# ISPRF: A Machine Learning Model to Predict the Immune Subtype of Kidney Cancer Samples by Four Genes

**Zhifeng Wang**
Henan Provincial People's Hospital
**Zihao Chen**
Southern Medical University
**Hongfan Zhao**
Charité-Universitätsmedizin
**Hao Lin**
Southern Medical University
**Junjie Wang**
Henan Provincial People's Hospital
**Ning Wang**
Henan Provincial People's Hospital
**Xiqing Li**
Henan Provincial People's Hospital
**Degang Ding** ( drdegang@126.com )
Henan Provincial People's Hospital

**Research Article**

**Keywords:** renal cell carcinoma, immune subtypes, machine learning, online website

**Posted Date:** February 11th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-184890/v1

# Abstract

### Background

Clear cell renal cell carcinoma (ccRCC) is the most common type of renal cell carcinoma. Immunotherapy, especially anti-PD-1, is becoming a pillar of ccRCC treatment. However, precise biomarkers and robust models are needed to select the appropriate patients for immunotherapy.

### Methods

A total of 831 ccRCC transcriptomic profiles were obtained from 6 datasets. Unsupervised clustering was performed to identify the immune subtypes among ccRCC samples based on immune cell enrichment scores. Weighted correlation network analysis (WGCNA) was used to identify hub genes distinguishing subtypes and related to prognosis. A machine learning model was established by random forest algorithm, and employed to an open and free online website to predict the immune subtype.

### Results

In the identified immune subtypes, subtype2 was enriched in immune cell enrichment scores and immunotherapy biomarkers. WGCNA analysis identified 4 hub genes related to immune subtype, CTLA4, FOXP3, IFNG, and CD19. The random forest model was constructed by mRNA expression of these four hub genes, and the value of areas under the curve of the receiver operating characteristic (AUC) was 0.78. Subtype2 patients in the independent validation cohort had a better drug response and prognosis for immunotherapy treatment. Moreover, an open and free website was developed by the random forest model (https://immunotype.shinyapps.io/ISPRF/).

### Conclusions

The current study constructs a model and provides a free online website that could identify suitable ccRCC patients for immunotherapy, and it is an important step forward to personalized treatment.


# 1. Introduction

Renal cell carcinoma (RCC) denotes accounts for over 90% of all kidney cancer, and clear cell renal cell carcinoma (ccRCC) is the most frequent histology subtype [1]. ccRCC affects around 300,000 patients worldwide and causes over 100,000 deaths annually [2]. The onset of symptoms of ccRCC is usually insidious, thus, the diagnosis of it occurs in the advanced stage [3]. Besides, ccRCC has a tendency to metastasis to distant organs such as lung and liver [4]. Owing to the resistance of ccRCC to radiotherapy and chemotherapy, the mortality rate of patients with metastatic ccRCC remains high [5]. Thus, it is important to provide novel therapeutic drugs.

Recently, the clinical trial results reported that immune checkpoint antibodies greatly improved patient survival in some types of cancer including ccRCC [6]. The Food and Drug Administration (FDA) approved Nivolumab (PD-1 antibody) in November 2015 for use in metastatic ccRCC patients who progressed on an angiogenesis inhibitor. The FDA made its decision based on findings from the phase 3 CheckMate025 trial, in which the PD-1 antibody improved the median overall survival (OS) and reduced the risk of death versus Everolimus (Afinitor) [7]. After that, FDA approved Pembrolizumab (PD-1 antibody) plus Axitinib for the first-line treatment of ccRCC patients by the benefit of OS in Keynote426 trial [8]. Besides, FDA approved Avelumab (PD-L1 antibody) in combination with Axitinib for first-line treatment of ccRCC patients in May 2019 [9].

Immunotherapies such as PD-1 or PD-L1 antibodies could greatly improve the prognosis of cancer patients, but the number of patients who showed consistent responses to the immunotherapy was limited [10]. Moreover, side effects and adverse toxicities caused by immune checkpoint antibodies are reported [10]. Thus, robust and reliable biomarkers/models which could select the appropriate patient for immune checkpoint antibodies are urgently needed. Currently, the clinical application of each FDA-approved PD-1/PD-L1 antibody is dependent on the PD-L1 immunohistochemistry assay results [11]. However, in the Checkmate025 study, the responses to Nivolumab (PD-1 antibody) had no correlations with PD-L1 level, and patients with the high level of PD-L1 had a worse prognosis [10]. Besides, PD-L1 seems to be a dynamic biomarker since its expression could be largely variable after

therapies such as mTOR inhibitors [11]. On the other hand, there are several different subtypes with distinct clinical behaviors and drug response rates, and a variety of genetic alterations among ccRCC patients [12]. Thus, identifying the potential immune subtypes might contribute to the personalized medicine, reduction in cost and improved survival rate in ccRCC patients.

In the current study, multiple genomic data of primary ccRCC samples were downloaded to identify immune subtypes that are more sensitive or resistant to the immunotherapy. An independent cohort that contained patients treated with immunotherapy was used to validate the correlation of immune subtypes with drug response and prognosis. After that, a web server based on a machine learning model was constructed to provide the prediction of the immune subtype of kidney cancer samples by the mRNA expression of four genes. Besides, novel therapeutic targets and drugs need to be provided for the subtype that is resistant to immunotherapy.

## 2. Materials And Methods

## 2.1 Data acquisition

In the current study, 7 independent cohorts were retrieved: (i) GSE15641 (32 ccRCC samples) [13], GSE36895 (29 ccRCC samples) [14], GSE40435 (101 ccRCC samples) [15], GSE46699 (67 ccRCC samples) [16], GSE53757 (72 ccRCC samples) [17] and TCGA (530 ccRCC samples) [18] were used for training datasets. (ii) IMvigor210 [19] study which contained 348 cancer patients treated with Atezolizumab (anti-PD-L1) were taken as the testing dataset. The expression matrix and the corresponding clinical information of these cohorts were downloaded.

## 2.2 Calculation of immune cell infiltration levels

The algorithm Single Sample Gene Set Enrichment Analysis (ssGSEA), an extension of Gene Set Enrichment Analysis (GSEA) method, could compute the specific cell enrichment scores by the cell-specific-genes such as immune cell-specific genes. In the current study, immune cell-specific marker genes were downloaded from the supplementary data of the previous article [20]. A total of 28 immune cell enrichment scores were calculated by the ssGSEA method from ‘GSVA’ package in R language [21]. Then, we normalized the immune cell enrichment scores by the equation x=(x−xmin)/(xmax−xmin), where, xmin and xmax denoted the minimum and maximum of the score.

## 2.3 The assignment of immune subtypes

Consensus clustering (CC) could find the potential clusters/subtypes within the RNA sequencing dataset and assess the stability of these subtypes [22]. The normalized immune cell enrichment score was eligible for CC analysis since the normalized data was necessary for CC analysis. In the current study, the package ‘ConsensusClusterPlus’ [23] in R language was implemented for CC analysis. The key parameters for were set as following: maxK=6, clusterAlg="hc", distance="pearson". Once CC analysis was performed and the final clusters (immune subtypes) were generated, the proper number of final clusters (K) could be estimated by commonly used methods including tracking plot, cumulative density function, and relative change in area under cumulative density function [24].

## 2.4 Differentially expressed genes (DEGs) and enrichment analysis

Based on ‘edgeR’ package, [25] fold-change (FC) and P.value for each mRNA were obtained. Then, the Benjamini and Hochberg method was used to calculate the adjusted P-values. The significant DEGs were characterized by FDR<0.05 and |Log2 (FC)|>1. The GSEA software was downloaded and the enrichment analysis [26] was conducted in TCGA dataset between immune subtypes. In the current study, enrichment analysis was completed on the reference gene sets (c2.cp.kegg.v6.1.symbols.gmt) that come from the Molecular Signatures Database.

## 2.5 Construction of a Co-expression Network

In order to identify the modules and genes which are highly associated with the obtained ccRCC immune subtypes, a co-expression network that contains the genes (points) and their correlations (lines) was built by WGCNA method from 'WGCNA' package of R language [27]. In the current study, only immune subtype DEGs (1136 genes) obtained in the last step were selected for WGCNA analysis since the necessary calculation resources would be reduced and the modules with higher correlations with the immune subtype would be found. The construction steps of the network in this study contained: 1) filtering outliers and bad samples; 2) selecting the β value to ensure a scale-free network; 3) calculating the correlation matrix; 4) setting the minimum size of a module; 5) calculating the relationships between modules and immune subtypes. The module that had the strongest association with the immune subtype was selected for further analysis.

## 2.6 Construction a random forest model for predicting immune subtypes

Random forest (RF) model, one of the most accurate supervised learning methods, was used to construct a model for prediction of immune subtypes. The data used for the random forest model was the hub genes expression matrix of TCGA samples. Step1: the input data was randomly separated into the training dataset (70%) and testing dataset (30%). Step2: the best parameters for random forest model were selected by 5-fold cross-validation in the training dataset. Step3: After setting the best parameter, the prediction accuracy of random forest model for the immune subtype prediction was tested in the testing dataset. Step4: An independent dataset (IMvigor210) was selected for the validation of the correlation of predicted immune subtype with the immunotherapy response rates and prognosis [19].

## 2.7 Construction a use-friendly website for immune subtype prediction

Shiny is a framework from R language and could build web applications. In the current study, the machine learning (Random forest) model was implemented in 'Rshiny' package in R language. The web application was named as Immune Subtype Prediction by Random Forest (ISPRF) and could be accessed via the URL (https://immunotype.shinyapps.io/ISPRF/). ISPRF App could be freely used by any user or organization without limitations. The ISPRF App has been tested in different environments (Linux, Windows, and Mac OS) and is also compatible with popular web browsers such as Chrome, Firefox, and Internet Explorer.

## 2.8 Potential drugs for the immunotherapy-resistant subtype

Drugs, targeting immunotherapy-resistant subtype hub genes, were selected by using the Drug-Gene Interaction Database (DGIdb; https://www.dgidb.org/) [28]. For the drugs from DGIdb, only the FDA-approved drugs were retained. Discovery Studio software could predict the pharmacologic properties of small molecules. These pharmacologic properties including aqueous solubility level, blood-brain barrier level, CYP2D6 binding, hepatotoxicity, human intestinal absorption level, and plasma protein binding properties, directly determine the viability of a drug candidate [29].

## 2.9 Statistical analysis

Overall survival (OS) plus progression-free survival (PFS) were selected to compare survival time between groups using Kaplan–Meier model in R package 'survival' [30]. We obtained immune cytolytic activity (CYT) according to the summation of two cytolytic effectors' expression value (GZMA and PRF1) [31]. For analyzing the levels of immunotherapy indicators in different immune subtypes, the Wilcoxon rank-sum test was used to compare the average value.

# 3. Results

## 3.1 Molecular immune subtypes in ccRCC patients

Six datasets, GSE15641 (32 ccRCC samples), GSE36895 (29 ccRCC samples), GSE40435 (101 ccRCC samples), GSE46699 (67 ccRCC samples), GSE53757 (72 ccRCC samples) and TCGA (530 ccRCC samples) were downloaded. The gene expression matrix of these six datasets was used to calculate the immune cells enrichment scores (ssGSEA score) by adopting ssGSEA method. The

survival analysis of immune cells enrichment scores in TCGA dataset showed that 6 immune cells including activated CD4 T cells, activated CD8 T cells were significant survival-related biomarkers, and the patients with high levels of them had worse OS than those in low levels group (Supplementary Fig. 1). The PCA results of these six datasets gene expression matrix indicated the obvious batch effects since these datasets displayed a significant difference (Supplementary Fig. 2A). But the PCA results of immune cells enrichment scores of these six datasets (normalized ssGSEA score) demonstrated that differences among datasets were eliminated (Supplementary Fig. 2B).

Consensus clustering of a total of 831 ccRCC patients from six datasets using immune cells enrichment scores were performed. Two main immune subtypes were identified and named as subtype1/subtype2 (Fig. 1A). The tracking plots showed that 2 was the best value of subtypes number (Fig. 1B) while cumulative distribution function (CDF) results indicated 3 (Supplementary Fig. 3A-B). Since the sample numbers in subtype3 to subtype6 were too small (Fig. 1B), two main immune subtypes were finally identified. The distribution of immune subtypes (subtype1 and subtype2) among different datasets was shown in Table 1. As shown in Fig. 1C-D, both in the overall survival (OS) and progression-free survival (PFS) analysis, differences in the survival curves between the subtype1 and subtype2 were statistically significant (P.value = 0.027 and P.value = 0.014, respectively). Patients in subtype1 had a better prognosis than subtype2 patients. Subsequently, ssGSEA scores indicated that subtype2 samples were highly infiltrated with innate and adaptive immune cells including B cells, CD8 T cells, CD4 T cells, macrophages, NK cells, and regulatory T cells (Tregs), while subtype1 samples only showed a high level of neutrophils (Fig. 2). Besides, some immune checkpoint therapy biomarkers such as CD8A, PDL1, PD1, and tumor mutational burden (TMB) were also enriched in subtype2 (Supplementary Fig. 4).

Table 1
The distribution of immune subtypes among different datasets.

| | GSE15641 | GSE36895 | GSE40435 | GSE46699 | GSE53757 | TCGA |
|---|---|---|---|---|---|---|
| ccRCC samples | 32 | 29 | 101 | 67 | 72 | 530 |
| Subtype1 | 18 | 15 | 46 | 38 | 36 | 301 |
| Subtype2 | 14 | 14 | 55 | 29 | 36 | 229 |

## 3.2 DEGs and enriched pathways between immune subtypes

DEGs between immune subtypes were analyzed. A total of 614 DEGs were found to be highly expressed in subtype2 and 522 DEGs were defined as down-regulated DEGs in subtype2. The volcano plot of TCGA cohort was shown in Supplementary Figure 5. Metabolism pathways such as *oxidative phosphorylation, fatty acid metabolism, retinol metabolism* and *tyrosine metabolism* were mostly enriched in subtype1 (Supplementary Table 1). On the other hand, Immune-related pathways, involving *natural killer cell mediated cytotoxicity*, *T cell receptor signaling pathway*, *antigen processing and presentation* were mostly enriched in subtype2 (Supplementary Table 2).

## 3.3 Construction of Co-Expression Network

The TCGA dataset was selected for WGCNA since the clinical information of other datasets were not available, and the expression matrix of DEGs from TCGA was used to construct the co-expression network. Based on the results of scale-free topology fitting indices R2 and mean connectivity (Figure 3A-B), the best value of β was 3 since it could construct a scale-free network. A total of 11 different modules, ranging in size from 30 to 525 genes, was provided by WGCNA results (Figure 3C). Among these modules, the turquoise module was selected since it had the highest correlation value with the immune subtype (correlation = 0.62; P.value < 0.01) (Figure 3D). Besides, the blue module was also selected since it had a significantly negative value of correlation with the immune subtype (correlation = -0.55; P.value < 0.01) (Figure 3D).

## 3.4 Identification of protein-protein interactions (PPIs) and hub genes

The PPI networks of the turquoise module and the blue module were individually retrieved from the STRING database and then visualized by Cytoscape software. Subsequently, ten hub genes (FOXP3, CTLA4, PTPRC, CD28, CD19, LCK, CD27, CD2, IFNG, and CD5) from the network of the turquoise module were selected with the cut-off value of Degree > 10, using the cytoHubba plug of Cytoscape (Figure 4A). The survival analysis results of these hub genes revealed that high expression levels of CTLA4, FOXP3, IFNG, and CD19 were associated with the worse overall survival (Supplementary Figure 6). Similarly, ten hub genes (AGTR1, CRP, G6PC, IGFBP1, MGAM, PCK1, PLG, REN, SLC5A1, and WT1) from the network of the blue module were identified with the cut-off value of Degree > 4 (Figure 4B). The survival analysis result revealed that high expression levels of three genes (CRP, IGFBP1, and WT1) and low expression levels of seven genes (AGTR1, G6PC, MGAM, PCK1, PLG, REN, and SLC5A1) were associated with the worse overall survival (Supplementary Figure 7).

## 3.5 Validation of immune subtypes in the independent cohort

The two subtypes were further validated in an external cohort of IMvigor210, using a random forest model. The hub genes (CTLA4, FOXP3, IFNG, and CD19) from turquoise module were selected to construct a random forest model for predicting immune subtype by gene expression. IMvigor210 contains 348 tumor patients who received treatment with the immune checkpoint inhibitor therapy (Atezolizumab). Clinical data of these 348 tumor patients were described in Supplementary Table 3.

The pipeline of machine learning (random forest) workflow was plotted in Figure 5. In the training phase, the input data (TCGA dataset samples with their subtype information and four genes expression matrix) was randomly separated into the training dataset (70%) and the testing dataset (30%). The parameter tuning results showed that 2 and 300 were the best value for ‘mtry’ and ‘ntree’ because of their highest value of AUC (Figure 6A-B). The random forest is then trained with the best parameter (mtry=2, ntree=300). In the testing phase, the AUC value in the testing dataset indicated a good prediction performance with 0.78 (Figure 6C). Subsequently, the immune subtype of patients from IMvigor210 cohort was predicted by their expression data profiles of four hub genes (CTLA4, FOXP3, IFNG, and CD19). Patients in subtype2 behaved better overall response rate to Atezolizumab, about 29%, whereas subtype1 worst ORR, about 16% (Figure 6D). The overall survival analysis results in IMvigor210 cohort confirmed that patients with subtype2 had better prognosis than subtype1 patients (Figure 6E). Consistent with results from TCGA cohort, subtype2 in the IMvigor210 cohort were characterized as high expression of various immunotherapy indicators (CD8A, PDL1, TIGIT, CTLA4, CYT, IFNG, LAG3, PDCD1, TMB) in Supplementary Figure 8.

## 3.6 Web tool development

Using the RStudio shiny package, a web application (https://immunotype.shinyapps.io/ISPRF/) was built for the prediction of immune subtypes. In this web application, expression profiles of four hub genes (CTLA4, FOXP3, IFNG, and CD19) were required as the input data (Figure 7A). The input data will be sent to the servers where the application pre-processes the data, including four steps: (1) combining the input data with the training dataset; (2) transforming the matrix into the one-hot matrix by the median value of each gene; (3) deleting the training dataset. 4) after pre-processing the input data, this application predicts the probability of immune subtypes using the random forest model. Then, the immune subtype that results in the highest probability is picked as the predicted immune subtype. In Figure 7B, the interface shows an example of predicting the immune subtype by four genes expression.

## 3.7 Identification of potential drugs for the immunotherapy-resistant subtype

Since patients from immune subtype1 might have a low response rate to immune checkpoint antibodies, some potential drugs are needed. Thus, hub genes in the blue module which had a negative correlation with the immune subtype could be the targets for the identification of potential drugs. According to the above significantly survival prognosis results, three hub genes from the blue module (CRP, IGFBP1, and WT1) were selected for further analysis because of their negative effect on the prognosis. A total of 6

small molecules, 1 monoclonal antibody, and 1 synthetic peptide for targeting these hub genes were provided by the DGIdb website that contained drug-gene interactions (Table 2). Pharmacologic properties of 6 small molecule drugs were unearthed under Discovery Studio 2019 software (Table 3): 1) the aqueous solubility results showed that no drug was characterized with low aqueous solubility ability; 2) only one drug was high penetrant for blood-brain barrier; 3) all small molecules drugs were non-inhibitor of CYP2D6 which was responsible for drug metabolism; 4) 2 drugs, ZINC150338696 and ZINC169294721, were non-toxic drugs based on hepatotoxicity prediction results; 5) 1 drug had the good intestinal absorption level; 6) as to plasma protein binding, 3 drugs were predicted to be absorbent strong. Based on the above results, ZINC169294721 was selected as the potential small molecule drug for immune subtype1 patients since it had the good aqueous-solubility ability and was non-toxic.

Table 2
The drugs from Drug-Gene Interaction Database (DGIdb).

| Gene | Drug | Zinc ID | Type | Interaction | PMID |
|---|---|---|---|---|---|
| CRP | Adalimumab | Not available | Monoclonal antibody | Inhibitor | 23517933 |
| CRP | Fenofibrate | ZINC584092 | Small Molecule | Inhibitor | 21939559 |
| CRP | Rosuvastatin | ZINC1535101 | Small Molecule | Inhibitor | 21641360 |
| WT1 | Sirolimus | ZINC169294721 | Small Molecule | Inhibitor | 18927120 |
| WT1 | Daunorubicin | ZINC3917708 | Small Molecule | Inhibitor | 30837363 |
| IGFBP1 | Buserelin | Not available | Synthetic peptide | Promoter | 1721621 |
| IGFBP1 | Octreotide | ZINC150338696 | Small Molecule | Inhibitor | 9604870 |
| IGFBP1 | Streptozocin | ZINC3995968 | Small Molecule | Promoter | 1698152 |

Table 3
The pharmacologic properties of drugs. BBB: blood-brain barrier; CYP2D6: cytochrome P-450 2D6; PPB: plasma protein binding.

| Compounds | Solubility level | BBB level | CYP2D6 | Hepatotoxicity | Absorption level | PPB level |
|---|---|---|---|---|---|---|
| ZINC3917708 | 2 | 4 | 0 | 1 | 3 | 0 |
| ZINC150338696 | 1 | 4 | 0 | 0 | 3 | 0 |
| ZINC3995968 | 4 | 4 | 0 | 1 | 3 | 0 |
| ZINC584092 | 2 | 1 | 0 | 1 | 0 | 1 |
| ZINC1535101 | 3 | 4 | 0 | 1 | 2 | 1 |
| ZINC169294721 | 3 | 4 | 0 | 0 | 3 | 1 |
| Aqueous-solubility level: 0, extremely low; 1, very low, but possible; 2, low; 3, good. | | | | | | |
| BBB level: 0, very high penetrant; 1, high; 2, medium; 3, low; 4, undefined. | | | | | | |
| CYP2D6 level: 0, noninhibitor; 1, inhibitor. | | | | | | |
| Hepatotoxicity: 0, nontoxic; 1, toxic. | | | | | | |
| Human-intestinal absorption level: 0, good; 1, moderate; 2, poor; 3, very poor. | | | | | | |
| PPB: 0, absorbent weak; 1, absorbent strong. | | | | | | |

## Discussion

Immunotherapies such as PD-1/PD-L1 antibodies are thought as promising tumor intervention methods since immunotherapies prolonged the overall survival time of ccRCC patients in different clinical trials [32]. However, the response rate of cancer patients to immunotherapy is still limited and unsatisfactory [33]. Since the tumor heterogeneity exists among cancer samples, identifying

the potential immune subtypes with different immunotherapy drug responses might contribute to the individualized immunotherapy treatment. Currently, some indicators including PD-L1 and tumor mutational burden (TMB) were recommended for selecting appropriate immunotherapy candidates [34]. However, PD-L1 is a dynamic biomarker since its expression could be remodeled by the use of antiangiogenic drugs [35]. TMB, defined as the total number of nonsynonymous mutations per coding area of a tumor genome, is determined using whole exome sequencing which is costive and needs long turnaround time [36]. Thus, robust biomarkers and prediction models for selecting the patients for immune checkpoint therapies are urgently needed.

In the current study, we pooled together data from TCGA and GEO datasets to enlarge our sample size, and used immune cells enrichment scores to successfully eliminate the batch effect among different datasets. Using the immune cells enrichment scores from multiple datasets (a total of 831 samples) and the consensus clustering method, we subdivided the ccRCC samples into two immune subtypes. These two immune subtypes were named as subtype1 and subtype2, demonstrating distinct prognosis. In the TCGA dataset, with the surgical treatment, subtype1 patients had a better prognosis than subtype2 patients. The immune-related characteristics or immunotherapy biomarkers, including T-cell cytolytic activity, immune checkpoints, and active IFN signaling, were significantly higher in subtype2. Thus, subtype2 patients were recommended to receive the treatment of immunotherapy.

Usually, the prediction of drug response requires robust models based on a substantial number of samples, effective biomarkers, and efficient computational tools. In the current study, we built a random forest model to predict the immune subtype by inputting the expression levels of only four mRNAs. The model indicated a good prediction performance in the testing dataset by the value of AUC (0.78). Moreover, the drug response and prognosis of subtype2 patients in the validation cohort that contains patients treated with immune checkpoint inhibitors, were better than subtype1 patients. We have developed the open and free online website of Immune Subtype Prediction by Random Forest (ISPRF) to make the random forest model available for organizations or individuals. The ISPRF offers an appropriate framework to employ machine learning algorithms on four mRNAs expression to predict the immune subtype of a patient, and thus provide the advice for the immunotherapy treatment choice.

To identify more potential drugs for immunotherapy-resistant subtype, 8 candidate drugs were obtained from the prediction of the DGIdb dataset depending on hub genes. Among the 8 drugs, Sirolimus was considered to be the most promising drug. Sirolimus (rapamycin) is a macrolide and is usually produced by Streptomyces hygroscopicus. Sirolimus could reduce the cell proliferative action by binding with FK-binding protein-12 and inhibiting mTORC1 [37]. Moreover, Sirolimus has the ability to inhibitor the growth of renal cancer cells [38]. For the target of Sirolimus, the expression of WT1 is extremely low in kidney normal epithelial cells but higher in kidney cancer cells [39]. Besides, WT1 has the ability to promote the survival of various cancer cells through anti-apoptotic functions [40].

Of note, the immune subtypes of ccRCC were constructed based on the TCGA and GEO cohorts which were treated with surgery and did not receive immune checkpoint therapies. Although we validated the identified immune subtypes in an independent cohort (IMvigor210), these two immune subtypes still require to be tested in clinical trials that concentrate on the correlation of immune subtypes and drug responses. Besides, the hub genes and the selected drug for immunotherapy-resistant subtype need validation by experiments, and it will be implemented in our future practice and research.

## Conclusion

We identified two ccRCC immune subtypes that have distinct clinical behavior and prognosis. Furthermore, a machine learning model to predict the ccRCC immune subtype by four mRNAs expression was constructed, and the model was also implemented in the online website and available for organizations and individuals. Our study has important clinical significance and the model could be taken as a reference by clinicians for individualized treatment.

## Abbreviations

ccRCC, Clear cell renal cell carcinoma; WGCNA, Weighted correlation network analysis; AUC, areas under the curve of the receiver operating characteristic; RCC, Renal cell carcinoma; FDA, Food and Drug Administration; OS, Overall Survival; IHC, Immunohistochemistry; TCGA, The Cancer Genome Atlas; ssGSEA, Single Sample Gene Set Enrichment Analysis; CC, Consensus clustering CC; FC, fold-change; MEs,Module Eigengenes; RF, Random forest; ISPRF, Immune Subtype Prediction by Random Forest;

DGIdb, Drug-Gene Interaction Database; ADME, Adsorption, Distribution, Metabolism, and Excretion; CYT, cytolytic activity; PPI, Protein-protein interactions; TMB, Tumor Mutational Burden.

## Declarations

# Ethical Approval and Consent to participate

All the expression data and clinical information were retrieved from publicly available datasets which were free to download and analyze without limitations. Investigators of each study obtained the approval from their local ethics committee and informed patient consent.

# Consent for publication

Not applicable.

# Availability of supporting data

The datasets generated and/or analysed during the current study are available in the github

Repository, https://github.com/degang123/ISPRF.

# Competing interests

The authors state that they have no conflicts of interest.

# Funding

The project was supported by Early diagnosis and recurrence monitoring of upper urothelial tumors based on non-invasive urine genomics (Science and Technology Research of Henan Provincial Health and Health Commission SBGJ202002002).

# Author contributions

Zhifeng Wang and Zihao Chen performed the study and wrote the manuscript; Hongfan Zhao Hao Lin, and Ning Wang contributed to data preparation; Junjie Wang and Xiqing Li performed technical modification; Degang Ding conceived the study. All authors read and approved the final manuscript.

# Acknowledgements

Not applicable.

# Authors' information

[1]Department of Urology, Henan Provincial People's Hospital, Zhengzhou, 450003, China. [2]Department of Urology, Nanfang Hospital, Southern Medical University, Guangzhou, 510515, China. [3]Department of Nephrology and Medical Intensive Care, Charité-Universitätsmedizin, Berlin, Germany. [4]Oncology Department, Henan Provincial People's Hospital, Zhengzhou, 450003, China.

## References

1. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, Heng DY, Larkin J, Ficarra V: **Renal cell carcinoma**. *NAT REV DIS PRIMERS* 2017, **3**(1):17009.
2. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2019**. *CA Cancer J Clin* 2019, **69**(1):7-34.
3. Waalkes S, Kramer M, Herrmann TR, Schrader AJ, Kuczyk MA, Merseburger AS: **Present state of target therapy for disseminated renal cell carcinoma**. *IMMUNOTHERAPY-UK* 2010, **2**(3):393-398.
4. Caceres W, Cruz-Chacon A: **Renal cell carcinoma: molecularly targeted therapy**. *P R HEALTH SCI J* 2011, **30**(2):73-77.
5. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2018**. *CA Cancer J Clin* 2018, **68**(1):7-30.
6. Barata PC, Rini BI: **Treatment of renal cell carcinoma: Current status and future directions**. *CA Cancer J Clin* 2017, **67**(6):507-524.
7. Motzer RJ, Escudier B, McDermott DF, George S, Hammers HJ, Srinivas S, Tykodi SS, Sosman JA, Procopio G, Plimack ER *et al*: **Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma**. *N Engl J Med* 2015, **373**(19):1803-1813.
8. Rini BI, Plimack ER, Stus V, Gafanov R, Hawkins R, Nosov D, Pouliot F, Alekseev B, Soulieres D, Melichar B *et al*: **Pembrolizumab plus Axitinib versus Sunitinib for Advanced Renal-Cell Carcinoma**. *N Engl J Med* 2019, **380**(12):1116-1127.
9. Motzer RJ, Penkov K, Haanen J, Rini B, Albiges L, Campbell MT, Venugopal B, Kollmannsberger C, Negrier S, Uemura M *et al*: **Avelumab plus Axitinib versus Sunitinib for Advanced Renal-Cell Carcinoma**. *N Engl J Med* 2019, **380**(12):1103-1115.
10. Shen X, Zhao B: **Efficacy of PD-1 or PD-L1 inhibitors and PD-L1 expression status in cancer: meta-analysis**. *BMJ* 2018, **362**:k3529.
11. Lopez-Beltran A, Henriques V, Cimadamore A, Santoni M, Cheng L, Gevaert T, Blanca A, Massari F, Scarpelli M, Montironi R: **The Identification of Immunological Biomarkers in Kidney Cancers**. *FRONT ONCOL* 2018, **8**:456.
12. Chen F, Zhang Y, Senbabaoglu Y, Ciriello G, Yang L, Reznik E, Shuch B, Micevic G, De Velasco G, Shinbrot E *et al*: **Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma**. *CELL REP* 2016, **14**(10):2476-2489.
13. Jones J, Otu H, Spentzos D, Kolia S, Libermann TA: **Gene Signatures of Progression and Metastasis in Renal Cell Cancer**. *AKTUEL UROL* 2006, **37**(S 1).
14. Pe A-Llopis S, Brugarolas J: **Simultaneous isolation of high-quality DNA, RNA, miRNA and proteins from tissues for genomic applications**. *Nature Protocol* 2013, **8**(11):2240-2255.
15. Wozniak MB, Le Calvez-Kelm F, Abedi-Ardekani B, Byrnes G, Durand G, Carreira C, Michelon J, Janout V, Holcatova I, Foretova L *et al*: **Integrative Genome-Wide Gene Expression Profiling of Clear Cell Renal Cell Carcinoma in Czech Republic and in the United States**. *PLOS ONE* 2013, **8**(3):e57886.
16. Eckel-Passow JE, Serie DJ, Bot BM, Joseph RW, Parker AS: **Somatic Expression of ENRAGE is Associated with Obesity Status among Patients with Clear Cell Renal Cell Carcinoma**. *CARCINOGENESIS* 2013, **35**(4):822.
17. Roemeling CAV, Radisky DC, Marlow LA, Cooper SJ, Grebe SK, Anastasiadis PZ, Tun HW, Copland JA: **Neuronal Pentraxin 2 Supports Clear Cell Renal Cell Carcinoma by Activating the AMPA-Selective Glutamate Receptor-4**. *CANCER RES* 2014, **74**(17):4796.
18. Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, Bowlby R, Gibb EA, Akbani R, Beroukhim R *et al*: **The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma**. *CELL REP* 2018, **23**(1):313-326.
19. Rosenberg, J., E., Galsky, M., D., Balar, A.: **Atezolizumab as first-line treatment in cisplatin-ineligible patients with locally advanced and metastatic urothelial carcinoma: a single-arm, multicentre, phase 2 trial (vol 389, pg 67, 2017)**. *The Lancet* 2017, **390**(10097):848.
20. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H, Trajanoski Z: **Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade**. *CELL REP* 2017, **18**(1):248-262.
21. Hänzelmann S, Castelo R, Guinney J: **GSVA: gene set variation analysis for microarray and RNA-Seq data**. *BMC BIOINFORMATICS* 2013, **14**(1):7.
22. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data**. *MACH LEARN* 2003, **52**(1-2):91-118.

23. Wilkerson MD, Hayes DN: **ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking**. *BIOINFORMATICS* 2010, **26**(12):1572-1573.
24. Y Ş, Michailidis G, Li JZ: **Critical limitations of consensus clustering in class discovery**. *SCI REP-UK* 2014, **4**:6207.
25. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *BIOINFORMATICS* 2010, **26**(1):139-140.
26. Subramanian, A.: **From the Cover: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. 2005, **102**(43):15545-15550.
27. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC BIOINFORMATICS* 2008, **9**:559.
28. Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, Wollam A, Spies NC, Griffith OL, Griffith M: **DGIdb 3.0: a redesign and expansion of the drug-gene interaction database**. *NUCLEIC ACIDS RES* 2018, **46**(D1):D1068-D1073.
29. Andrade EL, Bento AF, Cavalli J, Oliveira SK, Schwanke RC, Siqueira JM, Freitas CS, Marcon R, Calixto JB: **Non-clinical studies in the process of new drug development - Part II: Good laboratory practice, metabolism, pharmacokinetics, safety and dose translation to clinical studies**. *BRAZ J MED BIOL RES* 2016, **49**(12):e5646.
30. Lin H, Zelterman D: **Modeling Survival Data: Extending the Cox Model**. *TECHNOMETRICS* 2000, **44**(1):85-86.
31. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N: **Molecular and genetic properties of tumors associated with local immune cytolytic activity**. *CELL* 2015, **160**(1-2):48-61.
32. Hahn AW, Drake C, Denmeade SR, Zakharia Y, Maughan BL, Kennedy E, Link CJ, Vahanian N, Hammers H, Agarwal N: **A Phase I Study of Alpha-1,3-Galactosyltransferase-Expressing Allogeneic Renal Cell Carcinoma Immunotherapy in Patients with Refractory Metastatic Renal Cell Carcinoma**. *ONCOLOGIST* 2020, **25**(2):121-213.
33. Bai R, Chen N, Li L, Du N, Bai L, Lv Z, Tian H, Cui J: **Mechanisms of Cancer Resistance to Immunotherapy**. *FRONT ONCOL* 2020, **10**:1290.
34. Spencer KR, Wang J, Silk AW, Ganesan S, Kaufman HL, Mehnert JM: **Biomarkers for Immunotherapy: Current Developments and Challenges**. *Am Soc Clin Oncol Educ Book* 2016, **35**:e493-e503.
35. Kammerer-Jacquet SF, Deleuze A, Saout J, Mathieu R, Laguerre B, Verhoest G, Dugay F, Belaud-Rotureau MA, Bensalah K, Rioux-Leclercq N: **Targeting the PD-1/PD-L1 Pathway in Renal Cell Carcinoma**. *INT J MOL SCI* 2019, **20**(7).
36. Melendez B, Van Campenhout C, Rorive S, Remmelink M, Salmon I, D'Haene N: **Methods of measurement for tumor mutational burden in tumor tissue**. *Transl Lung Cancer Res* 2018, **7**(6):661-667.
37. Sehgal SN: **Sirolimus: its discovery, biological properties, and mechanism of action**. *Transplant Proc* 2003, **35**(3 Suppl):7S-14S.
38. Bissler JJ, McCormack FX, Young LR, Elwing JM, Chuck G, Leonard JM, Schmithorst VJ, Laor T, Brody AS, Bean J *et al*: **Sirolimus for angiomyolipoma in tuberous sclerosis complex or lymphangioleiomyomatosis**. *N Engl J Med* 2008, **358**(2):140-151.
39. Campbell CE, Kuriyan NP, Rackley RR, Caulfield MJ, Tubbs R, Finke J, Williams BR: **Constitutive expression of the Wilms tumor suppressor gene (WT1) in renal cell carcinoma**. *INT J CANCER* 1998, **78**(2):182-188.
40. Hastie ND: **Wilms' tumour 1 (WT1) in development, homeostasis and disease**. *DEVELOPMENT* 2017, **144**(16):2862-2872.

# Figures

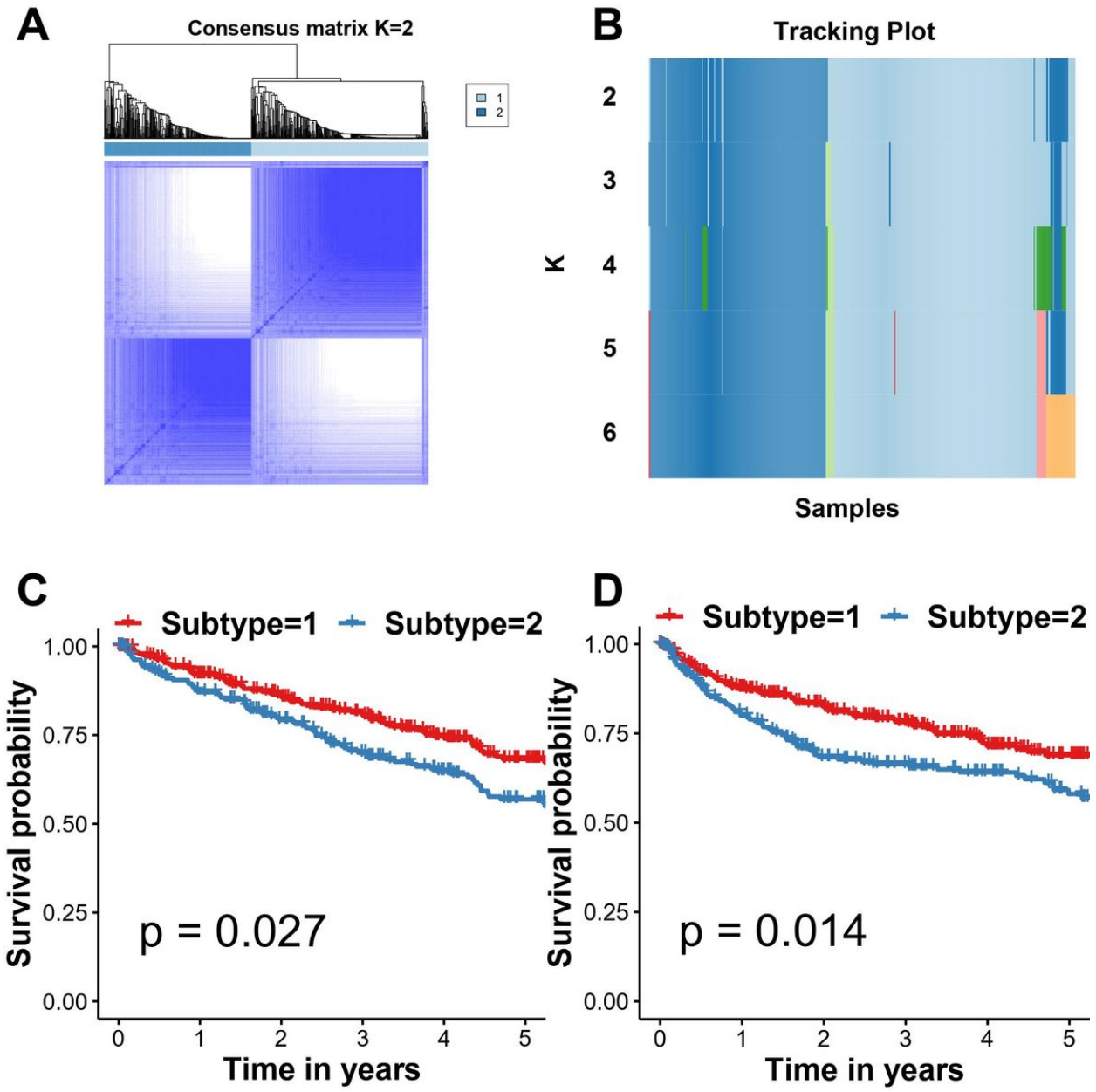
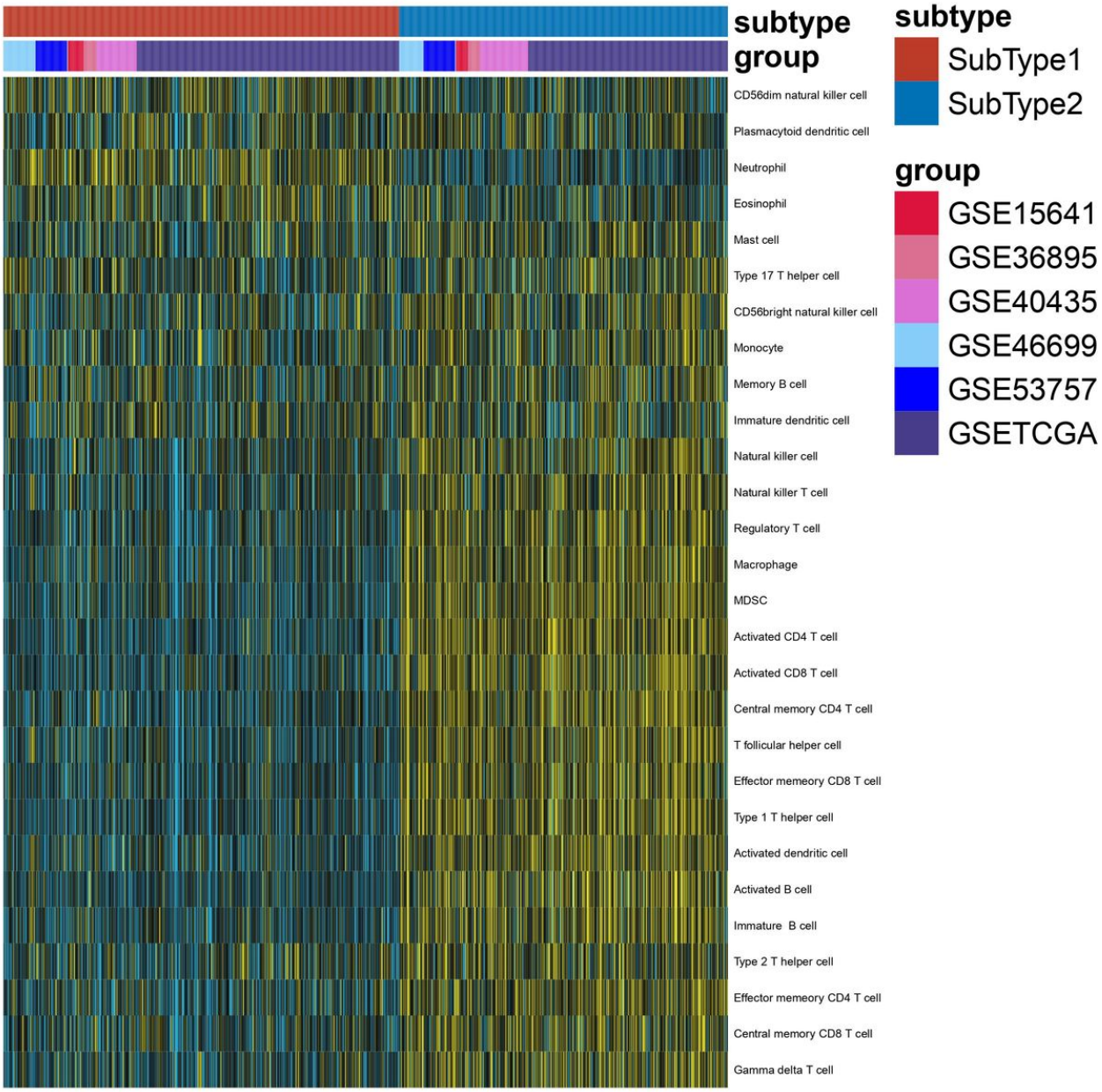



**Figure 1**

Consensus clustering for the ccRCC by combining 6 datasets GSE15641, GSE36895, GSE40435, GSE46699, GSE53757, and TCGA). (a) Consensus matrix heatmap plots when k=2. (b) Tracking plot for k=2 to 6. In the Tracking plot, the colors in each row represented the samples in different subtypes (c) Five-year Kaplan–Meier curves for OS of ccRCC patients stratified by the immune subtypes. (d) Five-year Kaplan–Meier curves for PFS of ccRCC patients stratified by the immune subtypes. P.value was calculated by the log-rank test among subtypes. Abbreviation: ccRCC, Clear cell renal cell carcinoma; TCGA, The Cancer Genome Atlas; OS, Overall Survival; PFS, Progression-free survival.




## Figure 2

The gene expression scores of 28 immune signatures in 2 subtypes are displayed by heatmap.

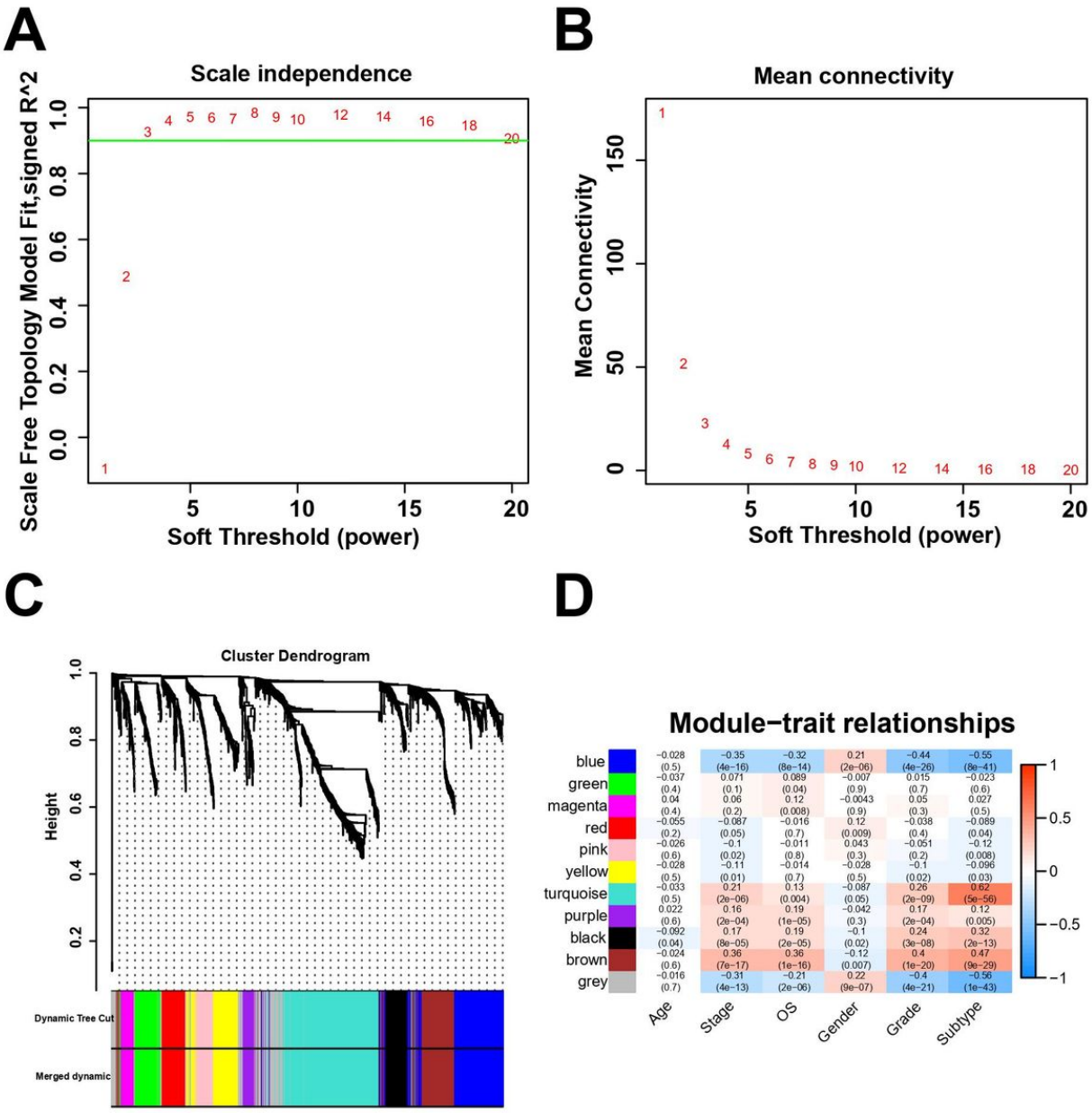



Figure 3

Identification of key modules connected with clinical features and immune subtypes through WGCNA. (a-b) The scale-free fit index and the mean connectivity for various soft-thresholding powers, respectively. When the soft-thresholding powers (β) equaled three, the average degree of connectivity was close to zero. (c) The cluster dendrogram of 5,000 module eigengenes from the TCGA dataset. Each branch in the figure represented one gene, and every color below represented one co-expression module. (d) Heatmap of the correlation between module eigengenes and clinical traits including molecular subtypes. The color of cells in the heatmap represented the correlation coefficients of different sizes. Specifically, red colors represented the positive correlations and green colors stood for the negative correlations. The figure without brackets in each cell indicated the clinical feature correlation coefficients. The corresponding p-value was shown below in parentheses. Abbreviation: WGCNA, Weighted correlation network analysis; TCGA, The Cancer Genome Atlas.

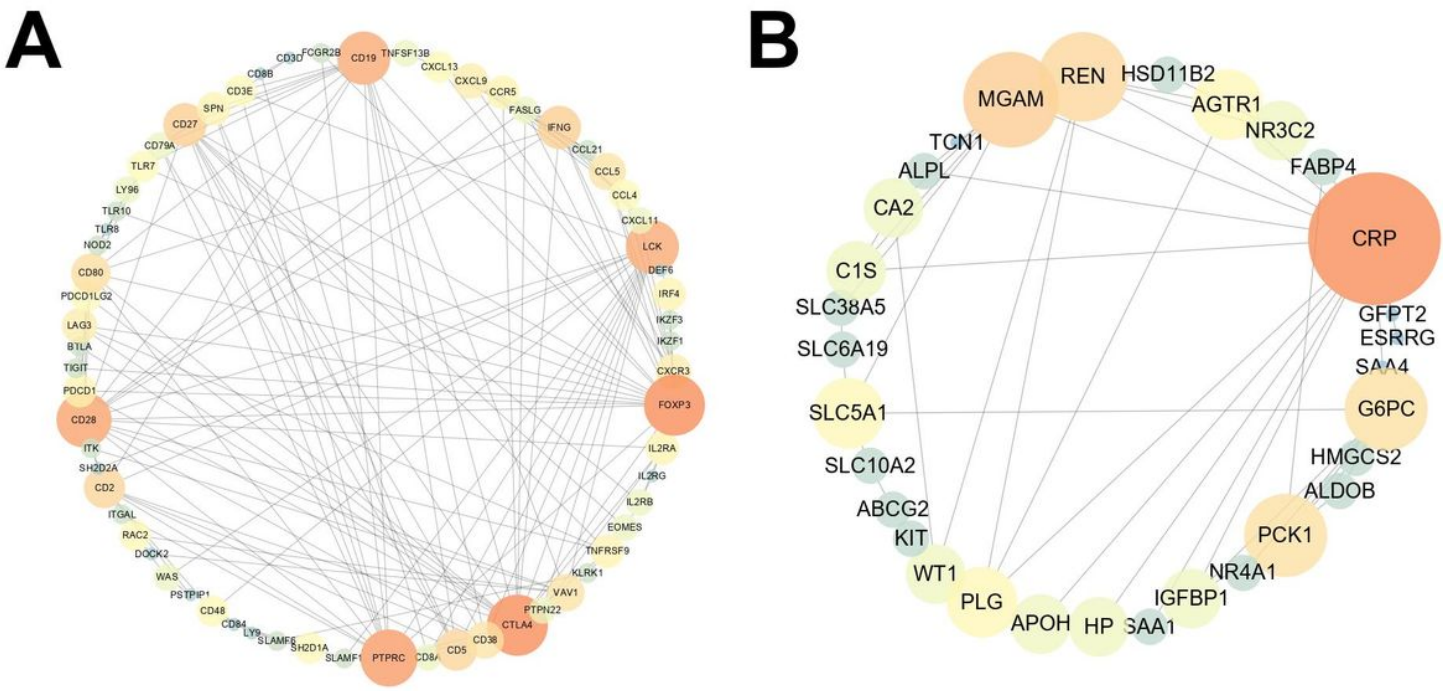



**Figure 4**

Protein–protein interaction network of genes in selected modules. The color intensity and the size of nodes were positively correlated with the degree score. (a) turquoise module; (b) blue module.

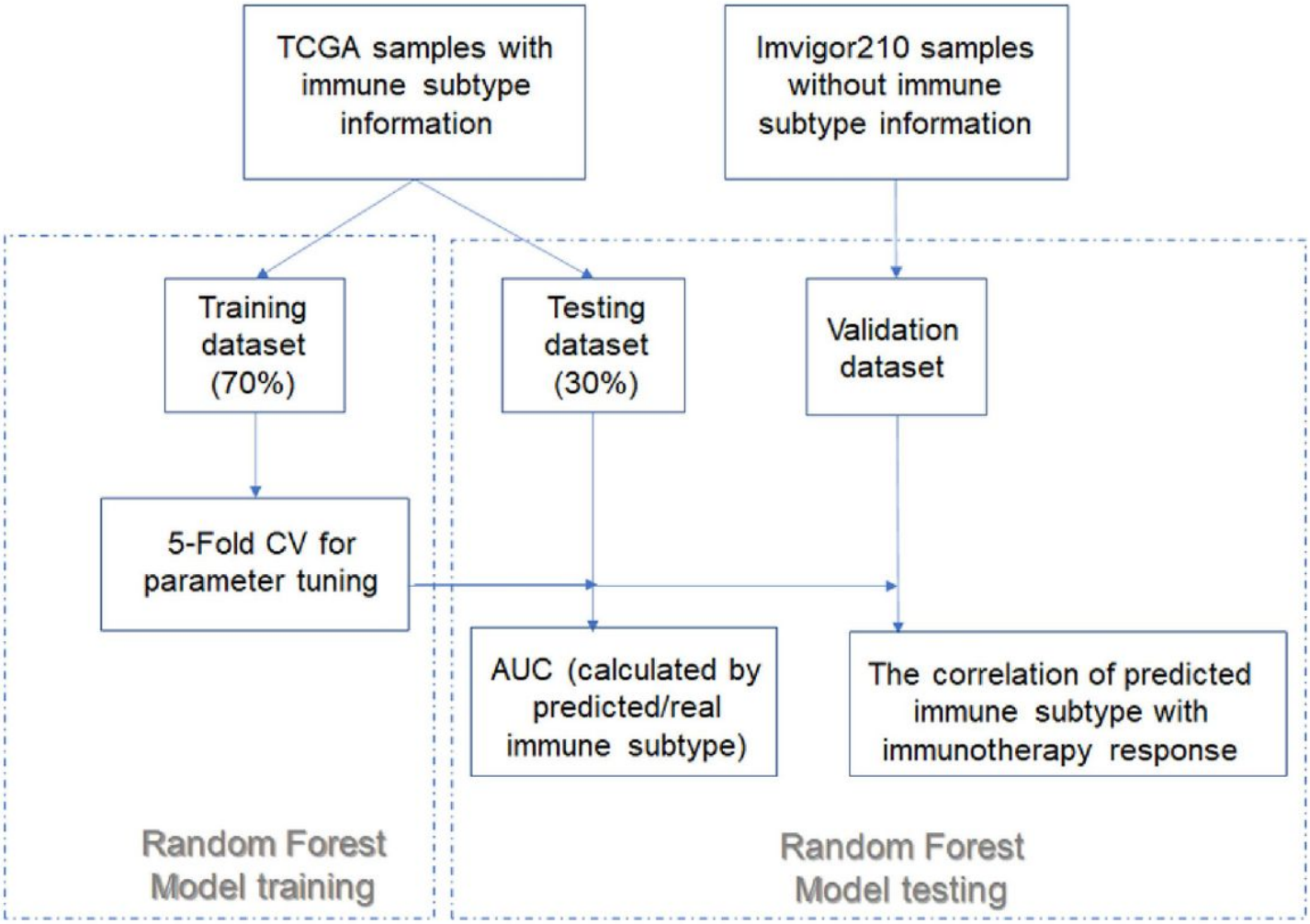
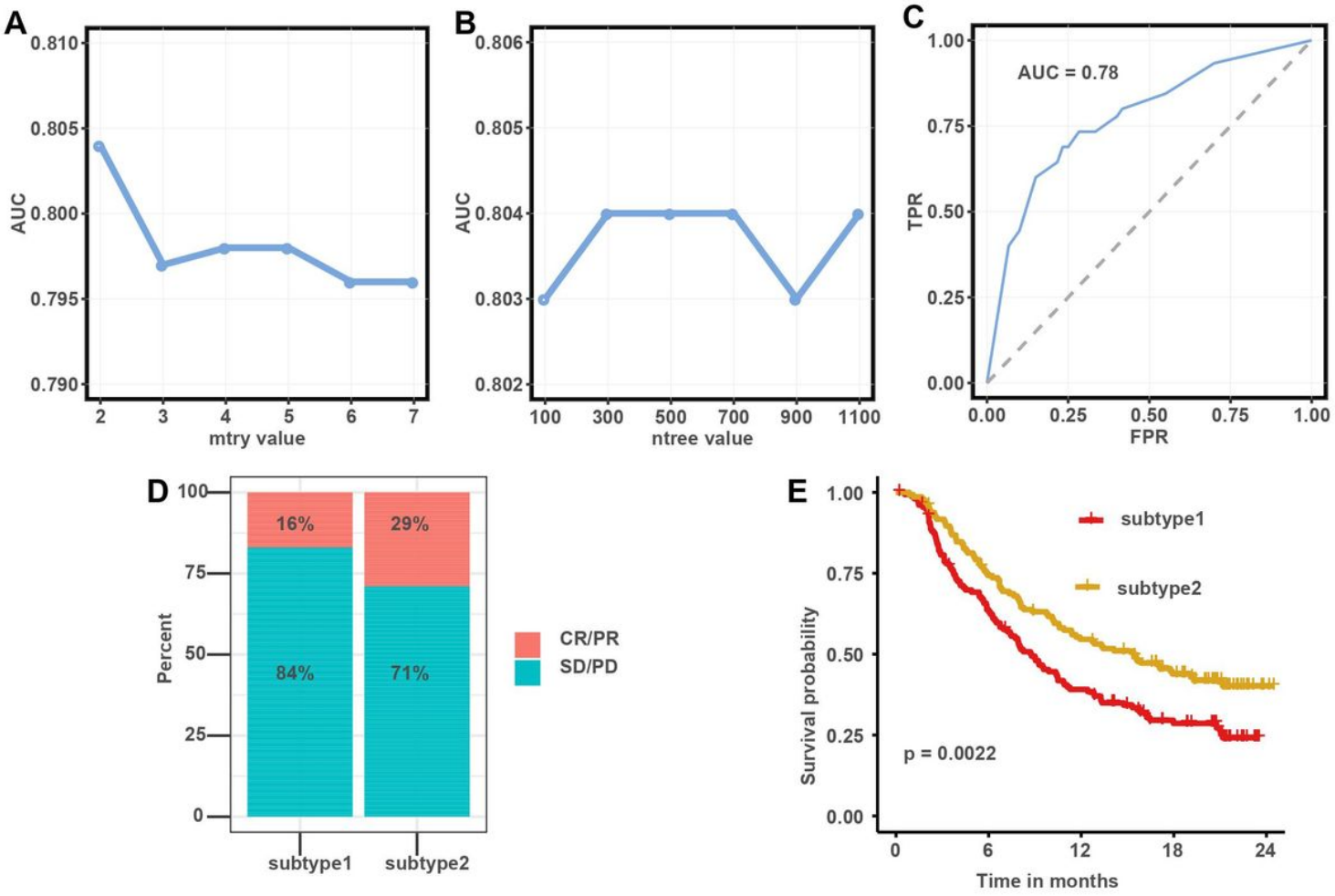



**Figure 5**

Pipeline of machine learning (random forest) workflow. Abbreviation: CV, cross-validation; AUC, Area under receiver operating characteristic Curve.




## Figure 6

Parameter tuning and model validation. (a) The 'mtry' with the highest of AUC was selected as the optimal value of the random forest algorithm. (b) The 'ntree' with the highest of AUC was selected as the optimal value of the random forest algorithm. (c) Validation of model in testing dataset. (d) The correlation of predicted immune subtype with the response rate to immunotherapy in IMvigor210 dataset. (e) The correlation of predicted immune subtype with the survival analysis in IMvigor210 dataset. Abbreviation: CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease;

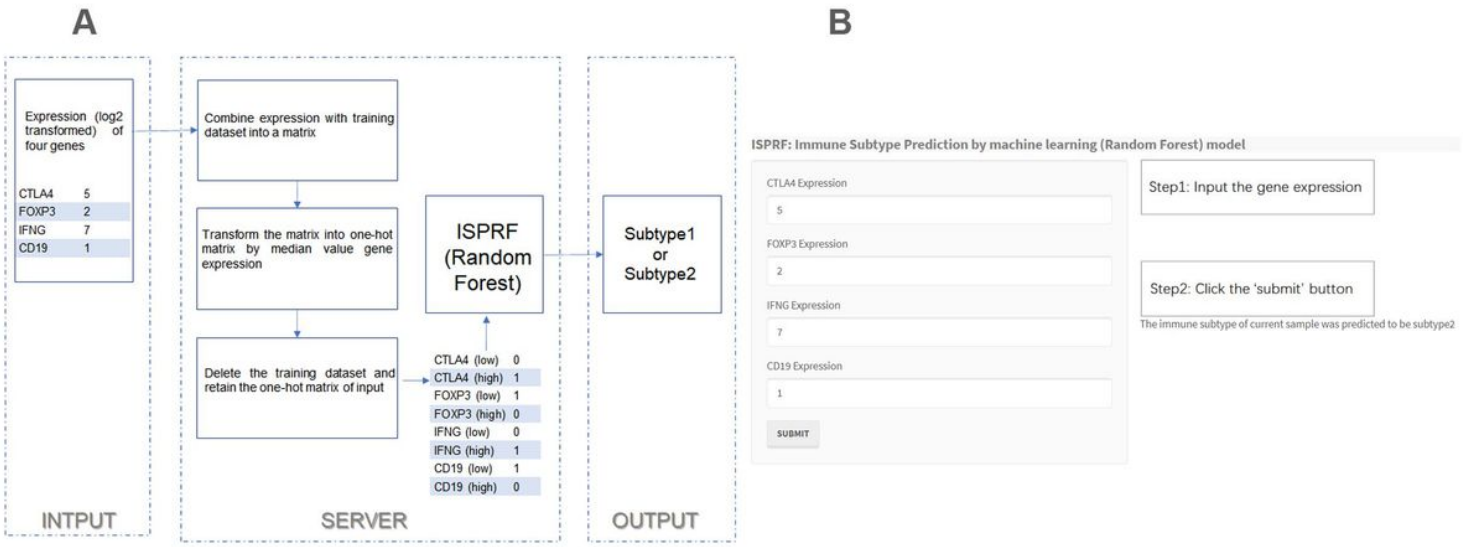



**Figure 7**

(a) The workflow of random forest model in Shiny APP (https://immunotype.shinyapps.io/ISPRF/). (b) The interface shows an example of predicting the immune subtype by four genes expression.

## Supplementary Files

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

- SupportinginformationforISPRFamachinelearningmodeltopredicttheimmunesubtypeofkidneycancersamplesbyfourgenes.docx