Integrated analysis of single-cell and transcriptome based RNA-seq multilayer network and WGCNA for construction and validation of an immune cell-related prognostic model in clear cell renal cell carcinoma

Guanlin Wu[1], Weiming Guo [2], Shuai Zhu[3], Gang Fan[3,4]

1. *Department of Pathology, School of Basic Medical Sciences, Fudan University, 200030 Shanghai, China.*
2. *The 2nd Affiliated Hospital of South China University, 421001 Hengyang, China.*
3. *Department of Urology, the Affiliated Cancer Hospital of Xiangya School of Medicine of Central South University, Hunan Cancer Hospital, 410013 Changsha, Hunan, China.*
4. *Department of Urology, Huazhong University of Science and Technology Union Shenzhen Hospital; the 6th Affiliated Hospital of Shenzhen University Health Science Center, 518060 Shenzhen, China.*

Correspondence: gang.fan.med@qq.com (G.F)

Abstract

Clear cell renal cell carcinoma (ccRCC) is the most common type of renal cancer (RCC). The increasing incidence and poor prognosis of ccRCC after tumour metastasis makes the study of its pathogenesis extremely important. Traditional studies mostly focus on the regulation of ccRCC by single gene, while ignoring the impact of tumour heterogeneity on disease progression. The purpose of this study is to construct a prognostic risk model for ccRCC by analysing the differential marker genes related to immune cells in the single-cell database for providing help in clinical diagnosis and targeted therapy. Single-cell data and ligand-receptor relationship pair data were downloaded from related publications, and ccRCC phenotype and expression profile data were downloaded from TCGA and CPTAC. The DEGs and marker genes of the immune cell were combined and then intersected with the ligand-receptor gene data, and the 981 ligand-receptor relationship pairs obtained were intersected with the target gene of the transcription factor afterwards; 7,987 immune cell multilayer network relationship pairs were finally observed. Then, the genes in the network and the genes in TCGA were intersected to obtain 966 genes for constructing a co-expression network. Subsequently, 53 genes in black and magenta modules related to prognosis were screened by WGCNA for subsequent analysis. Whereafter, using the data of TCGA, 28 genes with significant prognostic differences were screened out through univariate Cox regression analysis. After that, LASSO regression analysis of these genes was performed to obtain a prognostic risk scoring model containing 16 genes, and CPTAC data showed that the effectiveness of this model was good. The results of correlation analysis between the risk score

and other clinical factors showed that age, grade, M, T, stage and risk score were all significantly different ($p < 0.05$), and the results of prognostic accuracy also reached the threshold of qualification. Combined with clinical information, univariate and multivariate Cox regression analyses verified that risk score was an independent prognostic factor ($p < 0.05$). A nomogram constructed based on a predictive model for predicting the overall survival was established, and internal validation performed well. Our findings suggest that the predictive model built based on the immune cells scRNA-seq will enable us to judge the prognosis of patients with ccRCC and provide more accurate directions for basic relevant research and clinical practice.

Keywords: clear cell renal cell carcinoma (ccRCC); tumour heterogeneity; immune cell; single-cell RNA-sequencing (scRNA-seq); risk prognostic model; bioinformatics; multilayer network; weighted gene co-expression network analysis (WGCNA); regression analysis; nomogram

Main content

ccRCC single-cell transcriptome data was downloaded from [1]. After standardized data processing, we obtained a total of 23 subgroup clusters, among which immune cells were mainly concentrated in subgroups 0, 1, 2, 3, 4, 5, 6, 7, 9, 12, 13, 15, 16, 18 and 22 **(Fig. S1, Table S1)**. Differentially expressed genes (DEGs) in the 23 clusters were listed in **(Fig. S2A, Table S2)**. The functional status of differentially expressed genes (DEGs) in immune cells were also performed **(Fig. S2B-F)**. Furthermore, a total of 42 immune cell marker genes related to ccRCC were downloaded from the CellMarker (http://bio-bigdata.hrbmu.edu.cn/CellMarker/) database and [2] performed differential expression analysis **(Fig. S2G)**.

In order to construct the ligand-receptor network, we first took the union of the differential genes of all immune cell clusters and the marker genes of all these immune cells. Afterwards, we intersected them with the ligand-receptor relationship pairs downloaded from [3]. Finally, a total of 981 pairs of ligand-receptor relationships were obtained **(Fig. S3A, Table S3)**. Afterwards, we downloaded the transcription factors and target genes data from the TRRUST (https://www.grnpedia.org/trrust/) database and merged the ligand-receptor network data with the transcription factors and target genes data, which is the intersection of the target genes of the transcription factors and the genes from the ligand-receptor network, we obtained 7,987 immune cell multifactor network relationship pairs **(Table S4)**. Then, 966 genes used for the immune cell multifactor network construction **(Fig. S3B)** were obtained by intersecting the 973 genes contained in the ligand-receptor network and the genes in TCGA (https://portal.gdc.cancer.gov/) dataset about ccRCC **(Table S5)**.

Through the WGCNA algorithm [4], the 966 genes in the immune cell multifactor network were used to construct a co-expression network to find the key modules related to overall survival (OS) and OS time. The appropriate

scale-free power value was screened out as 4 **(Fig. S4A)** when the average connectivity degree was relatively higher [5]. All constructed modules are painted with different colours, and the cluster trees of genes are also shown in **Fig. S4B**. The black and magenta modules were chosen as the key modules, since they had the highest correlations with OS and OS time of ccRCC **(Fig. S4C-D)**. The correlations between module eigengenes (MEs) and clinic traits are shown in **Fig. S4E**. There were 53 genes in these two modules **(Table S6)**. For a deeper understanding about the biofunctions of these modules, genes in these modules were carried out to conduct gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses, the top terms were visualized **(Fig. S4F-I)**.

Using the "survival" R package to perform univariate Cox regression analysis on the 53 genes contained in the key modules of WGCNA, 28 genes with *P* value < 0.05 were obtained. Then, the 28 genes with significant prognostic differences were subjected to LASSO regression analysis. The specific steps were to first adopt the Cox proportional hazard model (family = "Cox") to calculate the hazard ratio (HR) and *P* values of these genes **(Fig. 1A)** and then randomly simulated 1,000 times (maxit = 1000) to find the most suitable λ value **(Fig. 1B)**. Finally, "lambda.min" was used to screen out 16 genes for constructing a risk scoring model from these 28 genes **(Fig. 1C)**.

Risk score = PAX2* (-0.00104) + NFKBIZ*0.11197 + TEAD4*0.212538 + HIPK2* (-0.09059) + CD14*0.051547 + COL1A1*0.084915 + NRG1* (-0.10501) + GPR182* (-0.30944) + ITGA6* (-0.08147) + HDAC1*0.225345 + HOXA9*0.078825 + E2F5*0.436804 + APP* (-0.14625) + FGF1* (-0.27736) + L1CAM*0.149607 + DDR1* (-0.11874)

According to the risk score of each patient given by the model, the median was taken as the cut-off value, and the samples were divided into high and low risk groups. The time-dependent receiver operating characteristic (ROC) curve was used to evaluate the predictive ability of the model's 1-, 3- and 5-year survival periods, and the survival curves of the high and low risk groups were also analysed. The CPTAC (https://cptac-data-portal.georgetown.edu/study-summary/S050) dataset for ccRCC was taken as the external validation database to verify the prognostic risk model. In the evaluation of the predictive efficacy of the prognosis model, Kaplan-Meier (K-M) survival analysis was performed on the high and low risk groups, and the difference was significant **(Fig. 1D)**. In the external CPTAC dataset, K-M survival analysis was performed on the high and low risk groups, and the

difference was also significant **(Fig. 1E)**. Moreover, the AUC values in their ROC curves validated their precision **(Fig. 1F-G)**.

In the correlation comparison between risk score and other clinical factors, the results showed that all factors other than age had a certain correlation with risk score **(Fig. S5A-G)**. From the ROC curve, it was also found that, except for the N stage, the AUC values of other ROC curves were above 0.5, and the AUC values of the risk score and stage were above 0.7 **(Fig. S5H-J)**.

The immune infiltration of genes in the model is shown in **(Fig. 6A).** The differential expression of them in TCGA **(Fig. 6B)** and CPTAC **(Fig. 6C)** are displayed. Univariate Cox regression analysis **(Fig. S6D)** and multivariate Cox regression analysis **(Fig. S6E)** for the independent analysis of the prognostic model were also done.

Additionally, we also developed a nomogram with predictors included age, gender, T, N, M, grade, risk score and stage **(Fig. S7A)**. The consistency index (C-index), which represents the discrimination, of the nomogram was 0.799. and our calibration chart revealed that the nomogram has a good predictive ability **(Fig. S7B-D)**. The AUC values indicated that the predictive precision of the nomogram combined with all factors was better **(Fig. S7E-G)**.

In summary, our research reveals that the immune cells in ccRCC have a significant impact on the prognosis of patients, and have obvious potential regulatory effects on the occurrence and development of ccRCC. The predictive model based on immune cell scRNA-seq will enable us to judge the prognosis of ccRCC patients and provide valuable guidance for clinical practice.

Abbreviation

RCC:	renal cell carcinoma
ccRCC:	clear cell renal cell carcinoma
OS:	overall survival
scRNA-seq:	single-cell RNA-sequencing
WGCNA:	weighted gene co-expression network analysis
TCGA:	The Cancer Genome Atlas
CPTAC:	Consortium for Clinical Proteome Cancer Analysis
DEGs:	differentially expressed genes
MEs:	module eigengenes
GO:	gene ontology
KEGG:	Kyoto Encyclopedia of Genes and Genomes
HR:	hazard ratio
ROC:	receiver operating characteristic
C-index:	concordance index
K-M:	KaplanMeier

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The data and materials can be obtained by contacting the corresponding author.

Declarations of interest

None.

Funding

This work was sponsored by the Guangdong Medical Research Foundation (A2021375).

Authors' contributions

G.W designed the research plan, analyzed datasets and wrote the manuscript. W.G and S.Z provided meaningful discussion of key points. G.F gave guidance in whole study and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Thanks to the assistance of Mugu lnc (Beijing) in bioinformatics.

Figure legends

Figure 1. Construction and assessment of disease prognosis risk model for ccRCC. **(A)** Forest plot for univariate regression analysis of 28 genes. **(B)** Select the appropriate λ value through LASSO regression analysis. **(C)** Scoring chart of risk model constructed by 16 genes. **(D, F)** Evaluation of predictive effectiveness of risk prognostic model. **(E, G)** Use of the data in the CPTAC database as an external dataset for verification.

Figure S1. Cell clustering. **(A)** Clustering of single-cell subpopulations. **(B)** The distribution of samples in clusters. **(C)** Annotation for all cell types.

Figure S2. (A) Heat map of the top five differential genes in each cluster. **(B)** B cell functional status analysis. **(C)** T cell functional status analysis. **(D)**

Monocyte functional status analysis. **(E)** Macrophage functional status analysis. **(F)** NK cells functional status analysis. **(G)** Heat map of immune cell marker genes.

Figure S3. (A) Ligand-receptor interaction network diagram, different colours and sizes represent different numbers of node connections. **(B)** Multilayer network diagram—green is the ligand gene, red is the receptor gene, and yellow is the transcription factor.

Figure S4. WGCNA and functional analysis in black and magenta modules. **(A)** Schematic diagram of threshold screening and determination. **(B)** Clustering dendrogram of all genes from last step. **(C)** Correlations of OS time with mean gene significance and errors in all modules. **(D)** Correlations of OS with mean gene significance and errors in all modules. **(E)** Heat map of the correlations between MEs and OS traits. **(F)** BP term in GO. **(G)** CC term in GO. **(H)** MF term in GO. **(I)** KEGG pathway analysis.

Figure S5. (A-G) Correlation analysis of risk score and other clinical factors. **(H-J)** The ROC curve was used to analyse the accuracy of risk scores and other clinical factors in predicting the 1-, 3-, and 5-year survival of patients with ccRCC.

Figure S6. (A) Cell infiltration analysis results. **(B)** The differential expression levels of genes in the risk prediction model obtained through the analysis of TCGA database. **(C)** The differential expression levels of genes in the risk prediction model obtained through the analysis of the CPTAC database. **(D)** Univariate association of the prognostic model and clinicopathological characteristics with overall survival. **(E)** Multivariate association of the prognostic model and clinicopathological characteristics with overall survival.

Figure S7. (A) Nomogram predicting 1-, 3- and 5- year OS for patients with ccRCC. **(B-D)** The calibration curve for predicting 1-, 3- and 5-year OS for patients with ccRCC. **(E-G)** Time-dependent ROC curve analysis evaluates the accuracy of the nomograms.

Tables

Table S1. Notes on cell clustering.

Table S2. Differential genes in each cell cluster.

Table S3. Ligand-receptor relationship pair.

Table S4. Immune cell multifactor network relationship pair.

Table S5. Intersection genes in immune cell multifactor network relationship pair and TCGA.

Table S6. Genes in black and magenta models of WGCNA.

References

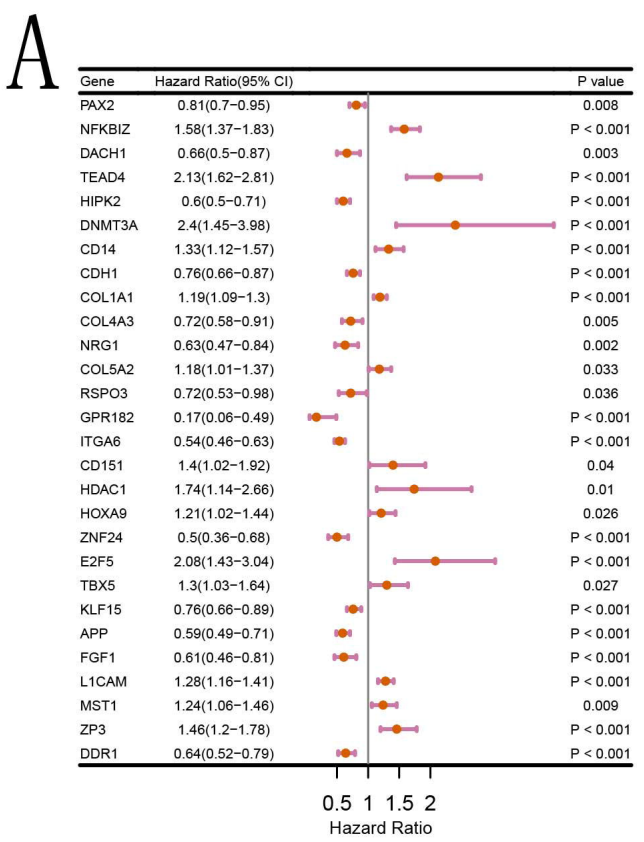
1. M. D. Young, T. J. Mitchell, F. A. Vieira Braga, et al., *Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors.* Science, 2018. **361**(6402): p. 594-599.

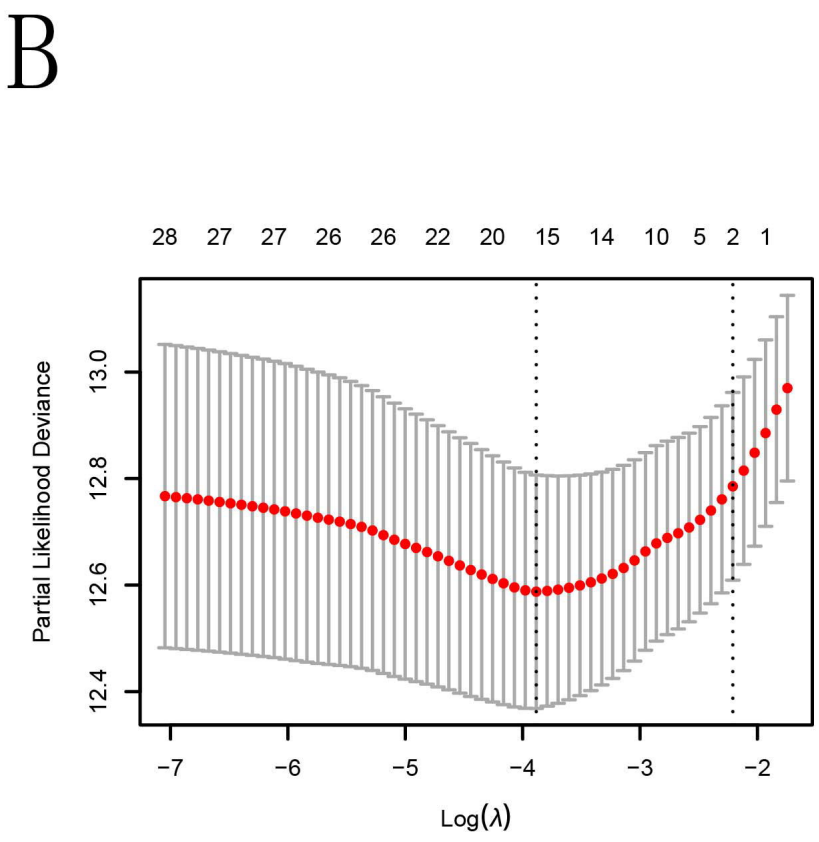
2. Ajaykumar Vishwakarma, Nicholas Bocherding, Michael S. Chimenti, et al., *Mapping the Immune Landscape of Clear Cell Renal Cell Carcinoma by Single-Cell RNA-seq.* bioRxiv, 2019: p. 824482.

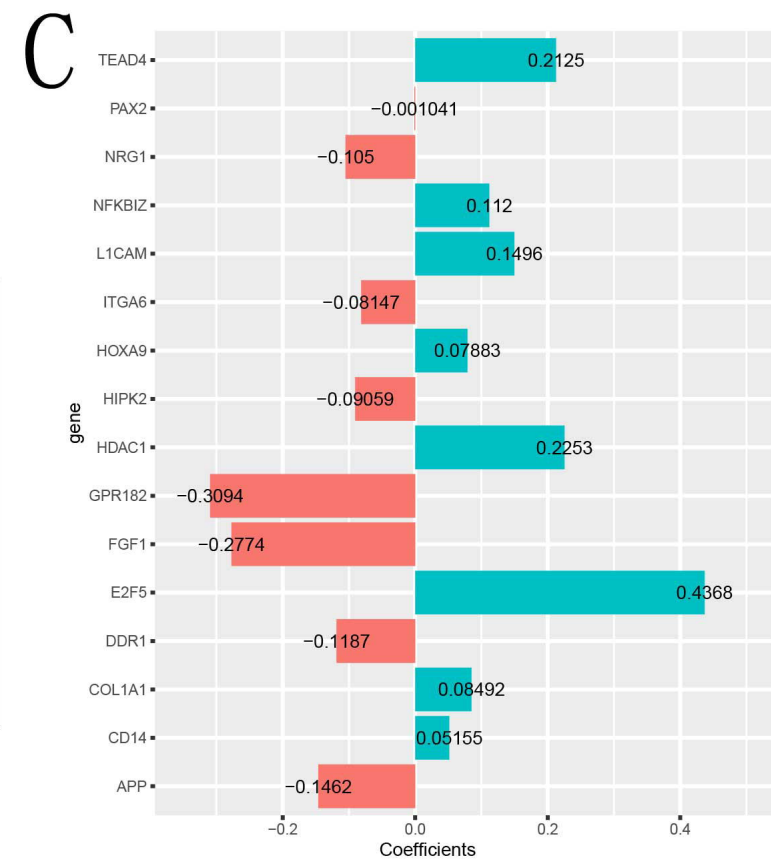
3. J. Zhang, M. Guan, Q. Wang, et al., *Single-cell transcriptome-based multilayer network biomarker for predicting prognosis and therapeutic response of gliomas.* Brief Bioinform, 2020. **21**(3): p. 1080-1097.

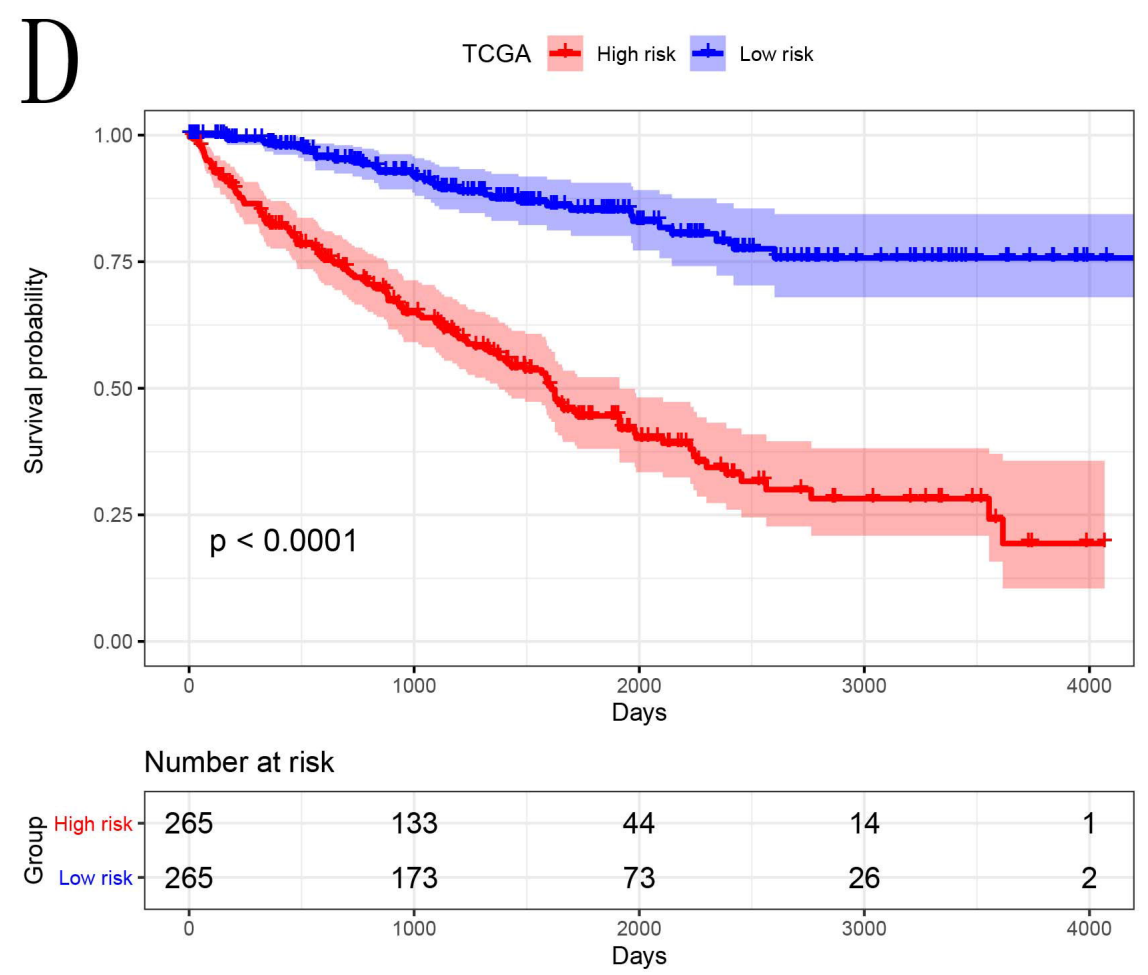
4. P. Langfelder and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.

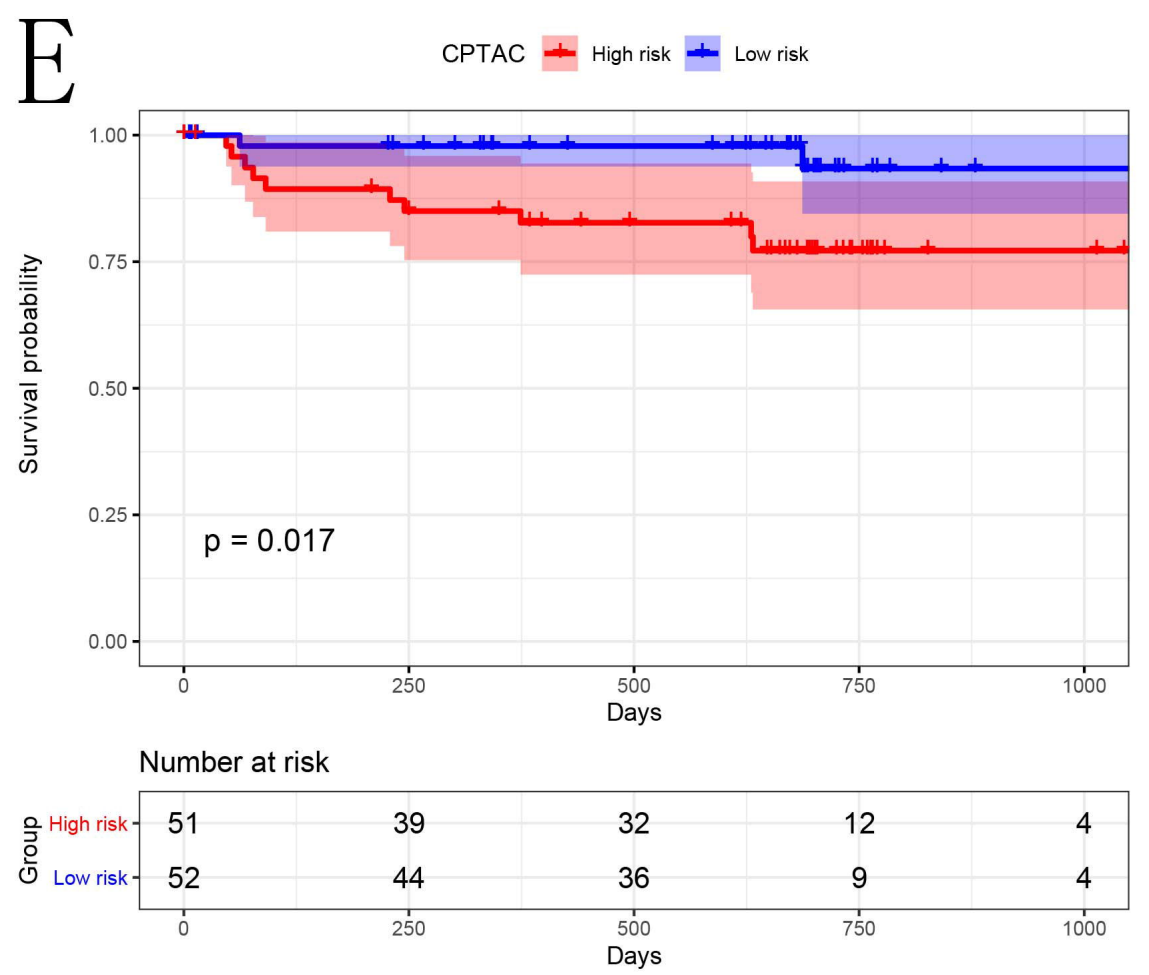
5. L. Chen, L. Yuan, Y. Wang, et al., *Co-expression network analysis identified FCER1G in association with progression and prognosis in human clear cell renal cell carcinoma.* Int J Biol Sci, 2017. **13**(11): p. 1361-1372.

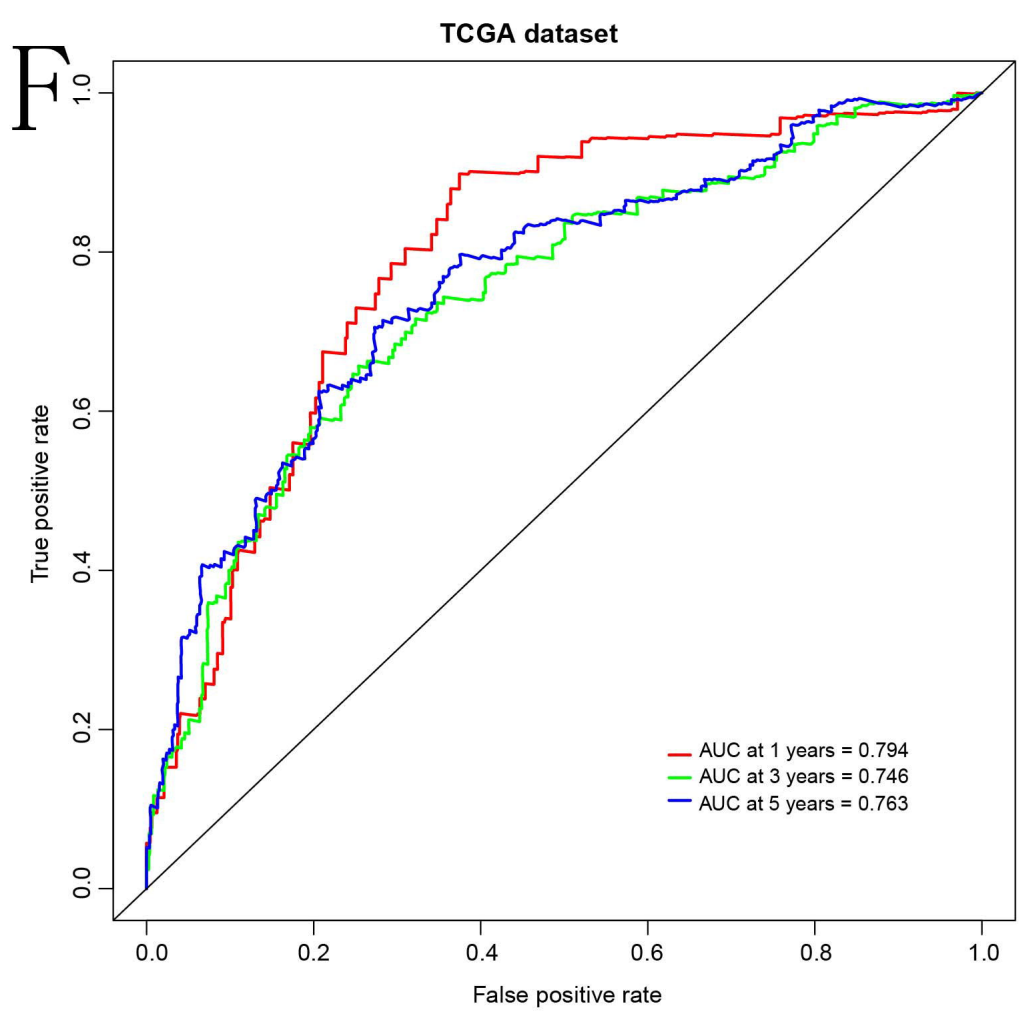

A
Gene
Hazard Ratio(95% CI)
P value
PAX2 0.81(0.7−0.95) 0.008
NFKBIZ 1.58(1.37−1.83) P < 0.001
DACH1 0.66(0.5−0.87) 0.003
TEAD4 2.13(1.62−2.81) P < 0.001
HIPK2 0.6(0.5−0.71) P < 0.001
DNMT3A 2.4(1.45−3.98) P < 0.001
CD14 1.33(1.12−1.57) P < 0.001
CDH1 0.76(0.66−0.87) P < 0.001
COL1A1 1.19(1.09−1.3) P < 0.001
COL4A3 0.72(0.58−0.91) 0.005
NRG1 0.63(0.47−0.84) 0.002
COL5A2 1.18(1.01−1.37) 0.033
RSPO3 0.72(0.53−0.98) 0.036
GPR182 0.17(0.06−0.49) P < 0.001
ITGA6 0.54(0.46−0.63) P < 0.001
CD151 1.4(1.02−1.92) 0.04
HDAC1 1.74(1.14−2.66) 0.01
HOXA9 1.21(1.02−1.44) 0.026
ZNF24 0.5(0.36−0.68) P < 0.001
E2F5 2.08(1.43−3.04) P < 0.001
TBX5 1.3(1.03−1.64) 0.027
KLF15 0.76(0.66−0.89) P < 0.001
APP 0.59(0.49−0.71) P < 0.001
FGF1 0.61(0.46−0.81) P < 0.001
L1CAM 1.28(1.16−1.41) P < 0.001
MST1 1.24(1.06−1.46) 0.009
ZP3 1.46(1.2−1.78) P < 0.001
DDR1 0.64(0.52−0.79) P < 0.001
0.5 1 1.5 2
Hazard Ratio

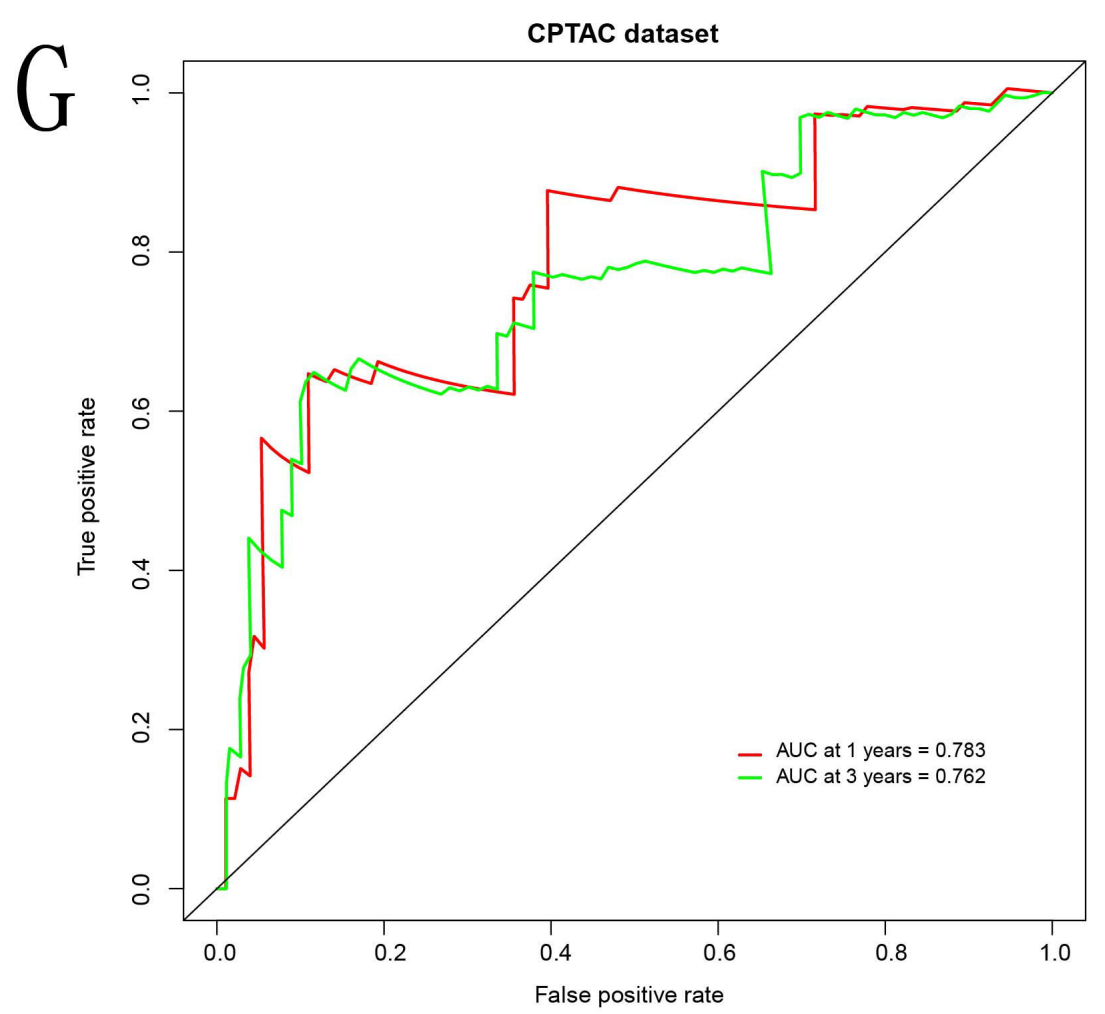

B
28 27 27 26 26 22 20 15 14 10 5 2 1
Partial Likelihood Deviance
12.4
12.6
12.8
13.0
−7 −6 −5 −4 −3 −2
Log(λ)


C
TEAD4 0.2125
PAX2 −0.001041
NRG1 −0.105
NFKBIZ 0.112
L1CAM 0.1496
ITGA6 −0.08147
HOXA9 0.07883
HIPK2 −0.09059
HDAC1 0.2253
GPR182 −0.3094
FGF1 −0.2774
E2F5 0.4368
DDR1 −0.1187
COL1A1 0.08492
CD14 0.05155
APP −0.1462
gene
−0.2 0.0 0.2 0.4
Coefficients


D
TCGA
High risk
Low risk
Survival probability
0.00 0.25 0.50 0.75 1.00
p < 0.0001
0 1000 2000 3000 4000
Days
Number at risk
Group
High risk 265 133 44 14 1
Low risk 265 173 73 26 2
0 1000 2000 3000 4000
Days


E
CPTAC
High risk
Low risk
Survival probability
0.00 0.25 0.50 0.75 1.00
p = 0.017
0 250 500 750 1000
Days
Number at risk
Group
High risk 51 39 32 12 4
Low risk 52 44 36 9 4
0 250 500 750 1000
Days


F
TCGA dataset
True positive rate
0.0 0.2 0.4 0.6 0.8 1.0
False positive rate
AUC at 1 years = 0.794
AUC at 3 years = 0.746
AUC at 5 years = 0.763


G
CPTAC dataset
True positive rate
0.0 0.2 0.4 0.6 0.8 1.0
False positive rate
AUC at 1 years = 0.783
AUC at 3 years = 0.762