was non significant related with ccRCC overall survival[22]. And since microsatellite instability (MSI-H and dMMR) is very rare in RCC patients, MSI is not supported by evidence-based medicine yet to be an effective immune biomarker[12, 23].

Above all, ICIs has been an promising treatment in clinical ccRCC, but the effective biomarkers for immune response prediction are still lacking, it is of great importance to keep exploring ccRCC genome and identifying new potential immune response biomarkers thus aiding more precise understanding of the disease and shedding promising light on further clinical immunotherapy application.

In the study, multi online ccRCC transcriptome profiles, our local hospital patients samples as well as various bioinformatic analysis tools were combine used to explore ccRCC genome data, identifying the survival related meanwhile immune regulation associated genes and constructing a potential immune signature, further necessary signaling mechanism was preliminary explored. The results shall provide meaningful insights to the unearth of potential new immune biomarkers and shed promising light on further ccRCC immune researches.

**Materials And Methods**

**Data source: ccRCC transcriptome data from GEO database**

From GEO online database, we widely screened ccRCC related profiles for exploring the changed genome information in cancer comparing to normal renal samples. The selection criteria of GEO profiles were set as: 1. the information of profiles were based on human tissues (not animal models of any species); 2. the samples type was solid tissue (not tumor cell lines); 3. the containing data were mRNA/ cDNA/ transcriptome sequencing data; 4. covering both ccRCC cancer and normal renal control samples; 5. each profile should contain at least 40 or more samples.

Based on above selection criteria, four ccRCC cDNA expression profiles namely GSE53000[24], GSE53757[25], GSE68417[26] and GSE71963[27] were selected for further analysis including a total of 186 cases of ccRCC samples and 108 normal kidney samples. (Table S1 for detailed information of the profiles including samples amount, contributors and accessed online website).
Data processing: identify the differently expressed meanwhile immune related genes in ccRCC comparing to normal control

The GEO transcriptome data were used to 1. explore the differently expressed genes in ccRCC comparing to normal kidney samples; 2. combine with IMMPORT immune database \cite{28} for collaborate identify the differently expressed meanwhile immune regulation related genes.

To reveal the aberrant differently expressed genes in ccRCC comparing to normal control samples, the four GEO profiles GSE53000, GSE53757, GSE68417 and GSE71963 were in succession analyzed with GEO2R which was provided pared with each GEO profile. The analysis criteria was set as adjusted P value < 0.05 meanwhile |log2FC| < 1, 1 ≤ |log2FC| < 2, 2 ≤ |log2FC| < 3 and |log2FC| ≥ 3 respectively, thus the genes expression change distribution namely the genes that were < 2 fold, 2 ~ 4 fold, 4 ~ 8 fold and > 8 fold different in cancer versus normal control would be preliminary understood.

Then, Venn diagram \cite{29} would be used to identify the immune related genes from all the high level differently expressed genes based on IMMPORT immune genes list, therefore, the genes that were preliminary supported to be both aberrant changed expressed and immune regulation related were selected as candidate genes for further analysis.

**Protein-protein interaction (PPI) network construction and function module analysis of the candidate genes**

After identifying the differently expressed meanwhile immune related candidate genes, STRING \cite{30}, which is short for Search Tool for the Retrieval of Interacting Genes was applied to construct the PPI network of the genes for observing the interaction between individual genes. Further, based on PPI network, Molecular Complex Detection (MCODE) function of Cytoscape 3.6.0 software \cite{31} was used to analyze the promising function modules (gene clusters sharing similar function) from the gene nest.

Further, Gene ontology analysis (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) \cite{32} were used to annotate basic biological attributes of the list of genes in each module including their main cellular location, involved biological processes, molecular functions and the signaling pathways they mainly enriched in. The module that was predicted to be most related with tumor immune modulation and possess the highest module score would be highly focused and identified as a potential gene cluster for further analysis.

**Univariate survival combine with Cox regression analysis of immune related gene cluster for hub genes**

Following the identification of the immune related gene cluster, each gene in the module would be in succession brought for firstly univariate survival analysis by UALCAN \cite{33} and GEPIA \cite{34}, which have been two effective online services for survival analysis. Then, the genes that were supported by both univariate analysis methods to be statistical significantly associated with ccRCC survival would be processed for multivariate COX regression analysis. Any gene that was indicated by all three analysis to be associating with patients survival would be identified as credible prognostic indicating hub genes and processed for next step interpretation.

**Estimation of hub genes’ physicochemical properties**

For understanding the basic information of the selected hub genes, ProtParam \cite{35}, ProtScale \cite{36}, and Human Protein Atlas \cite{37} were combine used. As for the physicochemical properties of genes, ProtParam and ProtScale were applied to understand the basic information of the genes encoding proteins including the aminoacid composition, estimated molecular weight and protein half life, computed protein instability index and theoretical isoelectric point, as well as hydrophobicity and hydrophilicity of proteins.

Besides ProtParam and ProtScale, Human Protein Atlas is also an effective and well used online service for interpreting certain proteins information, in the study, it was applied to predict the cellular location of the selected hub genes for the convenience of further clinical test.
Additionally, UALCAN as well as GEPIA, which have been two resourceful web services constructed based on TCGA and GTEx programs were in succession accessed to observe the expression difference of hub genes in broad-spectrum human cancers comparing to corresponding normal control samples, especially in ccRCC versus normal renal tissues.

**Construction and clinical features analysis of an prognosis related immune gene signature**

To maximum the clinical utilization of selected hub gene indicators, an prognosis-related immune signature was constructed based on LASSO algorithm performed with glmnet R package, thus an unique regression coefficient was assigned to each gene indicator which multiplies the gene expression. Based on the final score of each case calculated according to the gene signature, ccRCC patients were classified as high-risk and low-risk groups (median score was set as cut off value).

Further, the clinical features of high-risk and low-risk groups of patients were analyzed based on TCGA data which contains resourceful ccRCC samples, despite the censoring data, an effective pool of over 533 cases of ccRCC patients information were applied for preliminary observing the clinical features association of the constructed signature.

**Preliminary prognosis validation of the gene signature**

To validate the survival relationship of the gene signature, a series of methods were combine used including firstly Kaplan-Meier survival which was used to compare the survival difference between high-risk and low-risk groups of patients, then AUC curve was performed to observe the 1, 3 and 5 years survival prediction ability of the signature. Further, univariate as well as multivariate Cox regression analysis were applied for testing the independence survival prediction ability of the signature together with other well accepted prognosis related clinical parameters. Furthermore, a nomogram combining the signature and these validated clinical parameters was constructed for together evaluating clinical ccRCC patients prognosis.

**Gene set enrichment analysis (GSEA) of two ccRCC patients subgroups based on gene signature**

After preliminary prognosis relationship analysis, the gene signature was next step processed for further immune association evaluating. Based on the constructed gene signature, ccRCC patients were divided as high-risk and low-risk subgroups, for determining how the immunological pathways and corresponding immune genes differ between the two ccRCC subgroups, GSEA[38] was performed for signaling enrichment analysis, and the threshold was set as P < 0.05.

**Difference of immunogenic cell death (ICD) between high-risk and low-risk groups of ccRCC patients**

ICD has been gradually accepted as a form of regulated biological cell death meanwhile supported by evidence-based medicine to be able to trigger cellular adaptive immune response through the emission of damage associated molecular patterns (DAMPs), thus potentially contributing to clinical immunotherapy. In the study, the expression distribution of 32 ICD related genes[39] which were identified based on literature studies were explored in high-risk and low-risk groups of ccRCC patients for preliminary evaluating the immune status difference between the two groups patients.

**Association between the gene signature and ccRCC estimated environment immune score**

ESTIMATE, which is short for Estimation of Stromal and Immune cells in Malignant Tumors using Expression data and has been a well accepted cancer immune evaluation tool in multiple cancers was applied to estimate the immune score of ccRCC samples based on TCGA genes expression data. The correlation between the signature and ESTIMATE algorithm based ccRCC immune score, stromal score as well as tumor purity were evaluated using R package for aiding the validation of immune relationship of the constructed gene signature.
Correlation between gene signature and the expression of immune checkpoint inhibitors

Immune checkpoints have been showing inspiring drug targeting effects in multiple cancers by reversing the tumor immuno suppressive microenvironment, and expression of immune checkpoints especially PD-L1, CTLA4, TIGIT, TIM-3 and LAG-3 have been well accepted as clinical biomarkers for selecting potential cancer patients that were most likely to benefit from immunotherapy. Therefore, the association between the gene signature and expression level of these immune checkpoints were assessed in the study, as well as the comparison of expression difference of these immune checkpoints between high-risk and low-risk groups of ccRCC patients.

Evaluation of relationships between gene signature and 22 tumor infiltrating immune cells (TICs)

For characterizing the microenvironment immune landscape between high-risk and low-risk groups of ccRCC patients, CIBERSORT algorithm[40] was performed to calculate the relative contents of 22 TICs based on TCGA profiles data, followed by analyzing the relationship between the 22 TICs and gene signature. Further, the survival analysis of 22 TICs especially the ones that relates with gene signature were conducted for identifying the specific immune cell infiltration that impacts patients prognosis.

Results

cCRCC transcriptome data identified 269 high level differently expressed meanwhile immune related genes in cancer versus normal renal samples

Four GEO profiles GSE53000, GSE53757, GSE68417 and GSE71963 were combine applied to explore the aberrant differently expressed genes in ccRCC comparing to normal renal samples. And in GSE53000, a total of 5559 genes were identified to be differently expressed including 4270 genes with the expression change ≤ 2 fold, 1028 genes that were 2 ~ 4 fold, 180 genes that were 4 ~ 8 fold and 81 genes whose expression were > 8 fold in ccRCC comparing to normal renal samples (Fig. 1A). And in GSE53757, a total of 28021 genes were identified, and the number was 21367, 4857, 1195 and 602 in ≤ 2 fold, 2 ~ 4fold, 4 ~ 8 fold and > 8 fold groups respectively (Fig. 1B). In GSE68417, a total of 10150 genes were identified, and the number was 8276, 1425, 290 and 159 in each group (Fig. 1C). Meanwhile, in GSE71963, a total of 10744 genes were identified, and the number was 6266, 3120, 841 and 517 genes in each group respectively (Fig. 1D, Table S2).

Considering the feasibility of further clinical medical use, we mainly focused on the high differently expressed genes (at least > 4 fold in cancer vs. normal). As a result, besides the genes that were shared in multiple profiles, the analysis of 4 GEO profiles indicated a total of 3095 genes that were high level changed expressed in ccRCC comparing to normal control samples (Fig. 1E). Further, the immune related genes list was obtained from IMMPORT immune database, and Venn diagram analysis result identified 269 genes from the 3095 genes that were both high level expression changed and immune regulation related for next step analysis (Fig. 1F, detailed 269 genes information was listed in Table S3).

PPI network of 269 genes highlighted a 46 genes-containing immune relating gene cluster

The PPI network of 269 differently expressed meanwhile potentially immune related genes was constructed (Fig. 2A), and based on the network we identified three promising gene clusters. The first gene cluster posses the highest computed module score and contains 46 genes with a big portion of them predicted to be related with immune system modulation (Fig. 2B). Meanwhile, the second and third gene modules contain 25 and 29 genes respectively, and genes were mostly related with CXCR4, PI3K, EGF and mTOR related signaling pathways (Fig. 2C, 2D).
Given the first module gene cluster possess the highest score and a big percentage of containing genes were immune system related which shows more potential for further clinical immune indicators selection, the 46 genes in the gene cluster were mainly focused for next step analysis.

Kaplan-Meier combine with Cox regression analysis of cluster genes identified 4 ccRCC prognosis related hub genes

Genes should be of more potential if they were both immune regulation and survival related, and promising immune biomarkers should also be prognosis related for potential clinical medical drug targeting use. To further analyze survival relationship of the 46 selected immune related candidate genes, univariate survival analysis including UALCAN and GEPIA, as well as multivariate Cox regression analysis were in succession performed, and the results supported four genes, namely MMP9 (Fig. 3A), NFKB1 (Fig. 3B), IRF7 (Fig. 3C) and HMOX1 (Fig. 3D) as independent prognostic indicators in ccRCC, all four genes not only relate with patients overall survival but also progress free survival indicating their high potential in clinical medical use (Table 1).

Table 1 Univariate combine with multivariate Cox Regression analysis result of the 4 hub genes used for signature construction

<table>
<thead>
<tr>
<th>OSCC parameters</th>
<th>P Value</th>
<th>B value</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate analysis</td>
<td>Multivariate analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UALCAN</td>
<td>GEPIA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMP9</td>
<td>0.001</td>
<td>0.016</td>
<td>0.002</td>
<td>0.256</td>
</tr>
<tr>
<td>NFKB1</td>
<td>0.013</td>
<td>2.9E-05</td>
<td>0.004</td>
<td>-0.216</td>
</tr>
<tr>
<td>IRF7</td>
<td>&lt;0.0001</td>
<td>9E-04</td>
<td>0.022</td>
<td>0.077</td>
</tr>
<tr>
<td>HMOX1</td>
<td>0.034</td>
<td>0.00062</td>
<td>&lt;0.001</td>
<td>-0.222</td>
</tr>
</tbody>
</table>

Basic physicochemical properties of the 4 selected hub genes

Basic physiochemical properties of MMP9, NFKB1, IRF7 and HMOX1 were preliminary interpreted before deeper scientific use of them (Table 2). As for MMP9, which is a member of matrix metalloproteinase (MMP) family, locates in 20q13.12 and encodes a protein composed of 707 amino acids with an estimated molecular weight of 78.5KD. The theoretical isoelectric point of the protein is estimated to be 5.69 and instability index to be 41.10, meanwhile, the grand average of hydrophobic value of the protein is -0.394 indicating MMP9 works as a cellular unstable and hydrophilic protein which locates in cellular cytoplasm or to be secreted in the extracellular region, and the related signaling pathways include collagen chain trimerization and apoptotic pathways in synovial fibroblasts (Fig. 3E).
Table 2
Basic physicochemical properties of the 4 hub genes used for gene signature construction

<table>
<thead>
<tr>
<th>Gene Property</th>
<th>MMP9</th>
<th>NFKB1</th>
<th>IRF7</th>
<th>HMOX1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formula</td>
<td>$C_{3517}H_{5298}N_{958}O_{1035}S_{28}$</td>
<td>$C_{4643}H_{7343}N_{1271}O_{1458}S_{33}$</td>
<td>$C_{2418}H_{3740}N_{678}O_{710}S_{19}$</td>
<td>$C_{1475}H_{2323}N_{405}O_{427}S_{8}$</td>
</tr>
<tr>
<td>Molecular Weight</td>
<td>78.46KD</td>
<td>105.36KD</td>
<td>54.28KD</td>
<td>32.82KD</td>
</tr>
<tr>
<td>Number of amino acids</td>
<td>707AA</td>
<td>968AA</td>
<td>503AA</td>
<td>288AA</td>
</tr>
<tr>
<td>Theoretical pI</td>
<td>5.69</td>
<td>5.20</td>
<td>5.89</td>
<td>7.89</td>
</tr>
<tr>
<td>Aliphatic index</td>
<td>65.13</td>
<td>84.74</td>
<td>72.07</td>
<td>83.02</td>
</tr>
<tr>
<td>Hydrophobic value</td>
<td>-0.394</td>
<td>-0.339</td>
<td>-0.367</td>
<td>-0.427</td>
</tr>
<tr>
<td>Estimated protein half life</td>
<td>30h</td>
<td>30h</td>
<td>30h</td>
<td>30h</td>
</tr>
<tr>
<td>Instability index</td>
<td>42.10</td>
<td>38.15</td>
<td>63.17</td>
<td>60.81</td>
</tr>
</tbody>
</table>

NFKB1, which is short for Nuclear Factor Kappa B Subunit 1 locates in 4q24 and encodes a protein composed of 968 amino acids and estimated to be weighting 105KD with computed theoretical isoelectric point as 5.20 and instability index as 38.15. Meanwhile, the grand average of hydrophobic value of protein is -0.339 indicating NFKB1 to be cellular stable and hydrophilic. NFKB1 is predicted to locates in nucleoplasm and cytoplasm, it has been reported as a transcription regulator that could be activated by various cellular stimuli such as cytokines, ultraviolet irradiation, and bacterial or viral products. Activated NFKB1 translocates into cell nucleus and stimulates the expression of genes involved in various biological functions (Fig. 3F).

IRF7 is short for Interferon Regulatory Factor 7 and it’s a member of the interferon regulatory factor (IRF) family, locating in 11p15.5 and encoding a protein composed of 503 amino acids including 56 negatively charged amino acid residues (ASP + Glu) and 49 positively charged amino acid residues (Arg + Lys). The estimated protein molecular weight is 54.2KD with theoretical isoelectric point computed to be 5.89. Meanwhile, the estimated instability index of the protein is 63.17 and grand average of hydrophobic value is -0.367, the cellular location of the gene is predicted to be in nucleoplasm or cytoplasm (Fig. 3G).

Meanwhile, HMOX1 is short for Heme Oxygenase1, and locates in 22q12.3, encoding a protein composed of 288 amino acids including 35 negatively charged amino acid residues (ASP + Glu) and 36 positively charged amino acid residues (Arg + Lys). The estimated protein molecular weight is 32.8KD with theoretical isoelectric point computed as 7.89. Moreover, the estimated instability index of the protein is 60.81 and grand average of hydrophobic value is -0.427 indicating HMOX1 works as a cellular unstable and hydrophilic protein which is consistent with the ProtScale analysis result of HMOX1 structure showing that the protein harbors more hydrophilic regions than hydrophobicity regions. HMOX1 is predicted to locates in cellular Golgi apparatus and plasma membrane, it has been reported to be associated with the development of heme oxygenase1 deficiency and pulmonary disease, as well as chronic obstructive (Fig. 3H).

Validation of the changed expression of selected hub genes in ccRCC versus normal renal tissues
Although the four selected hub genes were obtained from the differently expressed gene clusters analyzed based on GEO data from the beginning, after preliminary interpretation of the basic physicochemical information, it's necessary to validate each of the gene's aberrant changed expression in ccRCC comparing to normal renal samples individually. In the study, two analysis databases including UALCAN and GEPIA were used, and the results revealed that as for MMP9 and IRF7 genes, they gain of expression in most of human cancers (Fig. 4A, 4C). And as for NFKB1 and HMOX1, their expression vary in different cancers (Fig. 4E, 4G), although in ccRCC, all four genes were indicated to be statistical significantly up regulated in cancers comparing to normal renal tissues (Fig. 4B, 4D, 4F, 4H).

**Construction of a 4 genes containing ccRCC prognosis related immune gene signature and clinical features analysis**

To maximum the clinical prediction value of the four selected genes, Cox-proportional hazards analysis based on LASSO algorithm was applied to construct a MMP9, NFKB1, IRF7 and HMOX1 four genes containing signature which weights the normalized expression level of each gene to the regression coefficient of multivariate Cox regression analysis. And the result revealed a formula: Risk Score = 0.256 * expression (MMP9) − 0.222 * expression (HMOX1) − 0.216 * expression (NFKB1) + 0.077 * expression (IRF7) as the best signature for combining the four differently expressed meanwhile immune related hub genes (Fig. 5A, 5B).

Based on the signature, the risk score for each patient was calculated followed by the patients being categorized into high-risk or low-risk groups according to the median risk score which was set as the cut off point for the signature (Fig. 5C). Further, the association between the gene signature and ccRCC clinical features was preliminary analyzed, which result revealed that higher risk score was positively related with older age (> 45 years old) and more advanced T, N and M stage, meanwhile, the low risk group of patients were tend to be younger (≤ 45 years old) female with lower TNM stage (Table 3).
Table 3
Association between the gene signature and ccRCC clinical features

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Gene signature</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low-risk group</td>
<td>High-risk group</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>112 (59.6%)</td>
<td>76 (40.4%)</td>
</tr>
<tr>
<td>male</td>
<td>154 (44.6%)</td>
<td>191 (55.4%)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 45</td>
<td>38 (64.4%)</td>
<td>21 (35.6%)</td>
</tr>
<tr>
<td>&gt; 45</td>
<td>228 (48.1%)</td>
<td>246 (51.9%)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>234 (50.6%)</td>
<td>228 (49.4%)</td>
</tr>
<tr>
<td>Yellow</td>
<td>4 (50.0%)</td>
<td>4 (50.0%)</td>
</tr>
<tr>
<td>Black</td>
<td>25 (44.6%)</td>
<td>31 (55.4%)</td>
</tr>
<tr>
<td>T stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>164 (60.01%)</td>
<td>109 (39.9%)</td>
</tr>
<tr>
<td>T2</td>
<td>31 (44.9%)</td>
<td>38 (55.1%)</td>
</tr>
<tr>
<td>T3</td>
<td>69 (38.3%)</td>
<td>111 (61.7%)</td>
</tr>
<tr>
<td>T4</td>
<td>2 (18.2%)</td>
<td>9 (81.8%)</td>
</tr>
<tr>
<td>N stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N0</td>
<td>115 (47.9%)</td>
<td>125 (52.1%)</td>
</tr>
<tr>
<td>N1</td>
<td>3 (18.8%)</td>
<td>13 (81.3%)</td>
</tr>
<tr>
<td>M stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M0</td>
<td>231 (54.7%)</td>
<td>191 (45.3%)</td>
</tr>
<tr>
<td>M1</td>
<td>22 (27.8%)</td>
<td>57 (72.2%)</td>
</tr>
</tbody>
</table>

High risk score based on the gene signature indicated worse ccRCC patients prognosis

All four genes used to construct the gene signature were previously supported to be high level differently expressed in ccRCC comparing to normal renal samples, but changed gene expression doesn't equal to survival association. To validate the survival relationship of the gene signature, a series of methods were used. Firstly, Kaplan-Meier survival analysis revealed that the high risk group of patients had a statistical significantly worse overall survival than their low risk counterparts (Fig. 5D). Then the ROC curve showed that the area under the ROC curve (AUC) of gene signature for overall survival was 0.747 at 1 year, 0.696 at 3 year and 0.705 at 5 years (Fig. 5E). Meanwhile, univariate Kaplan-Meier survival as well as multivariate Cox regression analysis were applied for testing the survival prediction ability of the signature, and the results supported the risk score calculated based on the gene signature works as an independent prognosis indicator for ccRCC patients together with some other well accepted prognosis related clinical parameters including patient T and M stage (Table 4). Further, a nomogram was constructed and in the nomogram, a point scale was assigned for each variable, the sum of all the
variables points equal to the final score of each patient, and the survival could be predicted by drawing a vertical line from the
total point axis downward to the outcome axis (Fig. 5F).

<table>
<thead>
<tr>
<th>Clinical parameters</th>
<th>P Value</th>
<th></th>
<th>Exp (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate analysis</td>
<td>Multivariate analysis</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>&lt; 0.001</td>
<td>0.026</td>
<td>1.627 (1.058–2.500)</td>
</tr>
<tr>
<td>Gender</td>
<td>0.693</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Race</td>
<td>0.719</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>T stage</td>
<td>&lt; 0.001</td>
<td>0.018</td>
<td>1.360 (1.054–1.755)</td>
</tr>
<tr>
<td>N stage</td>
<td>&lt; 0.001</td>
<td>0.480</td>
<td>1.290 (0.636–2.615)</td>
</tr>
<tr>
<td>M stage</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>2.794 (1.728–4.517)</td>
</tr>
<tr>
<td>Signature risk score</td>
<td>&lt; 0.001</td>
<td>0.002</td>
<td>1.871 (1.299–2.693)</td>
</tr>
</tbody>
</table>

**Table 4**

Survival prediction value of the gene signature included ccRCC clinical parameters

**High-risk group of ccRCC cases were more enriched in immune related phenotype**

After preliminary demonstrating the association between constructed gene signature and ccRCC prognosis, the influence of
gene signature on cancer immune profiles was to be investigated. And in the first step, GSEA was utilized to analyze the
immune-related biological processes linked to the signature, and the analysis result showed that the high-risk group cases
were significantly enriched in multiple biological processes, of which 4 immune-related processes were identified including
HUMORAL_IMMUNE_RESPONSE (NES = 1.733, Nominal p value = 0.0), REGULATION_OF_T_CELL_MIGRATION (NES = 1.762,
Nominal p value = 0.0), REGULATION_OF_LYMPHOCYTE_CELL_MIGRATION (NES = 1.743, Nominal p value = 0.003),
POSITIVE_REGULATION_OF_IMMUNOGLOBULIN_PRODUCTION (NES = 1.754,Nominal p value = 0.0). Meanwhile, the low-risk
group cases were not indicated to be enriched in any immune-related biological processes (Fig. 5G).

**High-risk and low-risk groups of ccRCC patients revealed disparate ICD expression levels**

Besides GSEA immune phenotype enrichment analysis, given the significant roles of ICD in antitumor immunological
responses, the connection between gene signature and ICD related genes were evaluated for additionally exploring the
immune status in high-risk and low-risk groups of ccRCC patients. And the results revealed that the expression of a large
portion of the 32 ICD related genes were statistical significantly different between the two groups of patients indicating the
diverse immune status in the microenvironment of two groups of patients (Fig. 6A).

**Risk score calculated based on the gene signature associated with ccRCC estimated environment immune score**

For further validating the immune association of the gene signature, ESTIMATE was performed to evaluate the immune,
stromal score and tumor purity of ccRCC samples. And the result revealed that although no significant correlation was found
between the gene signature and ccRCC stromal score, a mediate correlation was revealed between the risk score which was
calculated based on the gene signature and tumor immune score as well as tumor purity (Fig. 6B). Meanwhile, the high risk-
group patients were tend to posses higher immune score and lower tumor purity, indicating the immune targeting potential of
the group of patients (Fig. 6C).
Multi immune checkpoints expressed higher in high risk group of ccRCC patients based on the gene signature

Besides estimation of immune score, the association between the gene signature and clinical promising immune checkpoints including PD-L1, CTLA4, TIGIT, TIM-3 and LAG-3 were evaluated (Fig. 6D). And a median association was revealed between the gene signature and CTLA4 expression. Moreover, mild correlation was indicated between gene signature and two of the immune checkpoints including LAG3 and TIGIT, meanwhile, no significant relation was found between the signature and PD-L1 or TIM-3 expression (Fig. 6E). An inspiring fact was that all CTLA4, LAG3 and TIGIT tend to express higher in high-risk group of patients which was categorized based on the gene signature (Fig. 6F), and the distribution shall be an additional support besides above ESMINATE immune score evaluation result for indicating the immune targeting potential for this group of ccRCC patients.

Evaluation of relationships between gene signature and 22 tumor infiltrating immune cells (TICs)

Previous analysis supported that the constructed gene signature was related to immunity, so we carried out analyses on 22 TICs whose distribution profiles were draw based on CIBERSORT algorithm to further study the interaction between the gene signature and ccRCC immune microenvironment. And the correlation analysis result found four types of TICs to be related with the gene signature including plasma cells, activated CD4(+) T memory cells, activated dendritic cells and resting mast cells (Fig. 7A-7C).

Further, the prognostic abilities of the 22 TICs were tested and the results revealed that of the four signature related TICs, CD4(+) T cell and resting mast cell were able to predict ccRCC patients prognosis. The resting and activated CD4(+) T memory cells played opposite roles in patients survival, namely the activated CD4(+) T memory cells were related with worse patients survival (Fig. 7D), meanwhile, the resting CD4(+) T memory cells predicting better patients survival (Fig. 7E). Also, the resting mast cells were correlated with positive patients prognosis (Fig. 7F).

Combining the analysis results, one inspiring deduction could be draw that CD4(+) T memory cells and resting mast cells not only are significantly related to the gene signature but also predict ccRCC patients prognosis, indicating these immune cells may play important roles in the immune regulation of the gene signature in ccRCC microenvironment.

Discussion

ICIs has been an increasing rising up clinical method and holds great promise for treating ccRCC[7], but the effective biomarkers for predicting immune response are still lacking, the well accepted immune prediction biomarkers in other cancers for instance PD-L1 expression, MSI status and TMB haven't been supported thoroughly by evidence-based medicine to be effective in ccRCC[41, 42]. For the clinical benefit from ICI therapy, it is of great importance to keep exploring ccRCC genome and identifying new potential biomarkers thus benefiting further clinical application of immunotherapy in the cancer.

In recent years, multiple genes containing signatures representative of caner immune status have been identified in several cancers and they have been showing inspiring clinical effects[43–46]. It’s of clinical feasibility to explore ccRCC genome information and develop meaningful immune prediction models which were also prognosis related to evaluate the immune status of ccRCC microenvironment and stratify patients into different groups for increasing the potential effectiveness of ICIs therapy. In the study, GEO transcriptome profiles were strictly screened for analyzing the potentially immune regulation related gene candidates.

Based on four different ccRCC cDNA expression profiles which were all selected by strict criteria as see in the Materials and Methods part, we identified the differential expression genes in ccRCC cancer vs. normal renal tissues and then divided them into 4 groups according to the difference level as < 2 fold, 2 ~ 4 fold, 4 ~ 8 fold and > 8 fold genes considering the potential unique functions and clinical use of each group, for example, an interesting phenomenon has been discovered that the more genes expression difference are, the more their cellular location tend to be far away from cell nuclear[47–49]. For the
feasibility of further clinical medical use, in the selection process of candidate genes, we mainly focused on the high level differently expressed namely at least > 4 fold genes that were more convenient to be tested by IHC experiment which has been a common method in clinical pathology diagnosis, considering that the genes shall harbor more chance to be translated into clinical use if they are suitable to be tested by IHC. Further, the intersection between GEO selected high level aberrant differently expressed genes and immune related gene list from IMMPORT database indicated 269 genes that were both high level expression changed in ccRCC and immune related as candidate genes for next step analysis.

As for the construction of multi genes containing signature, LASSO algorithm has been widely accepted as an effective tool that is suitable to construct gene models basing on large numbers of correlated covariate. But instead of constructing an gene signature directly from the 269 candidate genes, we further in succession performed module analysis as well as multiple survival analysis to scale down the candidate genes and identify the promising “unique key genes” during ccRCC development, and only used LASSO for estimating the coefficient of signature genes. This is for the considering that the signature and consist genes should be of more clinical potential if they were not only immune regulation but also survival related for clinical medical drug targeting use. Therefore, the PPI network of the 269 genes was constructed followed by genes module analysis which highlighted a 46 genes containing cluster, and further survival analysis including GEPIA and UALCAN univariate survival as well as multivariate Cox Regression analysis of each of the 46 genes supported four genes: MMP9, NFKB1, IRF7 and HMOX1 to be associated with patients survival and worked as independent prognostic indicators in ccRCC development.

Interestingly, no direct relationship has yet been discovered among the four genes. MMP9 is a member of matrix metalloproteinase (MMP) family which is well known to be involved in the breakdown of extracellular matrix during multiple normal physiological and diseases processes, and MMP9 has been reported to be able to degrades type IV and V collagen which are important microenvironment elements. NFKB1 is a transcription regulator that could be activated by cellular stimuli such as cytokines, ultraviolet irradiation, bacterial and viral products, and inappropriate activation of the gene has been known to associate with a number of inflammatory diseases, while persistent inhibition of NFKB1 leads to inappropriate immune cell development or delayed cell growth. IFR7 is predicted to locates in nucleoplasm and cytoplasm and it has been reported to play roles in innate immune response against DNA and RNA viruses. Meanwhile, HMOX1 is predicted to locates in cellular Golgi apparatus and plasma membrane and it has been reported to be associated with the development of heme oxygenase1 deficiency and pulmonary disease, as well as chronic obstructive. The together identification of the four genes and a gene signature combine all of them indicating the elaborate collaboration network of various genes in cellular activities, opening up further cancer researches of unlimited possibilities.

Based on the selected four genes and coefficient calculated with LASSO algorithm for each gene, an immune meanwhile prognosis related gene signature was constructed. Survival relationship validation including Kaplan-Meier survival, Cox proportional-hazards model and ROC curve supported the signature worked as an prognostic factor after combining the four genes in one equation, proving the effectiveness of apply the gene signature in ccRCC prognosis prediction. Since many clinical parameters especially tumor TNM stage as been well known as critical survival related aspects, we proposed a nomogram assessment that combines the signature and other clinical features. Although current result has not supported the signature to be a better prognosis factor than TNM stage, the construction of the nomogram shall work as a complementary perspective on individual tumour and aiding the comprehensive evaluation of clinical ccRCC patients prognosis.

Although survival relation was an important part, the main aim of the signature was for potential immune prediction. Immune escape has been one of the major characteristics in malignant tumors involving multiple probable mechanisms[50, 51], for example the increasing immune suppressive cells including Treg cells and tumour-associated macrophages (TAM) in tumor microenvironment[43], and the up-regulated expression of immunosuppressive molecules for instance cytotoxic T lymphocyte associated antigen-4 (CTLA-4), also decreasing expression of cancer antigens which results the inactivation of tumor killing CD8 + T cells[52–55]. Therefore, we explored the probable relation between the gene signature and immune suppressive mechanisms. And the results revealed that the difference expression of ICD related genes in high-risk and low-risk groups of patients which were categorized based on the constructed gene signature, as well as the statistical significant correlation
between the signature and ESTIATE immune score supported the signature was immune modulation associated. Further, we investigated the expression of immune checkpoints including PD-L1, CTLA4, TIGIT, TIM-3 and LAG-3 between the high-risk and low-risk groups of patients, and the results showed the high-risk patients had higher expression of CTLA4, LAG3 and TIGIT than the low-risk patients indicating the immune targeting potential for this group of ccRCC patients.

As the results of the relation analysis between the signature and 22 TICs indicated that the signature was significantly related with CD4(+) T memory cells and resting mast cells infiltration, and not only the two immune cells were related with the signature, but also they were able to predict ccRCC patients prognosis. CD4(+) T memory cells have already been reported to confer vital functions on malignancy immune regulation, including participating in the activation of CD8 + T and NK killing cells, involving in the tumour immunological reactions[56, 57]. And mast cells were reported to be able to not only influence tumor expansion via inducing angiogenesis and changing tumor extracellular matrix composition, but also could influence the infiltration and activity of dendritic cells, tumor-associated macrophages and lymphocytes, promoting pro-inflammatory reactions in tumor microenvironment[43, 58, 59]. The association between the signature and immune checkpoints expression as well as different immune cells infiltration, suggesting the stronger immunosuppressive environment in high-risk groups of patients comparing to low-risk group, highlighting the potential of this group to benefit from further clinical immunotherapy.

**Limitation Of The Study**

In the study, based on ccRCC genes expression profiles from GEO online database, we screened the differently expressed genes in cancer comparing to normal renal tissues, followed by intersection with IMMPORT database immune genes list for identifying the differently expressed meanwhile immune related candidate genes. Further survival analysis of these genes highlighted four genes: MMP9, NFKB1, IRF7 and HMOX1, based on which a ccRCC prognosis related immune signature was constructed.

Although we did performed multiple analysis for validating the association between gene signature and immune factors including GSEA analysis, ICD related genes expression, immune checkpoints expression as the TICs infiltration, but the results were mostly based on retrospective data analysis, rigorous prospective studies performed on animal models or clinical trials are needed before proving the effectiveness of the signature application in clinical immunotherapy use. In addition, none obvious direct relation has been found among the four signature containing genes yet, more delicate functional and mechanistic studies of the four genes individually and in combination should be conducted for better understanding the mechanism behind their regulation on malignancy immune status.

**Conclusion**

The present study defined a four genes containing signature based on ccRCC genes expression information, the signature was not only closely associated with patients survival, but also immune regulation related. Multiple in vitro experiments data analysis supported the association between signature and ccRCC microenvironment immune aspects including immune checkpoints expression and various types of immune cells infiltration. Although the current result is not yet enough to support the application of the signature in clinical medical immunotherapy, the results shall provide meaningful insight into better understanding of the disease and shed lighting on further ccRCC immune regulation researches.

**Abbreviations**

ccRCC Clear cell renal cell carcinoma  
nccRCC Non clear cell Renal Cell Carcinoma  
GEO Gene Expression Omnibus  
PPI Protein-protein interaction network
ICls           Immune checkpoint inhibitors
TMB           Tumor mutation burden
ICD           Immunogenic cell death
TIC           Tumor infiltrating immune cells
TAM           Tumour-associated macrophages

Declarations

Ethnic approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
All data generated or analyzed during this study are included in this published article.

Competing interests
All of the authors agreed the publication of the paper and declare no conflicts of interests.

Funding
The work was supported by the China central government funds for guiding local scientific and technological development (YDZJSX2021A042), the fund of Shanxi Medical key scientific research project (2021XM34) and the fund of Natural Science Foundation of ShanXi Province in China (201901D211498).

Authors' contributions
NW, ZG and JD designed the study and together drafting the manuscript, contributed equally to the whole study. NS and ZY performed the data collecting and analysis. HY, ZZ, WL and ZL participated in the data interpretation and study design, RW were involved in the drafting and critical revision of the manuscript. As the corresponding author, both WM and CW had full access to all data of the manuscript, CW made the final decision to submit the article for publication. All authors read and approved the final manuscript.

Acknowledgements
We sincerely appreciate the researchers for providing their GEO databases information online which were important data resources for the study, it is our pleasure to acknowledge their contributions.

References


Figures
Figure 1

Differential expression genes in ccRCC comparing to normal renal samples identified from GEO profiles

From GEO profiles (A) GSE53000, (B) GSE53757, (C) GSE68417 and (D) GSE71963, the up-regulated (right side) and down-regulated (left side) differential expression genes in ccRCC were identified, and the genes were divided into four groups based on the expression difference level as: <2 fold genes (orange-colored spots), 2~4 fold genes (red-colored spots), 4~8 fold genes (green-colored spots) and >8 fold genes (black-colored spots). (E) The intersection of the genes in different GEO profiles, the genes that were shared in different profiles were identified. (F) The intersection of differential expressed genes revealed by GEO profiles and immune related genes from IMMPORT database, the genes that were both differential expressed and immune related were revealed.
Figure 2

PPI network construction of 269 differential expressed meanwhile immune related genes in ccRCC and function modules analysis

(A) The PPI network of the 269 differential expression meanwhile immune related genes in ccRCC which were identified based on GEO and IMMPORT datasets. (B) The first, (C) second and (D) third genes function modules were analyzed based on the PPI network, each module was shown with a diagrammatic sketch (left diagram) and the detailed module information (right table) including the computed module score, module description and detailed involving genes. (*The first module with the highest module score meanwhile immune regulation related was focused for further analysis).
Figure 3

Survival analysis and basic physicochemical properties exploration of four selected signature comprised genes

The overall survival (left) and disease free survival (right) analysis of (A) MMP9 gene, (B) NFKB1 gene, (C) IRF7 gene and (D) HMOX1 gene in ccRCC. The predicted cellular location (left) and computed hydrophilicity/hydrophobicity property of (E) MMP9 protein, (F) NFKB1 protein, (G) IRF7 protein and (H) HMOX1 protein.
Figure 4

The differential expression of four selected signature comprised genes in ccRCC included human cancers

UALCAN prediction of (A) MMP9 gene, (C) IRF7 gene, (E) NFKB1 gene and (G) HMOX1 gene expression in broad spectrum human cancers. GEPIA analysis of (B) MMP9 gene, (D) IRF7 gene, (F) NFKB1 gene and (H) HMOX1 gene in ccRCC comparing to normal renal samples.
Figure 5

Construction of a four genes containing meanwhile immune and prognosis related ccRCC gene signature

(A, B) LASSO analysis to calculate the coefficient and likelihood deviance for constructing a immune meanwhile prognosis related signature which was comprised of four genes. (C) TCGA ccRCC patients were divided into high-risk and low-risk groups based on the calculated signature score. (D) Survival analysis of the high-risk and low-risk groups of ccRCC patients. (E) ROC curve of the gene signature to predict ccRCC patients survival of 1 year, 3 years and 5 years respectively. (F) ccRCC patients prognosis prediction nomogram constructed based on genes signature and clinical parameters which were supported by Cox Regression to be independently related with patients survival. (G) Significant enrichment of immune-related phenotype in high-risk group of ccRCC patients compared with that in low-risk group patients.
Figure 6

Correlation between gene signature and ccRCC immune microenvironment landscape

(A) Relative expression of ICD related genes in high-risk and low-risk groups of ccRCC patients. (B) Correlation between gene signature and ccRCC computed immune score, stromal score and tumor purity calculated using ESTIMATE algorithm. (C) Estimated immune score, stromal score and tumor purity distribution in high-risk and low-risk ccRCC groups respectively. (D) Association between gene signature and immune checkpoints expression. (E) Correlation between gene signature and PD-L1, LAG-3, TIGIT and CALT-4 expression respectively. (F) Relative expression of five immune checkpoints including PD-L1, LAG-3, TIGIT, TIM3 and CALT-4 expression in high-risk and low-risk ccRCC groups respectively.
Figure 7

Correlation between gene signature and 22 immune cells infiltration in ccRCC

(A) Relative distribution of 22 immune cells in high-risk and low-risk groups of ccRCC patients. (B, C) Correlation between gene signature and various immune cells infiltration in ccRCC. Association between (D) T cells CD4 memory activated, (E) T cells CD4 memory resting and (F) Mast cells resting microenvironment infiltration and ccRCC patients survival.