*Research Article*

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

Guanlin Wu[1#], Miao Liu[2#], Weifeng Yang[3], Shuai Zhu[4], Weiming Guo[5], Gang Huang[6], Lei Xie[3], Yu Xie[4*], Gang Fan[3,4*]

1. *Department of Pathology, School of Basic Medical Sciences, Fudan University, 200030 Shanghai, China.*
2. *Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.*
3. *Department of Urology, Huazhong University of Science and Technology Union Shenzhen Hospital; the 6th Affiliated Hospital of Shenzhen University Health Science Center, 518060 Shenzhen, China.*
4. *Department of Urology, the Affiliated Cancer Hospital of Xiangya School of Medicine of Central South University, Hunan Cancer Hospital, 410013 Changsha, Hunan, China.*
5. *The 2nd Affiliated Hospital of South China University, 421001 Hengyang, China.*
6. *The 3rd People's Hospital of Chengdu; the Affiliated Hospital of Southwest Jiaotong University, 610014 Chengdu, China.*

**Running title:** Single-cell based prognostic model in clear cell renal cell carcinoma

#G.W and M.L contributed equally to this work and considered as joint first authors. *Y.X and G.F contributed equally to this work and considered as joint corresponding authors.

**Correspondence:** wuguanlin105109@gmail.com (G.W) and gang.fan.med@qq.com (G.F)

**Abstract**

Clear cell renal cell carcinoma (ccRCC) is the most common type of renal cancer (RCC). The increasing incidence and poor prognosis of ccRCC after tumour metastasis makes the study of its pathogenesis extremely important. Traditional studies mostly focus on the regulation of ccRCC by single gene, while ignoring the impact of tumour heterogeneity on disease progression. The purpose of this study is to construct a prognostic risk model for ccRCC by analysing the differential marker genes related to immune cells in the single-cell database for providing help in clinical diagnosis and targeted therapy. Single-cell data and ligand-receptor relationship pair data were downloaded from related publications, and ccRCC phenotype and expression profile data were downloaded from TCGA and CPTAC. The DEGs and marker genes of the immune cell were combined and then intersected with the ligand-receptor gene data, and the 981 ligand-receptor relationship pairs obtained were intersected with the target gene of the transcription factor afterwards; 7,987 immune cell multilayer network relationship pairs were finally observed. Then, the genes in the network and the genes in TCGA were intersected to obtain 966 genes for constructing a co-expression network. Subsequently, 53 genes in black and magenta modules related to prognosis were screened by WGCNA for subsequent analysis. Whereafter, using the data of TCGA, 28 genes with significant prognostic differences were screened out through univariate Cox regression analysis. After that, LASSO regression analysis of these genes was performed to obtain a prognostic risk scoring model containing 16 genes, and CPTAC data showed that the effectiveness of this model was good. The results of correlation analysis between the risk score and other clinical factors showed that age, grade, M, T, stage and risk score were all significantly different ($p < 0.05$), and the results of prognostic accuracy also reached the threshold of qualification. Combined with clinical information, univariate and multivariate Cox regression analyses verified that risk score was an independent prognostic factor ($p < 0.05$). A nomogram constructed based on a predictive model for predicting the overall survival was established, and internal validation performed well. Our findings suggest that the predictive model built based on the immune cells scRNA-seq will enable us to judge the prognosis of patients with ccRCC and provide more accurate directions for basic relevant research and clinical practice.

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

## Introduction

Renal cell carcinoma (RCC) is a common malignant tumour of the urinary system. According to the recent cancer statistics report, the number of newly diagnosed cases in the United States has increased to 65,000 each year, resulting in approximately 15,000 deaths each year, making it the sixth most common tumour [1]. Clear cell renal cell carcinoma (ccRCC) accounts for around 80% of renal cancer pathological type, and its survival results were poorer than other subtypes of kidney tumours (such as papillary renal cell carcinoma and chromophobe renal cell carcinoma) [2]. Nearly 20% of ccRCC cases progress to an advanced stage at the beginning of diagnosis, and the five-year overall survival (OS) rate of metastatic cases dropped to about 10% [3]. With the development of immunotherapy, radiotherapy and surgical intervention, combined strategies have greatly promoted carcinoma control. However, the actual clinical efficacy still needs to be improved, and 30% of patients with local ccRCC inevitably experience cancer-related progression and recurrence [4]. Recently, although targeted therapy has been shown to prolong the survival time of patients with metastases, the median survival time is still less than 3 years [5]. In addition, drug resistance and economic burden are unavoidable major problems in clinical practice [6]. Therefore, exploring the molecular mechanism of ccRCC pathogenesis and new therapeutic targets are still challenging issues.

Comprehensive heterogeneity is a critical characteristic of carcinomas, and it may lead to different responses of patients to the same treatment. Despite many efforts to clarify tumour heterogeneity, so far, the understanding of it is still mainly limited to the level of tumour cells [7]. Previously, it has also been proved that stromal cells and tumour-infiltrating immune cells exhibit heterogeneity [8]. In addition, the tumour microenvironment (TME) is gradually regarded as a potential solution for drug treatment target [9]. In addition to anti-PD-1/PD-L1 or anti-CTLA-4 treatment strategies, tumour-associated macrophages (TAMs) [10] and cancer-associated fibroblasts (CAFs) [11] have also been previously reported as potential strategies for cancer treatment research. The successful implementation of these treatment plans require a deeper insight of intratumoural heterogeneity.

Since traditional bulk RNA sequencing relies on the assumption that each gene is equally expressed in each cell, it is obviously impossible to study intratumoural heterogeneity at the cellular level. However, through the application of single-cell RNA sequencing (scRNA-seq), it is possible to conduct single-cell transcriptome analysis. The latest progress in scRNA-seq has facilitated the transcriptional classification of many malignant tumour cell types, including breast cancer, lung cancer and pancreatic ductal adenocarcinoma [12] [13]. Moreover, scRNA-seq is expected to have clinical utility in refractory cancer cases and is a non-invasive method for monitoring circulating cancer cells, analysing intratumoural heterogeneity and sensitively estimating recurrent tumours [14].

In order to investigate the intratumour heterogeneity in ccRCC, we performed a series of bioinformatics analyses by using data in other publications about

scRNA-seq and combined them with ligand-receptor network analysis, immune cell multilayer network analysis and weighted gene co-expression network analysis (WGCNA) to figure out hub genes for building an immune cell-related prognostic model, which would be several potential targets for ccRCC therapy. Moreover, we also investigated the prognostic value of immune cell clusters by correlating the scRNA-seq data with the data from The Cancer Genome Atlas (TCGA) and Consortium for Clinical Proteome Cancer Analysis (CPTAC) datasets. Our work will help elucidate the biology of ccRCC and promote the improvement of clinical diagnosis and treatment strategies for patients with ccRCC.

## Methods

### Raw data acquisition

ccRCC single-cell transcriptome data was downloaded from [15]. The datasets of RNA sequencing profiles and the related patient clinical traits of ccRCC were downloaded from TCGA (https://portal.gdc.cancer.gov/) and CPTAC (https://cptac-data-portal.georgetown.edu/study-summary/S050), separately. Ligand and receptor data for building the multilayer network was acquired from [16].

### Data processing

For single-cell data, "limma", "Seurat", "dplyr", and "magrittr" R packages were used for analysis. Data filtering criteria included: 1. The number of genes in the data sample was controlled between 200 and 5,000; 2. The number of gene sequences was mainly distributed between 0 and 20,000; and 3. The percentage of mitochondria was controlled below 5%. Then, the log was taken for standardization, and the first 2,000 genes with the larger coefficient of variation between cells were selected for analysis. Next, principal component analysis (PCA) dimensionality reduction was performed, the data were standardized before dimensionality reduction, and finally, significant dimensions were selected for subsequent analysis.

Since the data set downloaded from TCGA is log2-(data+1), log processing is not required and can be directly standardized. The data downloaded from the CPTAC database needs to be processed by log2 (data+1) before standardization. The standardization was carried out through the "limma" R package.

### Cell type recognition and clustering analysis

The recognition of different cell types was based on the "limma", "Seurat", "dplyr", and "magrittr" R packages. We used the 20 principal components (PCs) selected in the previous step to perform TSNE clustering. Subsequently, the cell type was annotated through the "singleR" R package. In order to facilitate the display of subsequent results, we have annotated both subpopulations and single cells.

**Identification of differentially expressed genes in each cluster**

We used several R packages, including “limma”, “Seurat”, “dplyr” and “magrittr” to analyse the genes contained in each cluster. The FindAllMarkers algorithm was used to screen and calculate the differentially expressed genes (DEGs) in each cluster. The screening standard is |logFC|> 0.5, and the *P* value after correction is < 0.05.

**Immune cell function status analysis**

In order to clarify the functional status of DEGs in immune cells, we referred to the marker genes of immune cell functional status provided by the CancerSEA (http://biocc.hrbmu.edu.cn/CancerSEA/home.jsp) database and used “GSVA” and “GSEABase” R packages to perform functional status analysis on samples annotated by single cell.

**Immune cell marker gene expression analysis**

The marker genes of immune cells in kidney cancer tissues were obtained from the CellMarker (http://bio-bigdata.hrbmu.edu.cn/CellMarker/) database. In addition, marker genes associated with macrophages and monocytes were acquired from [17]. The expression levels of these marker genes were analysed and displayed through a heat map.

**Construction of ligand-receptor and immune cell multilayer networks**

The construction of the ligand-receptor network was carried out using the "igraph" R package. We first took the union of the differential genes in all immune cell clusters and the marker genes of all included immune cells and then took the intersection with the genes in the ligand-receptor table provided in the literature [16] to ensure that the ligand genes and related receptor genes were all in the gene set we took the union. Afterwards, we downloaded the transcription factors and target genes data from the TRRUST (https://www.grnpedia.org/trrust/) database and merged the ligand-receptor network data with the transcription factors and target genes data, which is the intersection of the target genes of the transcription factors and the genes from the ligand-receptor network.

**Weighted gene co-expression network analysis**

Through the WGCNA algorithm [18], the genes in the immune cell multifactor network were used to construct a co-expression network to find the key modules related to OS and OS time. An appropriate soft threshold value was determined by an R software package (http://www.r-project.org/) implement according to the WGCNA algorithm. The gradient method was used to test different power values (ranging from 1 to 20) in both the scale independence degree and the module’s average connectivity. The most suitable power value could be fixed when the independence degree was above 0.9, as well as when the average connectivity degree was relatively higher [19]. The WGCNA algorithm was also implemented in the construction of scale-free gene

co-expression networks. Meanwhile, the corresponding gene sequencing information in each module was extracted. To assess modular feature associations, correlations between module eigengenes (MEs) and clinical features were applied. MEs are the most important components in the PCA of each gene module. The determination of relevant modules needs to be based on the calculation of the correlation strength between MEs and clinical features. The correlation was assessed by gene significance (GS = lgP), where the *P* value was derived from the linear regression analysis of gene expression and clinical information. The key module takes the highest correlation coefficient among all modules, which was selected out for the next step.

**Key modules functional enrichment analysis**

The sequencing information of genes in the key modules from WGCNA was utilized by using the “clusterProfiler” R package to perform gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses. Among them, GO is to annotate biological processes (BP), molecular function (MF), and cellular components (CC). The criteria for screening in GO term is *P* value < 0.05. The screening criteria for the KEGG pathway are minGSSize = 5, maxGSSize = 500, and qvalueCutoff = 0.05.

**Selection of candidate prognostic related genes**

Univariate Cox regression analysis was performed through the "survival" R package to screen prognostic characteristic genes from the previous OS-related WGCNA key modules. When the *P* value is less than 0.05, that is, the differential expression of these genes has a significant impact on the patient's OS, these genes can be regarded as potential prognostic-related genes. Data in this step was from ccRCC cancer samples in TCGA.

**Construction, evaluation and validation of disease prognosis risk model**

For the candidate prognostic-related genes, combined with their expression in TCGA and again univariate Cox regression analysis was used to obtain genes with more significant risk, and then LASSO dimensionality reduction with 1,000 iterations was performed to screen out redundant genes to obtain more precise prognostic-related genes with high hazard ratio (HR) to construct a risk prognosis model. The following formula was used to calculate the risk score for each patient by using a linear combination of specific features that were weighted by their respective coefficients from LASSO:

$$Risk\ score = \sum_{i=1}^{n} exp_i * ß_i,$$

where n is the number of prognostic genes, $exp_i$ is the expression value of the i-th gene, and $ß_i$ is the regression coefficient of the i-th gene in the LASSO algorithm. According to the risk score of each patient given by the model, the median was taken as the cut-off value, and the samples were divided into high and low risk groups. The time-dependent receiver operating characteristic

(ROC) curve was used to evaluate the predictive ability of the model's 1-, 3- and 5-year survival periods, and the survival curves of the high and low risk groups were also analysed. The CPTAC dataset was taken as the external validation database to verify the prognostic risk model.

**Analysis of the correlation between risk score and other clinical factors and their prognostic accuracy**

Through the “ggpubr” R package, we utilized the wilcox.test method to analyse the correlation between risk score and clinical factors (age, gender, grade, stage, T, N, and M) and applied the “survivalROC” package to perform an analysis of prognostic accuracy on these clinical factors and risk score.

**Immune cell abundance analysis**

To estimate the population-specific immune infiltration about the proportion and difference of cells in the high and low risk groups, we used the “CIBERSORT” algorithm to represent the degree of absolute enrichment of a gene set in each sample within a given dataset. Normalized enrichment scores were calculated for each immune category.

**Expression level of prognostic-related genes**

We analysed the levels of expression of prognostic-related genes between ccRCC tumour samples and paracancerous tissues in TCGA and CPTAC expression data by using the “DEseq2” R package. The threshold of different expression was set as |log- 2FoldChange| > 1 and adjusted *P* value < 0.05.

**Independence of the prognostic model compared with other clinical characteristics**

In order to determine whether the predictive power of the prognostic model was independent of other clinical variables (including gender, age, T, N, M, and grade) in patients with ccRCC, univariate and multivariate Cox regression analyses were performed, in which other traditional clinical features were taken as independent variables, and OS was used as the dependent variable. The HR and 95% confidence intervals were calculated. And all *P* values were two-sided.

**Construction and assessment of a predictive nomogram**

Nomograms are widely used to predict the prognosis of cancer patients, mainly because they can simplify statistical prediction models into a single numerical estimate of OS probability tailored to individual patient conditions. In this study, a prognostic model was used to construct a nomogram to assess the probability of OS in patients with ccRCC at one, three, and five years. Subsequently, a discrimination and calibration were carried out. The discrimination of the nomogram was calculated by the bootstrap method using the consistency index (C-index), with 1,000 resamples. The value of the

C-index is between 0.5 and 1.0, where 1.0 means that the results of the model can be correctly distinguished and 0.5 means random chance. The calibration curve of the nomogram is evaluated graphically by plotting the relationship between the predicted probability of the nomogram and the observed rate. Overlap with the reference line indicates that the model is exactly the same. In addition, we also compared the predictive accuracy between nomogram built only with risk score and nomogram combined with all factors using ROC analysis.

## Results

### Pre-treated data

The single-cell data of the downloaded ccRCC was pre-processed as described in the Methods section, and we obtained 30,092 cells in total. In addition, we found that the correlation between the sequencing depth and the detected genes was 0.95, indicating that the deeper the sequencing depth was, the more genes were detected. Subsequently, we selected 2,000 genes with large variances for PCA analysis. The differences of all 20 PCs were extremely significant, indicating that the theoretical value and the actual value are quite different, which can be used for subsequent analysis.

There were 607 samples in the KIRC expression profile data of TGCA, of which 72 were paracancerous samples, 535 were cancer samples, and one sample with incomplete clinical information, which was removed. Finally, the data of 534 ccRCC tumour tissue samples were used for subsequent relative analysis. Among the data downloaded from the CPTAC database, there were 110 cancer samples, 75 paracancerous samples, and only 103 cancer samples contained clinical information. As a result, the data of these 103 samples were eventually used for analysis.

### ccRCC heterogeneity

For cluster analysis of single-cell data, we obtained a total of 23 subgroup clusters. After annotating by cell type, we found that immune cells were mainly concentrated in subgroups 0, 1, 2, 3, 4, 5, 6, 7, 9, 12, 13, 15, 16, 18 and 22 **(Figure 1, Supplementary Table 1)**. Specifically, CD8+ T cells were distributed in clusters 0, 2, 3, 7, 12, and 18. NK cells were only annotated in cluster 1. Monocytes assembled in clusters 4, 5, 13 and 22. Clusters 6, 9 and 15 were annotated to macrophages. B cells annotated only cluster 16.

### Differentially expressed genes and functional enrichment in different immune cell subgroups

We performed differential expression analysis on the genes in the 23 clusters obtained in the above step and displayed the first five genes in each cluster **(Figure 2A, Supplementary Table 2)**. According to the results of gene differential expression, we analysed the functional status of the annotated

immune cell clusters. In each immune cell type, the enrichment degree of hypoxia and quiescence was relatively high. Besides, the enrichment levels of EMT, invasion and stemness in B cells were also relatively high **(Figure 2B-F)**.

**Identification of immune cell marker genes expression**

A total of 42 immune cell marker genes related to ccRCC were downloaded from the CellMarker database and [17] performed differential expression analysis. The results are shown in the heat map **(Figure 2G)**.

**Ligand-receptor network**

In order to construct the ligand-receptor network, we first took the union of the differential genes of all immune cell clusters and the marker genes of all these immune cells. Afterwards, we intersected them with the ligand-receptor relationship pairs downloaded from [16]. Finally, a total of 981 pairs of ligand-receptor relationships were obtained **(Figure 3A, Supplementary Table 3)**.

**Immune cell multifactor network based on ligand-receptor network combined with transcription factors**

Intersecting genes in 981 ligand-receptor relationship pairs with transcription factor target genes, we obtained 7,987 immune cell multifactor network relationship pairs **(Supplementary Table 4)**. Then, 966 genes were obtained by intersecting the 973 genes contained in the network and the genes in TCGA dataset about ccRCC **(Supplementary Table 5)**. Because there are many relationship pairs, **Figure 3B** only shows a network diagram of partial genes.

**Co-expression network**

The construction of co-expression modules included 966 genes from the immune cell multifactor network. The appropriate scale-free power value was screened out as 4 **(Figure 4A)**. All constructed modules are painted with different colours, and the cluster trees of genes are also shown in **Figure 4B**. The black and magenta modules were chosen as the key modules, since they had the highest correlations with OS and OS time of ccRCC **(Figure 4C, 4D)**. The correlations between MEs and clinic traits are shown in **Figure 4E**. There were 53 genes in these two modules **(Supplementary Table 6)**. For a deeper understanding about the biofunctions of these modules, genes in these modules were carried out to conduct GO and KEGG pathway analyses by using the “clusterProfiler” R package. According to the *P* value of each term, the top terms in the GO and KEGG pathways were extracted out and visualized **(Figure 4F-I)**.

**Prognostic risk scoring model**

Using the "survival" R package to perform univariate Cox regression analysis on the 53 genes contained in the key modules of WGCNA, 28 genes with *P* value < 0.05 were obtained. **Figure 5A-D** show the survival analysis results of

four genes among them. Then, the 28 genes with significant prognostic differences were subjected to LASSO regression analysis. The specific steps were to first adopt the Cox proportional hazard model (family = "Cox") to calculate the HR and *P* values of these genes **(Figure 5E)** and then randomly simulated 1,000 times (maxit = 1000) to find the most suitable λ value **(Figure 5F)**. Finally, "lambda.min" was used to screen out 16 genes for constructing a risk scoring model from these 28 genes **(Figure 5G)**.

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

**Prognostic model prediction effectiveness evaluation and external data set verification**

In the evaluation of the predictive efficacy of the prognosis model, Kaplan-Meier (K-M) survival analysis was performed on the high and low risk groups, and the difference was significant **(Figure 6A)**. Moreover, in its ROC curve, the one-year AUC value was 0.794, the three-year AUC value was 0.746, and the five-year AUC value was 0.763 **(Figure 6C)**. In the external CPTAC dataset, K-M survival analysis was performed on the high and low risk groups, and the difference was also significant **(Figure 6B)**. In addition, the one-year AUC value in its ROC curve was 0.783, and the three-year AUC value was 0.762 **(Figure 6D)**. Because the external data does not have five-year survival data, only one-year and three-year ROC analysis was performed.

**Correlation between risk score and other clinical factors and their prognostic accuracy**

In the correlation comparison between risk score and other clinical factors, the results showed that there was no difference in risk score among the three age groups **(Figure 6E)**. In the grade grouping, the risk scores of G4 and G3 were significantly different from other stages, while the scores of G1 and G2 were not significantly different from other stages **(Figure 6F)**. In the stage grouping, the scores of stages 3 and 4 were significantly different from other stages, while the scores of stages 1 and 2 did not show significant differences **(Figure 6G)**. In the grouping comparison based on T stages, the differences in each stage were significant **(Figure 6H)**. In the group comparison based on the M stages, there were significant differences between M1 and M0, as well as M1

and MX **(Figure 6I)**. In a group comparison based on N stages, the results showed significant differences between N1 and N0, as well as between N1 and NX **(Figure 6J)**. There were also significant differences between gender-based risk scores **(Figure 6K)**. From the ROC curve, it was also found that, except for the N stage, the AUC values of other ROC curves were above 0.5, and the AUC values of the risk score and stage were above 0.7 **(Figure 6L-N)**.

**Immune infiltration level of high and low risk groups**

Analysis of the level of immune infiltration between the high and low risk groups revealed that there were significant differences in plasma cells, T cells CD8, T cells CD4 memory resting, T cells CD4 memory activated, T cells follicular helper, T cells regulatory, NK cells resting, Macrophages M0, Macrophages M1, Macrophages M2, Dendritic cells resting and Mast cells resting **(Figure 7A)**.

**Validation of gene expression in prognostic risk scoring model**

The difference in expression of genes in the risk prediction model between normal tissues and cancer tissues included in TCGA database is shown in **Figure 7B**. **Figure 7C** uncovers that genes in the risk prediction model are significantly different between cancer tissues and tumour-adjacent tissues in the CPTAC database. From the results of these two figures, we found that, except for PAX2 and HIPK2, the other genes contained in the model were DEGs in the two databases.

**The prognostic model is independent of regular clinical factors**

Univariate and multivariate Cox regression analyses were performed to assess the independent predictive value of the risk model in patients with ccRCC with intact clinical data from TCGA. The prognostic model and clinical covariates of age, grade, T, M and stage had some prognostic value with the univariate Cox regression analysis **(Figure 7D)**. After the multivariate Cox regression analysis, age and the prognostic model were found to be independent prognostic factors associated with OS **(Figure 7E)**.

**Predictive nomogram**

For the purpose of building a clinically applicable method to estimate the survival possibility of patients with ccRCC, we developed a nomogram to predict the probability of 1-, 3- and 5-year OS based on the data in TCGA. The predictors of the nomogram included age, gender, T, N, M, grade, risk score and stage **(Figure 8A)**. The C-index for the model for evaluation of OS was 0.799. The visual calibration chart was used to evaluate the performance of the nomogram. When the angle of the line is 45°, it represents the best prediction result. Thus, our calibration chart indicated that the nomogram has a good predictive ability **(Figure 8B-D)**. The AUC values of the nomograms combined all factors were greater than the nomograms only built with risk

score, in spite of their values were all more than 0.7, which indicated that the predictive precision of the nomogram combined with all factors was better **(Figure 8E-G)**.

## Discussion

The emergence of Next Generation Sequencing (NGS) has provided a feasible and cost-effective way to determine the transcriptional landscape of human cancers, including both bulk and single-cell resolution with unprecedented base-pair precision [20] [21] [22]. It has been established that cancer is attributed to dysregulated evolution [23] [24] in acquiring inheritable genetic/epigenetic traits [25] [26] [27]. However, the presence of tumour heterogeneity poses substantial challenges in the diagnosis and treatment of tumours [28] [29] [30] [31]. Tumour heterogeneity exerts a vital role in various aspects (e.g. intertumour, intratumour and intermetastases heterogeneity, interdisease and interpatient heterogeneity, epigenetic and metabolic heterogeneity, TME heterogeneity, tumour-intrinsic genetic heterogeneity, etc. [32] [33] [33] [34] [35]. A landmark paper has demonstrated ccRCC is a heterogeneous disease with marked genetic intermetastases and intratumour heterogeneity (G-IMH and G-ITH) [36]. Further studies have elucidated whether somatic mutation landscape and genetic heterogeneity influence the clinical outcomes of ccRCC tumour management [37]. Because of this, we adopted a series of bioinformatics methods to try to use the ccRCC single-cell data in published articles and the ccRCC-related data in public databases to study whether immune cell-related genes can construct a predictive prognostic model for patients with ccRCC, which may be helpful for a further understanding of the intratumour heterogeneity of ccRCC, and provide corresponding support for related basic research and clinical applications in the future.

Since there are many genes used to annotate a certain cell, it is usually difficult to determine which of these genes are critical. As a result, we built some networks, hoping to better find key genes related to our target clinical traits to construct a risk prediction model. Researchers have traditionally been concerned with a few or linear pathways between different cells. Identifying the signalling network of communication within different cell types is invaluable in the diagnosis and treatment of ccRCC tumours. Furthermore, a complete network of cell-cell signalling includes not only intercellular signalling pathways but also intracellular signalling transduction and gene expression [38]. Thus, a complete network of molecular signalling mechanisms is required to show the interaction between the TME and related cell types. One study has proved a potential scRNA-seq transcriptome-based multilayer network approach, which can be considered as a useful technique to identify the interplay between the TME and tumour cells, further predicting the prognostic and predictive signatures of cancer patients [16]. In addition to the multilayer network, we

also applied the WGCNA algorithm to explore the hub genes in key modules associated with the OS and OS time. The WGCNA algorithm explores the relationship between co-expression modules and clinical traits, which provides an opportunity to identify the hub genes in a module but not a downstream gene; thus, it possesses the distinct advantage of making the results more reliable and higher in biological significance [39]. In our study, we found all the genes related to the immune cells in the ccRCC samples through multilayer networks, then divided these genes into multiple modules using the WGCNA method, and used the genes in the modules with the strongest correlation with OS and OS time as the candidate genes for risk scoring model construction. Through the survival difference analysis of the genes in the key modules of the WGCNA algorithm, the genes with significant prognostic significance were found and the genes used to construct the risk prediction model were confirmed after LASSO dimensionality reduction processing. Subsequently, we verified the feasibility and effectiveness of the model for assessing the prognosis of patients with ccRCC through relevant dataset verification, DEGs used in the construction of the risk model, univariate regression analysis, multivariate regression analysis, nomogram and other approaches, which also showed that the immune cells in ccRCC do have an impact on the prognosis of patients.

Some genes in the prognostic risk scoring model have been proven to exert various effects on the regulation of certain tumours or diseases. Previous study has found that an imbalance of APP may be involved with the negative correlation between cancer and Alzheimer disease [40]. As a vital target in the TLR signalling pathway, CD14 exerts a dual effect on oncogenesis, which can initiate several tumour-related signalling pathways or alter the immune microenvironment in the tumour [41]. COL1A1 was considered to be associated with the pathogenesis of COL1A1-PDGFB fusion uterine sarcoma [42]. DDR1 was reported as attending the development of cancer and fibrotic diseases [43]. Regulating E2F5 is of great significance in maintaining genome stability and the cell cycle [44]. Study has shown that, if certain signal mutations cause the destruction of FGF1, it is likely to give rise to cancer [45]. The dysregulation of HDAC1, a chromatin modifier, may cause harmful effects on cell fate and function, which could lead to cancer [46]. HIPK2, a multitalented protein, utilizes its kinase activity to regulate several pathways to limit the proliferation and differentiation of tumour cells and induce positive responses to therapies [47]. HOXA9, a homeodomain-containing transcription factor, exerts a vital role in the proliferation of hematopoietic stem cells and is commonly negatively affected in acute leukaemias [48]. Recent study has shown that ITGA6 can be a useful biomarker for early detection of colorectal cancer cells in a non-invasive assay and as a prognostic factor [49]. L1CAM has been found in various types of human cancers, which indicates a bad prognosis [50]. NFKBIZ is a psoriasis susceptibility gene that encodes the functions of TRAF6 signalling players, especially in terms of positive and

regulatory immune controls by interactions between immune cells and epithelial cells [51]. Oncogenic gene fusion involving NRG1 contribute to the activation of ErbB-mediated pathways, representing a potential target for tumour management [52]. PAX2 has been found in epithelial tumours of the kidney and the female genital tract [53]. TEAD4 binds with YAP, TAZ, VGLL, and other transcription factors to modulate various tumour-related processes, including tumour cell proliferation, cell survival, tissue regeneration, and stem cell maintenance, in cancer via its transcriptional output [54]. The above-reported functions and mechanisms of these genes could help elucidate their correlations with ccRCC and provide reliable evidence for further determination of diagnostic and therapeutic methods.

Although our study only used published data and information in public databases, as well as did not directly use clinical samples for experimental testing and analysis, it is still sufficient to show that the data obtained through single-cell sequencing is able to provide an effective support to predict the prognosis of patients with ccRCC. Additionally, our research can also provide ideas for clinical diagnosis and treatment. For example, the genes in the risk prediction model we have established are more likely to become marker genes for clinical screening of ccRCC or as therapeutic targets for metastatic ccRCC. Furthermore, our methods and results would enhance theoretical assistance for other researchers to explore other cancers related to tumour heterogeneity in the future.

**Abbreviation**

| | |
|---|---|
| RCC: | renal cell carcinoma |
| ccRCC: | clear cell renal cell carcinoma |
| OS: | overall survival |
| TME: | tumor microenvironment |
| TAMs: | tumor-associated macrophages |
| CAFs: | cancer-associated fibroblasts |
| scRNA-seq: | single-cell RNA-sequencing |
| WGCNA: | weighted gene co-expression network analysis |
| TCGA: | The Cancer Genome Atlas |
| CPTAC: | Consortium for Clinical Proteome Cancer Analysis |
| PCA: | principal component analysis |
| PC: | principal components |
| DEGs: | differentially expressed genes |
| MEs: | module eigengenes |
| GO: | gene ontology |
| KEGG: | Kyoto Encyclopedia of Genes and Genomes |
| BP: | biological processes |
| MF: | molecular function |
| CC: | cellular components |
| HR: | hazard ratio |
| ROC: | receiver operating characteristic |
| C-index: | concordance index |
| DCA: | decision curve analysis |
| K-M: | KaplanMeier |

NGS: Next Generation Sequencing

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Availability of data and material

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

### Competing interests

All the authors declare that there are no competing interests.


### Funding

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


### Authors' contributions

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


### Acknowledgements

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


### Figure legends

**Figure 1.** Cell clustering. (A) Clustering of single-cell subpopulations. (B) The distribution of samples in clusters. (C) Annotation for all cell types.

**Figure 2.** (A) Heat map of the top five differential genes in each cluster. (B) B cell functional status analysis. (C) T cell functional status analysis. (D) Monocyte functional status analysis. (E) Macrophage functional status analysis. (F) NK cells functional status analysis. (G) Heat map of immune cell marker genes.

**Figure 3.** (A) Ligand-receptor interaction network diagram, different colours and sizes represent different numbers of node connections. (B) Multilayer network diagram—green is the ligand gene, red is the receptor gene, and yellow is the transcription factor.

**Figure 4.** WGCNA and functional analysis in black and magenta modules. (A) Schematic diagram of threshold screening and determination. (B) Clustering dendrogram of all genes from last step. (C) Correlations of OS time with mean gene significance and errors in all modules. (D) Correlations of OS with mean gene significance and errors in all modules. (E) Heat map of the correlations between MEs and OS traits. (F) BP term in GO. (G) CC term in GO. (H) MF term in GO. (I) KEGG pathway analysis.

**Figure 5.** Construction of disease prognosis risk model. (A-D) Survival analysis of four genes among the candidate genes used to construct the risk prediction model. (E) Forest plot for univariate regression analysis of 28 genes. (F) Select the appropriate λ value through LASSO regression analysis. (G) Scoring chart of risk model constructed by 16 genes.

**Figure 6.** Assessment of the prognostic model. (A, C) Evaluation of predictive effectiveness of risk prognostic model. (B, D) Use of the data in the CPTAC database as an external dataset for verification. (E-K) Correlation analysis of risk score and other clinical factors. Cell infiltration analysis results. (L-N) The ROC curve was used to analyse the accuracy of risk scores and other clinical factors in predicting the 1-, 3-, and 5-year survival of patients with ccRCC.

**Figure 7.** (A) Cell infiltration analysis results. (B) The differential expression levels of genes in the risk prediction model obtained through the analysis of TCGA database. (C) The differential expression levels of genes in the risk prediction model obtained through the analysis of the CPTAC database. (D) Univariate association of the prognostic model and clinicopathological characteristics with overall survival. (E) Multivariate association of the prognostic model and clinicopathological characteristics with overall survival.

**Figure 8.** (A) Nomogram predicting 1-, 3- and 5- year OS for patients with ccRCC. (B-D) The calibration curve for predicting 1-, 3- and 5-year OS for patients with ccRCC. (E-G) Time-dependent ROC curve analysis evaluates the accuracy of the nomograms.

**Tables**

Supplementary Table 1. Notes on cell clustering.

Supplementary Table 2. Differential genes in each cell cluster.

Supplementary Table 3. Ligand-receptor relationship pair.

Supplementary Table 4. Immune cell multifactor network relationship pair.

Supplementary Table 5. Intersection genes in immune cell multifactor network relationship pair and TCGA.

Supplementary Table 6. Genes in black and magenta models of WGCNA.

## References


1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2018.* CA Cancer J Clin, 2018. **68**(1): p. 7-30.

2. Hakimi, A.A., et al., *Transcriptomic Profiling of the Tumor Microenvironment Reveals Distinct Subgroups of Clear Cell Renal Cell Cancer: Data from a Randomized Phase III Trial.* Cancer Discov, 2019. **9**(4): p. 510-525.

3. Mitchell, T.J., et al., *Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal.* Cell, 2018. **173**(3): p. 611-623 e17.

4. Miao, D., et al., *Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma.* Science, 2018. **359**(6377): p. 801-806.

5. Greef, B. and T. Eisen, *Medical treatment of renal cancer: new horizons.* Br J Cancer, 2016. **115**(5): p. 505-16.

6. Shih, Y.C., et al., *Economic burden of renal cell carcinoma in the US: Part II--an updated analysis.* Pharmacoeconomics, 2011. **29**(4): p. 331-41.

7. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.

8. Papalexi, E. and R. Satija, *Single-cell RNA sequencing to explore immune cell heterogeneity.* Nat Rev Immunol, 2018. **18**(1): p. 35-45.

9. Albini, A. and M.B. Sporn, *The tumour microenvironment as a target for chemoprevention.* Nat Rev Cancer, 2007. **7**(2): p. 139-47.

10. Mantovani, A., et al., *Tumour-associated macrophages as treatment targets in oncology.* Nat Rev Clin Oncol, 2017. **14**(7): p. 399-416.

11. Hanley, C.J., et al., *Targeting the Myofibroblastic Cancer-Associated Fibroblast Phenotype Through Inhibition of NOX4.* J Natl Cancer Inst, 2018. **110**(1).

12. Peng, J., et al., *Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma.* Cell Res, 2019. **29**(9): p. 725-738.

13. Chung, W., et al., *Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer.* Nat Commun, 2017. **8**: p. 15081.

14. Navin, N.E., *The first five years of single-cell cancer genomics and beyond.* Genome Res, 2015. **25**(10): p. 1499-507.

15. Young, M.D., et al., *Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors.* Science, 2018. **361**(6402): p. 594-599.

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

17. Vishwakarma, A., et al., *Mapping the Immune Landscape of Clear Cell Renal Cell Carcinoma by Single-Cell RNA-seq.* bioRxiv, 2019: p. 824482.

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

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

20. Stratton, M.R., *Exploring the genomes of cancer cells: progress and promise.* Science, 2011. **331**(6024): p. 1553-8.

21. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. **472**(7341): p. 90-4.

22. Tirosh, I., et al., *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.* Science, 2016. **352**(6282): p. 189-96.

23. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-8.

24. Yates, L.R. and P.J. Campbell, *Evolution of the cancer genome.* Nat Rev Genet, 2012. **13**(11): p. 795-806.

25. Vogelstein, B. and K.W. Kinzler, *The Path to Cancer --Three Strikes and You're Out.* N Engl J Med, 2015. **373**(20): p. 1895-8.

26. Gerlinger, M., et al., *Cancer: evolution within a lifetime.* Annu Rev Genet, 2014. **48**: p. 215-36.

27. Heppner, G.H., *Tumor heterogeneity.* Cancer Res, 1984. **44**(6): p. 2259-65.

28. Marusyk, A., V. Almendro, and K. Polyak, *Intra-tumour heterogeneity: a looking glass for cancer?* Nat Rev Cancer, 2012. **12**(5): p. 323-34.

29. Burrell, R.A., et al., *The causes and consequences of genetic heterogeneity in cancer evolution.* Nature, 2013. **501**(7467): p. 338-45.

30. Alizadeh, A.A., et al., *Toward understanding and exploiting tumor heterogeneity.* Nat Med, 2015. **21**(8): p. 846-53.

31. Longo, D.L., *Tumor heterogeneity and personalized medicine.* N Engl J Med, 2012. **366**(10): p. 956-7.

32. Jamal-Hanjani, M., et al., *Translational implications of tumor heterogeneity.* Clin Cancer Res, 2015. **21**(6): p. 1258-66.

33. Voss, M.H. and J.J. Hsieh, *Therapeutic Guide for mTOuRing through the Braided Kidney Cancer Genomic River.* Clin Cancer Res, 2016. **22**(10): p. 2320-2.

34. Bedard, P.L., et al., *Tumour heterogeneity in the clinic.* Nature, 2013. **501**(7467): p. 355-64.

35. Andor, N., et al., *Pan-cancer analysis of the extent and consequences of intratumor heterogeneity.* Nat Med, 2016. **22**(1): p. 105-13.

36. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-892.

37. Hsieh, J.J., et al., *Overcome tumor heterogeneity-imposed therapeutic barriers through convergent genomic biomarker discovery: A braided cancer river model of kidney cancer.* Semin Cell Dev Biol, 2017. **64**: p. 98-106.

38. Sun, X., et al., *Cytokine combination therapy prediction for bone remodeling in tissue engineering based on the intracellular signaling pathway.* Biomaterials, 2012. **33**(33): p. 8265-76.

39. Chou, W.C., et al., *Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer.* BMC Genomics, 2014. **15**: p. 300.

40. Galvao, F., Jr., et al., *The amyloid precursor protein (APP) processing as a biological link between Alzheimer's disease and cancer.* Ageing Res Rev, 2019. **49**: p. 83-91.

41. Wu, Z., et al., *CD14: Biology and role in the pathogenesis of disease.* Cytokine Growth Factor Rev, 2019. **48**: p. 24-31.

42. Croce, S., I. Hostein, and W.G. McCluggage, *NTRK and other recently described kinase fusion positive uterine sarcomas: A review of a group of rare neoplasms.* Genes Chromosomes Cancer, 2021. **60**(3): p. 147-159.

43. Yeh, Y.C., H.H. Lin, and M.J. Tang, *Dichotomy of the function of DDR1 in cells and disease progression.* Biochim Biophys Acta Mol Cell Res, 2019. **1866**(11): p. 118473.

44. Dominguez-Brauer, C., et al., *Tumor suppression by ARF: gatekeeper and caretaker.* Cell Cycle, 2010. **9**(1): p. 86-9.

45. Jiang, X., et al., *The roles of fibroblast growth factors in the testicular development and tumor.* J Diabetes Res, 2013. **2013**: p. 489095.

46. Laugesen, A. and K. Helin, *Chromatin repressive complexes in stem cells, development, and cancer.* Cell Stem Cell, 2014. **14**(6): p. 735-51.

47. D'Orazi, G., C. Rinaldo, and S. Soddu, *Updates on HIPK2: a resourceful oncosuppressor for clearing cancer.* J Exp Clin Cancer Res, 2012. **31**: p. 63.

48. Collins, C.T. and J.L. Hess, *Role of HOXA9 in leukemia: dysregulation, cofactors and essential targets.* Oncogene, 2016. **35**(9): p. 1090-8.

49. Beaulieu, J.F., *Integrin alpha6beta4 in Colorectal Cancer: Expression, Regulation, Functional Alterations and Use as a Biomarker.* Cancers (Basel), 2019. **12**(1).

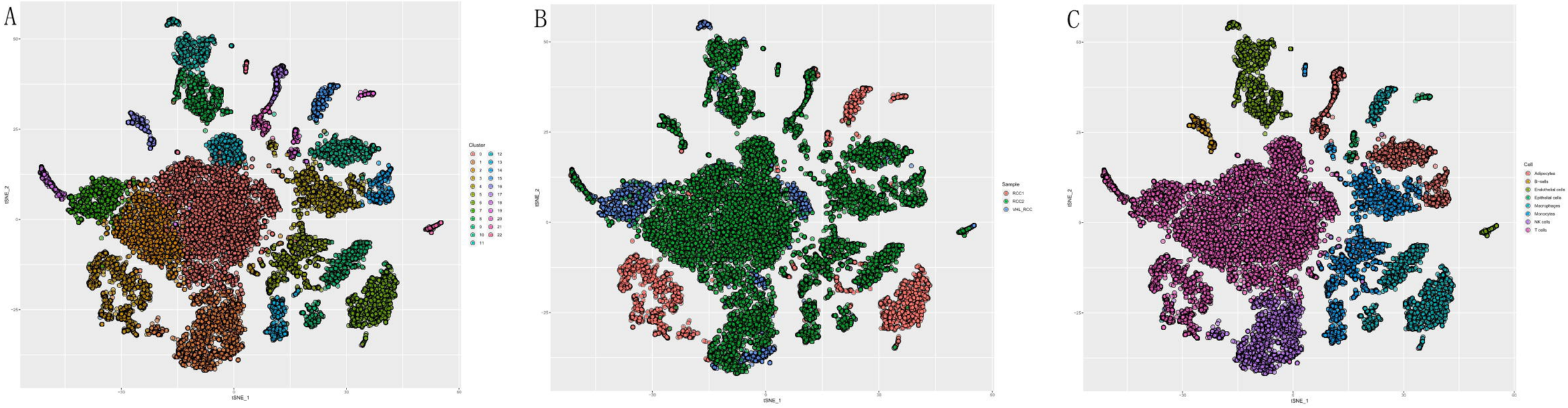
50. Altevogt, P., K. Doberstein, and M. Fogel, *L1CAM in human cancer.* Int J Cancer, 2016. **138**(7): p. 1565-76.

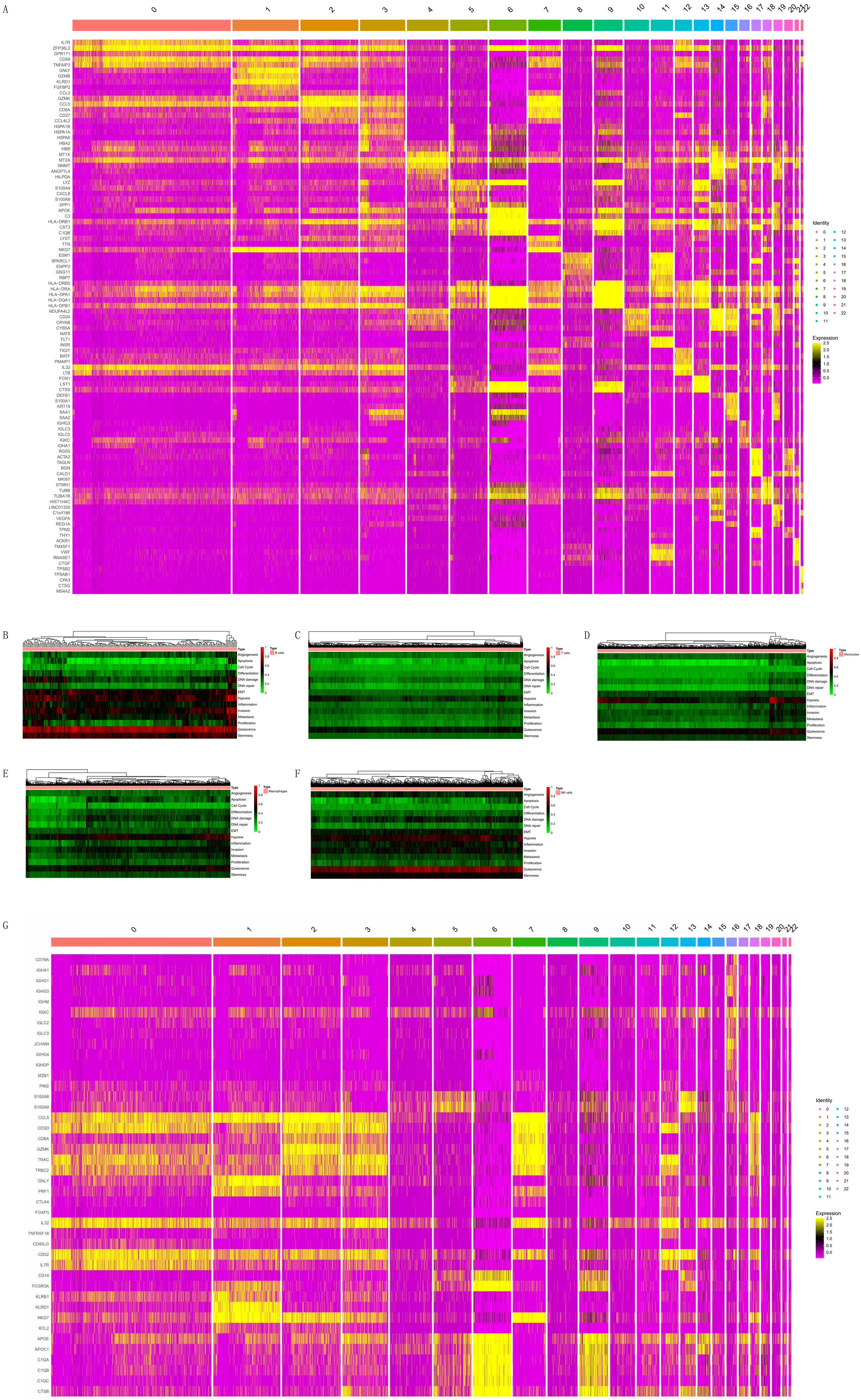
51. Dainichi, T., et al., *Immune Control by TRAF6-Mediated Pathways of Epithelial Cells in the EIME (Epithelial Immune Microenvironment).* Front Immunol, 2019. **10**: p. 1107.

52. Laskin, J., et al., *NRG1 fusion-driven tumors: biology, detection, and the therapeutic role of afatinib and other ErbB-targeting agents.* Ann Oncol, 2020. **31**(12): p. 1693-1703.

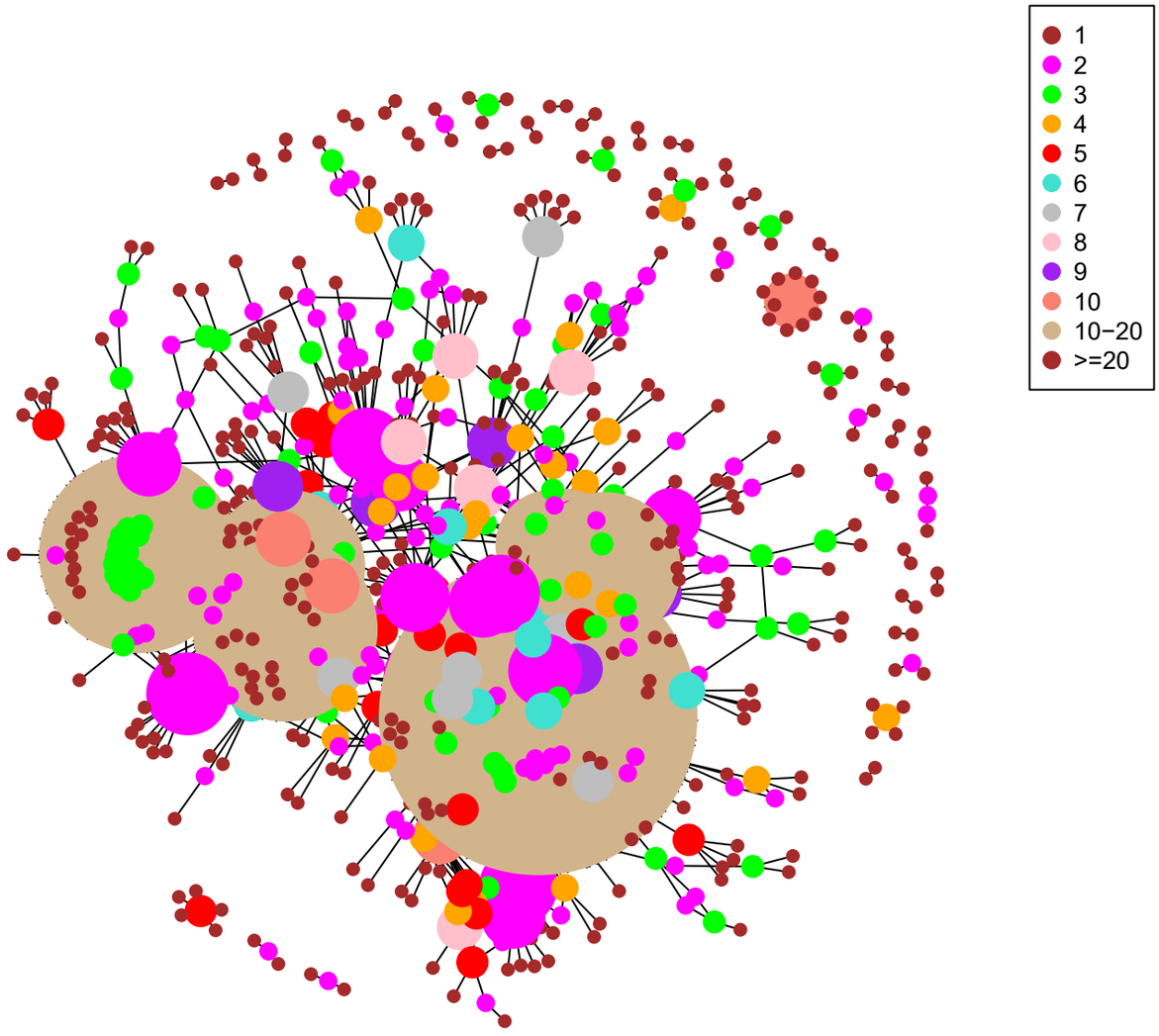
53. Ordonez, N.G., *Value of PAX2 immunostaining in tumor diagnosis: a review and update.* Adv Anat Pathol, 2012. **19**(6): p. 401-9.

54. Chen, M., et al., *Structural and Functional Overview of TEAD4 in Cancer Biology.* Onco Targets Ther, 2020. **13**: p. 9865-9874.

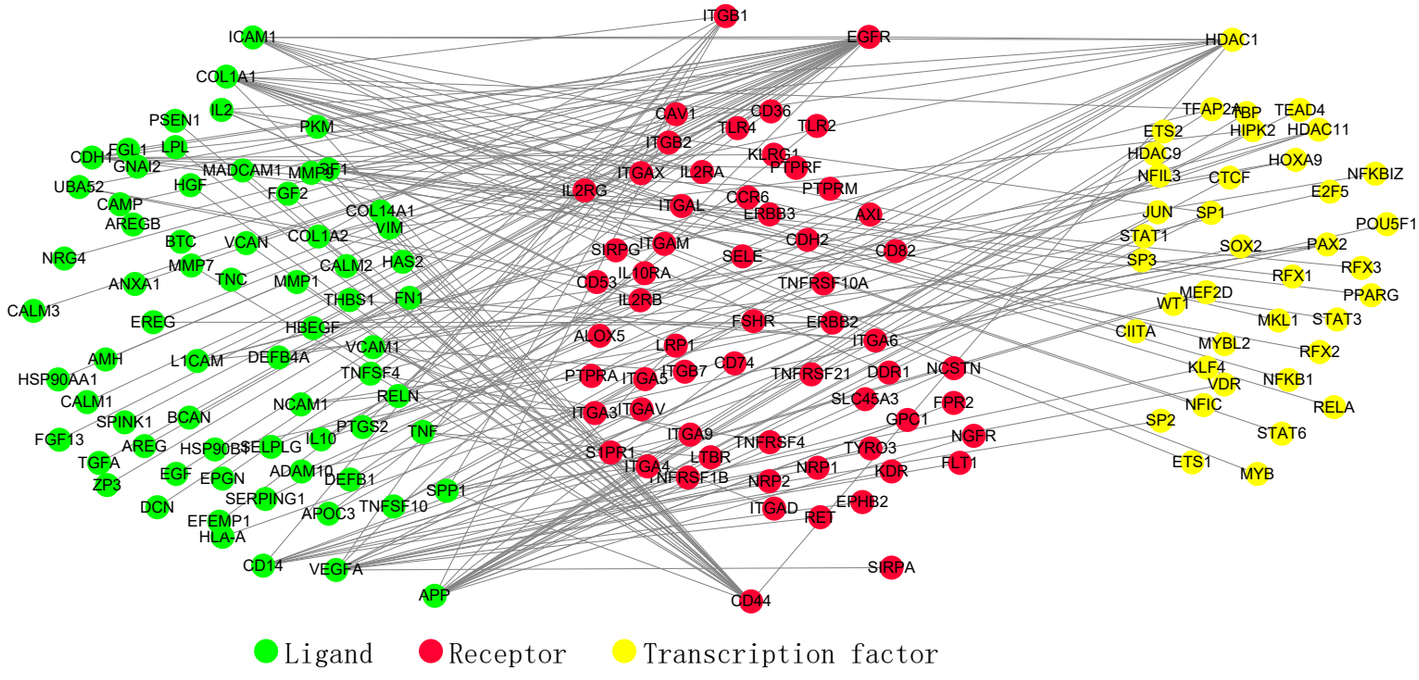

A
B
C
tSNE_1
tSNE_2
Cluster
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Sample
RCC1
RCC2
VHL_RCC
Cell
Adipocytes
B-cells
Endothelial cells
Epithelial cells
Macrophages
Monocytes
NK cells
T cells

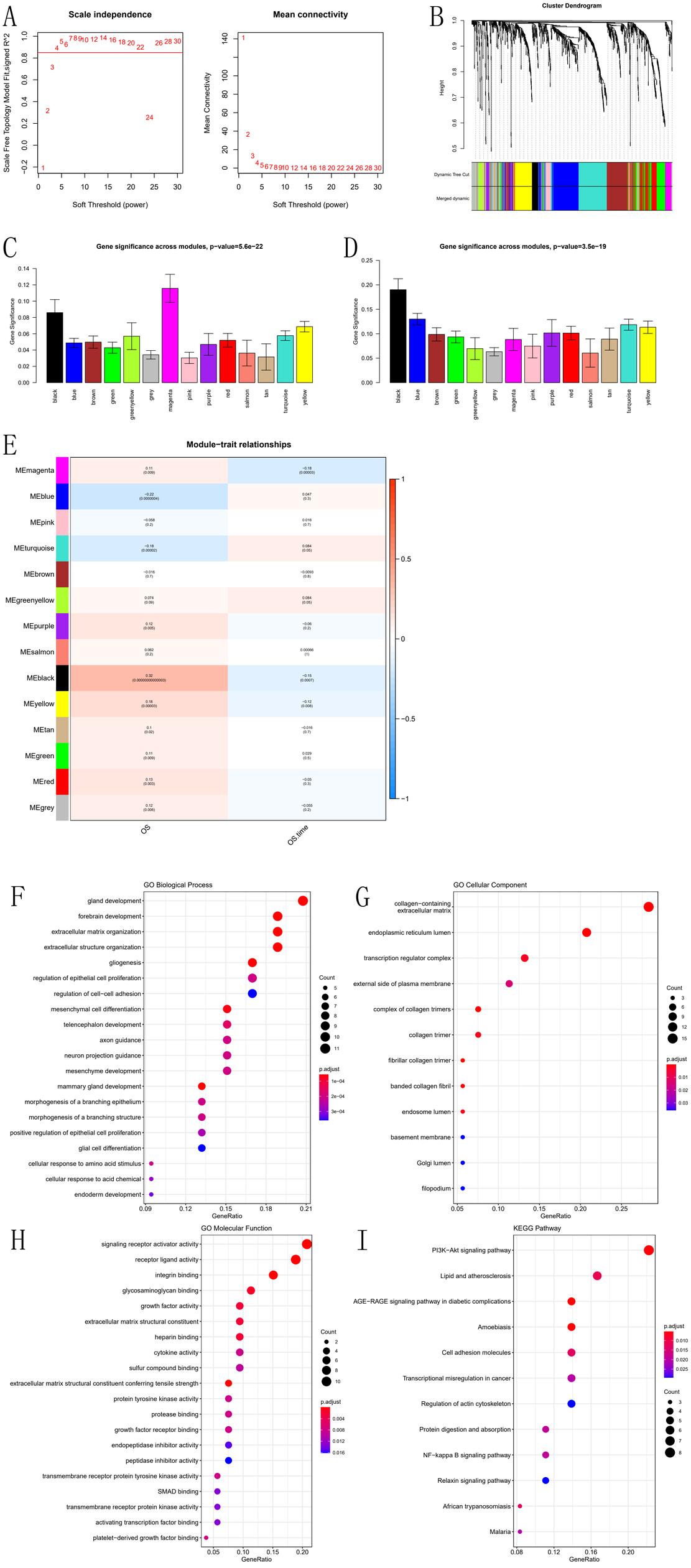

A
B
C
D
E
F
G
Identity
Expression
Type
B cells
T cells
Monocytes
Macrophages
NK cells
Angiogenesis
Apoptosis
Cell Cycle
Differentiation
DNA damage
DNA repair
EMT
Hypoxia
Inflammation
Invasion
Metastasis
Proliferation
Quiescence
Stemness

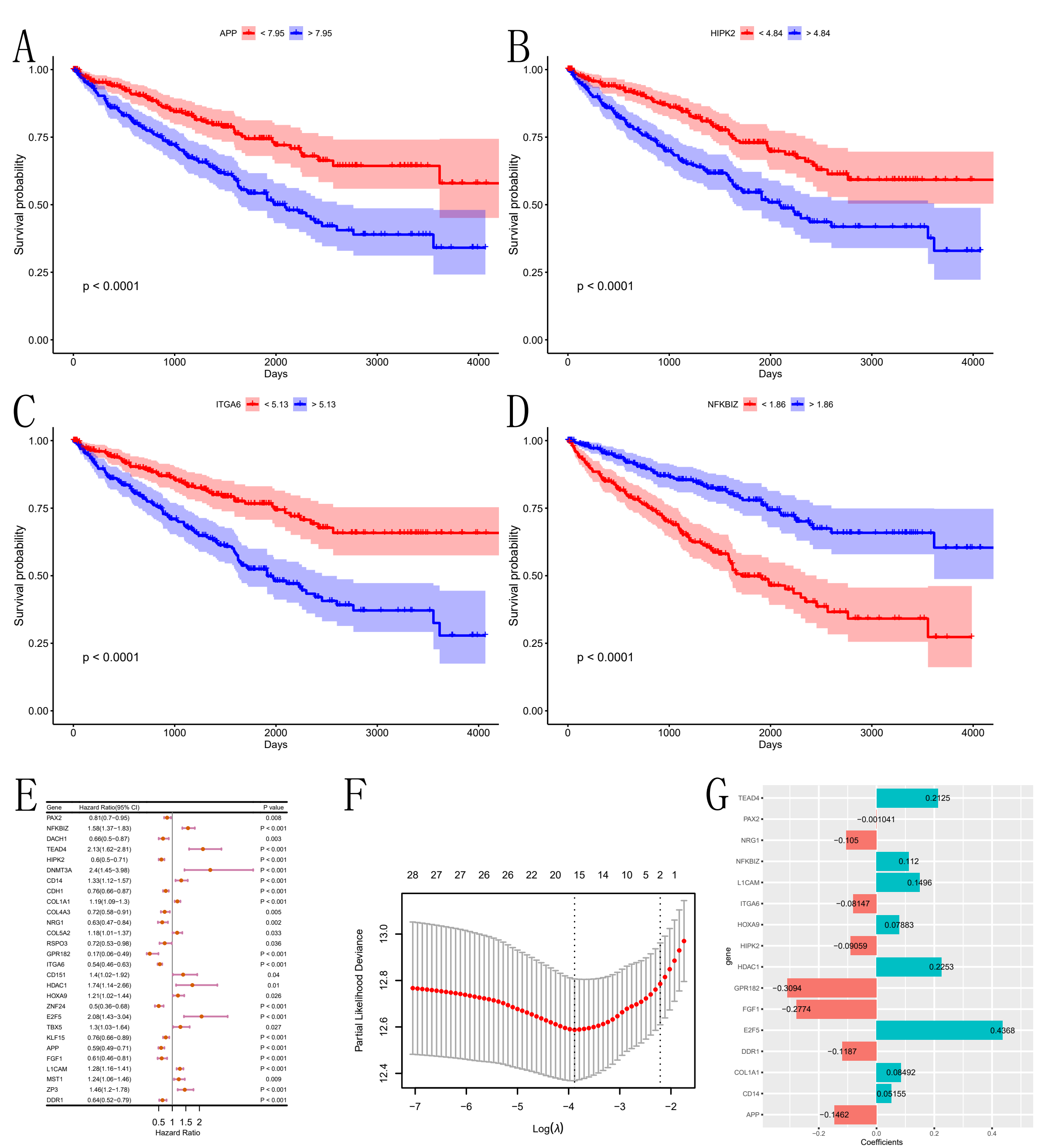
A

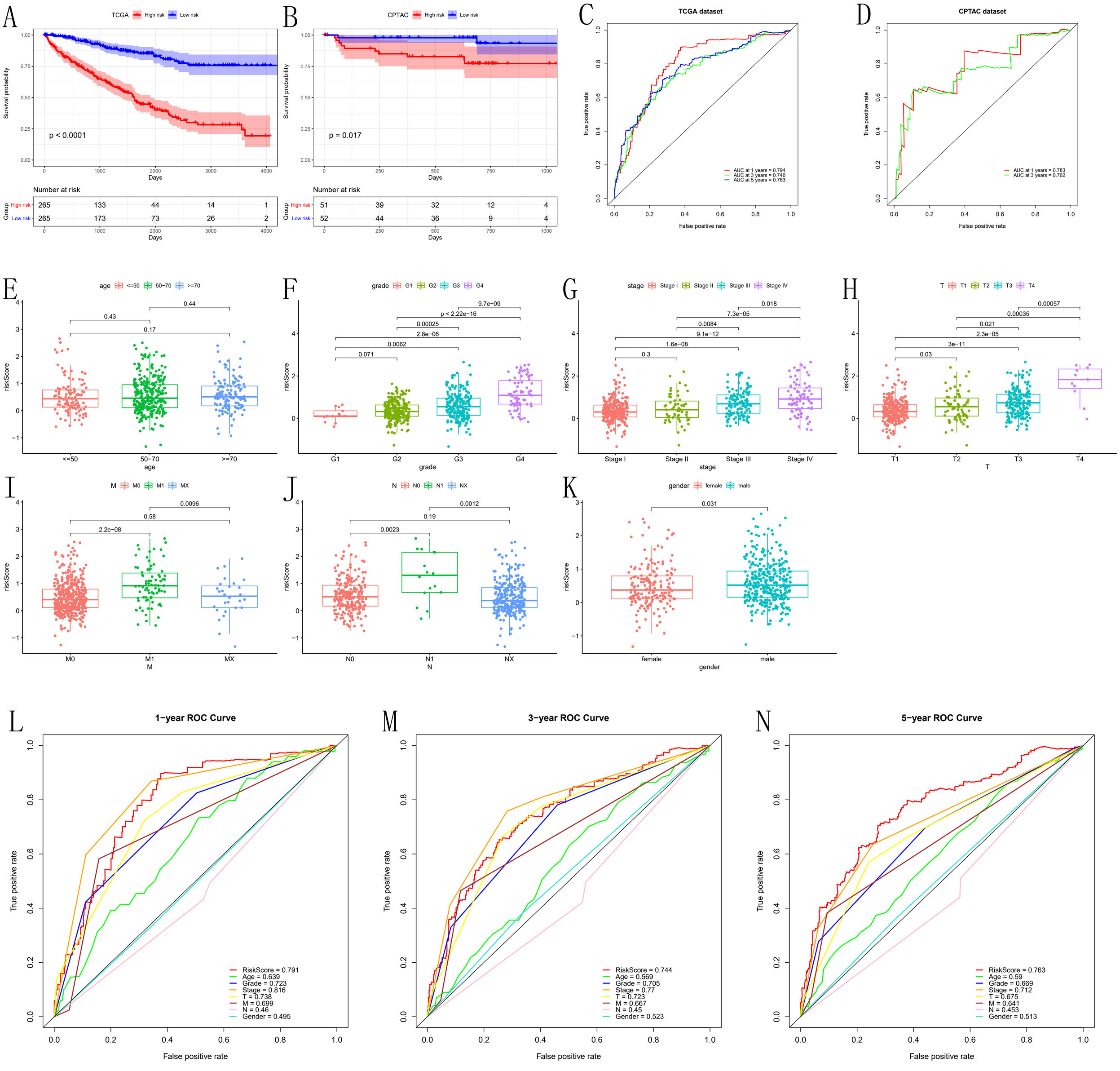

1
2
3
4
5
6
7
8
9
10
10−20
>=20

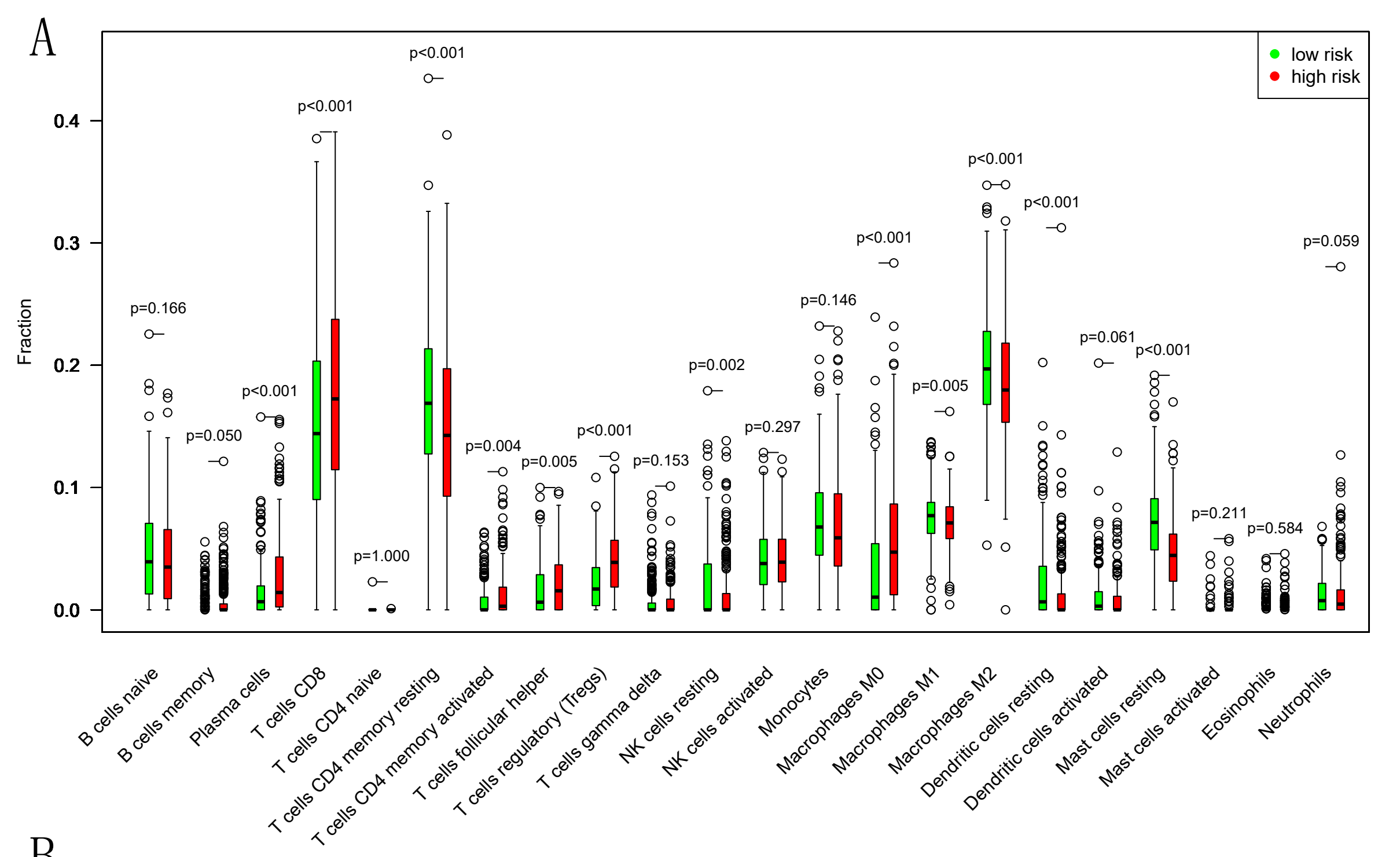
B

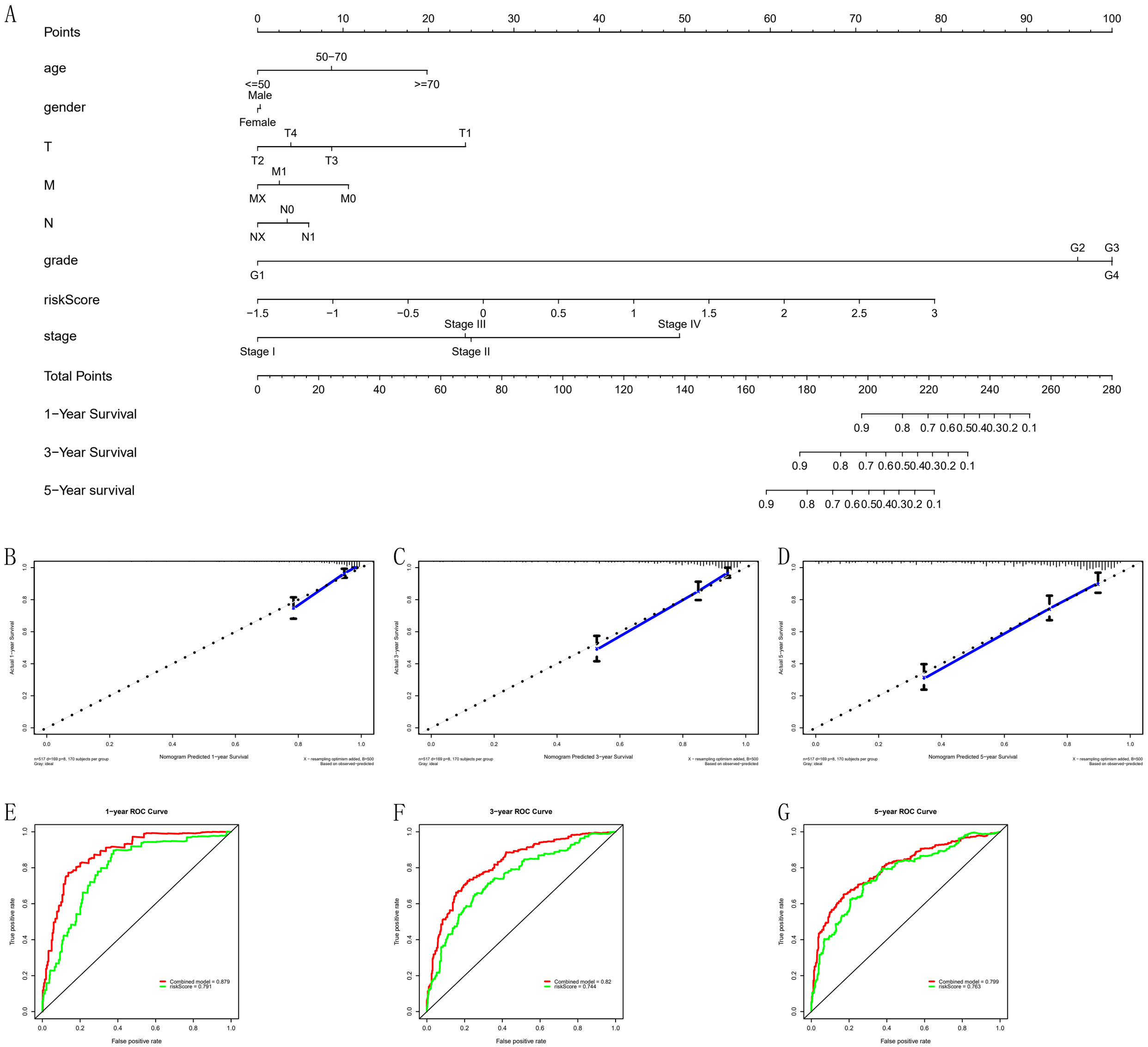

ITGB1
ICAM1
EGFR
HDAC1
COL1A1
IL2
PSEN1
PKM
CDH1
FGL1
LPL
GNAI2
MADCAM1
MMP3
SPF1
UBA52
HGF
FGF2
CAMP
AREGB
COL14A1
VIM
BTC
VCAN
COL1A2
NRG4
MMP7
CALM2
HAS2
ANXA1
TNC
MMP1
THBS1
FN1
CALM3
EREG
HBEGF
AMH
L1CAM
DEFB4A
VCAM1
HSP90AA1
TNFSF4
CALM1
RELN
SPINK1
BCAN
NCAM1
FGF13
PTGS2
IL10
TNF
AREG
HSP90B
SELPLG
TGFA
ADAM10
EGF
EPGN
DEFB1
ZP3
SERPING1
SPP1
DCN
APOC3
TNFSF10
EFEMP1
HLA-A
CD14
VEGFA
APP
CAV1
CD36
TLR4
TLR2
ITGB2
KLRG1
ITGAX
IL2RA
PTPRF
IL2RG
PTPRM
CCR6
ITGAL
ERBB3
AXL
ITGAM
CDH2
SIRPG
SELE
CD82
CD93
IL10RA
TNFRSF10A
IL2RB
FSHR
ERBB2
ALOX5
LRP1
ITGA6
CD74
PTPRA
ITGA5
ITGB7
TNFRSF21
DDR1
NCSTN
ITGA3
ITGAV
SLC45A3
FPR2
GPC1
ITGA9
TNFRSF4
NGFR
S1PR1
LTBR
TYRO3
ITGA4
TNFRSF1B
NRP1
FLT1
NRP2
KDR
ITGAD
EPHB2
RET
SIRPA
CD44
TFAP2A
TBP
TEAD4
ETS2
HIPK2
HDAC11
HDAC9
HOXA9
NFIL3
CTCF
NFKBIZ
E2F5
JUN
SP1
POU5F1
STAT1
SOX2
PAX2
SP3
RFX1
RFX3
MEF2D
PPARG
WT1
MKL1
STAT3
CIITA
MYBL2
RFX2
KLF4
VDR
NFKB1
NFIC
RELA
SP2
STAT6
ETS1
MYB
Ligand
Receptor
Transcription factor


A
Scale independence
Scale Free Topology Model Fit,signed R^2
Soft Threshold (power)
Mean connectivity
Mean Connectivity
Soft Threshold (power)
B
Cluster Dendrogram
Height
Dynamic Tree Cut
Merged dynamic
C
Gene significance across modules, p−value=5.6e−22
Gene Significance
black
blue
brown
green
greenyellow
grey
magenta
pink
purple
red
salmon
tan
turquoise
yellow
D
Gene significance across modules, p−value=3.5e−19
Gene Significance
E
Module−trait relationships
MEmagenta
MEblue
MEpink
MEturquoise
MEbrown
MEgreenyellow
MEpurple
MEsalmon
MEblack
MEyellow
MEtan
MEgreen
MEred
MEgrey
OS
OS.time
F
GO Biological Process
gland development
forebrain development
extracellular matrix organization
extracellular structure organization
gliogenesis
regulation of epithelial cell proliferation
regulation of cell−cell adhesion
mesenchymal cell differentiation
telencephalon development
axon guidance
neuron projection guidance
mesenchyme development
mammary gland development
morphogenesis of a branching epithelium
morphogenesis of a branching structure
positive regulation of epithelial cell proliferation
glial cell differentiation
cellular response to amino acid stimulus
cellular response to acid chemical
endoderm development
GeneRatio
Count
p.adjust
G
GO Cellular Component
collagen−containing extracellular matrix
endoplasmic reticulum lumen
transcription regulator complex
external side of plasma membrane
complex of collagen trimers
collagen trimer
fibrillar collagen trimer
banded collagen fibril
endosome lumen
basement membrane
Golgi lumen
filopodium
GeneRatio
Count
p.adjust
H
GO Molecular Function
signaling receptor activator activity
receptor ligand activity
integrin binding
glycosaminoglycan binding
growth factor activity
extracellular matrix structural constituent
heparin binding
cytokine activity
sulfur compound binding
extracellular matrix structural constituent conferring tensile strength
protein tyrosine kinase activity
protease binding
growth factor receptor binding
endopeptidase inhibitor activity
peptidase inhibitor activity
transmembrane receptor protein tyrosine kinase activity
SMAD binding
transmembrane receptor protein kinase activity
activating transcription factor binding
platelet−derived growth factor binding
GeneRatio
Count
p.adjust
I
KEGG Pathway
PI3K−Akt signaling pathway
Lipid and atherosclerosis
AGE−RAGE signaling pathway in diabetic complications
Amoebiasis
Cell adhesion molecules
Transcriptional misregulation in cancer
Regulation of actin cytoskeleton
Protein digestion and absorption
NF−kappa B signaling pathway
Relaxin signaling pathway
African trypanosomiasis
Malaria
GeneRatio
p.adjust
Count


A
APP < 7.95 > 7.95
Survival probability
p < 0.0001
Days
B
HIPK2 < 4.84 > 4.84
Survival probability
p < 0.0001
Days
C
ITGA6 < 5.13 > 5.13
Survival probability
p < 0.0001
Days
D
NFKBIZ < 1.86 > 1.86
Survival probability
p < 0.0001
Days
E
Gene Hazard Ratio(95% CI) P value
PAX2 0.81(0.7−0.95) 0.008
NFKBIZ 1.58(1.37−1.83) P < 0.001
DACH1 0.66(0.5−0.87) 0.003
TEAD4 2.13(1.62−2.81) P < 0.001
HIPK2 0.6(0.5−0.71) P < 0.001
DNMT3A 2.4(1.45−3.98) P < 0.001
CD14 1.33(1.12−1.57) P < 0.001
CDH1 0.76(0.66−0.87) P < 0.001
COL1A1 1.19(1.09−1.3) P < 0.001
COL4A3 0.72(0.58−0.91) 0.005
NRG1 0.63(0.47−0.84) 0.002
COL5A2 1.18(1.01−1.37) 0.033
RSPO3 0.72(0.53−0.98) 0.036
GPR182 0.17(0.06−0.49) P < 0.001
ITGA6 0.54(0.46−0.63) P < 0.001
CD151 1.4(1.02−1.92) 0.04
HDAC1 1.74(1.14−2.66) 0.01
HOXA9 1.21(1.02−1.44) 0.026
ZNF24 0.5(0.36−0.68) P < 0.001
E2F5 2.08(1.43−3.04) P < 0.001
TBX5 1.3(1.03−1.64) 0.027
KLF15 0.76(0.66−0.89) P < 0.001
APP 0.59(0.49−0.71) P < 0.001
FGF1 0.61(0.46−0.81) P < 0.001
L1CAM 1.28(1.16−1.41) P < 0.001
MST1 1.24(1.06−1.46) 0.009
ZP3 1.46(1.2−1.78) P < 0.001
DDR1 0.64(0.52−0.79) P < 0.001
Hazard Ratio
F
28 27 27 26 26 22 20 15 14 10 5 2 1
Partial Likelihood Deviance
Log(λ)
G
TEAD4 0.2125
PAX2 −0.001041
NRG1 −0.105
NFKBIZ 0.112
L1CAM 0.1496
ITGA6 −0.08147
HOXA9 0.07883
HIPK2 −0.09059
HDAC1 0.2253
GPR182 −0.3094
FGF1 −0.2774
E2F5 0.4368
DDR1 −0.1187
COL1A1 0.08492
CD14 0.05155
APP −0.1462
gene
Coefficients


A
TCGA High risk Low risk
Survival probability
p < 0.0001
Days
Number at risk
High risk 265 133 44 14 1
Low risk 265 173 73 26 2
B
CPTAC High risk Low risk
Survival probability
p = 0.017
Days
Number at risk
High risk 51 39 32 12 4
Low risk 52 44 36 9 4
C
TCGA dataset
True positive rate
False positive rate
AUC at 1 years = 0.794
AUC at 3 years = 0.746
AUC at 5 years = 0.763
D
CPTAC dataset
True positive rate
False positive rate
AUC at 1 years = 0.783
AUC at 3 years = 0.762
E
age <=50 50−70 >=70
riskScore
0.43
0.17
0.44
F
grade G1 G2 G3 G4
riskScore
0.071
0.0062
2.8e−06
0.00025
p < 2.22e−16
9.7e−09
G
stage Stage I Stage II Stage III Stage IV
riskScore
0.3
1.6e−08
9.1e−12
0.0084
7.3e−05
0.018
H
T T1 T2 T3 T4
riskScore
0.03
3e−11
2.3e−05
0.021
0.00035
0.00057
I
M M0 M1 MX
riskScore
2.2e−08
0.58
0.0096
J
N N0 N1 NX
riskScore
0.0023
0.19
0.0012
K
gender female male
riskScore
0.031
L
1−year ROC Curve
True positive rate
False positive rate
RiskScore = 0.791
Age = 0.639
Grade = 0.723
Stage = 0.816
T = 0.738
M = 0.699
N = 0.46
Gender = 0.495
M
3−year ROC Curve
True positive rate
False positive rate
RiskScore = 0.744
Age = 0.569
Grade = 0.705
Stage = 0.77
T = 0.723
M = 0.667
N = 0.45
Gender = 0.523
N
5−year ROC Curve
True positive rate
False positive rate
RiskScore = 0.763
Age = 0.59
Grade = 0.669
Stage = 0.712
T = 0.675
M = 0.641
N = 0.453
Gender = 0.513


A
low risk
high risk
Fraction
0.0
0.1
0.2
0.3
0.4
p=0.166
p=0.050
p<0.001
p<0.001
p=1.000
p<0.001
p=0.004
p=0.005
p<0.001
p=0.153
p=0.002
p=0.297
p=0.146
p<0.001
p=0.005
p<0.001
p<0.001
p=0.061
p<0.001
p=0.211
p=0.584
p=0.059
B cells naive
B cells memory
Plasma cells
T cells CD8
T cells CD4 naive
T cells CD4 memory resting
T cells CD4 memory activated
T cells follicular helper
T cells regulatory (Tregs)
T cells gamma delta
NK cells resting
NK cells activated
Monocytes
Macrophages M0
Macrophages M1
Macrophages M2
Dendritic cells resting
Dendritic cells activated
Mast cells resting
Mast cells activated
Eosinophils
Neutrophils
B
Normal
Tumor
gene expression
0
2
4
6
8
10
12
14
p<0.001
p=0.079
p=0.006
p<0.001
p=0.213
p<0.001
p=0.002
p<0.001
p=0.002
p<0.001
p<0.001
p<0.001
p<0.001
p<0.001
p<0.001
p<0.001
PAX2
NFKBIZ
TEAD4
HIPK2
CD14
COL1A1
NRG1
GPR182
ITGA6
HDAC1
HOXA9
E2F5
APP
FGF1
L1CAM
DDR1
C
Normal
Tumor
gene expression
0
2
4
6
8
p<0.001
p<0.001
p<0.001
p=0.662
p<0.001
p<0.001
p<0.001
p<0.001
p=0.001
p=0.005
p<0.001
p<0.001
p<0.001
p<0.001
p<0.001
p=0.006
PAX2
NFKBIZ
TEAD4
HIPK2
CD14
COL1A1
NRG1
GPR182
ITGA6
HDAC1
HOXA9
E2F5
APP
FGF1
L1CAM
DDR1
D
pvalue
Hazard ratio
age <0.001 1.029(1.016−1.043)
grade <0.001 2.297(1.872−2.819)
M <0.001 2.189(1.736−2.761)
N 0.189 0.903(0.776−1.052)
T <0.001 1.926(1.632−2.274)
gender 0.724 1.058(0.773−1.449)
stage <0.001 1.901(1.664−2.173)
riskScore <0.001 3.298(2.663−4.083)
0
1
2
3
4
Hazard ratio
E
pvalue
Hazard ratio
age <0.001 1.033(1.018−1.048)
grade 0.211 1.168(0.916−1.491)
M 0.889 1.029(0.691−1.532)
N 0.338 0.926(0.791−1.084)
T 0.100 0.759(0.546−1.054)
gender 0.617 0.919(0.661−1.279)
stage <0.001 1.905(1.400−2.591)
riskScore <0.001 2.503(1.946−3.220)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Hazard ratio


A
Points
age
gender
T
M
N
grade
riskScore
stage
Total Points
1−Year Survival
3−Year Survival
5−Year survival
50−70
<=50
>=70
Male
Female
T4
T1
T2
T3
M1
MX
M0
N0
NX
N1
G2
G3
G1
G4
Stage III
Stage IV
Stage I
Stage II
B
Actual 1−year Survival
Nomogram Predicted 1−year Survival
n=517 d=169 p=8, 170 subjects per group
Gray: ideal
X − resampling optimism added, B=500
Based on observed−predicted
C
Actual 3−year Survival
Nomogram Predicted 3−year Survival
D
Actual 5−year Survival
Nomogram Predicted 5−year Survival
E
1−year ROC Curve
True positive rate
False positive rate
Combined model = 0.879
riskScore = 0.791
F
3−year ROC Curve
Combined model = 0.82
riskScore = 0.744
G
5−year ROC Curve
Combined model = 0.799
riskScore = 0.763