Discovering Cure Hub Genes of Chromophobe Cell Renal Carcinoma Using weighted gene co-expression network analysis (WGCNA) and the Cure Defective Models

Maryam Ahmadian[a], Zahra Molavi[b], Ahmad Reza Baghestani[c], Ali Akbar Khadem Maboudi[a], Takoua Jendoubi[d]

[a]Department of Biostatistics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Science, Tehran, Iran.

[b] Proteomics Research Center, Shahid Beheshti University of Medical Science, Tehran, Iran.

[c]Department of Biostatistics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Science, Tehran, Iran.

[d]Department of Statistics Science, UCL, UK

Abstract

Renal cell carcinoma (RCC) is a common type of malignant tumor in the urinary system, with a high incidence rate and poor prognosis. It is frequently observed that a substantial proportion of patients who respond favorably to treatment appear to be free of symptoms in long-term follow-up and may be regarded as 'cured.' This study aimed to identify key genes and pathways associated with RCC using WGCNA and the 3-parameter defective Gompertz model as a cure model for identifying targeted therapeutic genes.

Six key modules and genes related to cancer stage, pathology, metastasis, status, and overall survival were identified using WGCNA, and hub genes were identified using the PPI network. Two methods, Cox regression and the defective Gompertz model, were used to identify survival-related genes (SRGs) and cured survival-related genes (CSRGs) in RCC.

Of the 14 hub genes identified, 10 were common between the defective 3-parameter Gompertz and Cox models. six genes (NCAPG, TTK, DLGAP5, TOP2A, BUB1B, and BUB1) of them exhibited the highest area under the curve (AUC) values. Additionally, the defective 3-parameter Gompertz model calculated the min and max expression values for the cure rate, and it was found that six genes (TTK, KIF20A, DLGAP5, BUB1, AURKB, and CDC45) exhibited a strong reduction in the cure rate at their maximum expression levels.

The results suggest that targeting these genes may hold promise as potential therapeutic strategies for improving treatment outcomes in RCC patients. The identified hub genes also showed excellent

prognostic value and potential use as diagnostic biomarkers. This study provides new insights into the molecular mechanisms underlying RCC and highlights the potential clinical significance of targeting specific genes and pathways using the defective 3-parameter Gompertz model.

Keywords: Renal cell carcinoma, defective 3-parameter Gompertz model, cure rate, prognostic biomarker

Introduction

Renal cell carcinoma (RCC) is a type of cancer that accounts for 5% and 3% of all oncological diagnoses worldwide in men and women, respectively (1). Men are almost twice as likely to develop kidney cancer as women, and the incidence rates of RCC increase consistently with advancing years, reaching a peak at about age 75 (2). About half of all RCC cases are identified before the age of 65 (1, 3). RCC is classified into four pathological phases based on tumor size, invasion extent, and metastasis. Most RCCs arise in the kidney's cortex, including the tubular apparatus, glomerulus, and collecting duct (4).

There are numerous genetic, clinical, and environmental risk factors associated with RCC, including smoking, drinking, high blood pressure, obesity, and occupational and environmental exposures to toxins such as cadmium, asbestos, and ionizing radiation. Inherited disorders, including Von Hippel-Lindau disease, hereditary leiomyomata's RCC, and hereditary papillary RCC, have also been linked to renal carcinoma (5-7). The cure rate for kidney cancer depends on various factors such as the stage of cancer, age, and treatments used. Early detection and prompt treatment lead to a high cure rate, with a five-year survival rate of 93% for stage I kidney cancer. However, if the cancer has spread, the cure rate drops significantly, with a five-year survival rate of around 12% for stage IV kidney cancer (8, 9).

.

The identification of biomarkers and potential therapeutic targets for RCC progression is a crucial area of research in cancer biology. In this study, the main focus was on identifying survival-related genes (SRGs) and cured survival-related genes (CSRGs) in RCC using two methods: Cox regression has been used to found SRGs and the defective Gompertz 3-parameter has been used to found CSRGs. In the standard survival models assume that all subjects are susceptible to the event of interest (such as recurrence or death from the disease). Cox regression is a standard survival model and a statistical technique used to examine how multiple variables influence the time until a specific event occurs. This method is considered semi-parametric since it does not make assumptions about the exact distribution of the data. (10, 11). who respond favorably to treatment appear to be free of symptoms in long-term follow up and may be regarded as 'cured' and these individuals will never experience the event of interest; these risk-free subjects. Cox method cannot survey this characteristic of subjects. So, we need models that monitor disease cure agents.

Existence of the cured fraction is indicated by a long flat tail which is not close to zero in Kaplan-Meier curves (12, 13); as it is shown in Figure 1.

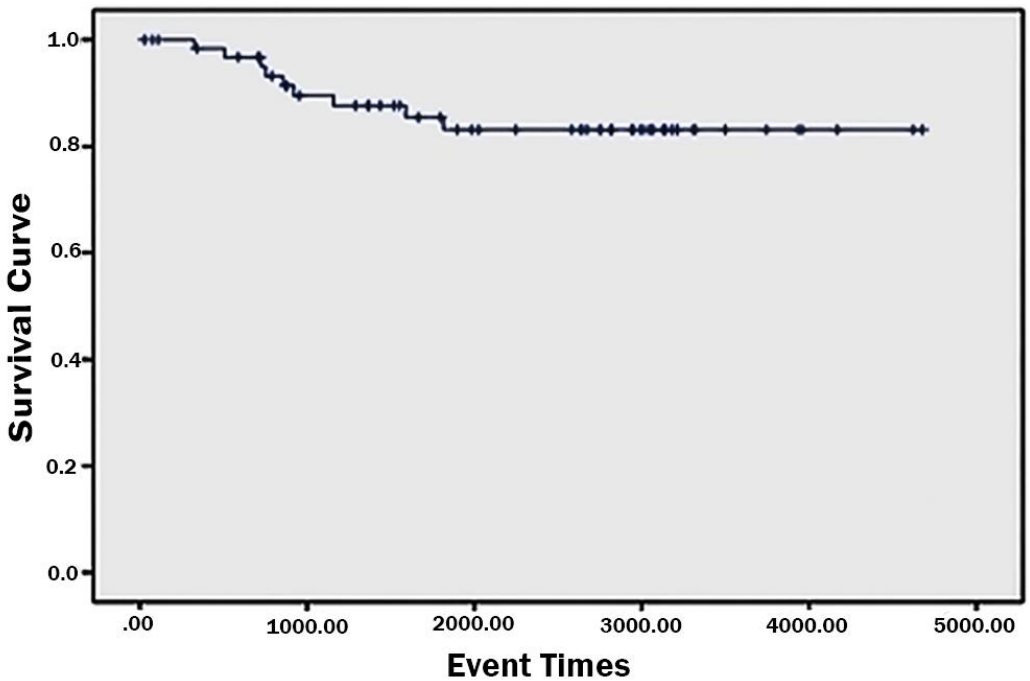


Fig.1. Kaplan-Meier curve of the kidney chromophobe cancer in TCGA data

Mixture models are commonly employed in modeling cure rates. A standard mixture model assumes a survival function of $S(t) = p + (1-p)s_0(t)$, where $p \in (0,1)$ and $s_0(t)$ is a typical survival function that declines to zero as time progresses. As time passes, $S(t)$ approaches p. Recently, defective distributions have been used for modeling cure rates. Defective distributions have the ability to act as cure rate models by changing the standard parameter domain without adding any extra parameters to the model. These changed parameters are known as defective parameters. The proportion of the cured population is determined by computing the limit of the survival function, which ranges from zero to one and is no longer a proper distribution. As a result, the survival function will cover to a value in range (0,1), which leads to a cure fraction. In the literature, two defective distributions have been proposed for cure rate modeling: the defective Gompertz and the defective Inverse Gaussian distributions (13, 14).

In this research we used the defective 3-parameter Gompertz model is proposed by Haji Zadeh and Baghestani for estimating cure fraction (15). It examines the impact of variables that result in risk-free subjects and cure rate. this issue is a reason for the difference between the common method of cox and the defective 3-parameter Gompertz model. It could be applied to kidney cancer

data to gain a better understanding of its effectiveness (13, 14). This model is important because the majority of studies have focused on identifying SRGs and hazard disease-related gene combinations, while CSRGs have been largely ignored. By incorporating this model into the analysis, researchers can potentially identify new biomarkers and therapeutic targets for kidney cancer treatment. this study is the first study that examines the gene that effected on the cure of patients and this issue has not been investigated in previous studies.

In summary, we applied WGCNA to the TCGA-KICH dataset to identify gene modules associated with survival outcomes in kidney cancer patients. Furthermore, both Cox proportional hazards regression and the defective Gompertz 3-parameters model were utilized to identify potential cures based on gene expression levels within these modules (figure 2).

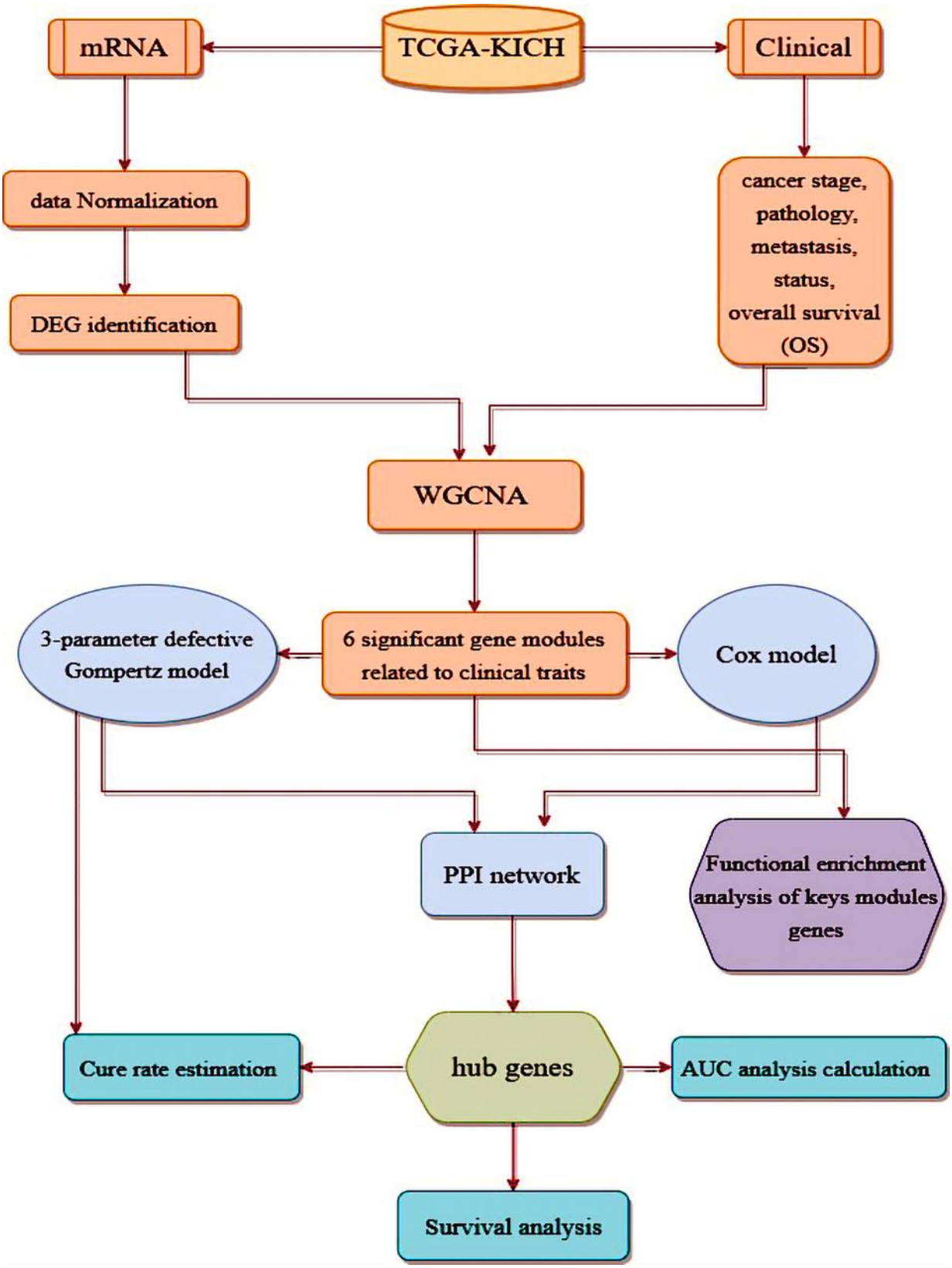


Fig.2. Flow chart of data preparation, processing, analysis, and validation.

Materials and methods

1. Data selection from TCGA and analysis

In this study, RNA-seq data and clinical information from 90 patients with kidney cancer were obtained from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). The RNA-seq data consisted of 25 normal cases and 65 cancer cases. The clinical information included data on the overall survival (OS) of patients with kidney cancer.

2. KICH differentially expressed genes (DEGs) identification

To reduce the impact of outliers on the overall data, 7940 genes were selected based on an average expression greater than 1 and the greatest MAD. In this study, the Limma R software package was utilized for identifying DEGs between different groups. First, the 'lmFit' function was used for multiple linear regression analysis on the expression spectrum dataset. The significance of DEGs was determined using the 'eBays' function. A sample cluster map was drawn using the R language to identify and remove any outliers. Finally, the gene expression profiles were normalized using log(x+1) transformation to ensure that upregulated or downregulated genes were evenly distributed around 0. Volcano plot of the DEGs were created using the 'ggplot2' function in R.

3. Calculation of gene Co-expressing using WGCNA

The process of calculating the Co-expression between different genes involved in kidney cancer were performed using the Weighted Gene Co-expression Network Analysis (WGCNA) package. This approach involves constructing a network of genes based on their expression patterns and then measuring the similarity or "correlation" between the genes in the network. It is a powerful tool for uncovering the complex interactions between genes and for identifying key players in the disease process (17).

The DEG data screened out in the TCGA KICH data were selected to establish a weighted gene co-expression network with WGCNA package (16) in R Studio 3.6.0 software(http://www.r-project.org/). Initially, then data were tested by to check if they were the good samples and good genes.

Then, a weighted adjacency matrix was constructed using a power function $A_{mn} = |C_{mn}|^{\beta}$ (C_{mn} = Pearson's correlation between genes m and n; A_{mn} = adjacency between genes m and n), and a soft-thresholding parameter was applied to emphasize strong correlations(β) while penalizing weak ones. After selecting the power parameter β, the adjacency matrix was converted into a topological overlap matrix (TOM), which measures the network connectivity of a gene as the sum of its adjacency with all other genes in the network. The corresponding

dissimilarity (1-TOM) was then calculated. To cluster genes with similar expression profiles into gene modules, average linkage hierarchical clustering was performed using the TOM-based dissimilarity measure. Gene dendrograms were constructed with a minimum size of 30 for gene groups. The module eigengenes dissimilarity measure was used to further analyze the gene modules. The cut line of the module dendrogram was selected and then some modules were merged to refine the analysis.

4. Identifying significant gene modules related to clinical traits

The WGCNA algorithm employs the module eigengene (ME) concept to assess the relationship between gene modules and clinical traits. ME is determined as the principal component obtained through a principal component analysis that captures the expression pattern of genes within a specific module. To identify the module highly correlated with RCC, the Pearson correlation coefficient between ME and clinical traits was computed. For intramodular analysis, gene significance (GS) and module membership (MM) were calculated. MM represents the correlation between ME and the gene expression profile and GS are the logarithmic transformation (log10) of the P-value (GS = lgP) indicating the correlation between gene expression and the clinical trait.

5. Functional enrichment analysis of keys module genes

To further understand the function of the selected genes, a functional enrichment analysis was performed using the online database Enrichr (https://maayanlab.cloud/Enrichr/). This involved performing functional enrichment analysis, specifically GO and KEGG pathway analysis, on all genes in the key modules. The analysis produced a list of the top 10 biological processes, cellular components, molecular functions and KEGG pathways, which were ranked based on their P-value. The results were then plotted to facilitate interpretation.

6. Using Cox model for identification of SRGs

The Cox model is a statistical method used in survival analysis to investigate the relationship between survival time and predictor variables. It allows for the estimation of hazard ratios to measure the relative risk of an event occurring in one group compared to another. In identifying hub genes, the Cox model assesses the association between gene expression and survival time of patients with a disease. By examining hazard ratios and p-values, it identifies genes strongly associated with survival, providing insights into disease processes. The Cox model has been applied to identify hub genes in diseases such as cancer and Alzheimer's, aiding in understanding their underlying biology and developing new treatments. we applied Cox proportional hazards regression analysis to extract genes of 6 key modules that are strongly associated with survival outcomes and other clinical variables.

7. Using defective 3-parameter Gompertz model for identification of CSRGs

In this study, we employed the defective 3-parameter Gompertz model to analyze the impact of variables leading to risk-free subjects in770 selected gene modules related to survival in RCC. The defective 3-parameter Gompertz model is a method for analyzing kidney cancer data and understanding the impact of various risk factors on the disease. By focusing on the identification of CSRGs, this model offers a unique perspective that has been largely overlooked in previous studies.

7. Hub genes identification

The genes selected modules, as well as the results of the Cox and defective Gompertz 3-parameter models, were uploaded separately into the STRING database. After filtering for a confidence value greater than 0.4, a protein-protein interaction (PPI) network was constructed. Three PPI networks were visualized using Cytoscape (v3.9.1). MCODE and the MCC and Degree algorithm of the CytoHubba app were used to identify the hub genes common to the three networks.

8. Estimation of cure rate of hub genes

The Gompertz model estimates the cure rate, which represents the proportion of patients who will not experience recurrence or death due to cancer. The minimum and maximum cure expression can help identify the optimal range of gene expression associated with better treatment outcomes, providing valuable insights for personalized treatment options. So, we calculated the minimum and maximum cure expression for identified hub genes.

The survival function of Gompertz 3-parameter distribution is as follows:

$$s(t) = \exp\{-\frac{\beta x^T}{\alpha\eta}\left(e^{\eta e^{\alpha t}} - e^{\eta}\right)\}$$

Here, α and η are shape parameters and belong to real values, T is survival time.

$x^T\beta = b_0 + b_1x + b_2x + \cdots + b_kx$ in which $\beta = (b_0, b_1, \ldots, b_k)$ indicates the coefficients vector to any gene and $x^T = (1, x_0, x_1, \ldots, x_k)$ indicates the covariates vector which we enter any time one gene as covariate and achieved model as univariate. when $\alpha<0$, we have the defective 3-parameter Gompertz (DG) model.

The cure fraction or cure rate of the DGD model is calculated based on the following formula:

$$p = \lim_{t\to\infty} s(t) = \exp\{-\frac{\beta x_i}{\alpha\eta}(e^{\eta} - 1)\}$$

where p is the cure rate parameter.

9. AUC analysis calculation on hub genes

We used AUC analysis to evaluate the performance of hub genes in predicting kidney cancer outcomes using the SurvivalROC package and KM method of R software. AUC, or area under the curve, is a statistical measure that reflects the accuracy of a diagnostic test or predictive model. The higher the AUC value, the more accurate the model is at distinguishing between positive and negative outcomes. In our study, we identified the top hub genes using a comprehensive network analysis approach and then performed AUC analysis on these genes to evaluate their prognostic power for kidney cancer prognosis.

10. Survival analysis of hub genes

The statistical analysis of overall survival (OS) performed using the "survival" package and event-free survival (EFS) performed using the programing in RStudio software (http://www.r-project.org/). To estimate the OS/EFS, Kaplan-Meier survival curves are utilized with the survft function, and the differences in OS/EFS between the high- and low-risk groups are computed using a log-rank test and risk score determined by median. The data subjected to univariate analysis using the Cox proportional hazards regression model through the coxph function in R. The significance of the survival association determined by comparing the p-values of the log-rank and Wald tests with a threshold of 0.05.

Results

1. Identification of DEGs

Following data pre-processing and quality assessment, expression matrices were generated from the 22420 mRNAs. A stringent threshold of adj-P value < 0.05 and $|\log2FC| > 1$ was applied to identify the DEGs. This resulted in a final set of 7940 DEGs, including 5182 up-regulated and 2758 down-regulated genes, which will be further analyzed in subsequent stages of the study (fig.2).

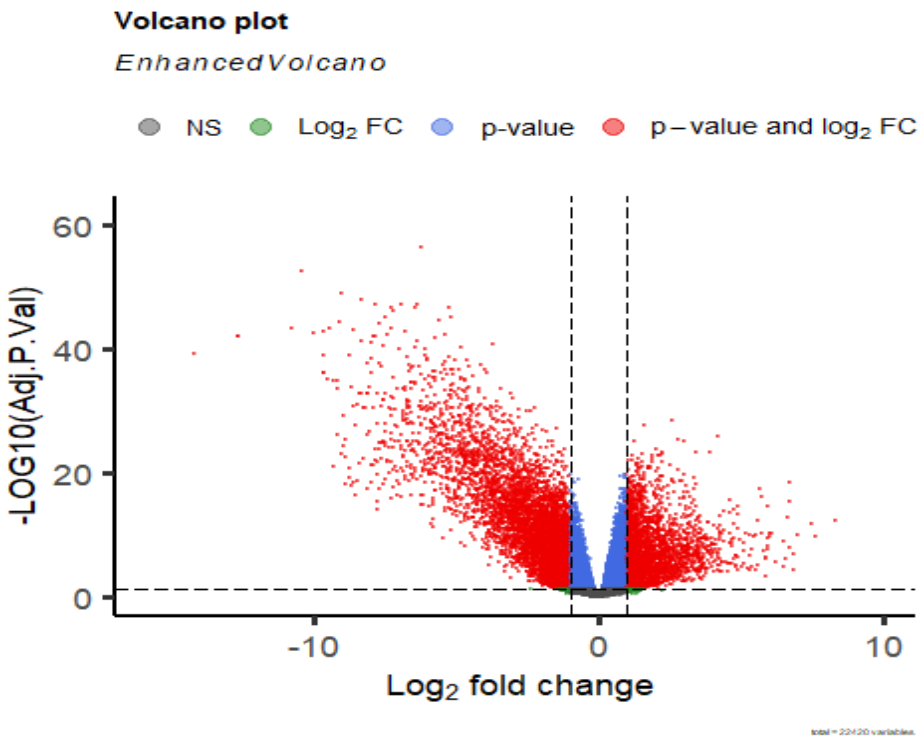


Fig.3.Visualizing Significant Differential Expression of Genes by Volcano Plot of DEGs

2. Co-Expression Network construction

To find the correlation coefficients of the expression level in each of the 90 given samples, we clustered similar expression patterns of DEGs using the 'WGCNA' package in R and removed one outlier sample (figure 3). We ensured a scale-free network by selecting a soft threshold of β=4 (with a scale-free R^2 = 0.90) (figure 4). The gene dendrograms and the corresponding module colors of the identified modules are shown in Figure 4. Eighteen modules were identified, and we used two methods to test the association of each module with the KICH progression.

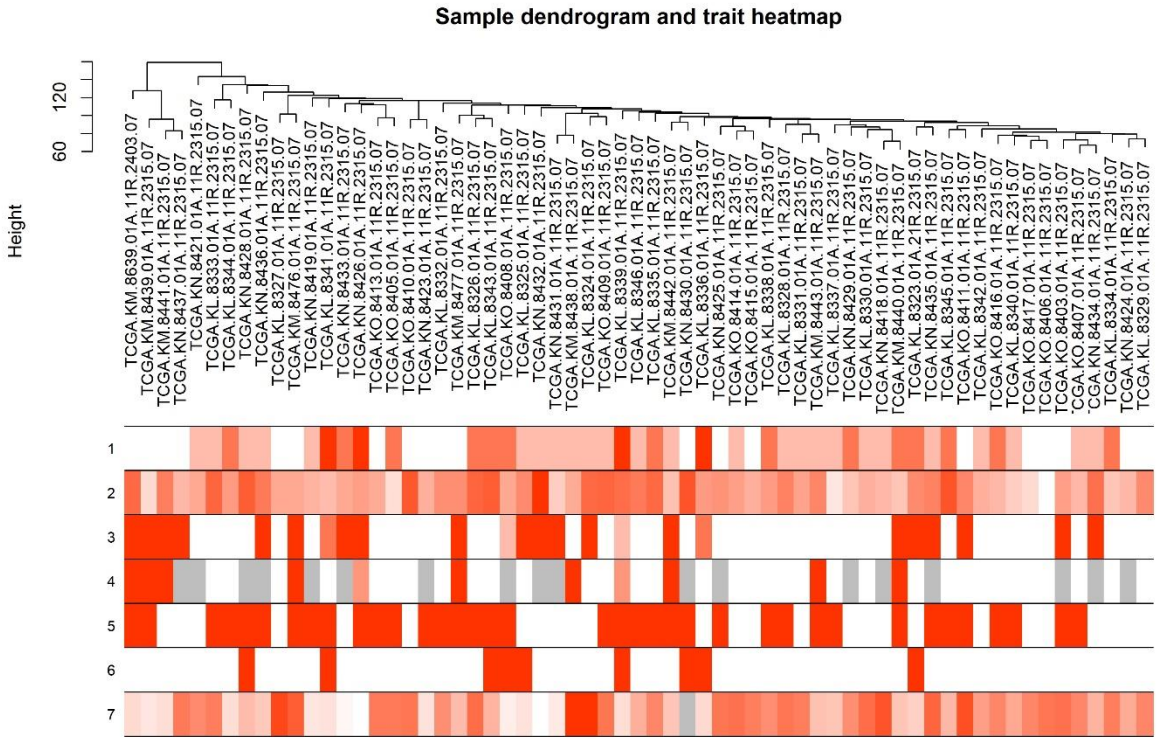


Fig.4. Dendrogram clustering of 90 KICH samples along with their clinical traits. Clustering based on expression data of genes differentially expressed between tumour and non-tumour KICH samples. Clinical traits (age, gender, cancer stage, pathology, metastasis, status, or overall survival (OS)) were proportional to color intensity.

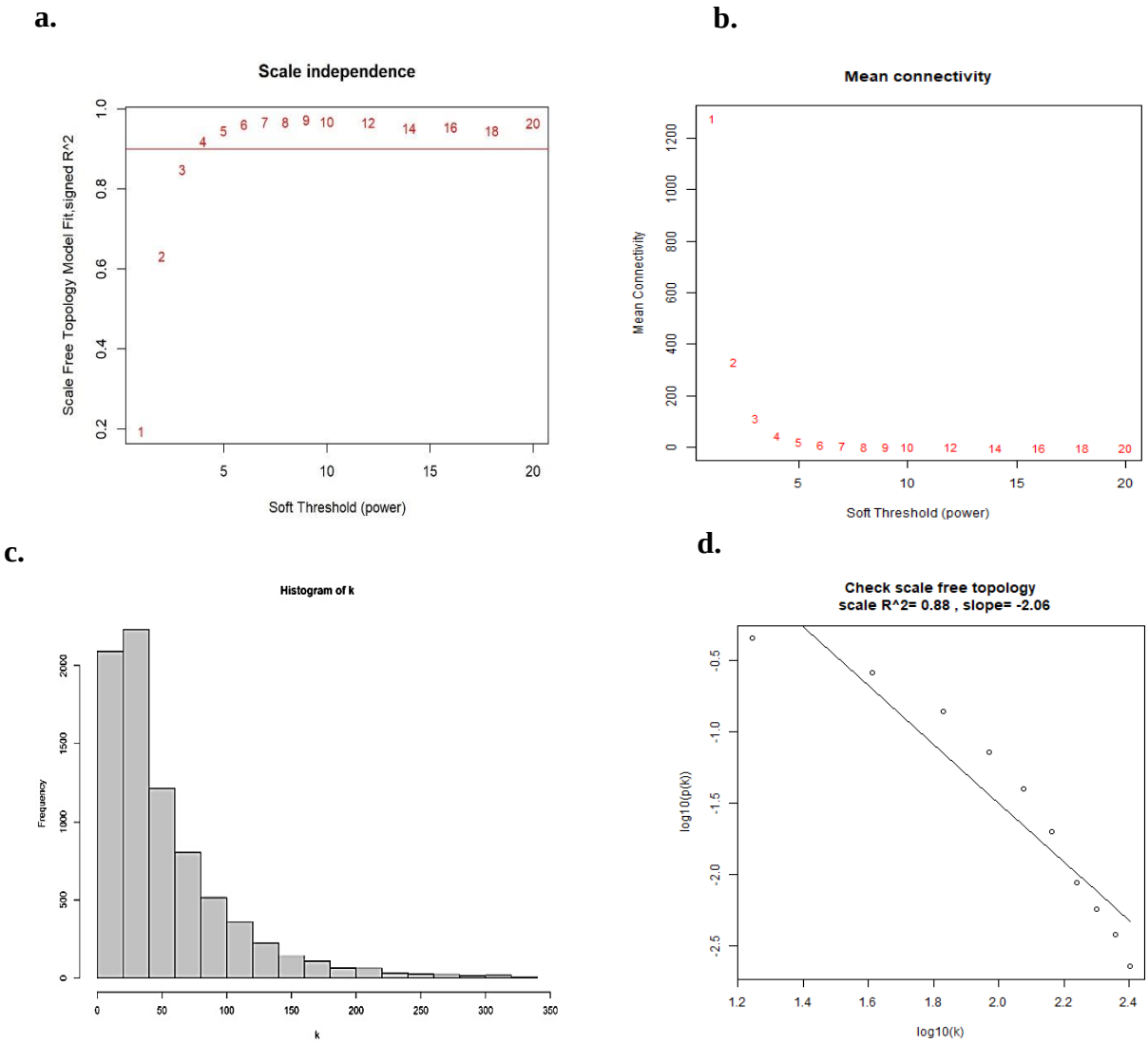


Fig.5. Highlights the process of determining the parameter β in the WGCNA algorithm for the adjacency function. a) analyzing the scale-free fit index for different soft thresholding powers, b) evaluating the mean connectivity for each thresholding power, c) checking the scale-free topology, and d) examining the connectivity distribution when β=4.

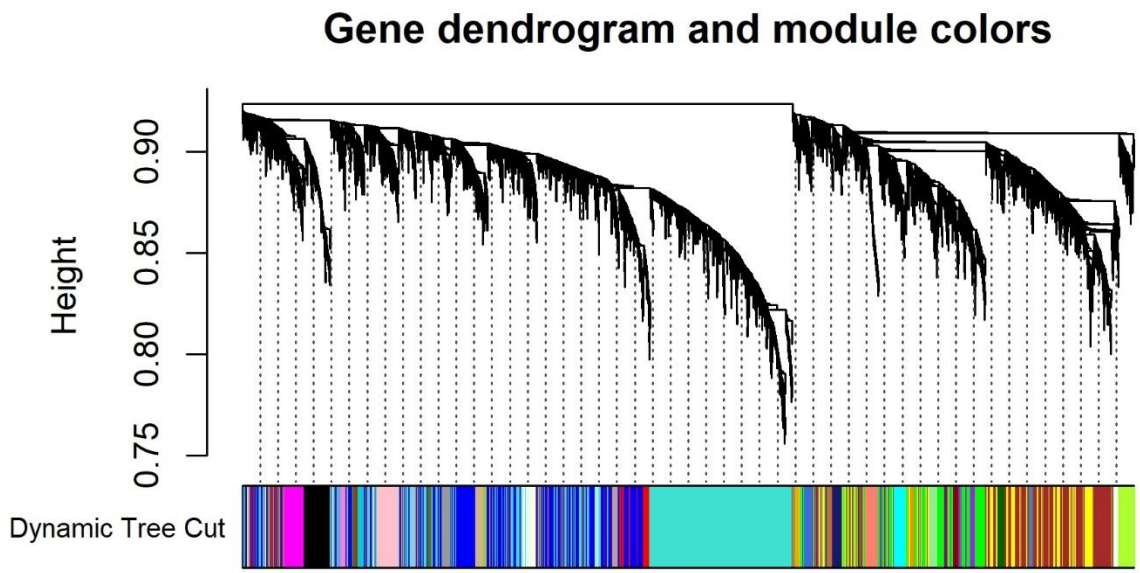


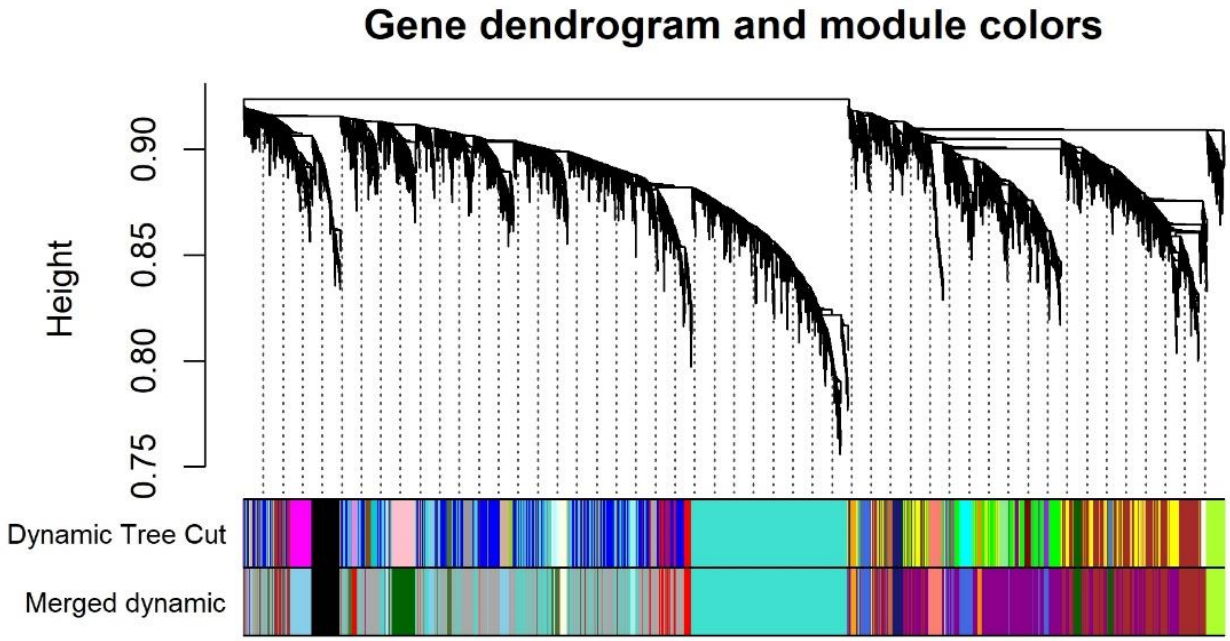
Fig.6. Dendrogram clustering of genes by hierarchical clustering adjacency-based dissimilarity.

3. Identifying the Clinically Significant Modules

Although most of the correlations were low to moderate ($R^2 < 0.90$), we found that the 6 modules had a higher MS. Furthermore, the ME in the 6 modules (salmon, red, Eskyblue, darkgrey, black, and brown modules) showed a stronger relationship with cancer stage, pathology, metastasis, status and overall survival (OS). Therefore, we identified the 6 modules as the clinically significant module and extracted them for further analysis (figure 6). 770 genes with high connectivity in 6 key modules were identified as hub genes (figure 7).

To identify the gene constitution of specific modules closely associated with overall survival, two unique features of the network - GS and MM - were utilized. Essentially, if the genes had highest absolute for both MM (0.7) and GS (0.75), they were considered particularly relevant to overall survival. Ultimately, 770 genes with high connectivity in 6 key modules were identified as hub genes. The red, brown, salmon, and black modules were found to have significant correlations between MM and GS, as shown in Figure 7. This suggests that these modules may play a particularly important role in the progression of kidney cancer.

a.


Gene dendrogram and module colors
Height
0.75
0.80
0.85
0.90
Dynamic Tree Cut
Merged dynamic

b.

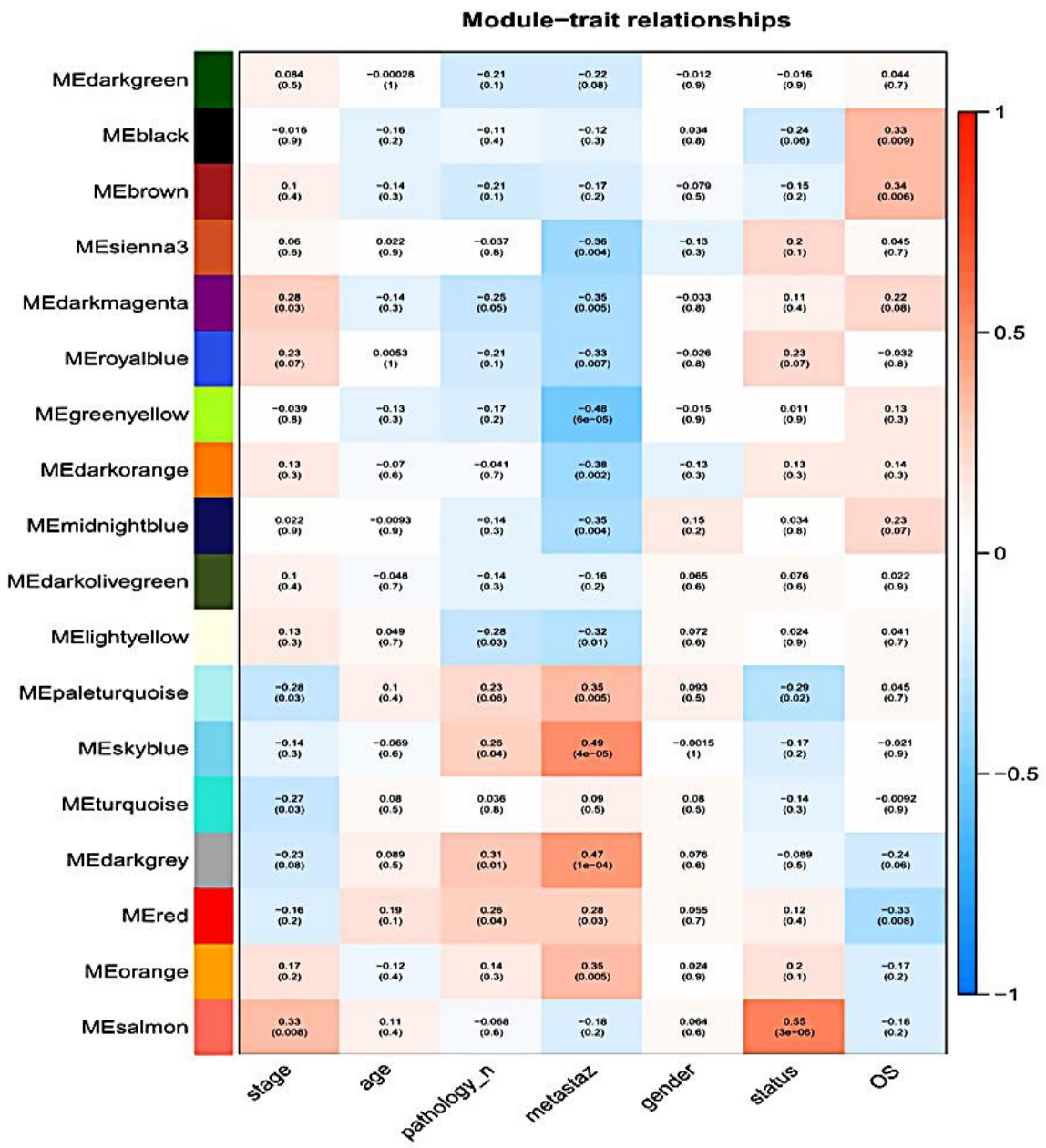
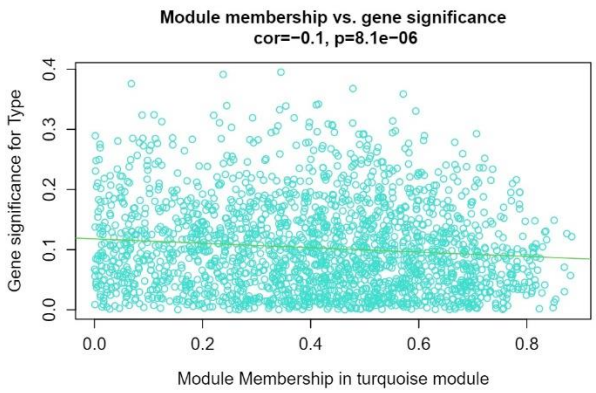
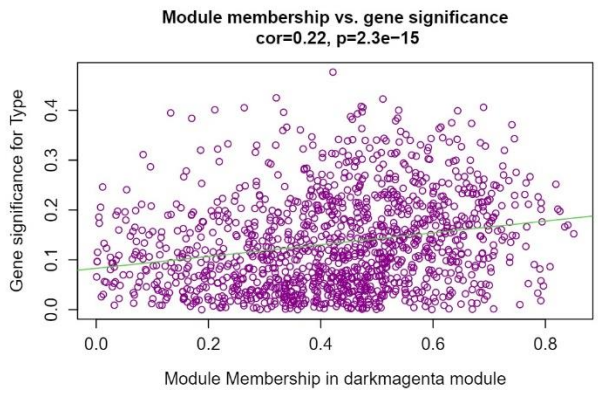
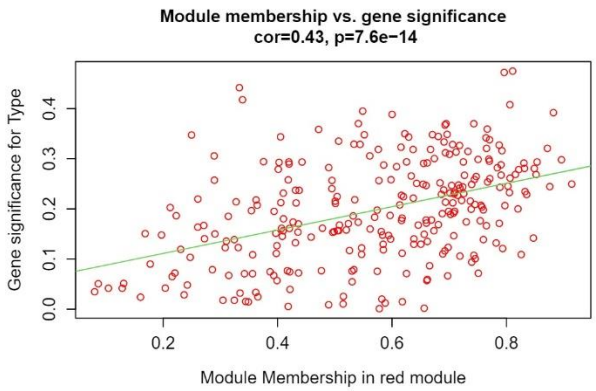
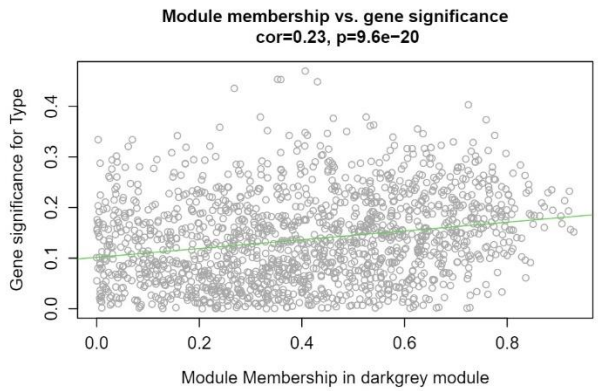


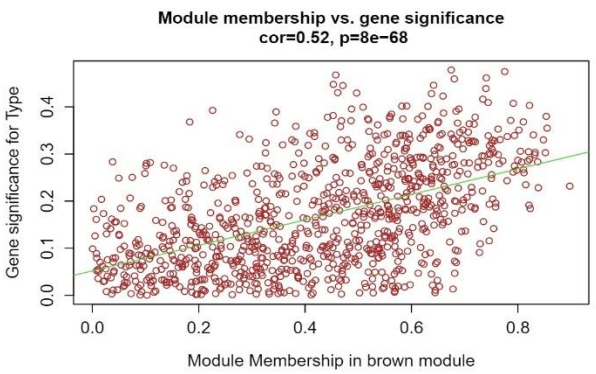
Fig.7. Identifying modules related to clinical traits. a) Dendrogram showing the clustering of all expressed genes that are different, using a measure of dissimilarity called "1-TOM". b) Heatmap indicating the correlation between module eigengenes and various clinical features of kidney cancer, such as age, gender, cancer stage, pathology, metastasis, status or overall survival (OS) (Significant p-value≤0.05). In the heatmap, the color salmon represents the association between eigengenes and cancer stage and status, while red, Eskyblue, and darkgrey indicate the correlation with metastasis, and black and brown signify the relationship with overall survival.

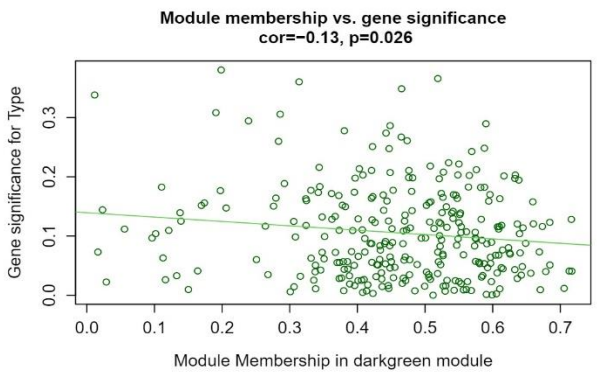

Module membership vs. gene significance
cor=−0.1, p=8.1e−06
Gene significance for Type
Module Membership in turquoise module


Module membership vs. gene significance
cor=0.22, p=2.3e−15
Gene significance for Type
Module Membership in darkmagenta module

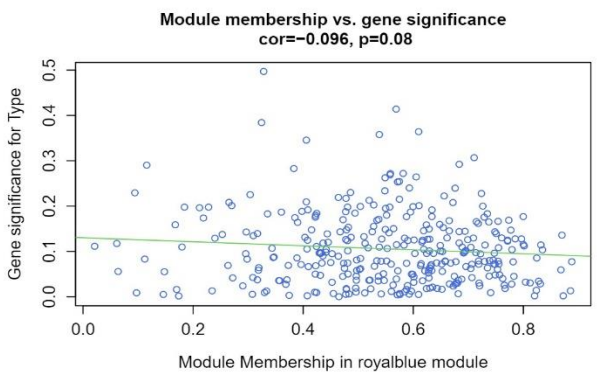

Module membership vs. gene significance
cor=0.43, p=7.6e−14
Gene significance for Type
Module Membership in red module

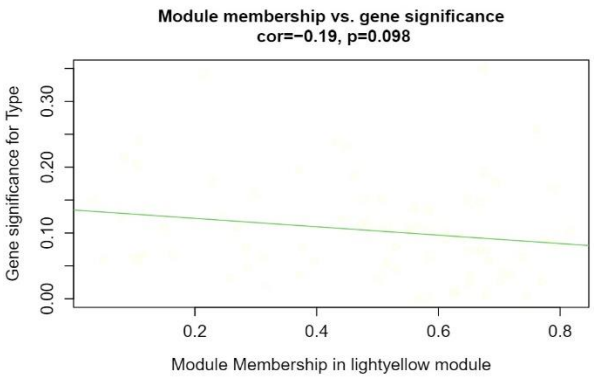

Module membership vs. gene significance
cor=0.23, p=9.6e−20
Gene significance for Type
Module Membership in darkgrey module

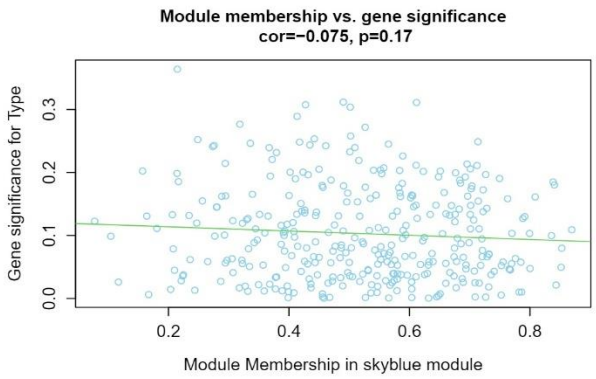

Module membership vs. gene significance
cor=0.52, p=8e−68
Gene significance for Type
Module Membership in brown module

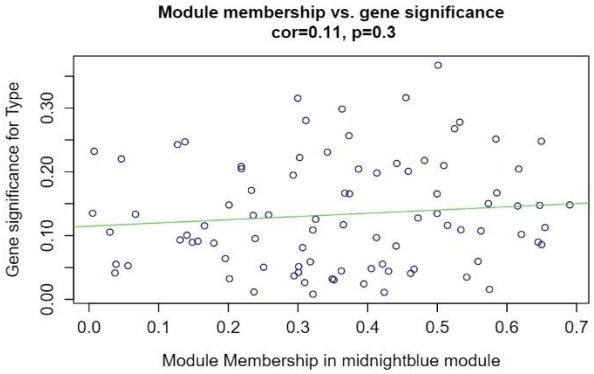

Module membership vs. gene significance
cor=−0.13, p=0.026
Gene significance for Type
Module Membership in darkgreen module

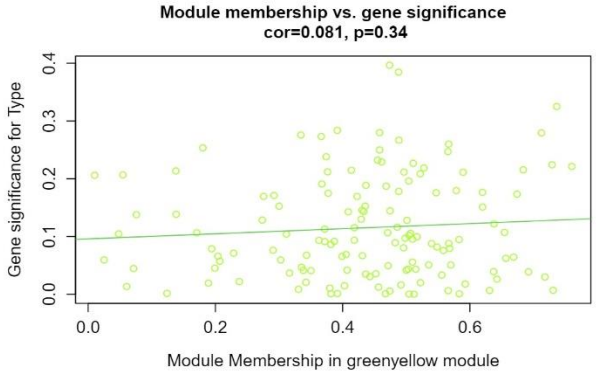
Module membership vs. gene significance
cor=0.37, p=5.5e−05
Gene significance for Type
Module Membership in salmon module

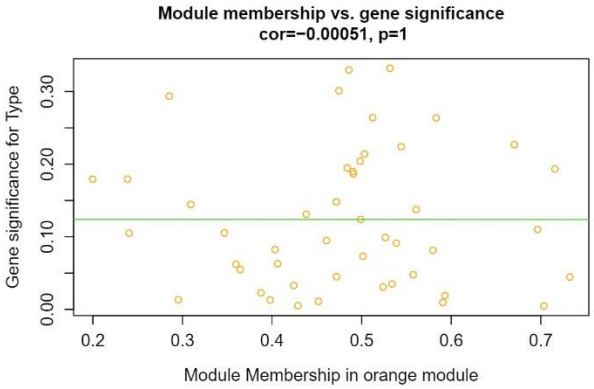

Module membership vs. gene significance
cor=−0.096, p=0.08
Gene significance for Type
Module Membership in royalblue module

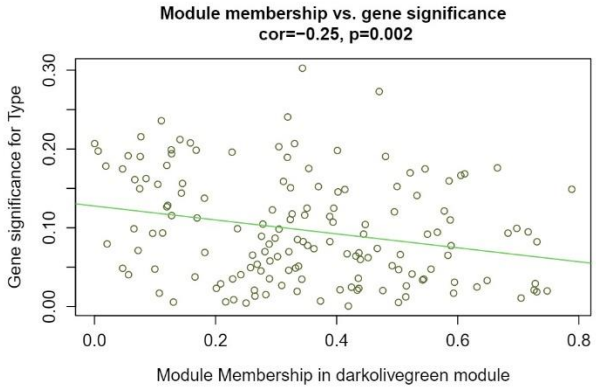

Module membership vs. gene significance
cor=−0.19, p=0.098
Gene significance for Type
Module Membership in lightyellow module

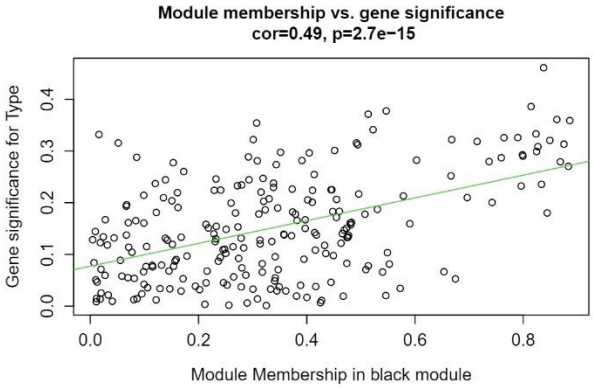

Module membership vs. gene significance
cor=−0.075, p=0.17
Gene significance for Type
Module Membership in skyblue module

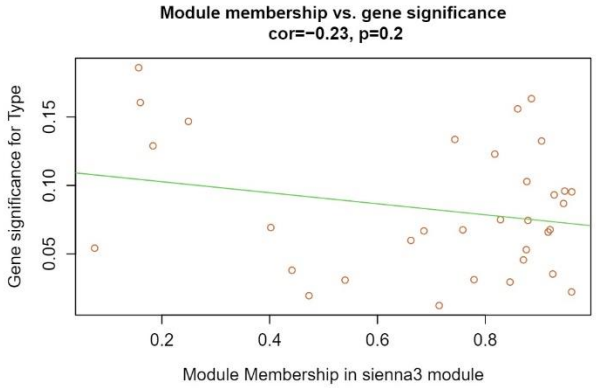

Module membership vs. gene significance
cor=0.11, p=0.3
Gene significance for Type
Module Membership in midnightblue module


Module membership vs. gene significance
cor=0.081, p=0.34
Gene significance for Type
Module Membership in greenyellow module


Module membership vs. gene significance
cor=−0.00051, p=1
Gene significance for Type
Module Membership in orange module


Module membership vs. gene significance
cor=−0.25, p=0.002
Gene significance for Type
Module Membership in darkolivegreen module


Module membership vs. gene significance
cor=0.49, p=2.7e−15
Gene significance for Type
Module Membership in black module


Module membership vs. gene significance
cor=−0.23, p=0.2
Gene significance for Type
Module Membership in sienna3 module

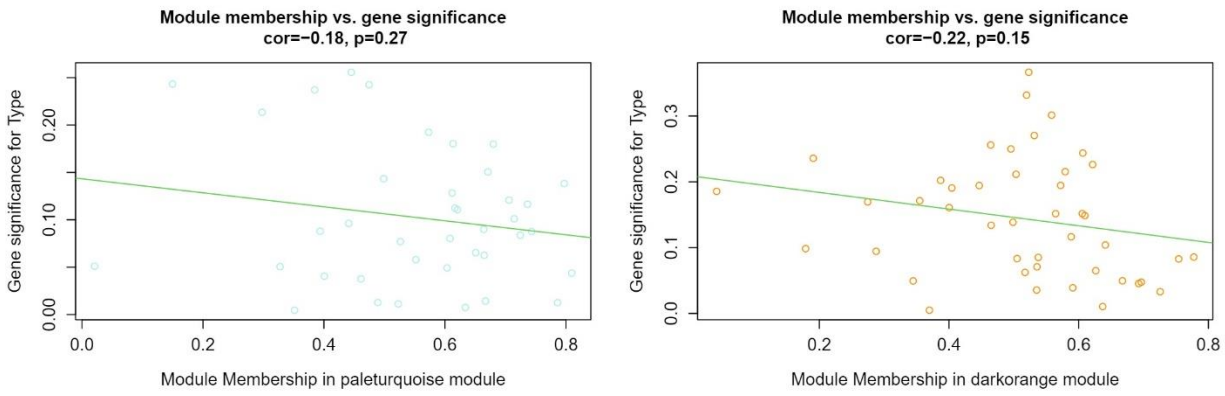
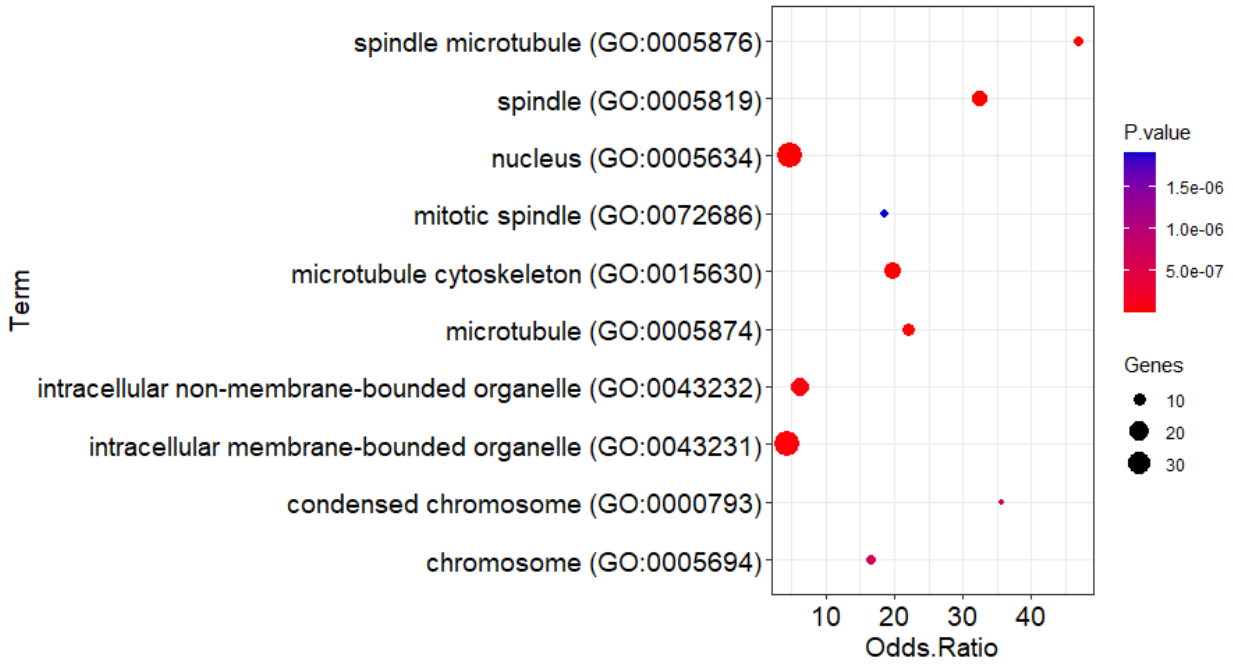


Fig.8. Module features of GS and MM. The relationship between the module's membership of the significance of genes with clinical traits in all modules.

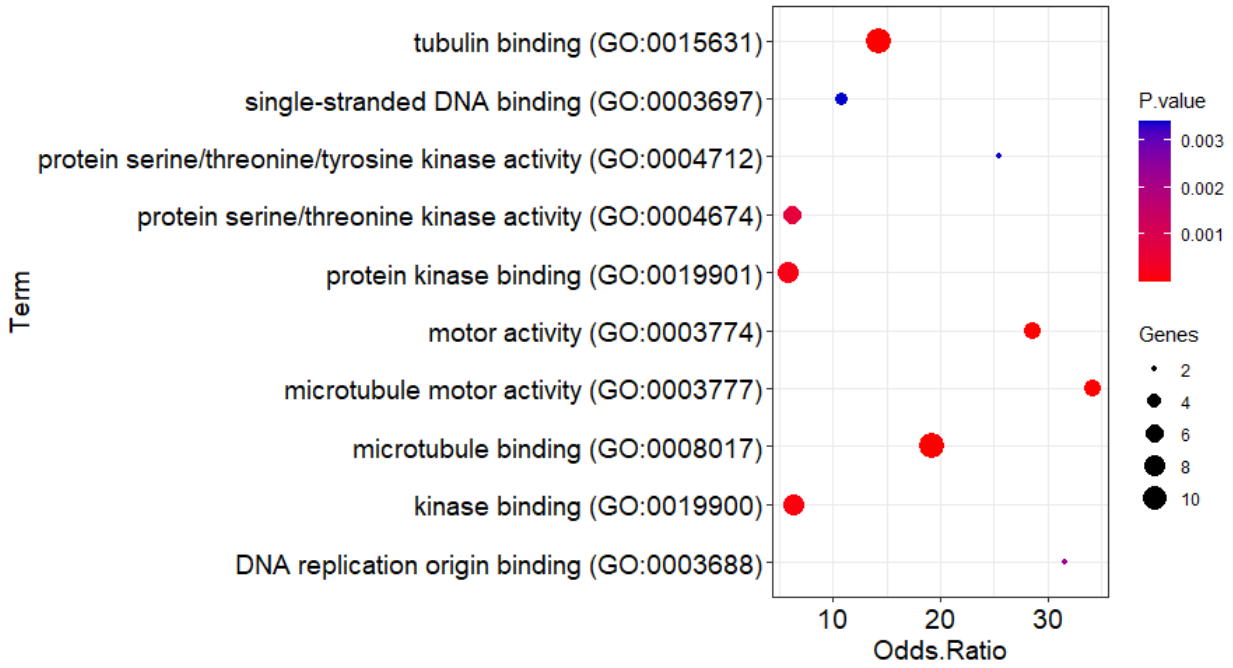
4. Gene Pathway Analysis

As mentioned above, six modules were identified as significant and selected for further study. A total of 770 genes were extracted from these modules and analyzed using GO and KEGG pathway analysis via the Enrichr database to explore potential mechanisms. The results are shown in Figure 8 and indicate that the genes within the six modules are involved in a range of physiological processes and diseases, including Parkinson's and Huntington's diseases, as well as gastric acid secretion. The KEGG pathway analysis also indicated that these genes are involved in metabolic processes such as ascorbate and adorate metabolism and glyoxylate and dicarboxylate metabolism. Moreover, pathways related to bacterial infections such as Vibrio cholerae infection and Helicobacter pylori infection were also noted. These pathways suggest that bacterial infections may have a role in the progression or development of kidney cancer.

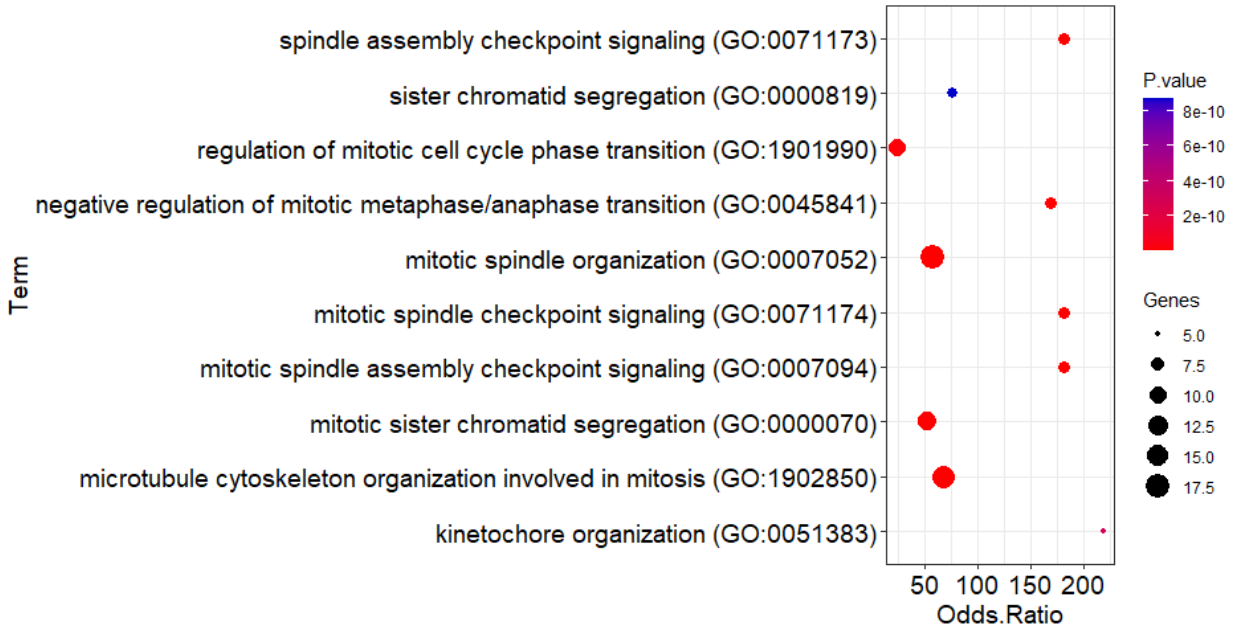
a.


spindle microtubule (GO:0005876)
spindle (GO:0005819)
nucleus (GO:0005634)
mitotic spindle (GO:0072686)
microtubule cytoskeleton (GO:0015630)
microtubule (GO:0005874)
intracellular non-membrane-bounded organelle (GO:0043232)
intracellular membrane-bounded organelle (GO:0043231)
condensed chromosome (GO:0000793)
chromosome (GO:0005694)
Term
10
20
30
40
Odds.Ratio
P.value
1.5e-06
1.0e-06
5.0e-07
Genes
10
20
30

b.


tubulin binding (GO:0015631)
single-stranded DNA binding (GO:0003697)
protein serine/threonine/tyrosine kinase activity (GO:0004712)
protein serine/threonine kinase activity (GO:0004674)
protein kinase binding (GO:0019901)
motor activity (GO:0003774)
microtubule motor activity (GO:0003777)
microtubule binding (GO:0008017)
kinase binding (GO:0019900)
DNA replication origin binding (GO:0003688)
Term
10
20
30
Odds.Ratio
P.value
0.003
0.002
0.001
Genes
2
4
6
8
10

c.



d.

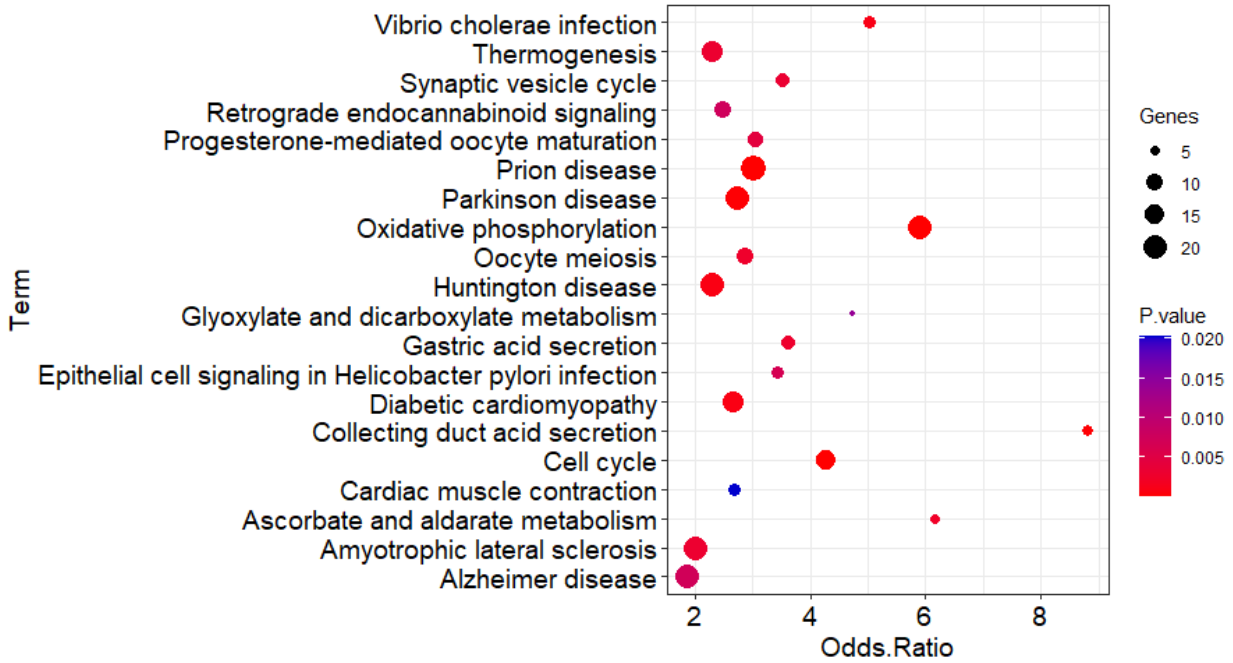


Fig.9. GO term and KEGG pathway analysis for genes of key modules. a) Cellular Component, b) Molecular Function, c) Biological Process and d) KEGG pathway enrichment analysis of key modules.

5. Outputs of Cox and 3-parameter defective Gompertz models

After analyzing 770 genes belonging to selected modules with both the Cox and 3-parameter defective Gompertz models separately, it was found that the output of both models was almost the same, with only 6 genes showing variation between the two. Specifically, the Cox model identified 173 genes associated with kidney cancer, while the 3-parameter defective Gompertz model identified 175 genes. The Venn diagram analysis revealed that two genes, GFER and AL590096.1, were uniquely identified by the Cox model, while four genes, IFT81, RAI14, CSNK1E, and CLIC4, were unique to the 3-parameter defective Gompertz model (Figure 9) (Supplementary fig.1). However, the majority of the genes were common to both models.

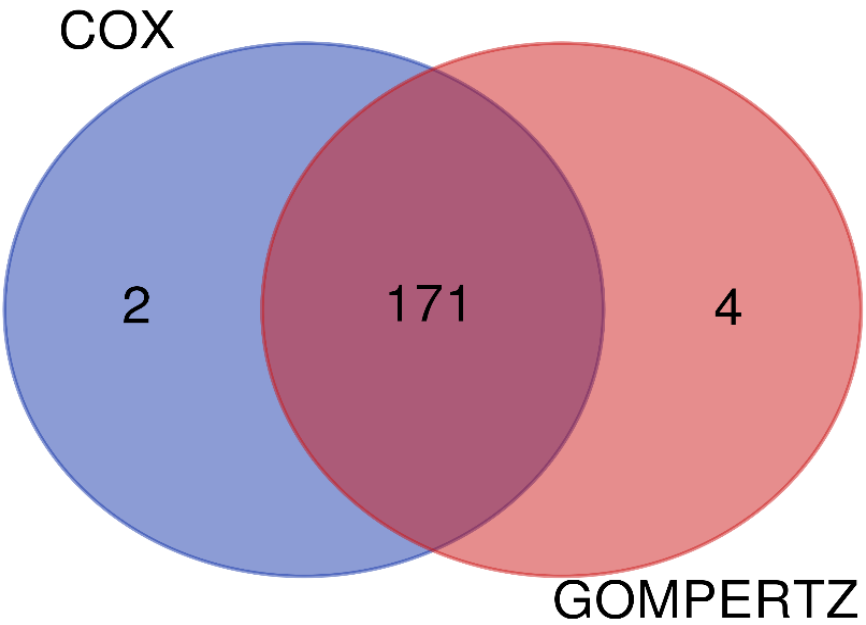
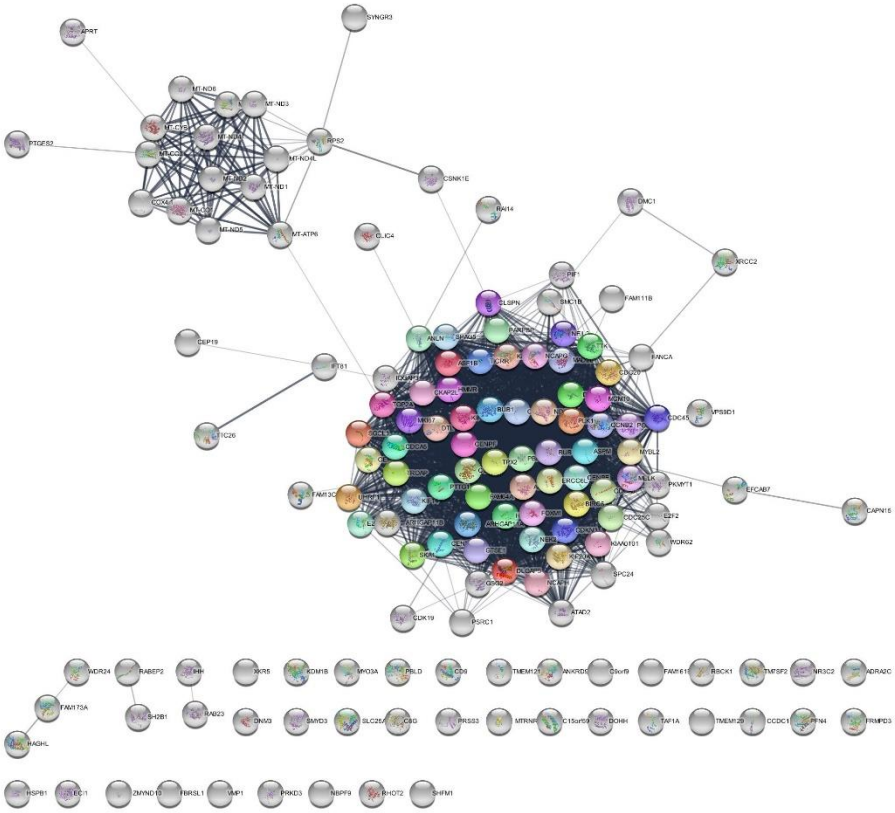


Fig.10. Venn diagram of the Cox/3-parameter defective Gompertz model output.

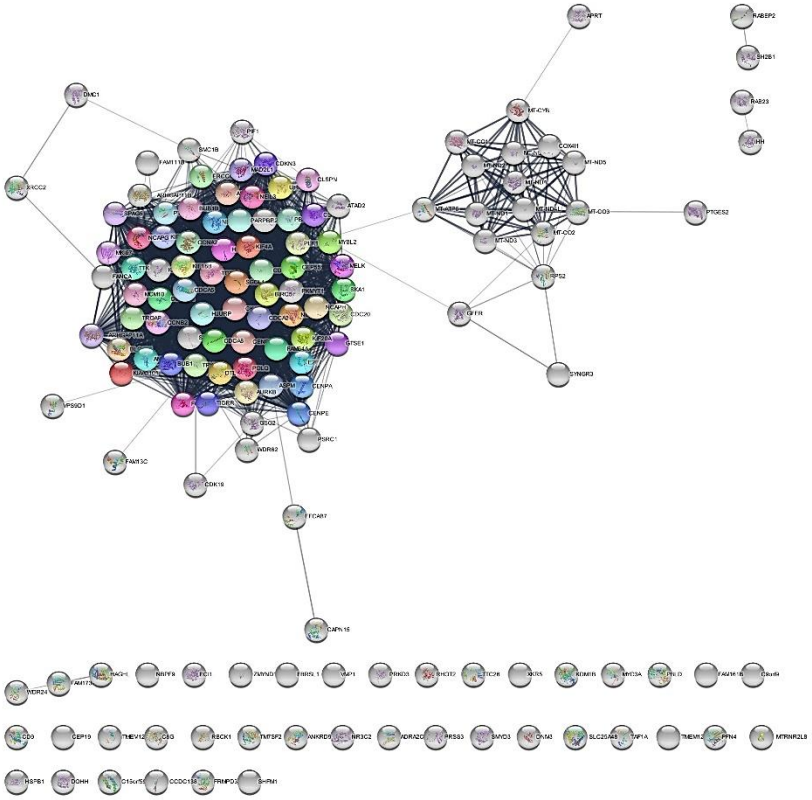
6. Identification of Hub genes in two PPI networks

To construct the PPI network, we examined the output of the Cox and 3-parameter defective Gompertz models (figure 10a, b). Using the MCODE plugin, we extracted several networks from all three networks, but identified one network that was exactly identical with 62 genes and a score of 57.9 between the three networks (figure 10c). We used MCC and degree of CytoHubba app for the top 10 genes to finding the hub genes present in two networks (figure 10d, e). Ultimately, we identified 10 hub genes (KIF20A, BUB1, AURKB, NCAPG, TOP2A, BUB1B, DLGAP5, TTK, TPX2, CDCA8) that were common in two networks. Meanwhile, 3 different hub genes (PLK1, CCNA2, CDC45) were observed in defective 3-parameter defective Gompertz and 4 different hub genes (PBK, CENPF, ASPM, CCNB2) in Cox PPI network (figure 11).

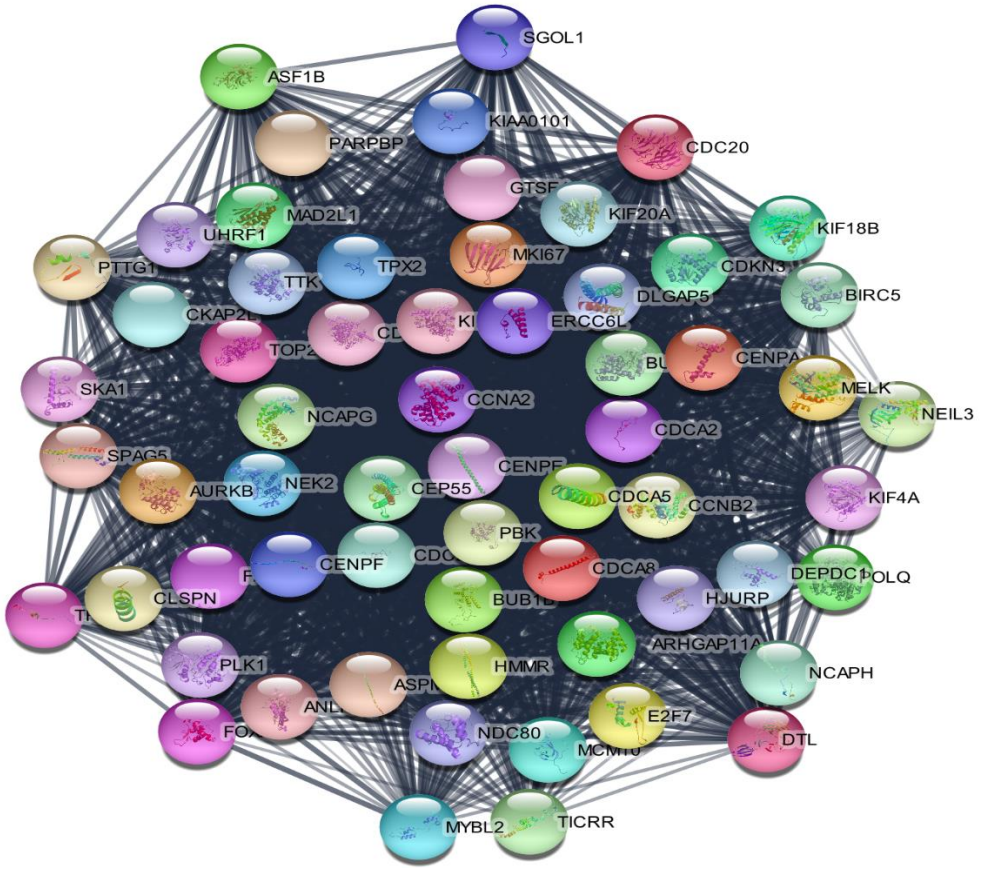
a.



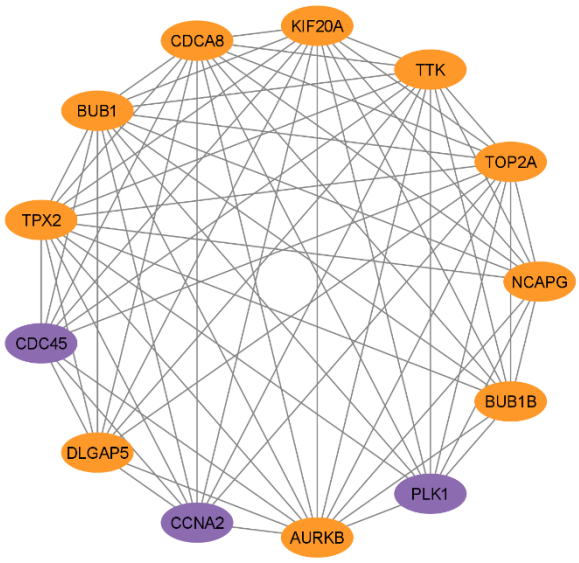
b.



c.



d.



e.

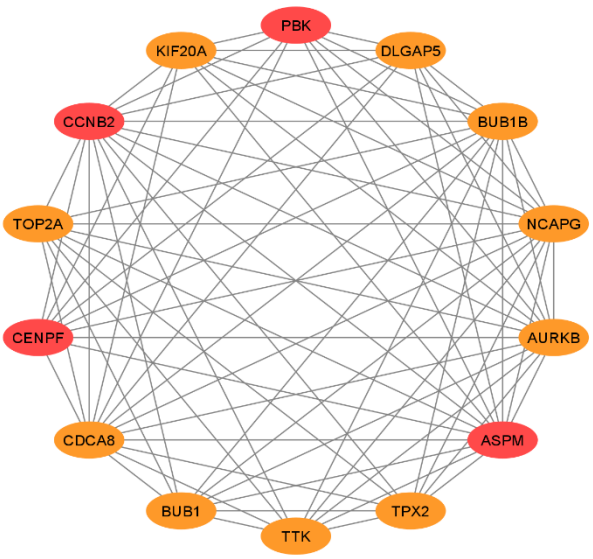


Fig.11. PPI network construction and screening for key genes. a) PPI network of defective Gompertz model output, b) PPI network of Cox model output c) MCODE cluster network of differentially expressed genes in cox and defective Gompertz output, d) merged networks of the

top 10 MCC and the degree of the defective Gompertz model, e) merged networks of the top 10 MCC and the degree of the Cox model output. Note: Orange color = common hub genes, purple color = different hub genes with Cox model, red color = different hub genes with defective Gompertz model.

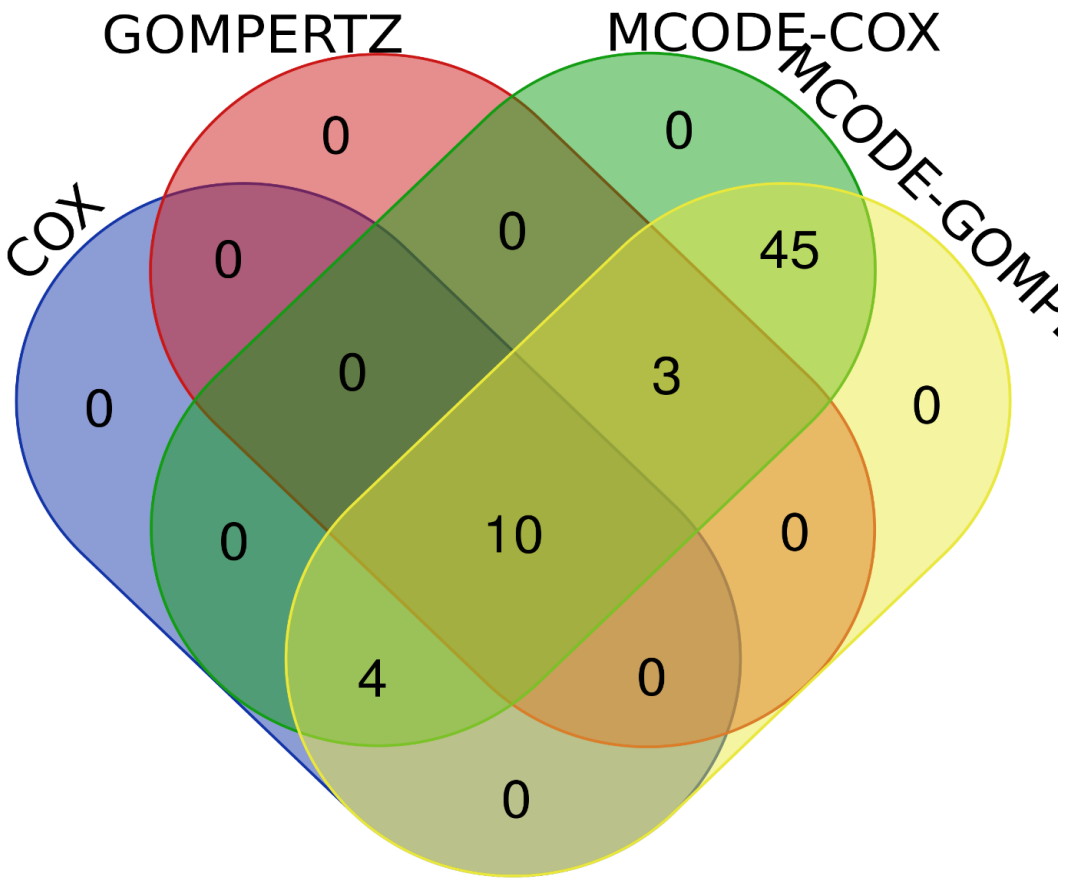


Fig.12. Venn diagram of two networks' hub genes and MCODE cluster

7. Cure rate estimation of hub genes

The findings of this analysis indicate a significant association between the expression levels of the identified hub genes and the cure rate in renal cell carcinoma. Specifically, the expression levels of six genes, namely TTK, KIF20A, DLGAP5, BUB1, AURKB, and CDC45, exhibited the strongest correlation with a reduction in the cure rate (table 1 and supplementary fig.1). These results suggest that targeting these genes may hold promise as a potential therapeutic strategy for improving treatment outcomes in RCC patients.

8. Prognostic value of hub genes

In this study, we identified a total of 14 hub genes in the Cox model and 13 hub genes in the Gompertz model, of which 10 hub genes were common between the two models. We assessed the prognostic value of these hub genes using the survivalROC package and KM method, with a cutoff point of 4676 days (12 years) for calculating gene expression in individuals' follow-up time. The results showed that almost all hub genes had an AUC value higher than 0.93 for one year, indicating excellent prognostic value. Additionally, hub genes with AUC values between 0.7 and 0.89 exhibited good diagnostic accuracy, indicating their potential use as diagnostic biomarkers (Table 1-3). According to our analysis, among the hub genes identified, six genes (NCAPG, TTK, DLGAP5, TOP2A, BUB1B, and BUB1) had the highest area under the curve (AUC) values.

Table 1. AUC and min and max cure expression of common hub genes in Cox and Gompertz model

symbol	Gene name	Beta in cox	p-value	HR	Beta in Gompertz	p-value	Cure in min exp	Cure in max exp	HR	AUC
ENSG00000088325.16	TPX2	0.792	0.000027	2.2086	0.784	0.000016	0.96	0.039	0.456462	0.781
ENSG00000109805.10	NCAPG	1.083	0.000007	2.9537	1.077	0.0000028	0.955	0.034	0.340747	0.809
ENSG00000112742.10	TTK	1.235	0.000002	3.4371	1.21	0.0000009	0.94	0.0005	0.447603	0.889
ENSG00000112984.12	KIF20A	0.945	0.000009	2.5734	0.994	0.00000065	0.91	0.00006	0.370214	0.786
ENSG00000126787.13	DLGAP5	0.745	0.000023	2.1066	0.784	0.0000077	0.81	0.0000003	0.456549	0.806
ENSG00000131747.15	TOP2A	0.711	0.000019	2.0366	0.704	0.000011	0.96	0.034	0.494858	0.808
ENSG00000134690.11	CDCA8	0.805	0.000044	2.2359	0.795	0.0000303	0.93	0.045	0.451452	0.762
ENSG00000156970.13	BUB1B	0.91	0.000033	2.4793	0.961	0.0000131	0.95	0.05	0.38265	0.830
ENSG00000169679.15	BUB1	0.834	0.0000448	2.3030	0.82	0.0000327	0.92	0.003	0.440453	0.857
ENSG00000178999.13	AURKB	1.081	0.00000516	2.9475	1.013	0.000000156	0.93	0.0000001	0.362972	0.77

Table 2. AUC and min and max cure expression of Gompertz hub genes

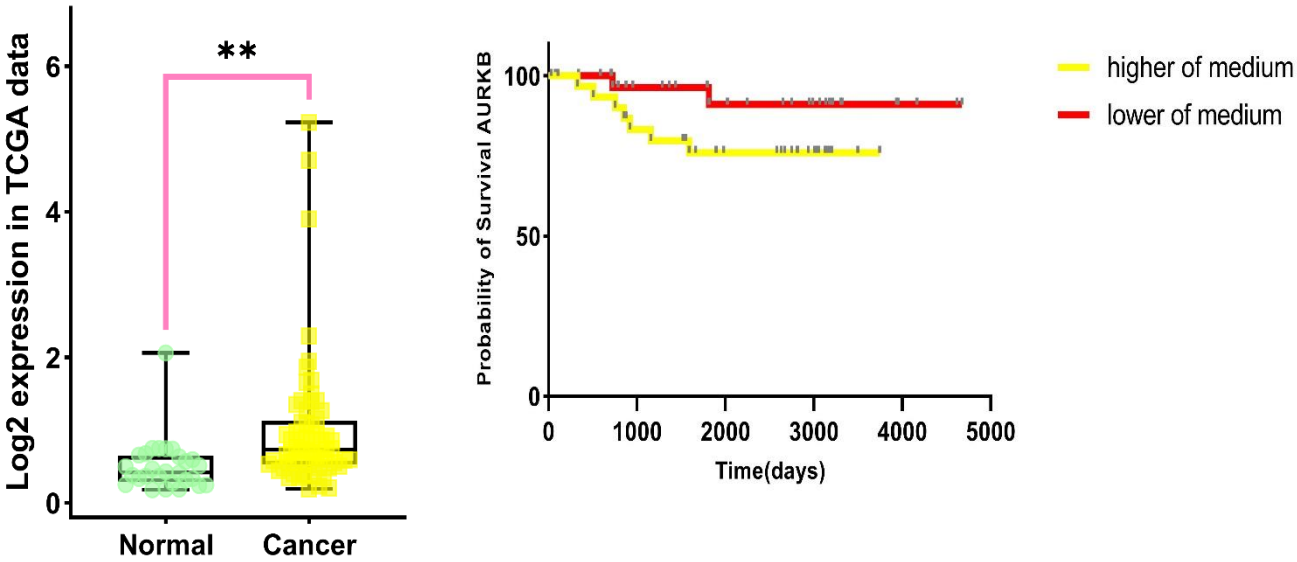
Symbol		Beta	p-value	Median in expression	Cure in min	Cure in max	HR	AUC
ENSG00000093009.11	CDC45	1.235	0.000003	0.960447	0.84	0.0000034	0.29067	0.774
ENSG00000145386.11	CCNA2	0.847	0.000132	1.599271	0.95	0.10	0.428625	0.778
ENSG00000166851.15	PLK1	0.855	0.0000228	1.319569	0.94	0.057	0.425247	0.736

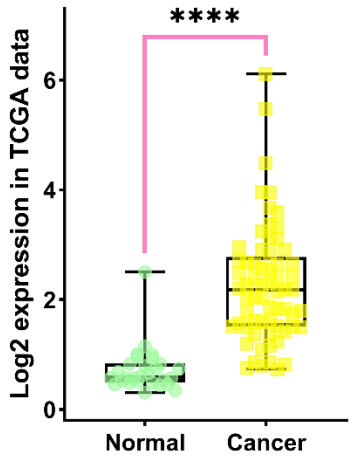
Table 3. AUC of Cox hub genes

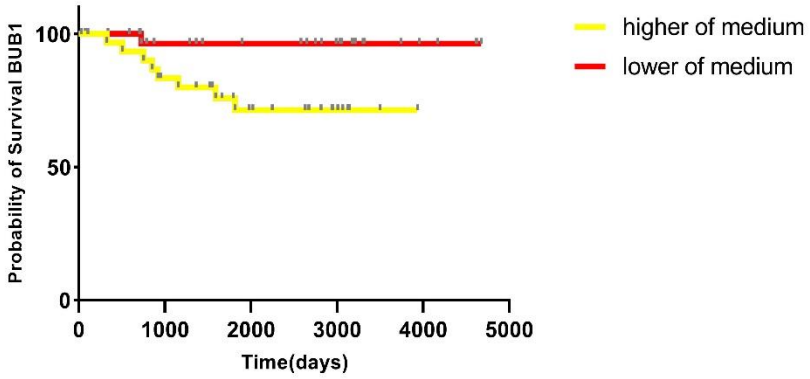
symbol	Gene name	Beta	p-value	Median in expression	AUC	HR
ENSG00000168078.10	PBK	0.815676	0.000038	1.366824	0.797	2.26
ENSG00000117724.13	CENPF	0.756063	0.0000177	2.380264	0.786	2.12
ENSG00000066279.18	ASPM	0.765976	0.000023	1.168842	0.783	2.15
ENSG00000157456.8	CCNB2	0.81976	0.000066	1.813181	0.787	2.27

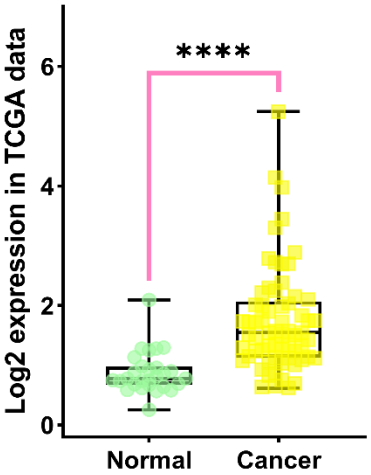
9. km-plots of hub genes

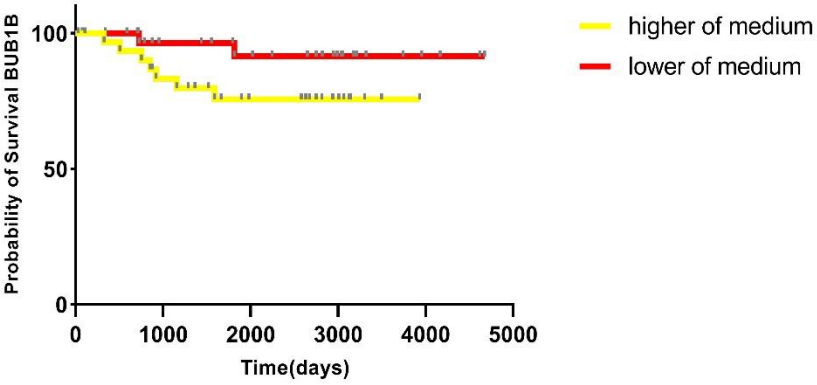
The findings from the survival analysis of hub genes indicate a significant association between gene expression and overall survival (OS) in renal cell carcinoma. The use of Kaplan-Meier survival curves and log-rank tests enabled the computation of differences in OS between the high- and low-risk groups. Notably, in all survival curves, the graph appeared as a straight line (except TPX2), indicating cure (figure 12).

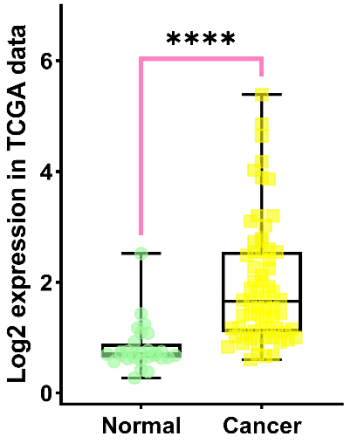


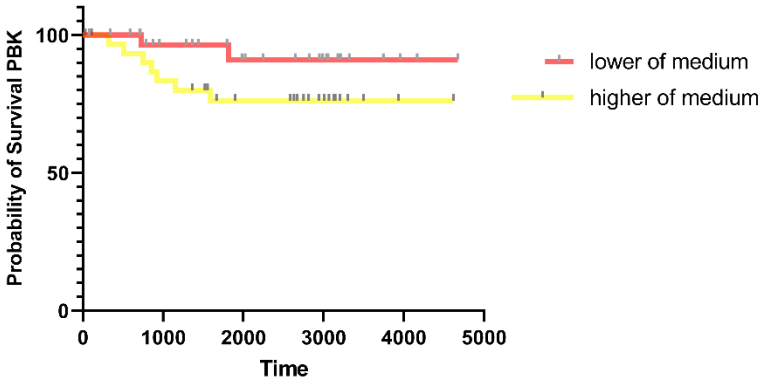

****
Log2 expression in TCGA data
0
2
4
6
Normal
Cancer

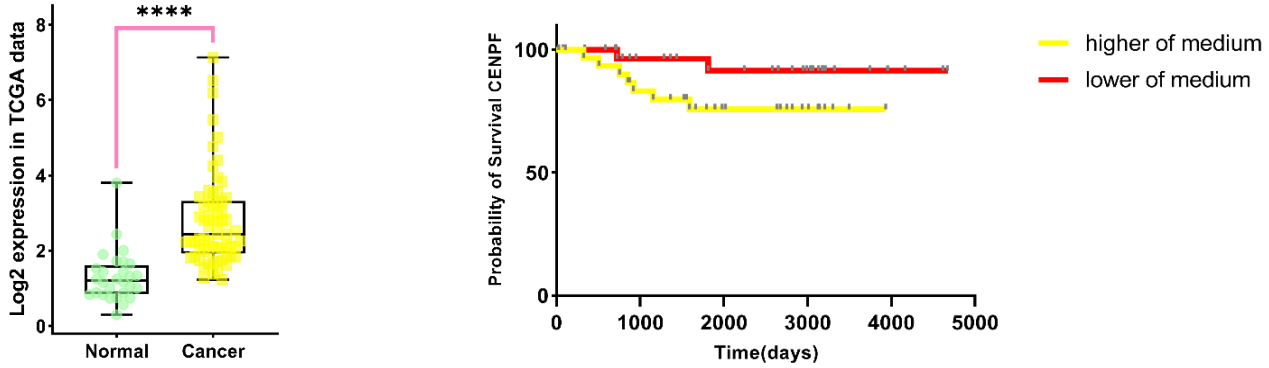

Probability of Survival BUB1
100
50
0
0
1000
2000
3000
4000
5000
Time(days)
higher of medium
lower of medium

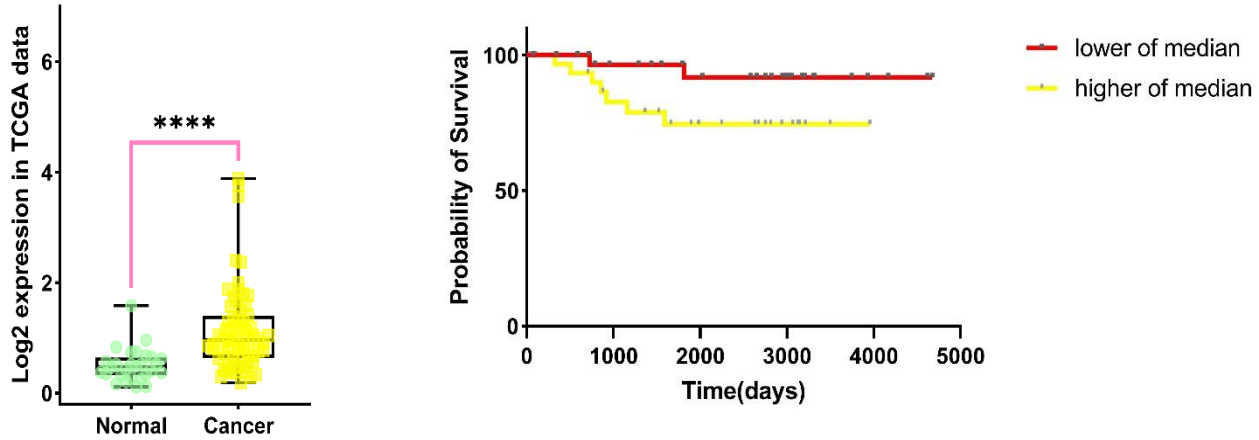

****
Log2 expression in TCGA data
0
2
4
6
Normal
Cancer


Probability of Survival BUB1B
100
50
0
0
1000
2000
3000
4000
5000
Time(days)
higher of medium
lower of medium


****
Log2 expression in TCGA data
0
2
4
6
Normal
Cancer


Probability of Survival PBK
100
50
0
0
1000
2000
3000
4000
5000
Time
lower of medium
higher of medium


****
Log2 expression in TCGA data
Normal
Cancer
Probability of Survival CENPF
Time(days)
higher of medium
lower of medium


****
Log2 expression in TCGA data
Normal
Cancer
Probability of Survival
Time(days)
lower of median
higher of median

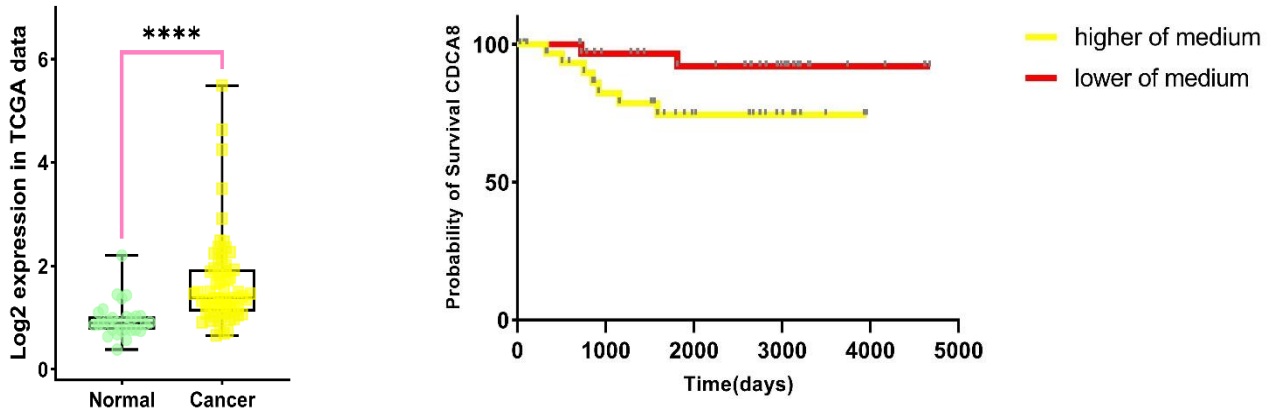

****
Log2 expression in TCGA data
Normal
Cancer
Probability of Survival CDCA8
Time(days)
higher of medium
lower of medium

Fig.13. Box plots (left) and km-plots (right) of hub genes In the box plots, the significance levels were denoted as ****, indicating a p-value of ≤0.00001, and ** representing a p-value of ≤0.001. In the Kaplan-Meier (KM) plots, all lines are represented by a long, flat tail that is noticeably distant from zero. This pattern suggests a significant survival advantage or potential for cure in kidney cancer patients.

Discussion

In this study, we used two different statistical models to analyze mRNA expression profiles in RCC cancer types from TCGA. The WGCNA technique was utilized to identify 6 key modules related to with cancer stage, pathology, metastasis, status and OS, and key genes were selected based on their Module Membership (MM) and Gene Significance (GS) values. To determine key pathways, we performed KEGG pathway enrichment analysis, and identified key genes using the PPI network method. Using both the Cox proportional hazards model and the 3-parameters defective Gompertz model, we identified SRGs, and observed differences in the results obtained from both statistical methods. The findings revealed that the hub genes identified in these two models were not entirely identical. Specifically, the 3-parameters defective Gompertz model identified three different hub genes, namely PLK1, CCNA2, and CDC45, while the Cox PPI network identified four different hub genes, namely PBK, CENPF, ASPM, and CCNB2.

PLK1, CCNA2, and CDC45 are all involved in cell cycle regulation and have been implicated in the development and progression of various cancers (19-21). PLK1 is a Ser/Thr kinase that is expressed predominantly during the G2/S and M phase of the cell cycle and is frequently overexpressed in various human cancers (19, 20). CCNA2 is a cyclin that regulates the G1/S transition of the cell cycle and is overexpressed in many types of cancer (22). CDC45 is a protein that is involved in DNA replication and is overexpressed in several types of cancer (23, 24). Additionally, these genes, as reported in Table 2, are highly relevant to the cure rate of RCC patients, as their maximum expression strongly reduces the cure rate.

PBK, CENPF, ASPM, and CCNB2 have also been implicated in cancer development and progression. PBK is a serine/threonine kinase that is overexpressed in many types of cancer, including kidney cancer, and is associated with poor prognosis (25, 26). CENPF is involved in cell division (27). Also, CENPF is involved in the centromere-kinetochore complex and has been found to impact cell proliferation and metastasis in multiple types of cancer, including kidney cancer(28-30). ASPM is a microtubule-associated protein that is involved in mitotic spindle assembly, and its overexpression has been observed in many cancers, including kidney cancer. CCNB2 is a regulatory subunit of CDK1, which plays a critical role in mitosis, and has been shown to be overexpressed in various cancers.

As reported in the results section, the expression levels of TTK, KIF20A, DLGAP5, BUB1, AURKB, and CDC45 demonstrated a significant association with decreased cure rate in renal cell

carcinoma. Notably, the highest expression levels of AURKB and DLGAP5 were associated with the lowest chance of cure, suggesting that the overexpression of these two genes may lead to a reduced likelihood of successful treatment. AURKB and DLGAP5 are involved in multiple pathways that have been shown to be critical in the development and progression of kidney cancer.

AURKB has been identified as a promising biomarker in clear cell renal carcinoma and has been found to play a key role in the tumorigenesis and progression of renal cell carcinoma (31). The AURKB gene is involved in the regulation of the cell cycle and is essential for the accurate segregation of chromosomes during cell division. Dysregulation of AURKB has been implicated in the development of various cancers, including lung cancer, colorectal cancer, prostate cancer, breast cancer, and liver cancer (32-35). AURKB has emerged as an attractive drug target leading to the development of small molecule inhibitors. Inhibition of AURKB activity has been shown to disrupt development and maintain proliferation and differentiation at later stages of development (36). However, it should be noted that AURKB mRNA expression levels were significantly up-regulated in kidney cancer patients compared to normal tissues.

DLGAP5 is a protein that is phosphorylated by AURKA, and it is one of its direct targets. DLGAP5 is involved in the formation of tubulin polymers which lead to tubulin sheets surrounding the microtubules. This protein plays an essential role in mitotic spindle formation during cell division. The phosphorylation of DLGAP5 by AURKA helps in stabilizing its association with the mitotic spindle (37). DLGAP5 has been found to be overexpressed in several cancers, including breast, ovarian, colorectal, and lung cancers, and is often associated with poor prognosis and disease progression. Also, its high area under the curve (AUC) value of 0.8 suggests its potential as a prognostic biomarker for kidney cancer.

Five genes (NCAPG, TTK, TOP2A, BUB1B, and BUB1) besides DLGAP5, have an area under the curve (AUC) above 0.8, indicating their potential as prognostic biomarkers for RCC. These genes are involved in critical cellular processes, such as cell division, DNA replication, and chromosomal segregation, making them potential targets for cancer therapy. For instance, TOP2A encodes a topoisomerase that plays a crucial role in DNA replication and chromosome segregation during mitosis, and its overexpression has been observed in various types of cancer (38). Similarly, BUB1 and BUB1B encode kinases that are involved in spindle checkpoint function, and their dysregulation has been implicated in the development and progression of cancer (39, 40). NCAPG and TTK are both involved in cell cycle regulation and have been found to be overexpressed in various cancers, including renal cell carcinoma (41-43). The identification of these genes as potential biomarkers can aid in the development of personalized treatment strategies for cancer patients.

Conclusion

In conclusion, the defective 3-parameters Gompertz model offers several advantages over the Cox proportional hazards model when analyzing mRNA expression profiles in cancer types such as renal cell carcinoma. One of the main benefits of the Gompertz model is that it provides more comprehensive information about the expression of survival-related genes, including the calculation of the minimum and maximum expression of cure rate. In contrast, the Cox model only calculates hazard ratios, which may not fully capture the complexity of gene expression patterns in cancer. This additional information is especially valuable in identifying potential therapeutic targets that may not be evident using the Cox model alone. Additionally, the Gompertz model's ability to identify survival-related genes (SRGs) that are not captured by the Cox model may lead to the discovery of new therapeutic targets for RCC. Overall, the Gompertz model's comprehensive approach to analyzing mRNA expression profiles in RCC provides a more nuanced understanding of the disease's molecular mechanisms and offers potential avenues for improving patient outcomes.

References

1. Capitanio U, Bensalah K, Bex A, Boorjian SA, Bray F, Coleman J, et al. Epidemiology of renal cell carcinoma. European urology. 2019;75(1):74-84.
2. Scelo G, Larose TL. Epidemiology and risk factors for kidney cancer. Journal of Clinical Oncology. 2018;36(36):3574.
3. Huang H, Zhu L, Huang C, Dong Y, Fan L, Tao L, et al. Identification of Hub Genes Associated With Clear Cell Renal Cell Carcinoma by Integrated Bioinformatics Analysis. Frontiers in Oncology. 2021;11.
4. Padala SA, Barsouk A, Thandra KC, Saginala K, Mohammed A, Vakiti A, et al. Epidemiology of renal cell carcinoma. World journal of oncology. 2020;11(3):79.
5. Pastore A, Palleschi G, Silvestri L, Moschese D, Ricci S, Petrozza V, et al. Serum and urine biomarkers for human renal cell carcinoma. Disease Markers. 2015;2015.
6. Chow W-H, Dong LM, Devesa SS. Epidemiology and risk factors for kidney cancer. Nature Reviews Urology. 2010;7(5):245-57.
7. Sims JN, Yedjou CG, Abugri D, Payton M, Turner T, Miele L, et al. Racial disparities and preventive measures to renal cell carcinoma. International journal of environmental research and public health. 2018;15(6):1089.
8. Pandey J, Syed W. Renal Cancer. 2020.

9. Sung W-W, Wang S-C, Hsieh T-Y, Ho C-J, Huang C-Y, Kao Y-L, et al. Favorable mortality-to-incidence ratios of kidney Cancer are associated with advanced health care systems. BMC cancer. 2018;18(1):1-7.

10. Wulandari I, Kurnia A, Sadik K, editors. Weibull regression and stratified cox regression in modelling exclusive breastfeeding duration. Journal of Physics: Conference Series; 2021: IOP Publishing.

11. Ihwah A. The use of Cox regression model to analyze the factors that influence consumer purchase decision on a product. Agriculture and Agricultural Science Procedia. 2015;3:78-83.

12. Rondeau V. Statistical models for recurrent events and death: Application to cancer events. Mathematical and Computer modelling. 2010;52(7-8):949-55.

13. Masud A, Tu W, Yu Z. Variable selection for mixture and promotion time cure rate models. Statistical methods in medical research. 2018;27(7):2185-99.

14. Balka J, Desmond AF, McNicholas PD. Review and implementation of cure models based on first hitting times for Wiener processes. Lifetime data analysis. 2009;15:147-76.

15. Hajizadeh N, Baghestani AR, Pourhoseingholi MA, Khadem Maboudi AA. Defective 3-parameter Gompertz model with frailty term for estimating cure fraction in survival data. Journal of Biopharmaceutical Statistics. 2023;33(1):90-113.

16. Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S. Weighted gene coexpression network analysis: state of the art. Journal of biopharmaceutical statistics. 2010;20(2):281-300.

17. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9(1):1-13.

18. Adamic LA, Huberman BA, Barabási A, Albert R, Jeong H, Bianconi G, editors. Power-law distribution of the world wide web. science2000: American Association for the Advancement of Science.

19. Gheghiani L, Wang L, Zhang Y, Moore XT, Zhang J, Smith SC, et al. PLK1 Induces Chromosomal Instability and Overrides Cell-Cycle Checkpoints to Drive TumorigenesisPLK1 Promotes Chromosomal Instability and Tumorigenesis. Cancer research. 2021;81(5):1293-307.

20. Iliaki S, Beyaert R, Afonina IS. Polo-like kinase 1 (PLK1) signaling in cancer and beyond. Biochemical pharmacology. 2021;193:114747.

21. Matthews HK, Bertoli C, de Bruin RA. Cell cycle control in cancer. Nature Reviews Molecular Cell Biology. 2022;23(1):74-88.

22. Jiang A, Zhou Y, Gong W, Pan X, Gan X, Wu Z, et al. CCNA2 as an immunological biomarker encompassing tumor microenvironment and therapeutic response in multiple cancer types. Oxidative Medicine and Cellular Longevity. 2022;2022.

23. Moyer SE, Lewis PW, Botchan MR. Isolation of the Cdc45/Mcm2–7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. Proceedings of the National Academy of Sciences. 2006;103(27):10236-41.

24. Srinivasan SV, Dominguez-Sola D, Wang LC, Hyrien O, Gautier J. Cdc45 is a critical effector of myc-dependent DNA replication stress. Cell reports. 2013;3(5):1629-39.

25. Wen H, Chen Z, Li M, Huang Q, Deng Y, Zheng J, et al. An integrative pan-cancer analysis of PBK in human tumors. Frontiers in Molecular Biosciences. 2021:1006.

26. Nagano-Matsuo A, Inoue S, Koshino A, Ota A, Nakao K, Komura M, et al. PBK expression predicts favorable survival in colorectal cancer patients. Virchows Archiv. 2021;479:277-84.

27. Varis A, Salmela A-L, Kallio MJ. Cenp-F (mitosin) is more than a mitotic marker. Chromosoma. 2006;115(4):288-95.

28. Li M-X, Zhang M-Y, Dong H-H, Li A-J, Teng H-F, Liu A-L, et al. Overexpression of CENPF is associated with progression and poor prognosis of lung adenocarcinoma. International Journal of Medical Sciences. 2021;18(2):494.

29. Huang Y, Chen X, Wang L, Wang T, Tang X, Su X. Centromere protein F (CENPF) serves as a potential prognostic biomarker and target for human hepatocellular carcinoma. Journal of Cancer. 2021;12(10):2933.

30. He R, Wang L, Li J, Ma L, Wang F, Wang Y. Integrated analysis of a competing endogenous RNA network reveals a prognostic signature in kidney renal papillary cell carcinoma. Frontiers in Cell and Developmental Biology. 2020;8:612924.

31. Wan B, Huang Y, Liu B, Lu L, Lv C. AURKB: a promising biomarker in clear cell renal cell carcinoma. PeerJ. 2019;7:e7718.

32. Tang A, Gao K, Chu L, Zhang R, Yang J, Zheng J. Aurora kinases: novel therapy targets in cancers. Oncotarget. 2017;8(14):23937.

33. Pohl A, Azuma M, Zhang W, Yang D, Ning Y, Winder T, et al. Pharmacogenetic profiling of Aurora kinase B is associated with overall survival in metastatic colorectal cancer. The pharmacogenomics journal. 2011;11(2):93-9.

34. Fang H, Niu K, Mo D, Zhu Y, Tan Q, Wei D, et al. RecQL4-Aurora B kinase axis is essential for cellular proliferation, cell cycle progression, and mitotic integrity. Oncogenesis. 2018;7(9):68.

35. Marima R, Hull R, Penny C, Dlamini Z. Mitotic syndicates Aurora Kinase B (AURKB) and mitotic arrest deficient 2 like 2 (MAD2L2) in cohorts of DNA damage response (DDR) and tumorigenesis. Mutation Research/Reviews in Mutation Research. 2021;787:108376.

36. Shaalan AK, Teshima TH, Tucker AS, Proctor GB. Inhibition of Aurora Kinase B activity disrupts development and differentiation of salivary glands. Cell Death Discovery. 2021;7(1):16.

37. Schneider MA, Christopoulos P, Muley T, Warth A, Klingmueller U, Thomas M, et al. AURKA, DLGAP5, TPX2, KIF11 and CKAP5: Five specific mitosis-associated genes correlate with poor prognosis for non-small cell lung cancer patients. International journal of oncology. 2017;50(2):365-72.

38. Uusküla-Reimand L, Wilson MD. Untangling the roles of TOP2A and TOP2B in transcription and cancer. Science advances. 2022;8(44):eadd4920.

39. Sekino Y, Han X, Kobayashi G, Babasaki T, Miyamoto S, Kobatake K, et al. BUB1B Overexpression Is an Independent Prognostic Marker and Associated with CD44, p53, and PD-L1 in Renal Cell Carcinoma. Oncology. 2021;99(4):240-50.

40. Jiang N, Liao Y, Wang M, Wang Y, Wang K, Guo J, et al. BUB1 drives the occurrence and development of bladder cancer by mediating the STAT3 signaling pathway. Journal of Experimental & Clinical Cancer Research. 2021;40(1):378.

41. Li H, Zheng P, Li Z, Han Q, Zhou B, Wang X, et al. NCAPG Promotes the Proliferation of Renal Clear Cell Carcinoma via Mediating with CDK1. Disease markers. 2022;2022:6758595.

42. Liu XD, Yao DW, Xin F. TTK contributes to tumor growth and metastasis of clear cell renal cell carcinoma by inducing cell proliferation and invasion. Neoplasma. 2019;66(6):946-53.

43. Zhang Q, Su R, Shan C, Gao C, Wu P. Non-SMC condensin I complex, subunit G (NCAPG) is a novel mitotic gene required for hepatocellular cancer cell proliferation and migration. Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics. 2018;26(2):269-76.