Diagnostic Accuracy Study

Medicine®

OPEN

Comprehensive assessment gene signatures for clear cell renal cell carcinoma prognosis

Peng Chang, MD[a,b,c], Zhitong Bing, PhD[c,d], Jinhui Tian, PhD[c,d], Jingyun Zhang, PhD[c,d], Xiuxia Li, PhD[c,d,e], Long Ge, MD[c,d], Juan Ling, MM[c,d], Kehu Yang, MD[a,d,*], Yumin Li, MD[a,b,*]

Abstract

There are many prognostic gene signature models in clear cell renal cell carcinoma (ccRCC). However, different results from various methods and samples are hard to contribute to clinical practice. It is necessary to develop a robust gene signature for improving clinical practice in ccRCC.

A method was proposed to integrate least absolute shrinkage and selection operator and multiple Cox regression to obtain mRNA and microRNA signature from the cancer genomic atlas database for predicting prognosis of ccRCC. The gene signature model consisted by 5 mRNAs and 1 microRNA was identified. Prognosis index (PI) model was constructed from RNA expression and median value of PI is used to classified patients into high- and low-risk groups.

The results showed that high-risk patients showed significantly decrease survival comparison with low-risk groups [hazard ratio (HR) =7.13, 95% confidence interval=3.71–13.70, $P < .001$]. As the gene signature was mainly consisted by mRNA, the validation data can use transcriptomic data to verify. For comparison of the performance with previous works, other gene signature models and 4 datasets of ccRCC were retrieved from publications and public database. For estimating PI in each model, 3 indicators including HR, concordance index , and the area under the curve of receiver operating characteristic for 3 years were calculated across 4 independent datasets.

The comparison results showed that the integrative model from our study was more robust than other models via comprehensive analysis. These findings provide some genes for further study their functions and mechanisms in ccRCC tumorigenesis and malignance, and may be useful for effective clinical decision making of ccRCC patients.

Abbreviations: AUC = area under the curve, ccRCC = clear cell renal cell carcinoma, CI = confidence interval, C-index = concordance index, DFS = disease-free survival, GEO = gene expression omnibus, GO = gene ontology, HR = hazard ratio, LASSO = least absolute shrinkage and selection operator, OS = overall survival, PI = prognosis index, RCC = renal cell carcinoma, ROC = receiver operating characteristic, TCGA = the cancer genome atlas, TF = transcription factor.

Keywords: clear cell renal cell carcinoma, Cox regression, gene regulatory network, least absolute shrinkage and selection operator, prognosis

1. Introduction

Renal cell carcinoma (RCC) is a frequent malignant tumor of the adult kidney. According to cancer statistics in 2018, 65,340 new kidney cancer patients and 14,970 deaths.[1] Comparison to 10 years ago (2007), 51,190 individuals suffered from kidney cancer and 12,890 died.[2] In 10 years, the morbidity and mortality of patients with renal cancer have not been significantly changed. One of the important reasons is that RCC is a highly heterogeneous set of disease. Of these subtypes of RCCs, The clear-cell renal cell carcinoma (ccRCC) is one of most common subtypes, accounting for approximately 70% to 80% of the whole RCC.[3] Thus, identification of robust biomarkers for ccRCC prognosis is necessary.

In an age of precision medicine, molecular subtype and gene signature can provide new insight for clinical strategy and drug development. High-throughput gene sequencing technology provides us with a powerful tool to find genetic differences among different patients. Therefore, different strategies can be used to treat the patient in molecular level. Nevertheless, different approaches using by different groups have produced many different prognostic biomarkers for ccRCC.[3–24] How do we decide which gene signature is effective against the current ccRCC, which gene signature is more universal? In this study, a least absolute shrinkage and selection operator (LASSO) penalized Cox regression analysis method was combined with multivariate Cox regression to obtain a set of gene biomarkers and compared it with other gene signature from publications. In addition, there are many studies in microRNA prognostic signature.[15,25–28] Thus, we also integrated microRNA and mRNA expression for predicting prognosis of ccRCC.

Editor: Bernhard Schaller.

PC and ZB contributed equally to this study.

Funding: This work is supported by Lanzhou Talent Innovation and Entrepreneurship Project (Grant No. 2016-RC-104) and Gansu Province Science Foundation (Grant No. 1606RJZA016).

The authors have no conflicts of interest to disclose.

[a] School of Life Sciences, Lanzhou University, [b] Lanzhou University Second Hospital, [c] Evidence Based Medicine Center, School of Basic Medical Science of Lanzhou University, [d] Key Laboratory of Evidence Based Medicine and Knowledge Translation of Gansu Province, [e] School of Public Health, Lanzhou University, Lanzhou, China.

[*] Correspondence: Kehu yang, 199 West Donggang Road, Lanzhou, Gansu 730000, China (e-mail: kehuyangebm2006@126.com), Yumin Li, 99 West Donggang Road, Lanzhou, Gansu 730000, China (e-mail: liyumin2018@126.com).

Medicine (2018) 97:44(e12679)

Received: 2 June 2018 / Accepted: 7 September 2018

http://dx.doi.org/10.1097/MD.0000000000012679

In this study, we select data from 2 common databases of the cancer genome atlas (TCGA) and gene expression omnibus (GEO) database as training data sets and validating datasets, respectively. Other 3 gene signatures were also tested with these 2 database sets. To exclude methodological heterogeneity, we selected 3 studies of gene prognostic biomarkers that were obtained using Cox regression. In the 3 studies, Yao et al[9] find 3 genes for prognosis of ccRCC, Boguslawska et al[22] find 10 genes and Zhan et al[21] find 5 genes. In our study, we found 5 mRNAs and 1 microRNA can predict prognosis of ccRCC. Interestingly, each of the gene signature has no intersection, but they performed well in validating the data set.

Here we demonstrate that the gene signature from LASSO combined multivariate Cox regression has more stability and universal in prognosis of ccRCC. These may be helpful for selecting high-risk ccRCC patients for better clinical decision making and provide useful biomarkers for downstream experimental.

2. Methods

2.1. Data collected and preprocessing

The microRNA expression proliferation and mRNA expression proliferation are collected TCGA database (https://portal.gdc.cancer.gov/). The mRNA dataset is performed by the Illumina Hiseq platform. And microRNA dataset is also performed by Illumina Hiseq platform. In ccRCC cohort of TCGA database, RNASeqV2 expression data contain 20,530 genes and microRNA expression data contain 1046 microRNAs. The patients without survival time and event information were excluded. The samples with mRNA expression contained 510 patients with primary tumor and 70 patients with solid tissue normal. The samples with microRNA expression contained 254 patients with primary tumor and 71 patients with solid tissue normal. We used both mRNA and miRNA samples as training dataset (n = 239). Validation data are collected from GEO database. GSE22541 contains 68 samples which include 24 primary and 44 metastasis samples.[29] And this dataset employed Affymetrix Human Genome U133 Plus 2.0 Array. This work did not directly use tissues from patients or animals.

2.2. Gene differential expression analysis

After differential gene expression analysis, there are 4205 mRNAs and 59 microRNAs with differential expression comparison of normal tissue. In this study, we selected differential expression genes with fold-change >1.5 and false discovery rate (FDR) <0.01 as candidate genes for next step. The microRNA and mRNA differential expression are assayed by R package by "limma"[30] from Bioconductor 2.14.

2.3. Univariate Cox regression gene test

Firstly, univariate Cox regression and survival analysis are applied to analyze clinical factors and each differential expression gene. For clinical factors, we employed univariate Cox regression to test hazard ratio (HR). And survival analysis using Kaplan-Meier is applied to analyze clinical factor significant difference by log-rank (2-sided test). For estimating the clinical factor in ccRCC patients, HR >1 is considered as risk increasing group and HR <1 is considered as risk-decreasing factor.

2.4. Multivariate Cox regression for clinical factor and LASSO Cox regression for RNAs

For univariate Cox regression clinical factor, we screened the factor with Wald test $P < .05$ as candidate for multivariate Cox regression.

For mRNA and miRNA expression, we filtered the RNA with Wald test $P < .05$ as candidate for LASSO Cox regression. By univariate Cox regression filtering, 2498 mRNAs and 18 miRNAs were selected to submit to LASSO. LASSO is performed by R package of "glmnet." LASSO is employed to filter 2498 mRNAs and 18 miRNAs, respectively. After 10,000 iterations and 10 folds cross-validation, 16 mRNAs and 9 miRNAs are obtained.

2.5. Integrative mRNAs and miRNAs for predicting survival of ccRCC

For investigating integrative model of miRNAs and mRNAs, ccRCC patients who both include mRNA and miRNA expression are selected. There are 239 samples including for assay the integrative model. We employed multivariate Cox regression for 25 RNAs (16 mRNAs and 9 microRNAs), and we obtained 5 mRNAs and 1 microRNA-independent predictors for ccRCC.

2.6. Literature reviews

Prognostic model search was employed by PubMed. The following terms were searched in PubMed: ({"clear cell renal cell carcinoma" OR "clear cell renal cell cancer" OR "clear cell renal cell tumor" OR "clear cell renal cell tumour" OR "clear cell kidney cell carcinoma" OR "clear cell kidney cell cancer" OR "clear cell kidney cell tumor" OR "clear cell renal cell tumour"} AND {"gene expression" OR "gene signature" OR "gene proliferation" OR "microarray" OR "high-throughput" OR "microRNA expression" OR "mRNA expression"} AND {"survival" OR "survivor" OR "outcome" OR "prognosis" OR "prognostic" OR "prediction"} AND {"risk score" OR "cox regression"}). After above filtering, 3 studies were included in this study.

2.7. Prognostic index construction

A prognosis index (PI) as an integrated indicator of candidate RNAs for each ccRCC patient was constructed. The PI was computed as a linear combination of the RNA expression value and weighted by LASSO Cox regression coefficients.

$$\mathrm{PI} = \sum\nolimits_i \beta_i X_i$$

where β_i is the regression coefficient of the ith variable. X_i is the value of the ith variable. In this study, X_i is the $\log_2$-transformed expression value of each RNA and β_i is the LASSO Cox regression coefficient of the ith RNA.

2.8. Validation datasets construction

For assessment performance of gene signature model, gene expression array (AgilentG4502A_07_3) and RNAseq (IlluminaHiSeq) were selected. The RNAseq data were divided into 2 parts. One part includes both miRNA and mRNA samples (n = 239) that were used to be training data in this

Table 1

Clinical factor information of 4 clear cell renal cell carcinoma cohort datasets.

	239 Patients		510 Patients		72 Patients		24 Patients	
Clinical factors	Patients (event/patients)	Median of survival (95% CI)	Patients (event/patients)	Median of survival (95% CI)	Patients (event/patients)	Median of survival (95% CI)	Patients (event/patients)	Median of survival (95% CI)
Age								
>60	41/114	2190(1588-NA)	103/257	1986(1639–2601)	6/33	NA	NA	NA
≤60	28/125	NA(2454-NA)	63/257	NA	8/39	NA	NA	NA
Sex								
Male	47/164	2454(2090-NA)	104/335	2454(1986-NA)	5/43	NA	10/13	82 (0-NA)
Female	22/75	NA(1714-NA)	62/179	2386(1964-NA)	9/29	2227(1610-NA)	7/11	63(45-NA)
Grade								
G1	0/6	NA	0/11	NA	0/4	NA	NA	NA
G2	13/94	NA(2454-NA)	42/220	NA	4/36	NA	13/18	71(6-NA)
G3	29/97	2830(2090-NA)	70/199	2299(1964-NA)	7/22	NA	4/6	76(0-NA)
G4	27/39	819(561–1714)	53/76	885(600–1588)	3/6	3227(1661-NA)	NA	NA
Gx	0/1	NA	1/5	NA	0/3	NA	NA	NA
Unknown	0/2	NA	0/3	NA	0/1	NA	NA	NA
Adjuvant treatment:								
Yes	3/8	1714(1714-NA)	9/18	1714(1034-NA)	0/1	NA	NA	NA
No	66/231	2830(2090-NA)	157/496	2454(2190-NA)	14/71	NA	NA	NA
Laterality								
Bilateral	0/1	NA	0/1	NA	0		NA	NA
Left	38/107	2299(1620-NA)	93/241	2227(1639–2752)	9/38	2227(1661-NA)	NA	NA
Right	31/131	NA	73/272	NA	5/34	NA	NA	NA
T stage								
TO	19/124	NA(2454-NA)	48/261	NA	5/41	NA	3/10	NA
T2	8/30	2830(NA)	18/67	2830(2256-NA)	1/14	2227(2227-NA)	9/9	6(0-NA)
T3	37/79	1567(1097-NA)	90/175	1337(1019–1724)	8/17	992(709-NA)	5/5	90(83-NA)
T4	5/6	1022(206-NA)	10/11	206(110-NA)	0		NA	NA
Stage								
Stage I	41/121	2090(1625-NA)	43/256	NA	4/40	NA	NA	NA
Stage II	5/26	NA(1417-NA)	10/55	2830(2256-NA)	0/13	NA	NA	NA
Stage III	14/48	2830(2454-NA)	49/122	1724(1417-NA)	5/14	1610(885-NA)	NA	NA
Stage IV	9/44	NA(2299-NA)	64/81	578(445–1034)	5/5	709(431-NA)	NA	NA
Tumor status								
With tumor	20/78	1238(932–1625)	109/157	1034(845–1371)	8/13	1610(709-NA)	NA	NA
Tumor free	48/155	NA(2830-NA)	51/341	NA	5/52	NA	NA	NA
Unknown	1/6	1964(1075-NA)	6/16	1964(953-NA)	1/7	NA	NA	NA

CI = confidence interval.

Table 2

Clinical factor log-rank and multivariate Cox regression test.

Clinical factors	Log-rank	HR (95% CI)	Multivariate *P* value	HR (95% CI)
Age: >60 vs ≤60	0.025*	1.72 (1.06–2.78)	.109	1.48 (0.92–2.40)
Sex: female vs male	0.858	0.95 (0.57–1.59)		
Grade				
G3–4 vs G1–2	2.36e-4*	2.95 (1.61–5.39)	.012*	2.20 (1.18–4.08)
Adjuvant treatment: yes vs no	0.275	1.89 (0.59–6.03)		
Laterality				
Right vs left	0.129	0.69 (0.43–1.12)		
T stage				
T3–4 vs TO–2	4.6e-7*	3.27 (2.00–5.31)	.0001*	2.68 (1.62–4.41)
Stage II vs stage I	0.585	0.76 (0.30–1.93)		
Stage III vs stage I	0.631	0.86 (0.46–1.58)		
Stage IV vs stage I	0.296	0.66 (0.32–1.36)		
Tumor status				
With tumor vs tumor free	0.481	0.93 (0.49–1.40)		

CI = confidence interval, HR = hazard ratio.
* Log-rank test showed significant difference.

Table 3

List of high- and low-risk candidate genes by Cox regression analysis with clear cell renal cell carcinoma (n = 239).

Gene symbol	HR (univariate) 95% CI	*P*	HR (multivariate) 95% CI	*P*	Description
Risky genes					
INTS8	3.708 (2.508–5.483)	5.15E-11	2.363 (0.904–6.170)	.079	Integrator complex subunit 8
GTPBP2	2.916 (2.274–3.739)	0	2.179 (0.941–5.043)	.069	GTP-binding protein 2
Protective genes					
ANK3	0.716 (0.658–0.778)	6.44E-15	0.770 (0.601–0.987)	.039	*Ankyrin-3*
SLC16A12	0.792 (0.749–0.839)	8.88E-16	0.864 (0.728–1.025)	.093	*Monocarboxylate transporter 12*
LIMCH1	0.559 (0.477–0.655)	6.60E-13	0.451 (0.291–0.698)	.0003	*LIM and calponin homology domains–containing protein 1*
Hsa-mir-374a	0.623 (0.393–0.988)	.044	0.506 (0.289–0.885)	.017	

CI = confidence interval, HR = hazard ratio.

study. Another part contained all ccRCC patients (n = 510). The gene expression array contained other ccRCC patients (n = 72). All ccRCC cohorts in TCGA data were downloaded from Cancer browser (https://xenabrowser.net/datapages/) which is built by UCSC. In addition, the samples in gene expression array has little overlap in training data samples (overlap n = 5). Thus, the microarray data can be used as an independent dataset. In addition, GSE22451 is an independent dataset which downloads from GEO database.

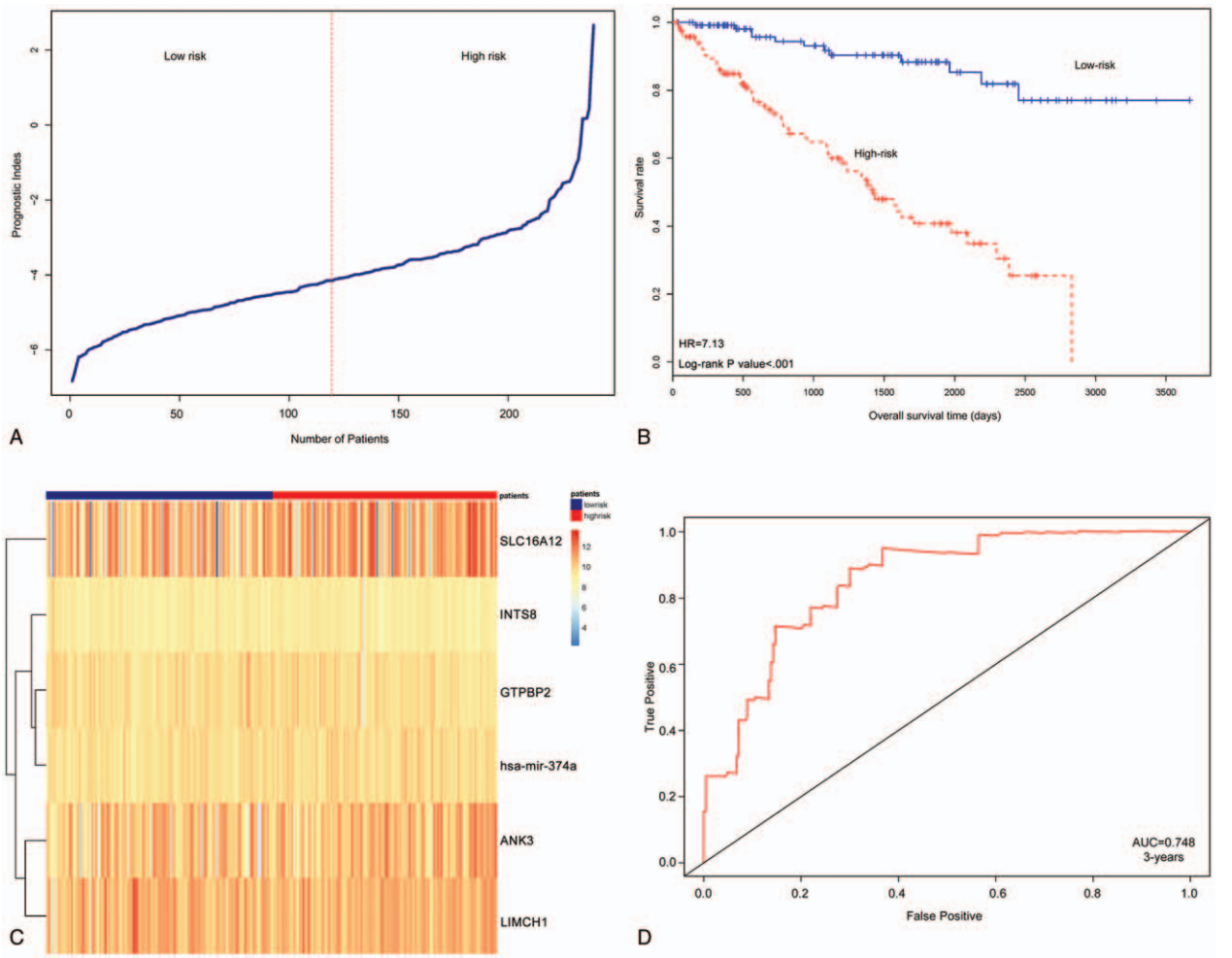

Figure 1. The integrative model for predicting outcome of clear cell renal cell carcinoma (ccRCC) in the cancer genome atlas (TCGA) cohort. A, Median of prognosis index (PI) value is as a sign for classification ccRCC patients into low-risk and high-risk groups. B, The heatmap of 6 RNAs in 329 patients. C, Survival analysis of low-risk and high-risk groups. D, Receiver operating characteristic (ROC) curve for estimating the effect of PI for classification of patients. CI = confidence interval, HR =hazard ratio.

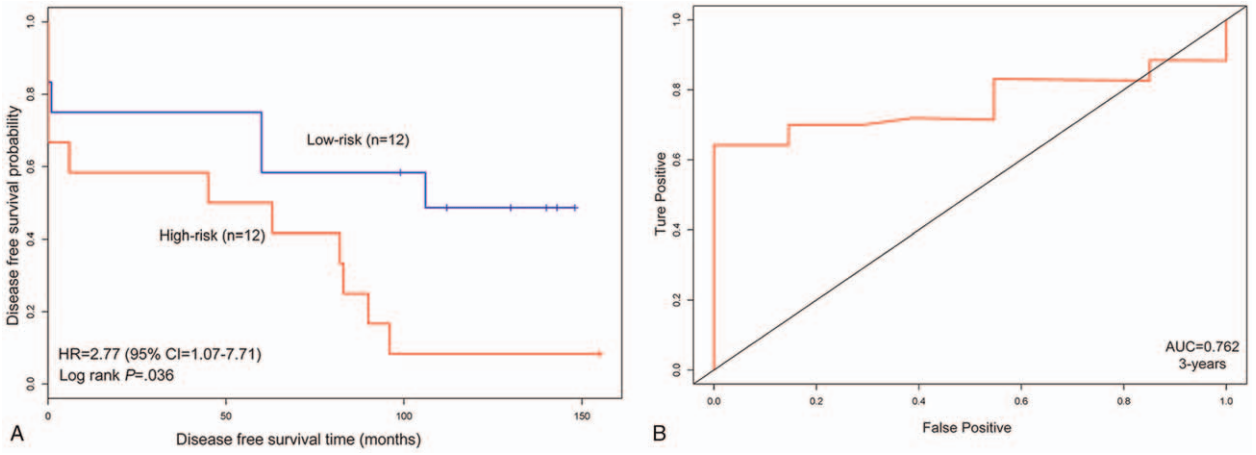

Figure 2. The integrative model for predicting outcome of clear cell renal cell carcinoma (ccRCC) in gene expression omnibus (GEO) cohort. A, Survival analysis of low-risk and high-risk groups in GEO dataset. B, Receiver operating characteristic (ROC) curve for estimating the effect of prognosis index (PI) for classification of patients. CI = confidence interval, HR =hazard ratio.

2.9. Estimating performance of the models

The ability and efficiency of each model to predict ccRCC patient outcome was estimated by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC), which was conducted using the survival ROC package in R software. Another indicator called concordance index (C-index) was conducted using "Hmisc" package.

2.10. MicroRNA target predicting

Many computational prediction approaches are available recently such as TargetScan, miRanda, PicTar, TarBase, RNAHybrid, etc, which are mostly based on complementarity, thermodynamics, or experimental validation. In this study, TargetScan (Release 7.1)[31] (http://www.targetscan.org/) and miRanda (Release 19) (http://www.microrna.org/) methods were employed to predict target.[32] Moreover, TargetScan and miRanda tool were used considering both conserved and nonconserved targets.

Table 4

The published 3 prognostic gene signature models of clear cell renal cell carcinoma.

Gene symbol	Coef	HR	Description
Yao et al, 2008 model			
VCAM1	−0.095	0.909	Vascular cell adhesion protein 1
EDNRB	−0.229	0.796	Endothelin receptor type B
RGS4	−0.181	0.834	Regulator of G-protein signaling 4
Zhan et al, 2015 model			
CKAP4	0.422	1.525	Cytoskeleton-associated protein 4
SLC40A1	−0.369	0.691	Solute carrier family 40 member 1
OTOF	0.330	1.391	Otoferlin
MAN2A2	0.551	1.735	Alpha-mannosidase 2x
ISPD	−0.443	0.642	D-ribitol-5-phosphate cytidylyltransferase
Boguslawska et al, 2015 model			
COL1A1	0.230	1.7	Collagen alpha-1(I) chain
COL5A1	0.279	1.9	Collagen alpha-1(V) chain
COL11A1	0.258	1.81	Collagen alpha-1(XI) chain
FN1	0.152	1.42	Fibronectin
THBS2	0.170	1.48	THBS2
ICAM1	0.248	1.77	Intercellular adhesion molecule 1
ITGAM	0.146	1.4	Integrin alpha-M
ITGAL	0.057	1.14	Integrin alpha-L
TIMP1	0.407	2.55	Metalloproteinase inhibitor 1
ITGB2	0.100	1.26	Integrin beta-2

HR = hazard ratio.

2.11. Gene regulation network and Gene Ontology enrichment

Each gene signature model contains a very small number of genes. It is difficult to enrich pathway through Gene Ontology (GO) analysis. Thus, the transcription factor (TF) of each gene in gene signature model is to predict and combines them to analyze to ClusterProfiler for GO analysis.[33] TF target genes were identified using the approach developed by Kathrin et al,[34] via defining the ±1000 bp sequence around transcription start sites as the promoter region. The genes with promoter regions completely overlapped with TF binding sites were considered as TF targets. To further enhance the reliability of TF, we calculated the correlation between predicted TF and target genes in TCGA dataset. Pearson correlation was used to estimate the relation between TFs and targets. Generally, TFs were considered to promote the expression of their targets. So, TFs expressions have positive correlations with their target genes ($r > 0.3$), and visualization of regulation network is used by Cytoscape software (version 3.5.1).[35]

3. Result

3.1. Demographic and clinical factors

In this study, 4 gene signature models were validated in 4 datasets, and baseline information of patients with ccRCC is listed in Table 1. The cohorts include 329 samples that both including miRNAs and mRNAs expression is used to train prognostic model. Others were employed to testing datasets. Of these datasets, TCGA with Agilent G450 platform was considered as an independent dataset which has little intersection from TCGA with Illumina Hiseq platform. For identification RNAs that significantly associated with overall survival (OS), the clinical factors of training dataset (329 samples) were analyzed. Eight clinical factors were assayed by univariate survival analysis (the 2-sided log-rank test), including age at initial diagnosis, sex,

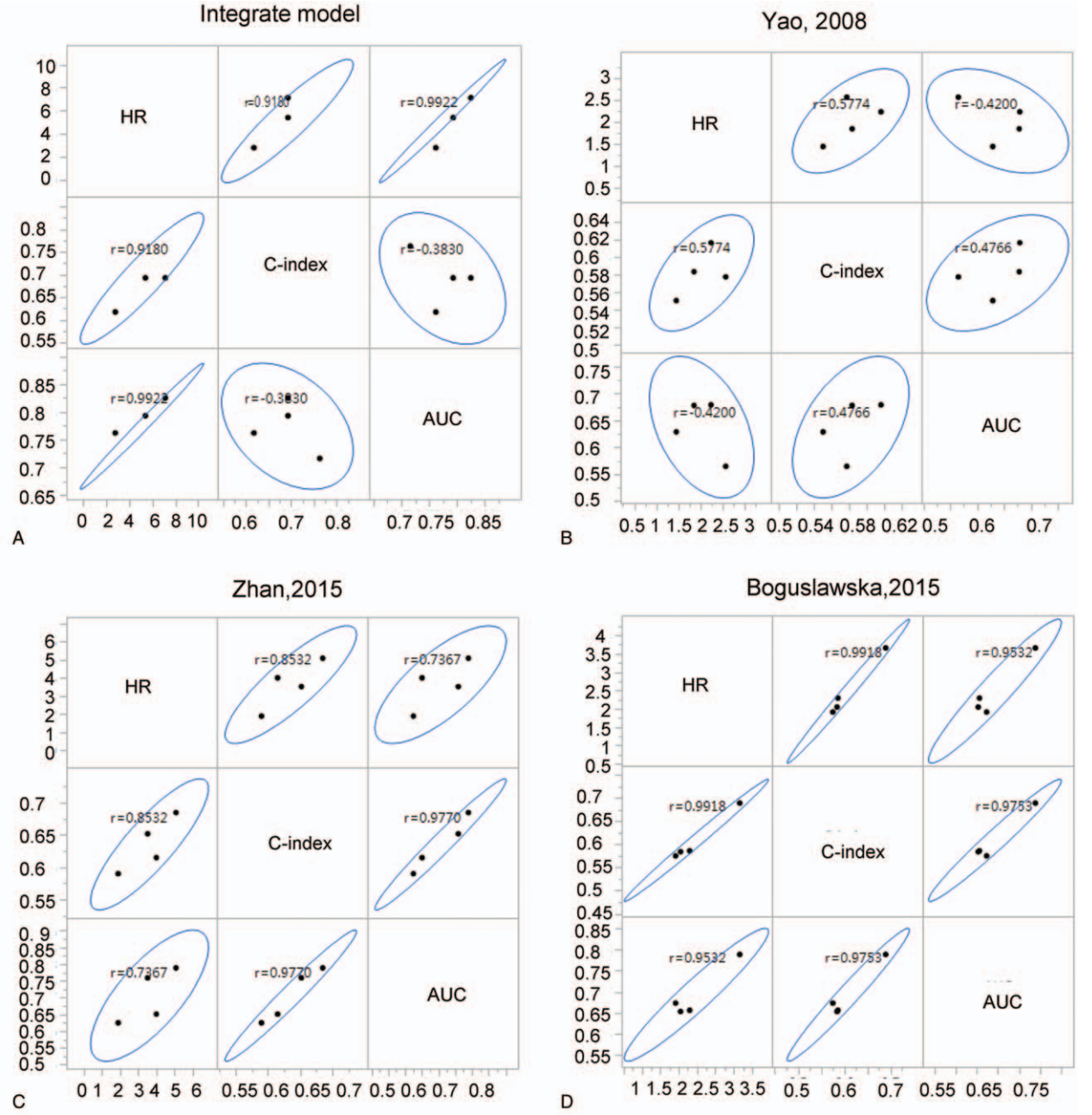

Figure 3. correlation analysis of 3 indicators. A, The correlation of 3 indicators in integrated model. B, The correlation of three indicators in Yao model. C, The correlation of 3 indicators in Zhan model. D, The correlation of 3 indicators in Boguslawska model. AUC = area under the curve, HR = hazard ratio.

grade, adjuvant treatment, laterality, T stage, Topography, Lymph Node and Metastasis (TNM) stage, and tumor status. The results of log-rank test showed that age, grade, and T stage were significantly associated with OS in ccRCC. Multivariate Cox regression analysis of these factors suggested that grade and T stage were independent factors correlated with OS (Table 2).

3.2. Integrative miRNA and mRNA model in TCGA ccRCC

With LASSO combined multivariate Cox regression, 6 RNAs were obtained, and the result is listed in Table 3. Of these 6 RNAs, 3 mRNA and 1 miRNA were protective RNAS (HRs < 1) and the other 2 mRNAs were risky RNAs (HRs > 1), and the coefficient of multivariate Cox regression is applied to calculate PI for ccRCC.

As a linear combination of the expression values of 6 RNAs, the PI was significantly associated with OS in ccRCC [HR = 7.13, 95% confidence interval (CI) = 3.71–13.70, $P < .001$]. The HR of PI was greater than HRs of grade (HR = 2.20, 95% CI = 1.18–4.08, $P = .012$) or T stage (HR = 2.68, 95% CI = 1.62–4.41, $P < .001$). The patients with ccRCC were ranked by PI value (Fig. 1A). The median of PI value as threshold can classify patients into high-risk group and low-risk group. The result showed that the gene signature can significantly classify survival time of ccRCC patients (Fig. 1B). The survival time of high-risk group is significant shorter than low-risk by log-rank test

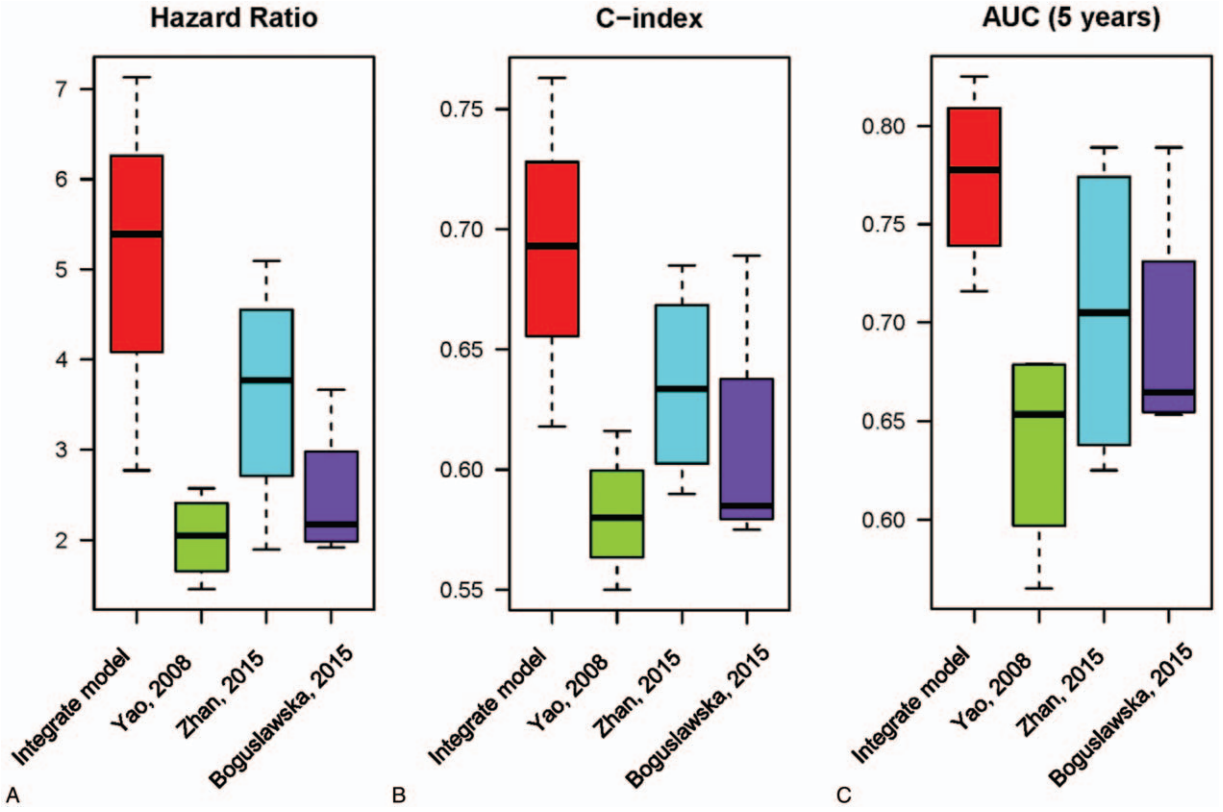

Figure 4. Box plot of 3 indicators [hazard ratio (HR), C-index, AUC) in different datasets. A, Box plot of HRs from 4 models distribute in different datasets. B, Box plot of C-indexes from 4 models distribute in different datasets. C, Box plot of AUCs from 4 models distribute in different datasets. AUC = area under the curve.

($P < .001$). The identified RNA expressions in high-risk and low-risk are listed in Figure 1C, and the value of AUC = 0.748 (3 years) demonstrated that the model performed well in predicting prognosis of ccRCC (Fig. 1D).

3.3. Validating the result in independent dataset

For validation of the result, the GSE22541 dataset is employed to be an independent data to test above result. The dataset contains 24 primary ccRCC tumor and disease-free survival (DFS) time of patients. Although this data set does not have OS time, DFS data can also reflect patient outcomes. We just employed mRNA data to validate the results due to lack of microRNA data. The validation result is shown in Figure 2.

From Figure 2, we find that 5 mRNAs can significantly classify 2 groups into high-risk and low-risk ($P = .03$). The PI was significantly associated with DFS in independent data of ccRCC (HR = 2.77, 95% CI = 1.07–7.71). The value of AUC = 0.762 (3 years) also indicated that the model performed well. Above results demonstrated that the integrative model could effectively classify patients.

3.4. Other gene signature of ccRCC performance in 4 datasets

For further validation the result, we tested the model in other ccRCC data in TCGA. Moreover, we also validate the other 3 models in 4 data sets. These 3 gene signature models that were published previously are listed in Table 4.

For estimating the performance of various gene signature models, 3 indicators (HR, C-index, and AUC) of prognostic models need to be calculated. These 3 indicators were analyzed correlation in each other. Thus, the relationship of these 3 indicators in 4 models was assayed (Fig. 3).

The 3 indicators of Boguslawska model showed the strongest collinearity (Fig. 3D). The collinearity represented the model has good generalization ability. In addition, 1 value of HR in our study is missing. The integrate model from our study showed null value in TCGA_GA450 dataset. Because low-risk group that classified by integrate model has no end event occurs.

For testing performance of gene signature models, the box plot was employed to test the variation among indicators. Therefore, we consider 3 indicators (HR, C-index, and AUC) to evaluate the effect of all models in the 4 datasets (Fig. 4). These 3 indicators usually indicate the capacity of model prediction and high level of these indicators represents better performance of the model. The box plot also indicated the dispersion of gene signature models in different data sets. The results showed that the integrate model from our work had higher HR, C-index, and AUC among all datasets.

3.5. Gene Ontology enrichment of 4 gene signature models

The results in Figure 4 show that 3 indicators of the integrated model were higher than those of other models. Thus, we try to analyze the GO enrichment and pathways in which these models involved in. Of these gene signature models, the number of genes

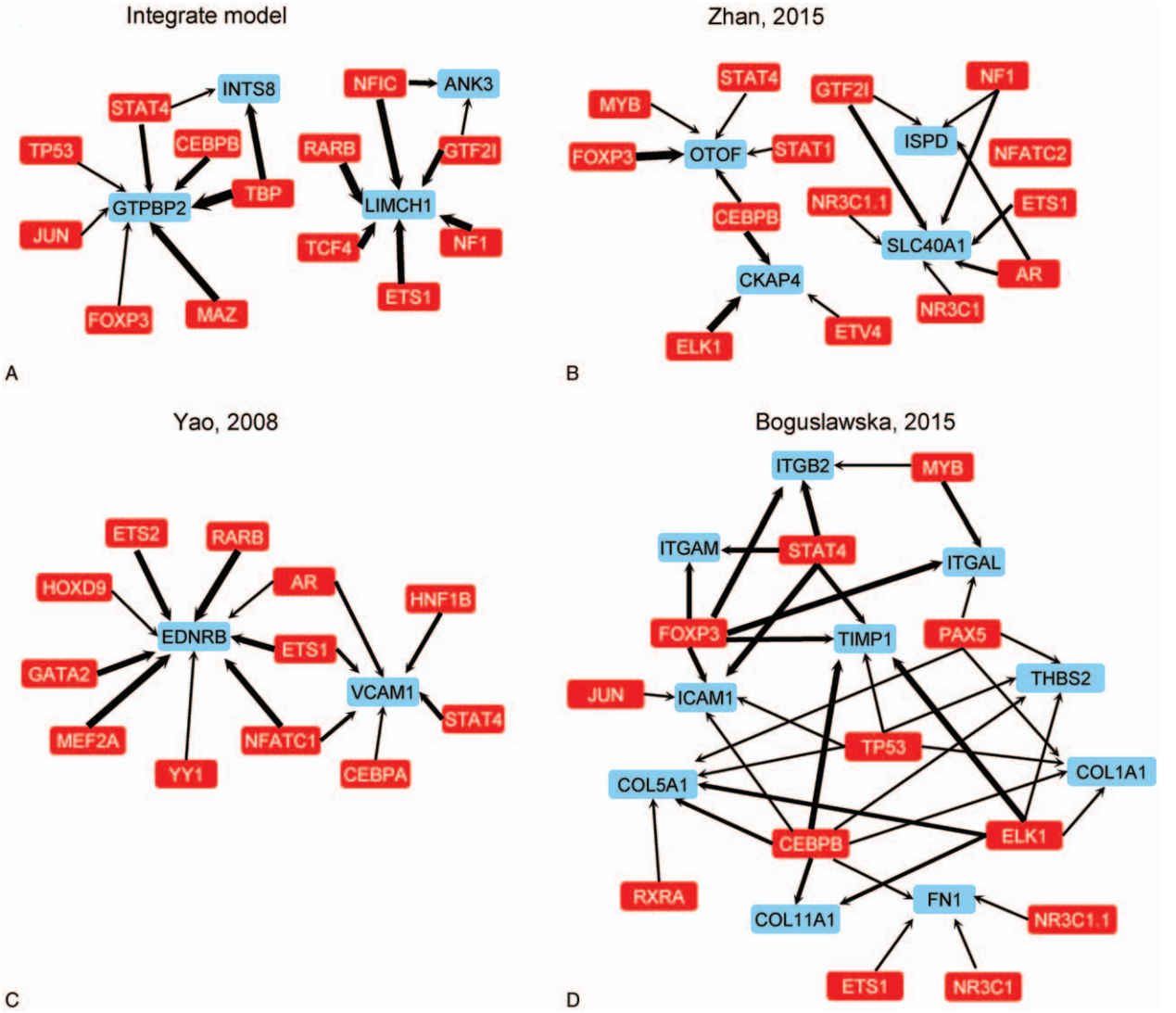

Figure 5. The gene regulation network of 4 gene signature models. The red blocks represent transcription factor (TF) and blue blocks represent target (genes in each gene signature). A, The gene regulation network and pathway analysis of integrate model in the cancer genome atlas (TCGA) dataset. B, The gene regulation network and pathway analysis of Zhan et al model in TCGA dataset. C, The gene regulation network and pathway analysis of Yao et al model in TCGA dataset. D, The gene regulation network and pathway analysis of Boguslawska et al model in TCGA dataset.

in a gene signature model is so small that it is difficult to enrich in GO analysis. Therefore, TF of these genes in gene signature was involved in pathway analysis. The regulation network of TF and genes was constructed by method section (Fig. 5). The integrate gene signature model from our work showed that 4 genes were regulated by 13 TFs (Fig. 5A). The width of lines represented the weighted of regulation by correlation coefficient of their expression level. Regulation network of other gene signature models are listed in Figure 5B, C, and D, respectively. The results showed that these genes shared some common TFs such as STAT4, ETS1, and FOXP3.

For further investigating the GO enrichment and pathway of these genes and TFs, ClusterProfiler package was employed to analyze 4 models. The above package can compare the results of biological process, cellular component, molecular function, and KEGG pathway in 4 models (Fig. 6). The results of biological process suggested that the 4 gene signature models share many similar processes (Fig. 6A). The molecular function of these gene signature models showed that integrate model was similar to model of Boguslawska. And model of Zhan was similar to model of Yao (Fig. 6B). The molecular function enrichment showed that 4 models were very similar (Fig. 6C). In KEGG pathway, the comparison results showed that the integrative model is involved in more cancer-associated pathways (Fig. 6D). The model of Boguslawska et al showed very complex and mainly involved in many signaling pathways associated with cancer. Although these gene signature models and TFs are very different, the biological process and pathways were very similar.

4. Discussion

Our present study combined LASSO and multivariate Cox regression to calculate a prognostic gene signature model from integrative microRNA and mRNA expression of TCGA dataset. The other platform of TCGA and GEO dataset as validation datasets were employed to validate the results. Previous study has

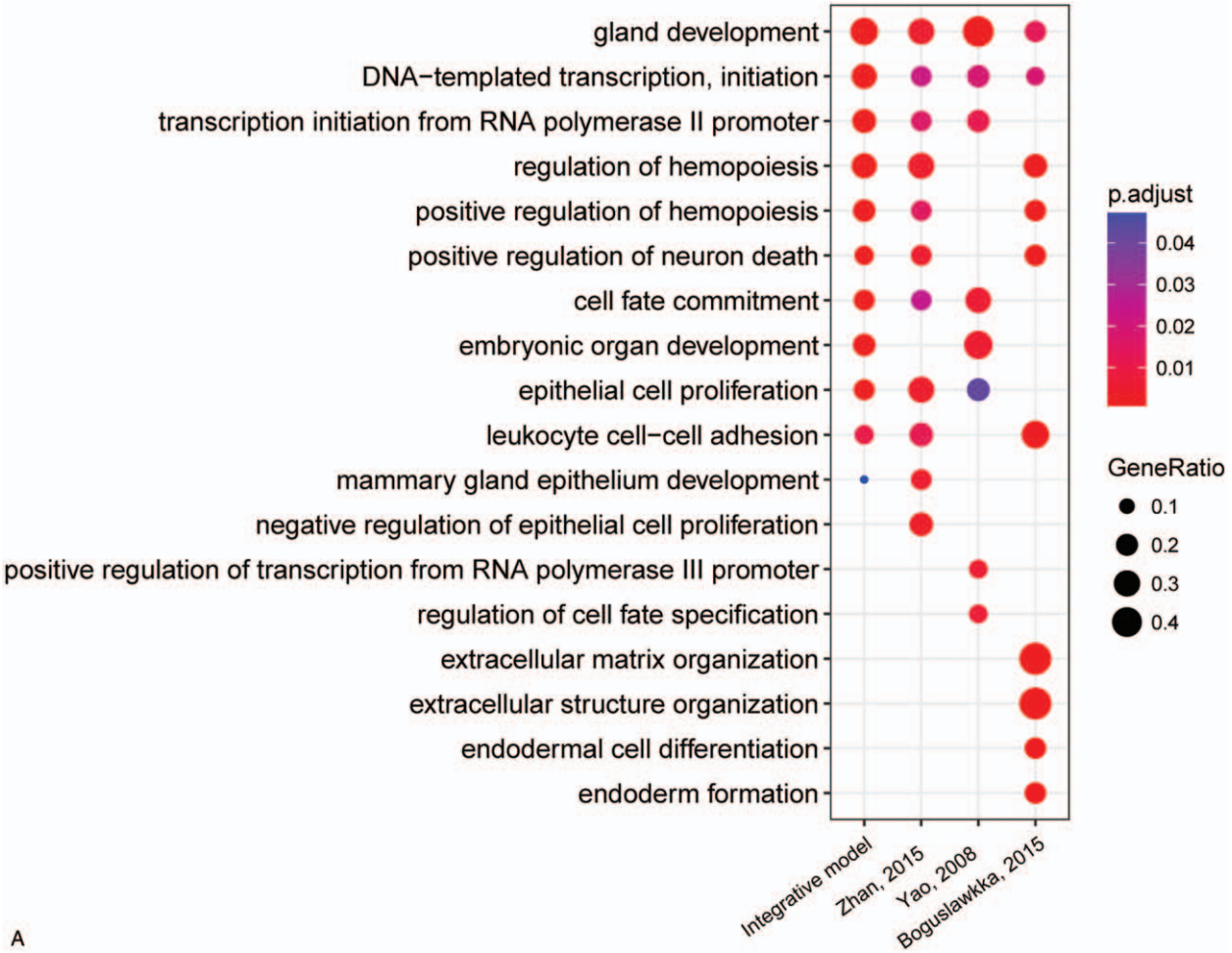

Figure 6. Gene Ontology (GO) and KEGG pathway enrichment comparison in 4 models. A, Biological process enrichment comparison in 4 models. B, Cellular component enrichment comparison in 4 models. C, Molecular function enrichment comparison in 4 models. D, KEGG pathways enrichment comparison in 4 models.

provided many biomarkers for predicting prognosis of ccRCC. In this study, we proposed 5 mRNAs and 1 microRNA (*INTS8*, *GTPBP2*, *ANK3*, *SLC16A12*, *LIMCH1*, and *hsa-mir-374a*) as robust gene signature model that could effectively predict the prognosis for ccRCC. In addition, we also found a regulation pair of hsa-mir-374a and *ANK3* from TargetScan.

Of these genes, *INTS8*, *ANK3*, and *LIMCH1* indicated that they are associated with renal cancer by previous publication.[36–38] To the best of our knowledge, we did not find the *GTPBP2* and *SLC16A12* associated with kidney cancer. Although the gene *hsa-mir-374a* is associated with cancer in many reports, there is no study on ccRCC. Previous studies have shown that hsa-mir-374a (HR = 0.64, 95% CI: 0.48–0.86) can reduce the risk of colorectal cancer.[39] Our findings in kidney cancer also showed similar results (HR = 0.51, 95% CI: 0.29–0.89), so we hypothesized that hsa-mir-374a could reduce the risk of death. These 6-gene signatures showed robust ability in predicting prognosis of ccRCC.

Generally, gene signature prediction for prognosis mainly derived from Cox regression. However, different data preprocess and steps for Cox regression might lead to different results. This study combined the genes with differential expression, univariate Cox regression, LASSO, and multivariate Cox regression method to obtain gene signature for prognosis of ccRCC. In addition, 3 indictors including HR, C-index, and value of AUC were employed to estimate all models in systems level (Fig. 5). The results showed that the integrate model had more advantages than others.

Although our results show more advantages, it does not mean that other models are not good. Among the various gene signature models previously proposed, prognosis is thought to be predictive. In fact, different gene signature has similar pathway and its own special function. The similar pathways are possible to perform similar functions that affect prognosis. The different pathways may represent the heterogeneity of ccRCC.

In the work, the integrate model mainly involved in viral infection and inflammatory bowel disease (IBD)-related pathways. From literature review, there are few reports about viral infection associated with ccRCC. However, there are many reports about the relationship between IBD and renal cancer[40,41]. Although this work could not reveal the relationship between IBD and prognosis of ccRCC, the result might provide a new insight for further study about the ccRCC.

In addition, the gene expression data and clinical data of available ccRCC are very limited, which results in difficulty to further verify. We just used different platforms of TCGA dataset

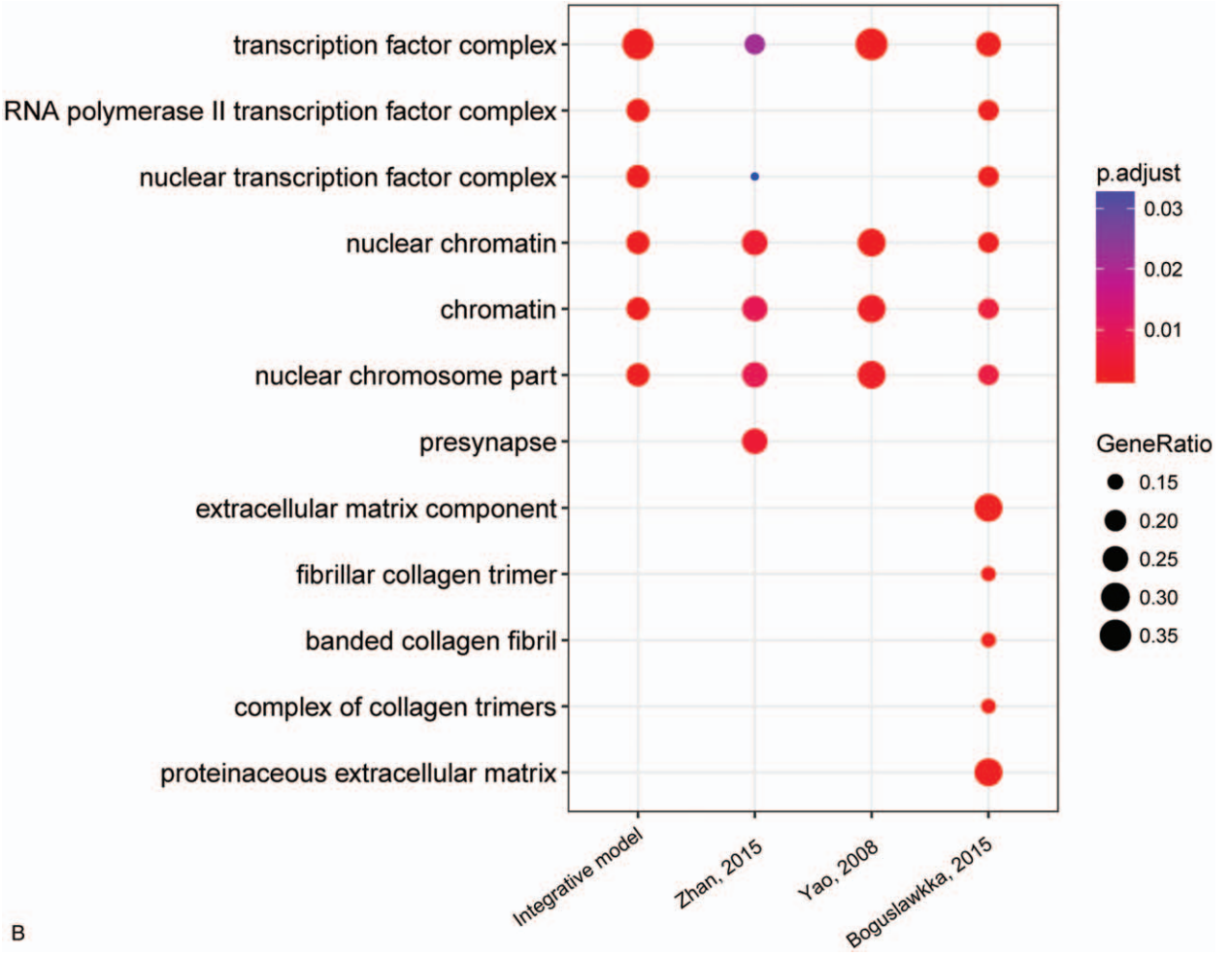

Figure 6. (Continued).

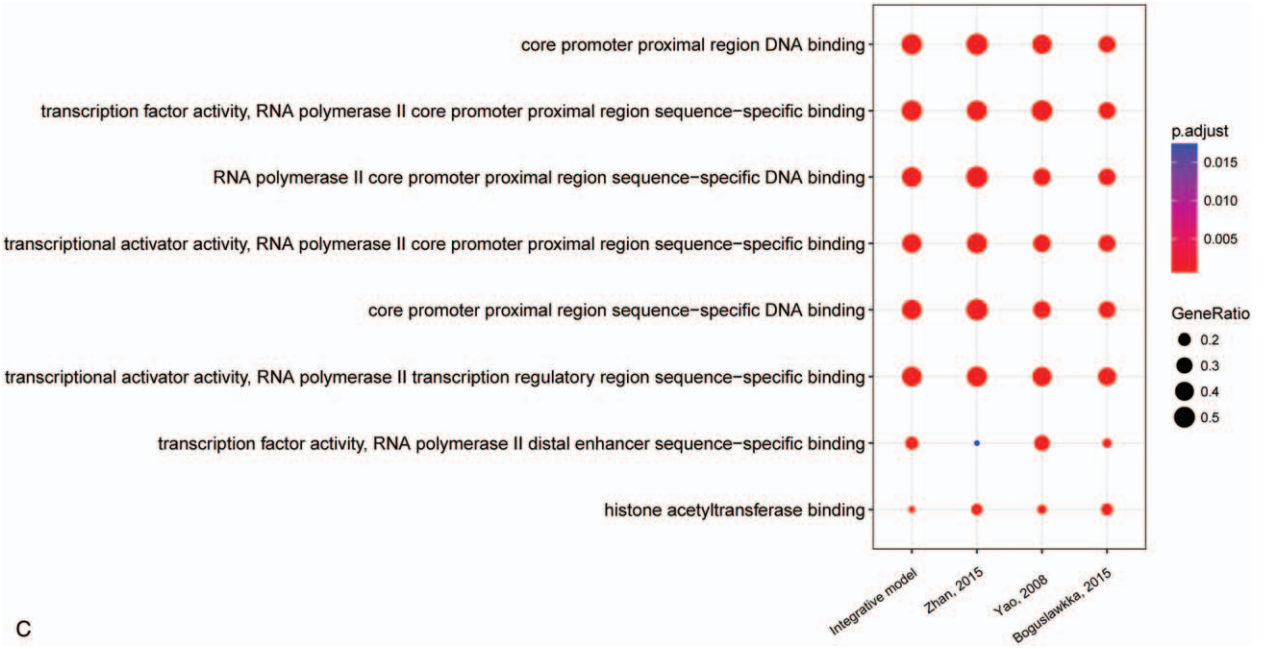

Figure 6. (Continued).

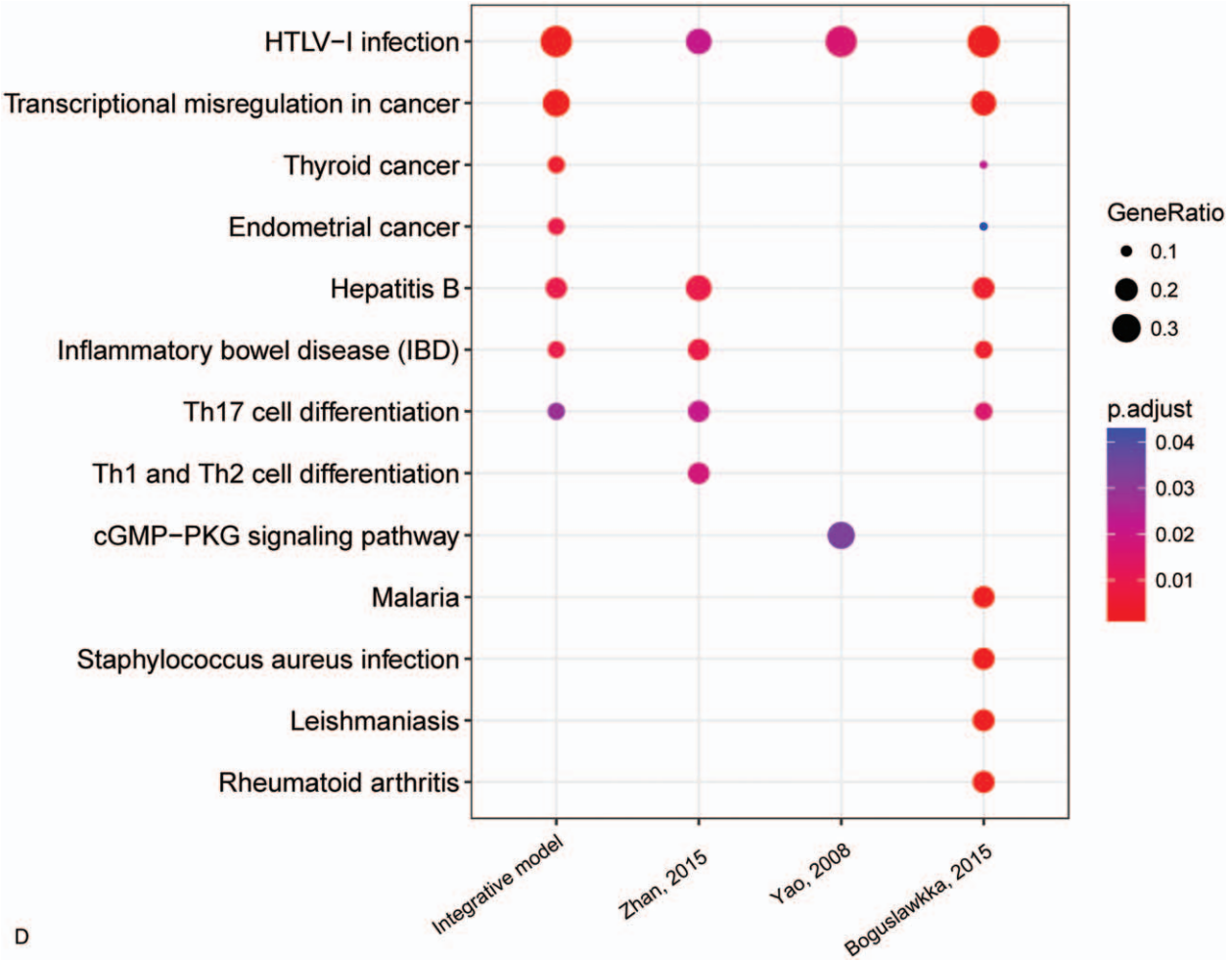

Figure 6. (Continued).

and GEO dataset as independent datasets for training and validation. For further validation of different, we test other 3 gene signature models (from Cox regression method) in different datasets. Moreover, the integrate model indeed showed greater stability and versatility in the TCGA and GSE22541 datasets.

Despite the limited data available, the data we obtained may have bias. However, the gene markers obtained by LASSO coupled multivariate Cox regression are indeed more stable in various public databases. In this study, we propose the optimization steps for analyzing gene prognostic markers by Cox regression. In addition, when gene markers are too scarce to enrich their functions by GO analysis, we can further analyze GO functional enrichment by predicting their TFs. We expect to find more and more stable genetic markers by this way to provide a more scientific reference for drug development and clinical decision-making.

Acknowledgment

The publications retrieval from staff in Evidence Based Medicine Center is appreciated by the authors.

Author contributions

Conceptualization: Peng Chang, Kehu Yang.
Data curation: Jingyun Zhang.
Formal analysis: Peng Chang, Juan Ling.
Funding acquisition: Zhitong Bing.
Investigation: Jinhui Tian, Xiuxia Li, Yumin Li.
Methodology: Peng Chang, Xiuxia Li.
Project administration: Juan Ling.
Resources: Jingyun Zhang, Long Ge.
Software: Zhitong Bing, Jinhui Tian, Jingyun Zhang, Long Ge.
Supervision: Kehu Yang.
Validation: Zhitong Bing.
Visualization: Jinhui Tian, Yumin Li.
Writing – original draft: Peng Chang, Zhitong Bing, Yumin Li.
Writing – review and editing: Peng Chang, Kehu Yang.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin 2018;68:7–30.
[2] Jemal A, Siegel R, Ward E, et al. Cancer Statistics, 2007. CA Cancer J Clin 2007;57:43–66.
[3] Christinat Y, Krek W. Integrated genomic analysis identifies subclasses and prognosis signatures of kidney cancer. Oncotarget 2015;6: 10521–31.
[4] Takahashi M, Rhodes DR, Furge KA, et al. Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. Proc Natl Acad Sci U S A Proc Natl Acad Sci U S A 2001; 98:9754–9.

[5] Kosari F, Parker AS, Kube DM, et al. Clear cell renal cell carcinoma: gene expression analyses identify a potential signature for tumor aggressiveness. Clin Cancer Res 2005;11:5128.
[6] Sültmann H, Heydebreck AV, Huber W, et al. Gene expression in kidney cancer is associated with cytogenetic abnormalities, metastasis formation, and patient survival. Clin Cancer Res 2005;11(2 pt 1):646.
[7] Yao M, Tabuchi H, Nagashima Y, et al. Gene expression analysis of renal carcinoma: adipose differentiation-related protein as a potential diagnostic and prognostic biomarker for clear-cell renal carcinoma. J Pathol 2005;205:377.
[8] Zhao H, Ljungberg B, Grankvist K, et al. Gene expression profiling predicts survival in conventional renal cell carcinoma. PLoS Med 2006; 3:e13.
[9] Yao M, Huang Y, Shioi K, et al. A three-gene expression signature model to predict clinical outcome of clear cell renal carcinoma. Int J Cancer 2008;123:1126–32.
[10] Mertz K, Demichelis FA, Hirsch M, et al. Association of cytokeratin 7 and 19 expression with genomic stability and favorable prognosis in clear cell renal cell cancer. Int J Cancer 2008;123:569.
[11] Heinzelmann J, Henning B, Sanjmyatav J, et al. Specific miRNA signatures are associated with metastasis and poor prognosis in clear cell renal cell carcinoma. World J Urol 2011;29:367–73.
[12] Cancer Genome Atlas Research NetworkComprehensive molecular characterization of clear cell renal cell carcinoma. Nature 2013;499:43–9.
[13] Brooks SA, Brannon AR, Parker JS, et al. ClearCode34: a prognostic risk predictor for localized clear cell renal cell carcinoma. Eur Urol 2014;66:77.
[14] Gulati S, Martinez P, Joshi T, et al. Systematic evaluation of the prognostic impact and intratumour heterogeneity of clear cell renal cell carcinoma biomarkers. Eur Urol 2014;66:936–48.
[15] Heinzelmann J, Unrein A, Wickmann U, et al. MicroRNAs with prognostic potential for metastasis in clear cell renal cell carcinoma: a comparison of primary tumors and distant metastases. Ann Surg Oncol 2014;21:1046–54.
[16] Fu H, Liu Y, Xu L, et al. Galectin-9 predicts postoperative recurrence and survival of patients with clear-cell renal cell carcinoma. Tumour Biol 2015;36:5791–9.
[17] Ge YZ, Wu R, Xin H, et al. A tumor-specific microRNA signature predicts survival in clear cell renal cell carcinoma. J Cancer Res Clin Oncol 2015;141:1291–9.
[18] Kim HL, Halabi S, Li P, et al. A molecular model for predicting overall survival in patients with metastatic clear cell renal carcinoma: results from CALGB 90206 (Alliance). EBioMedicine 2015;2:1814–20.
[19] Rini B, Goddard A, Knezevic D, et al. A 16-gene assay to predict recurrence after surgery in localised renal cell carcinoma: development and validation studies. Lancet Oncol 2015;16:676–85.
[20] Tang K, Xu H. Prognostic value of meta-signature miRNAs in renal cell carcinoma: an integrated miRNA expression profiling analysis. Sci Rep 2015;5:10272.
[21] Zhan Y, Guo W, Zhang Y, et al. A five-gene signature predicts prognosis in patients with kidney renal clear cell carcinoma. Comput Math Methods Med 2015;2015:1–7.
[22] Boguslawska J, Kedzierska H, Poplawski P, et al. Expression of genes involved in cellular adhesion and ECM-remodelling correlates with poor survival of renal cancer patients. J Urol 2016;195:1892–902.
[23] Dai J, Lu Y, Wang J, et al. A four-gene signature predicts survival in clear-cell renal-cell carcinoma. Oncotarget 2016;7:82712.
[24] de Velasco G, Culhane AC, Fay AP, et al. Molecular subtypes improve prognostic value of international metastatic renal cell carcinoma database consortium prognostic model. Oncologist 2017; 22:286–92.
[25] Ge YZ, Wu R, Xin H, et al. A tumor-specific microRNA signature predicts survival in clear cell renal cell carcinoma. J Cancer Res Clin Oncol 2015;141:1291.
[26] Wu X, Weng L, Li X, et al. Identification of a 4-microRNA signature for clear cell renal cell carcinoma metastasis and prognosis. PLoS One 2012;7:e35661.
[27] Liang B, Zhao J, Wang X. A three-microRNA signature as a diagnostic and prognostic marker in clear cell renal cancer: an in silico analysis. PLoS One 2017;12:e0180660.
[28] Ran L, Liang J, Deng X, et al. miRNAs in prediction of prognosis in clear cell renal cell carcinoma. BioMed Res Int 2017;2017:1–6.
[29] Wuttig D, Zastrow S, Füssel S, et al. CD31, EDNRB and TSPAN7 are promising prognostic markers in clear-cell renal cell carcinoma revealed by genome-wide expression analyses of primary tumors and metastases. Int J Cancer 2012;131:E693–704.
[30] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
[31] Agarwal V, Bell GW, Nam J-W, et al. Predicting effective microRNA target sites in mammalian mRNAs. eLife 2015;4:e05005.
[32] Enright AJ, John B, Gaul U, et al. MicroRNA targets in Drosophila. Genome Biol 2003;5:R1.
[33] Yu G, Wang LG, Han Y, et al. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS 2012; 16:284–7.
[34] Kathrin P, Jan S, Michaela N, et al. How microRNA and transcription factor co-regulatory networks affect osteosarcoma cell proliferation. PLoS Comput Biol) 2013;9:e1003210.
[35] Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:1–6.
[36] Federico A, Rienzo M, Abbondanza C, et al. Pan-cancer mutational and transcriptional analysis of the integrator complex. Int J Mol Sci 2017;18:936.
[37] Morris MR, Ricketts CJ, Gentle D, et al. Genome-wide methylation analysis identifies epigenetically inactivated candidate tumour suppressor genes in renal cell carcinoma. Oncogene 2011;30:1390–401.
[38] Eckel-Passow JE, Serie DJ, Bot BM, et al. ANKS1B is a smoking-related molecular alteration in clear cell renal cell carcinoma. BMC Urol 2014;14:14.
[39] Slattery ML, Herrick JS, Mullany LE, et al. An evaluation and replication of miRNAs with disease stage and colorectal cancer-specific mortality. Int J Cancer 2015;137:428–38.
[40] Tsianos EV, Katsanos KH, Christodoulou D, et al. The epidemiological profile of inflammatory bowel disease in different parts of North-West Greece. Ann Gastroenterol 2005;18:434–40.
[41] Fialho A, Fialho A, Shabbir A, et al. Su1812 renal cancer is associated with the use of immunomodulators in patients with inflammatory bowel disease. Gastroenterology 2016;150:S559–60.