


# SurviveAI: Long Term Survival Prediction of Cancer Patients Based on Somatic RNA-Seq Expression

Omri Nayshool<sup>1,2</sup> , Nitzan Kol<sup>1</sup>, Elisheva Javaski<sup>1</sup>, Ninette Amariglio<sup>1</sup> and Gideon Rechavi<sup>1,2</sup>

<sup>1</sup>Bioinformatics Unit, Sheba Cancer Research Center and Wohl Institute for Translational Medicine, Sheba Medical Center, Tel HaShomer, Israel. <sup>2</sup>Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Israel.

Cancer Informatics  
Volume 21: 1–9  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/11769351221127875



## ABSTRACT

**MOTIVATION:** Prediction of cancer outcome is a major challenge in oncology and is essential for treatment planning. Repositories such as The Cancer Genome Atlas (TCGA) contain vast amounts of data for many types of cancers. Our goal was to create reliable prediction models using TCGA data and validate them using an external dataset.

**RESULTS:** For 16 TCGA cancer type cohorts we have optimized a Random Forest prediction model using parameter grid search followed by a backward feature elimination loop for dimensions reduction. For each feature that was removed, the model was retrained and the area under the curve of the receiver operating characteristic (AUC-ROC) was calculated using test data. Five prediction models gave AUC-ROC bigger than 80%. We used Clinical Proteomic Tumor Analysis Consortium v3 (CPTAC3) data for validation. The most enriched pathways for the top models were those involved in basic functions related to tumorigenesis and organ development. Enrichment for 2 prediction models of the TCGA-KIRP cohort was explored, one with 42 genes (AUC-ROC = 0.86) the other is composed of 300 genes (AUC-ROC = 0.85). The most enriched networks for both models share only 5 network nodes: DMBT1, IL11, HOXB6, TRIB3, PIM1. These genes play a significant role in renal cancer and might be used for prognosis prediction and as candidate therapeutic targets.

**AVAILABILITY AND IMPLEMENTATION:** The prediction models were created and tested using Python SciKit-Learn package. They are freely accessible via a friendly web interface we called surviveAI at <https://tinyurl.com/surviveai>.

**KEYWORDS:** Cancer survivors, supervised machine learning, gene expression, classification, molecular targeted therapy

**RECEIVED:** October 25, 2021. **ACCEPTED:** September 5, 2022.

**TYPE:** Software or Database Review

**FUNDING:** The author(s) received no financial support for the research, authorship, and/or publication of this article.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: GR is a consultant to Pangea Biomed.

**CORRESPONDING AUTHORS:** Omri Nayshool, Sheba Cancer Research Center, Sheba Medical Center, Tel HaShomer, Derech Sheba 2, Ramat Gan 52621, Israel. Email: omri.nayshool@sheba.gov.il

Gideon Rechavi, Sheba Cancer Research Center, Sheba Medical Center, Tel HaShomer, Derech Sheba 2, Ramat Gan 52621, Israel. Email: Gidi.Rechavi@sheba.health.gov.il

## Introduction

Cancer is a leading cause of death, and accounts for 17% of mortality worldwide. According to a report from the International Agency for Research on Cancer (IARC), in 2018 a total of 9.5 million deaths and 18 million new cases of cancer were reported worldwide. Interestingly, incidence and mortality rates are higher in men and in the developed world.<sup>1</sup> While some types of cancers are treated based on biomarkers and specific genetic mutations,<sup>1,2</sup> most cases are still treated according to specific guidelines by surgery, chemotherapy, and/or radiotherapy based on data integrating the clinical, histopathological, details of therapy, imaging, and outcome information of the patients.

Accurate prediction of prognosis of the various subtypes of cancer may improve tailoring of therapy by allowing to take into consideration the expected outcome versus therapy choice, intensity, risk, side effects, and late complications.

In the last decade, large OMICS databases were created that contain data generated from thousands of cancer samples. The largest one, The Cancer Genome Atlas (TCGA), a repository

that contains genomic, epigenomic, transcriptomic, proteomic, and clinical data, characterizing 33 types of tumors from over 20,000 patients, is considered to be one of the largest sources for cancer OMICS data. Many groups have tried to use TCGA data to predict the prognosis of patients affected by various tumors using machine learning approaches, with varying levels of success.<sup>3–8</sup>

Random Forest<sup>9</sup> is a simple yet effective Machine Learning algorithm that proved to be a successful predictor when using structured data such as RNA expression analysis.<sup>10</sup> It has low overfitting and a simple feature importance scoring function that is based on the Mean Decrease in Impurity function (Gini Importance). This allows refinement of prediction models and adds important insights into the biological role of each feature in cancer development and prognosis.

Cancer outcome prediction using OMICS-related data evolved in the last 2 decades starting with the use of gene-expression microarrays.<sup>11,12</sup> The accumulation of data from various OMICS technologies calls for the development of advanced cancer outcome prediction tools.



Here we describe a robust and simple analysis prediction tool using the Random Forest algorithm on 5 tumor types using the TCGA database.

## Methods

### Data

All RNA-seq datasets were downloaded from Genomic Data Commons (GDC). Clinical data was downloaded from the firehose data portal. The RNA-seq FPKM-UQ normalized data for cancer types of the TCGA projects were downloaded from National Cancer Institute's Genomic Data Commons data portal. The samples in each project were divided into 2 groups. The first group included samples from patients who were tumor-free for over 3 years (Tumor-Free samples), the second group included samples from patients that succumbed to the disease at any time point (Deceased group). We only used projects where the ratio between group size and the total number of samples was between 20% and 80% (Table 1). Validation of the models was done using 2 datasets from Clinical Proteomic Tumor Analysis Consortium v3(CPTAC3): Clear Cell Renal Cell Carcinoma (CPTAC3-ccRCC) and Uterine Corpus Endometrial Carcinoma (CPTAC3-UCEC).

### Software

We used python 3.7.6 and dependencies for full data analysis. Random Forest classifiers were created using scikit-learn 0.23.2. Data parsing and analysis were done using pandas 1.1.1. Ingenuity Pathway Analysis was used for network enrichment assessments. The webapp was created based on Flask 1.1.2 and Jinja2 2.11.2.

### Random forest model training and testing

Dimension reduction for the model required several steps as illustrated in Figure 1. The first RF model for each TCGA project was created using all 65483 mRNA features. The model parameters were selected using GridSearchCV module from scikit-learn, which tests all possible combinations from the provided list as detailed in Supplemental Table 2. After the model training, the features were scored and sorted using the model's property *feature\_importances\_*. The features with no importance (score 0) were removed. With the rest of the features, we have built a second model using a second GridSearchCV parameter search. The features were sorted and scored, and again features with no importance (score 0) were removed. We have continued dimensionality reduction using a backward feature elimination loop,<sup>13</sup> until only one feature was left. For each reduction (N-1), each TCGA model was trained using 70% of the samples. The Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC) was calculated using the predictions made on the testing data (the remaining 30% of the TCGA samples). Figure 2 shows that the optimal

model that was selected was the one in where the minimal features provided the average AUC-ROC of the last 500 modes.

The code for the models creation pipelines can be downloaded from <https://github.com/omrin/surviveai>.

### Results

*AUC-ROC mean of over 80% was achieved in 5 projects.* Out of 26 RNA-Seq TCGA-tumor type projects, only 14 had the required ratio between group size and the total number of samples (20%-80%) and had a minimum of 30 samples.

An average AUC-ROC score was calculated for the last 500 models (features range from 1 to 500). Out of the 15 cancer types, 5 tumor groups had an average AUC-ROC of over 80% TCGA-LGG (low grade glioma) 0.92 AUC-ROC, TCGA-COAD (colon adenocarcinoma) 0.84 AUC-ROC, TCGA-SARC (sarcoma) 0.86 AUC-ROC, TCGA-CESC (cervical squamous cell carcinoma and endocervical adenocarcinoma) 0.8 AUC-ROC, and TCGA-KIRP (kidney renal papillary cell carcinoma) 0.88 AUC-ROC. Detailed results and statistics for each TCGA project can be found at Table 1.

The model with the minimal features that most closely predicted the calculated AUC-ROC average was selected (See Figure 2). Each selected model used dozens of dimensions: a maximum of 90 features for TCGA-LGG and minimum 12 features for TCGA-COAD.

*Prediction of the top 5 models highly correlates with sample tumor origin.* The top 5 models were tested on all 15 TCGA project sample datasets. AUC-ROC scores were calculated for each dataset using the predictions of each sample and the known final results, see heatmap in Table 3. As expected, the scores for the training samples that were used to create prediction models were high and close to 1. For other datasets, the predictions were almost without correlation to the true condition of the samples (score .5)

Interestingly, a high negative correlation was found between the prediction models TCGA-CESC, TCGA-LGG, and TCGA-SARC and the predictions for the samples of the TCGA-READ project.

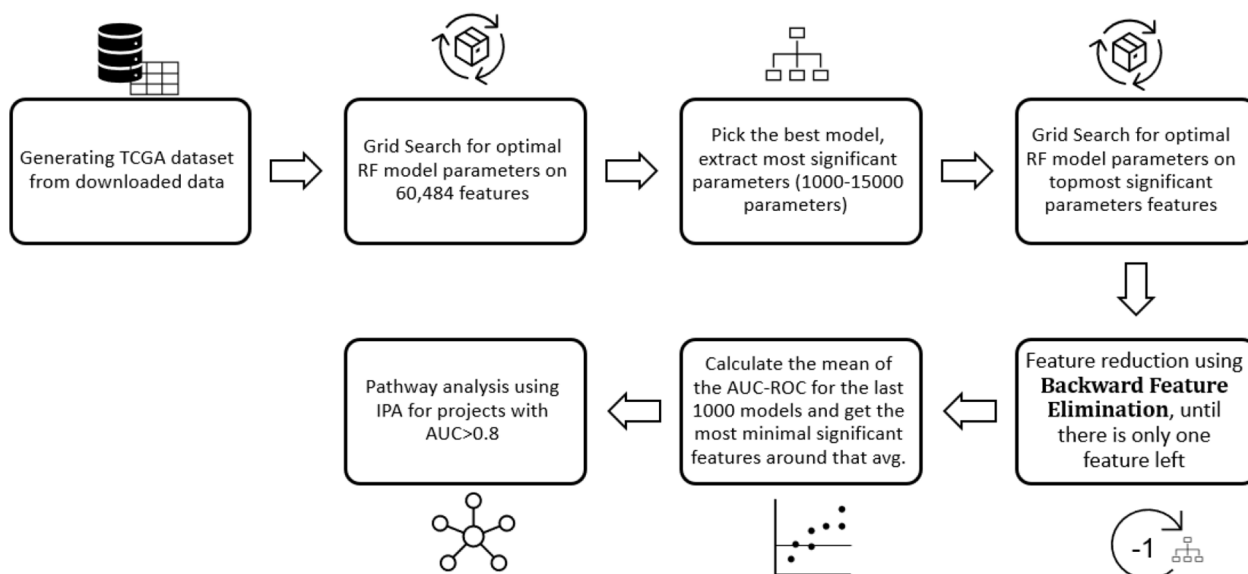
*Correlation of outcome predictions of TCGA dataset analyses with the validation datasets within tumor types.* In addition to the validation of the predictions obtained by analysis of the TCGA training sets using the testing sets, we further validated the models using 2 independent datasets that served for measuring the robustness of the models. We chose CPTAC3-ccRCC renal tumor dataset for the validation of a model developed for the same tissue of origin tumor which had a high prediction score, and CPTAC3-UCEC uterine tumor as an independent dataset for the testing of our model for the analysis of a tumor type where specific prediction model had a lower score.

The AUC-ROC obtained by the application of the TCGA-KIRP based model for the analysis of the CPTAC3-ccRCC

**Table 1.** Samples summary for each TCGA project for total samples in the cohort and samples with RNA-seq data. The Area Under the Curve of Receiver Operating Characteristic curve (AUC-ROC) mean for the last 500 models (500 to 1 features) was calculated for each project. The bold lines are the models that scored averaged AUC-ROC of above 0.8. The CPTAC3-ccRCC and CPTAC3-UCEC data were tested on the selected model with the minimal number of features for each project and the AUC-ROC was calculated respectively.

Project	Cancer type	TCGA SAMPLES			SAMPLES WITH AVAILABLE RNA-SEQ DATA			AUC-ROC AVERAGE (TOP 500 MODELS)	AUC-ROC CPTAC3-CCRCC	AUC-ROC CPTAC3-UCEC	NUMBER OF FINAL SELECTED MODEL FEATURES
		Tumor-free	Deceased	Tumor-free/Total	Tumor-free	Deceased	Tumor-free/Total				
TCGA-HNSC	Head and neck squamous cell carcinoma	50	170	0.23	47	166	0.221	0.50	0.34	30	
<b>TCGA-LGG</b>	<b>Brain lower grade glioma</b>	<b>28</b>	<b>92</b>	<b>0.23</b>	<b>27</b>	<b>91</b>	<b>0.229</b>	<b>0.52</b>	<b>0.45</b>	<b>90</b>	
TCGA-LUSC	Lung squamous cell carcinoma	49	161	0.23	49	157	0.238	0.53	0.42	29	
TCGA-LUAD	Lung adenocarcinoma	39	127	0.23	36	125	0.224	0.62	0.30	47	
TCGA-READ	Rectum adenocarcinoma	3	9	0.25	3	8	0.273				
<b>TCGA-COAD</b>	<b>Colon adenocarcinoma</b>	<b>20</b>	<b>57</b>	<b>0.26</b>	<b>19</b>	<b>56</b>	<b>0.253</b>	<b>0.64</b>	<b>0.69</b>	<b>12</b>	
TCGA-LIHC	Liver hepatocellular carcinoma	32	91	0.26	32	88	0.267	0.66	0.49	99	
<b>TCGA-SARC</b>	<b>Sarcoma</b>	<b>33</b>	<b>76</b>	<b>0.30</b>	<b>33</b>	<b>74</b>	<b>0.308</b>	<b>0.85</b>	<b>0.74</b>	<b>36</b>	
TCGA-BLCA	Bladder urothelial carcinoma	49	109	0.31	48	107	0.310	0.54	0.50	50	
TCGA-SKCM	Skin cutaneous melanoma	84	156	0.35	84	154	0.353	0.51	0.43	63	
<b>TCGA-CESC</b>	<b>Cervical squamous cell carcinoma and endocervical adenocarcinoma</b>	<b>53</b>	<b>60</b>	<b>0.47</b>	<b>53</b>	<b>59</b>	<b>0.473</b>	<b>0.69</b>	<b>0.60</b>	<b>83</b>	
TCGA-KIRC	Kidney renal clear cell carcinoma	159	162	0.50	157	157	0.500	0.79	0.65	96	
<b>TCGA-KIRP</b>	<b>Kidney renal papillary cell carcinoma</b>	<b>39</b>	<b>32</b>	<b>0.55</b>	<b>37</b>	<b>32</b>	<b>0.536</b>	<b>0.88</b>	<b>0.71</b>	<b>42</b>	
TCGA-BRCA	Breast invasive carcinoma	193	104	0.65	191	104	0.647	0.71	0.34	99	
TCGA-UCEC	Uterine corpus endometrial carcinoma	94	45	0.68	93	44	0.679	0.72	0.63	19	

## Algorithm workflow for each TCGA project



**Figure 1.** Model generation workflow for each TCGA project.

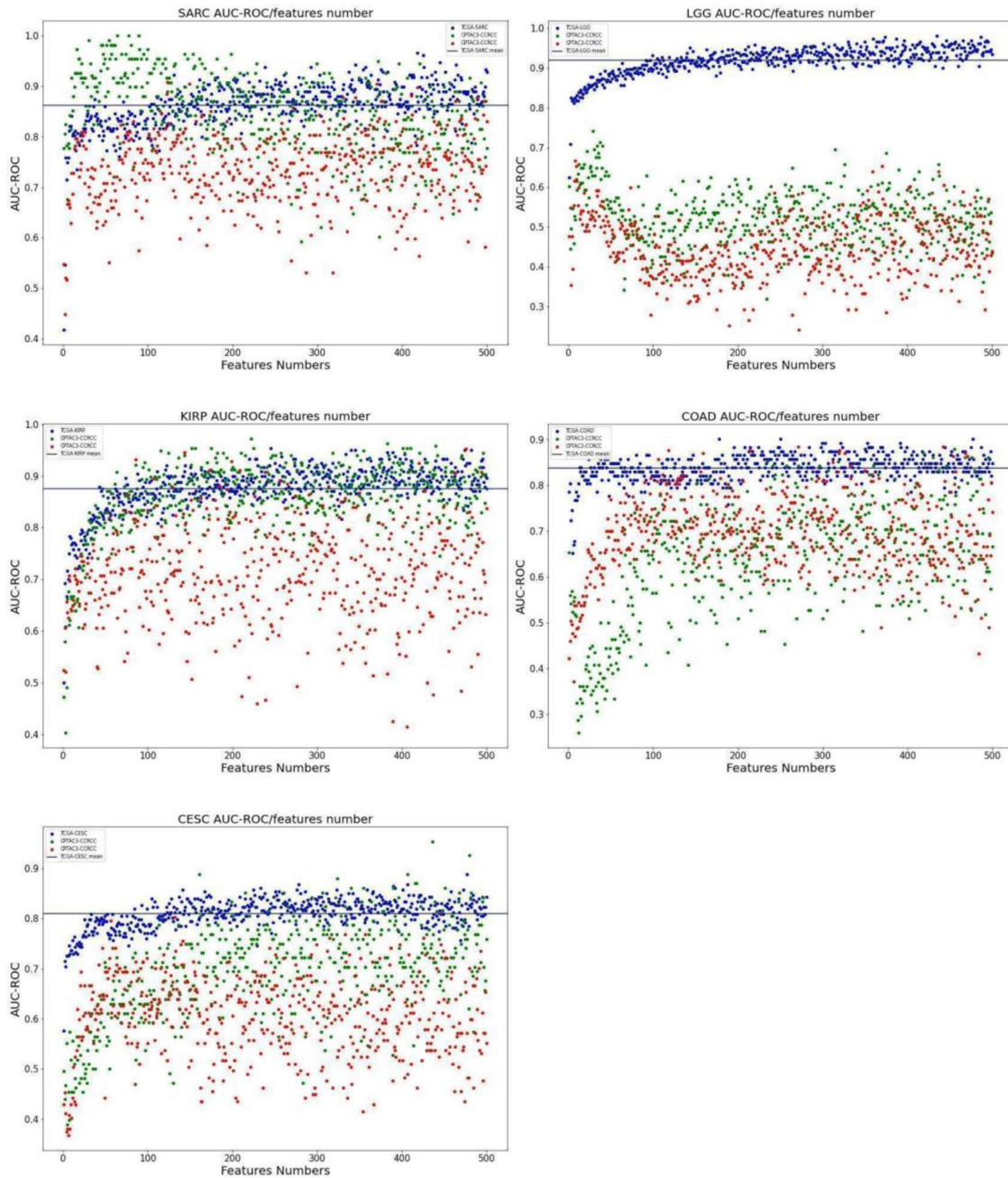
data was 0.86, very similar to the value 0.88 of the TCGA-KIRP test group. Interestingly, TCGA renal cell tumors have 2 sub-types: Kidney renal clear cell carcinoma (TCGA-KIRC) and Kidney renal papillary cell carcinoma (TCGA-KIRP). We analyzed these subtypes separately and the predictions were 0.79 and 0.88 AUC-ROC for the test groups, respectively. When we applied these 2 prediction models for validation cohort CPTAC3-ccRCC, which contains clear cell renal cell tumors, the predictions of both models were similar, 0.77 and 0.86 AUC-ROC, respectively. The prediction of the less efficient TCGA-UCEC model for the CPTAC3-UCEC data indeed gave a low predictions score, 0.63. Surprisingly the tissue discordant TCGA-KIRP prediction model for the analysis of uterine the CPTAC3-UCEC data set over performed the prediction of the tissue concordant TCGA-UCEC model and scored AUC-ROC 0.663. Finally, we have tested all the TCGA predictions models using renal CPTAC3-ccRCC dataset. For most of the models, the results were below 0.7, except for the TCGA-SARC model which was 0.85. When we used the uterine CPTAC3-UCEC dataset on all the TCGA prediction models, all the scores were very low except for the TCGA-SARC model which was 0.74.

*TCGA-KIRP model accurately predicted the prognosis of CPTAC3-ccRCC samples, but on a different scale.* We analyzed the RNA-SEQ data of CPTAC3-ccRCC samples using the selected TCGA-KIRP final model (created using 42 features as described in Table 2). The mean scores for the Deceased and Tumor-free groups were significantly different as shown in Figure 3, however, the scale by which each group was measured also differed. For the TCGA-KIRP samples, the model produced scores between 0.025 and 0.95 while the

CPTAC3-ccRCC sample scores were 0.18 to 0.425 (before normalization to 1). The model prediction AUC-ROC score for the CPTAC3-ccRCC was 0.86, almost identical to the TCGA-KIRP testing set.

*Pathway analysis revealed enrichment for cancer and cancer-related canonical pathways.* We used the Ingenuity Pathway Analysis (IPA) to analyze the genes selected in the final model for enrichment of related pathways. The top pathways were those involved in basic functions related to tumorigenesis and organ development. For example, in the TCGA-KIRP tumor prognosis prediction model, the pathways of Cell Cycle, Connective Tissue Development and Function, and Renal and Urological System Development and Function were the most highly enriched with a  $P$ -value of  $10^{-38}$ . In the TCGA-CESC, the highly represented pathways were Inflammatory Diseases as well as Inflammatory Response and Organismal Injury and Abnormalities. Due to the fact that cervical cancer is usually related to the effect of the human papillomavirus, it is not surprising that the genes associated with an inflammatory response may influence the prognosis of the cancer in those patients.<sup>14</sup> The TCGA-SARC model genes are enriched in Cell Cycle, Cell Death and Survival, and Cellular Development pathways which are relevant to cancer progression and prognosis. The LGG model genes are enriched in pathways involved in Organismal Injury and Abnormality, which are related to tumor microenvironment inflammation which was recently linked to these tumors.<sup>15</sup> A full list of the features of the models used, pathway nodes, scores, and pathways functions can be seen in Table 2.

As expected, functional pathways related to the tissues of origin such as Renal and Urological System Development and



**Figure 2.** AUC-ROC results as a function of the features numbers for 3 datasets: TCGA selected model data, CPTAC3-ccRCC, and CPTAC3-UCEC. The datasets were tested on each model and AUC-ROC score was calculated. The blue line represents the average AUC-ROC for all 500 results of the TCGA dataset.

Function, Reproductive System Disease, and Connective Tissue Disorders correlated to the primary tissue of the TCGA prediction models: kidney, cervix and glial cells, respectively.

*Comparison of gene enrichment of 2 prediction models for the same samples cohort.* We developed, based on the analysis of the same set of TCGA-KIRP samples, 2 separate models based on 300-feature model and 42-feature model. The predictions of the 2 models were 0.85 and 0.86, respectively. The 42 genes that comprised the smaller prediction model are included in the 300 genes model. Analysis of the 300 features TCGA-KIRP model by the Ingenuity Pathways Analysis software

matched 7 networks that are significantly enriched by those genes (see Supplemental Table 3 for specific molecules in each network and  $P$ -values). The most enriched network ( $P$ -value =  $10^{-45}$ ) that was selected by the IPA is composed of 35 components, out of them 26 are shared with the 300 features of the model. The network is related to the following functions: Cancer, Organismal Injury and Abnormalities, and Reproductive System Disease. In the 42 feature TCGA-KIRP model on the other hand the most enriched network ( $P$ -value =  $10^{-38}$ ) is also composed of 35 genes, out of them 15 genes are shared with the 42 features of the model. Cell Cycle, Connective Tissue Development and Function, and Renal and

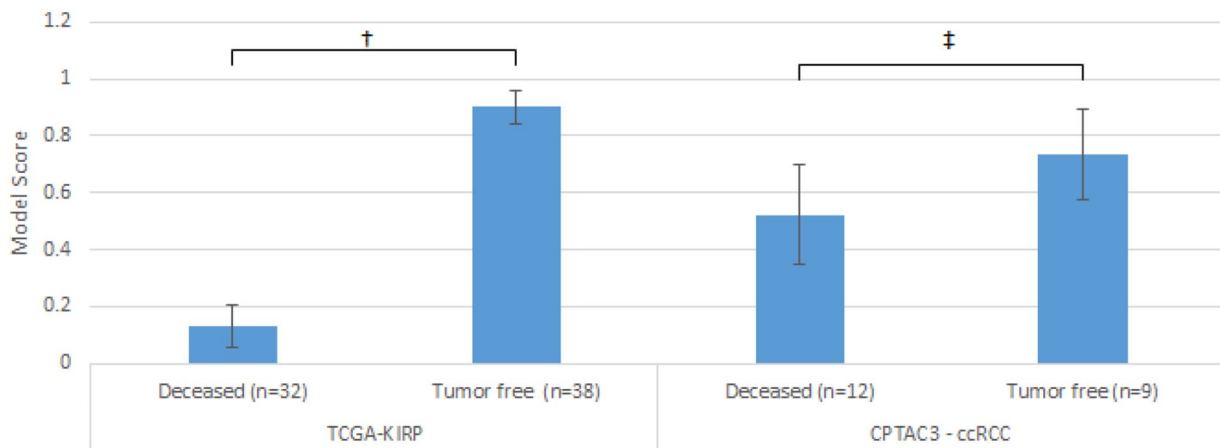
**Table 2.** The selected TCGA modes features were analyzed using Ingenuity Pathway Analysis (IPA) software for enriched networks. Only significant results ( $P$ -value  $< 10^{-20}$ ) are shown. The networks molecules and observed function are shown. For the TCGA-CESC features list, there were 2 matched networks with a different but related function.

PROJECT MODEL	MODELS FEATURES (ORDER BY SIGNIFICANCE)	PATHWAY NODES	PATHWAY SCORE	MOLECULES IN PATHWAYS	PATHWAY FUNCTION
TCGA-KIRP	DMBT1, RHEBL1, AC1448313, RP11_665C166, TOX2, IL11, GBPPI1, FAM83D, NLRP9P1, PLCB3, PROX1-AS1, RP3_337H46, HOXB6, FGD5, RP11_63N96, RP11_395L1417, LOC105375267, RP11_627K111, CUX1, ZSWIM1, RP11_214N155, TBL1XR1, MDS2, IL20RB, TPM2, RP11_134K12, LINC01108, RP11_181B111, AC0921714, KRT8P5, PLBD2, RP11_427P53, LZTS3, RP11_466P246, E2F8, TRIB3, LINC01358, GABBR2, RAP2C-AS1, NPAS1, PIM1, RP11_553N16,	Akt, ANGPT4, CCND1, CDKN1B, CSTB, CUX1, DMBT1, E2F8, ERK, ERK1/2, FAM83D, GABBR2, GBPPI1, HAPLN3, Histone h3, HOXB6, IL11, IL20RB, Lh, LRRTM2, MOSPD2, NFKB (complex), PIM1, PLCB3, SMARCA4, STAB2, STAT1, TBL1XR1, TBX21, TMEM204, TOX2, TPM2, TRIB3, USPL1, WNT8B	10 <sup>-38</sup>	15	[Cell cycle, connective tissue development and function, renal and urological system development and function]
TCGA-CESC	PIP4P2, LOC102724050, P4HA2, LINC01152, HENMT1, RP5_882O71, PLAAT2, RPS19P1, RBM38, PTMAP10, DNAJC9-AS1, DAAM2-AS1, mir-210, EREG, FNDC4, TMEM253, ANKRD37, ARHGFE25, ESM1, NPY1R, SLC10A3, EEF1E1, MMP8, RP11_447H194, RP11_45A174, SERPINH1, ITGA5, FOXC2, CTB, 193M123, BAIAP2L1, FUT11, VEGFA, ANXA5, ANKRD20A11P, ENPEP, RP11_89F32, FUNDCC2P1, RPSAP5, TMEM120B, UBAC1, COL4A2, LATS2, RP11_378A132, RP11_598F75, PTTG3P, EPN2, SLN, TMEM138, ABHD1, SEPSecs-AS1, SLC19A3, BCORL1, ZNF686, TXNP6, WASHC2A/WASHC2C, CHST14, MATN3, MPRIP, KCND2, RP11_107F64, MMP1, BCO1, SLC35A4, LINC00460, H3P36, APEX2, ANKRD34B, CTRB2, GCOM1, LRRN4, SPON1, RP11_455F54, MMP3, LOC105378645, PTPRB, ITGAD, ITGAV, SCN2A, GLUL, ANGPTL6,	ARHGFE25, BAIAP2L1, CPS1, E2F3, EBAG9, EEF1E1, ENPEP, EPN2, FUT1, FUT11, GLUL, HENMT1, HLA-J, HNF1A, HNM1, ITGAD, ITGB2, LATS2, mir-210, MLF2, MMP8, MPRIP, NINJ1, NR3C1, PLEKHF1, PPP1R14C, RHOA, SGO2, SERPINH1, SMARCA4, SON, SRC, TNF, TP53, YWHAG	10 <sup>-28</sup>	14	[Cardiovascular system development and function, organismal injury and abnormalities, reproductive system disease]
TCGA-SARC	NMU_B4GALT2, SHOC2, RNU2-22P, MON1B, HNRNPR, ARHGAP28, NDS2, ZFYVE28, ZNF146, SFTA2, RBM48, ARMCX3, ADCY1, QSER1, POR, BTFL3L4, LINC01121, ER18, RP11_680G244, CPG1, BLOC1S6, RAB5B, ELOVL2, DPP9-AS1, HPS6, ISCU, RPL29P19, TEN2, ARMH3, JRKL, NKX6-1, AC241585.1, FEM1A2P, NFKB2, ACOX1,	ABLIM, Akt, ANXA5, CG, COL4A2, EREG, ERK, ERK1/2, ESM1, estrogen receptor, FOXC2, FSH, GNRH, IRS, ITGA5, ITGAV, Jnk, Lh, MAP4K4, MMP1, MMP3, MT3, NFKB (complex), P38 MAPK, P4HA2, Pdgf (complex), PDGF BB, PDXK, POPS, PTPRB, RBM38, REXO5, SRD5A2, TLK1, VEGFA	10 <sup>-25</sup>	13	[Dermatological diseases and conditions, inflammatory disease, inflammatory response]
TCGA-SARC	NMU_B4GALT2, SHOC2, RNU2-22P, MON1B, HNRNPR, ARHGAP28, NDS2, ZFYVE28, ZNF146, SFTA2, RBM48, ARMCX3, ADCY1, QSER1, POR, BTFL3L4, LINC01121, ER18, RP11_680G244, CPG1, BLOC1S6, RAB5B, ELOVL2, DPP9-AS1, HPS6, ISCU, RPL29P19, TEN2, ARMH3, JRKL, NKX6-1, AC241585.1, FEM1A2P, NFKB2, ACOX1,	ACOX1, ARHGAP28, BLOC1S6, CCND1, CLEC11A, CPG1, DHRS2, ERK1/2, HNRNPR, HPS6, HSPA1L, HSPA2, ISCU, JKAMP, KHRBS1, LATS, miR-515-3p (and other miRNAs w/seed AGUGCCU), MLF2, MYB, NFKB (complex), NFKB2, NMU, POR, RAB22A, RPTOR, SH3GLB2, SHOC2, SLC9B2, STAB2, TFAP2A, TGM2, TIMM8A, TP53, TRIM6, ZNF146	10 <sup>-28</sup>	12	[Cell cycle, cell death and survival, cellular development]
TCGA-LGG	RP11_13N135, RP11_10C243, LOC101929494, SP7, RP11_54O71, MAN1B1-DT, ACTRIA, SPATA17, LOC105372974, AQP7, RP11_394B25, RP11_14C103, HOMER2, LINC01521, RP11_330H66, BEX4, SORT1, RP11_141O111, LOC100287042, BOK-AS1, MGC16275, CTC_498J121, STARD9, TMEM67, LOC101928982, SPINK5, HINT3, INPPL1, NUTM2A-AS1, KLF2, RPL39P36, RP1_196A121, RNU6-453P, AC1456762, LINC00624, ZNF655, LINC02198, RP3_337O189, CNTNAP4, RP11-517H2.6, MTND5P16, ACAD10, LOC728975, FKBP3, PPP2R2B, RP11_299G205, LOC400710, RP11_429P32, RP11_497D63, LINC01270, IGFBP3, DNPEP, RP1_125I34, RP11_312J185, TPST2, MYL2, TAGAP, MIDN, RP11_111F52, FAM171A1, LINC00671, CLDN6, RP11_53164, RNU6-1196P, RP11_299G202, ZNF514, SDCBP2P1, FAM204A, AP0002556, AC0040143, PRR26, MLLT6, MARCHF5, MAPK8IP1, ADAMTS12, RP11_631N164, SUDS3P1, FIRRE, RP11_9E171, DNAJA3, LOC101929592, DACT3-AS1, TNF, RP11_514P82, TNFRSF11B, IKB1	ADCYAP1, ADGRB1, BCAR3, caspase, CCL27, CLEC11A, DNAJA3, ERK, ERK1/2, FBXO31, FIRRE, FXN, HCG11, Histone h3, Hsp90, HTR4, IGFBP3, IL17RD, KLF2, MAP4K4, MAPK8IP1, NEU1, P2RY6, P38 MAPK, PPP2R2B, RBM17, SLC8A1, SND1-BRAF, SP7, STK10, SYK, SYNPO, TAGAP, TNF, TNFRSF11B	10 <sup>-20</sup>	10	[Connective tissue disorders, organismal injury and abnormalities, skeletal and muscular disorders]
TCGA-COAD	ZNF266, IZUMO1, ORM2, RPSAP4, DDX50P1, PABPC1P3, ZNF767P, RP5-1056L3.3, U4792429, RALGAPB, RP11_15B245, RP4_665J23,				No significant network

**Table 3.** Each data set was tested on top 5 models. The AUC-ROC score was calculated based on the predictions rate for each dataset.

		PREDICTION MODEL				
		CESC	COAD	KIRP	LGG	SARC
Samples data set	BLCA	0.569	0.357	0.456	0.574	0.494
	BRCA	0.447	0.469	0.459	0.440	0.511
	CESC	0.979	0.391	0.540	0.601	0.544
	COAD	0.540	0.951	0.467	0.625	0.566
	HNSC	0.571	0.384	0.482	0.478	0.434
	KIRC	0.561	0.625	0.671	0.555	0.646
	KIRP	0.539	0.502	0.981	0.508	0.616
	LGG	0.566	0.570	0.584	0.990	0.470
	LIHC	0.552	0.513	0.594	0.602	0.525
	LUAD	0.649	0.438	0.562	0.506	0.517
	LUSC	0.539	0.406	0.458	0.489	0.455
	READ	0.074	0.519	0.667	0.148	0.222
	SARC	0.532	0.437	0.586	0.465	0.988
	SKCM	0.507	0.429	0.475	0.440	0.591
	UCEC	0.549	0.553	0.571	0.460	0.582

Red indicates opposite prediction correlation and intensity ranges between 0 to 0.5. Green indicates direct prediction correlation and ranges from 0.5 to 1. White indicates that there is no correlation.

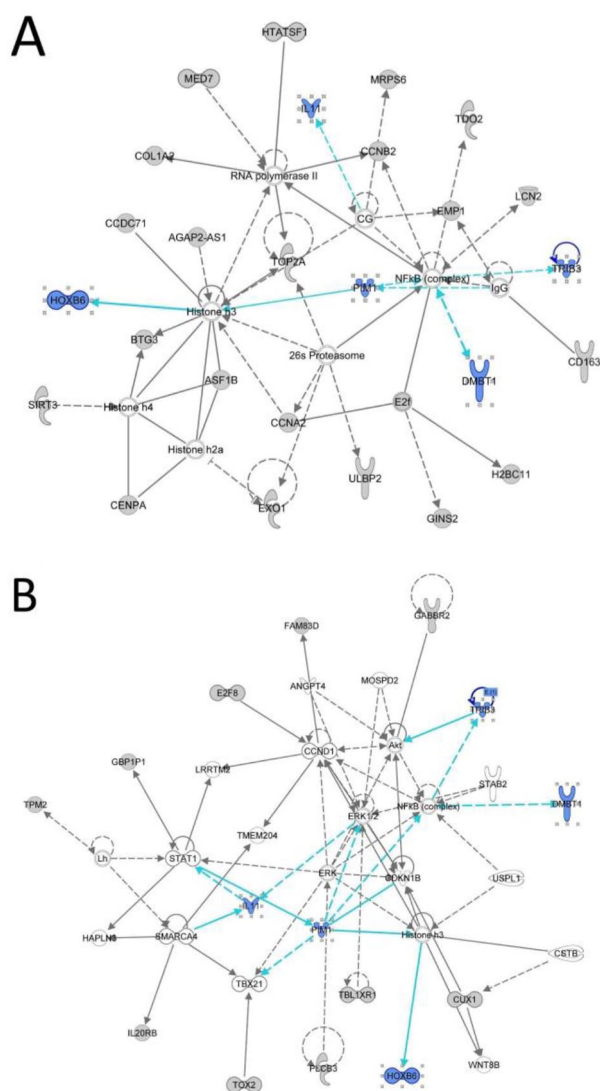


**Figure 3.** Mean score results from TCGA-KIRP model on TCGA-KIRP and CPTAC3-ccRCC groups, deceased, and tumor free. †P-value=2.23 × 10<sup>-54</sup>. ‡P-value=.01.

Urological System Development and Function pathways categories were significantly enriched in this network. As depicted in Figure 4, only 5 genes from the 42 feature TCGA-KIRP model were characterized as IPA network nodes in the larger network: DMBT1, IL11, HOXB6, TRIB3, PIM1. Those genes appear in both networks and might play a significant role in the disease renal cancer prognosis.

DMBT1<sup>16</sup> (Deleted In Malignant Brain Tumors 1) is a tumor suppressor gene. Deletions in this gene play a role in the progression of many human cancers, including brain, lung,

esophageal, gastric, and colorectal tumors. IL11, as part of KRT8-IL11 axis activation upregulation,<sup>17</sup> promotes tumor metastasis and is predictive of a poor prognosis in renal cell carcinoma. It was also suggested as a potential therapeutic target in cancer treatment.<sup>18</sup> HOXB6 (Homeobox B6) was found to play different roles in several cancer pathways,<sup>19,20</sup> including in methylation-driven genes related to prognosis in renal cell carcinoma.<sup>21</sup> TRIB3 (Tribbles pseudokinase 3) has many biological functions. However, high expression of TRIB3 was correlated with both advanced tumor stage and unfavorable



**Figure 4.** Shared features between top networks of TCGA-KIRP prediction models: (A) The top network from IPA prediction for the TCGA-KIRP 300 features model. That network is associated with Cancer, Organismal Injury and Abnormalities, Reproductive System Disease pathways. The gray nodes are the nodes from the model feature list (26 out of 35 network nodes,  $P$ -value =  $10^{-42}$ ). (B) The top network from IPA prediction for the TCGA-KIRP 42 features model. That network is associated with Cell Cycle, Connective Tissue Development and Function, Renal and Urological System Development and Function. The gray nodes are the nodes from the model feature list (15 out of 35 network nodes,  $P$ -value =  $10^{-38}$ ). The blue nodes are the shared genes between the networks that are also features in both models: DMBT1, IL11, HOXB6, TRIB3, PIM1.

prognosis.<sup>22</sup> High expression of TRIB3 in other cancer types, such as hepatocellular carcinoma and lung cancer, also correlated with poor survival rate.<sup>23,24</sup> PIM1 is a proto-oncogene belonging to the Ser/Thr protein kinase family. It was recently found that when overexpressed in human renal cell carcinoma tissues and cell lines, it positively correlated with disease progression.<sup>25</sup> PIM1 was found to be involved in Smad2, Smad3, and c-Myc<sup>26</sup> phosphorylation and was suggested as a potential therapeutic target for renal cell carcinoma patients.

*SurviveAI webapp.* An interactive free software based on the models was created using Flask 1.1.2. It enables physicians and researchers to get clinical predictions (for research purposes only) for RNA-Seq cancer multiple samples. The easy to use interface allows one to insert specific gene lists with FPKM-UQ values for each gene and to get predicted survival scores for 5 cancer types: Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), kidney renal papillary cell carcinoma (KIRP), brain lower grade glioma (LGG) and sarcoma (SARC). The tool uses scikit-learn's *predict\_proba(X)* method in the chosen TCGA-project model, and provides the probability for the given sample to match each group. Higher scores correlated with higher survival likelihood, while lower scores suggest poor survival prognosis. That being said, it is essential to calibrate the prediction scores scale for each cohort before using the tool as a predictor for specific samples, due high batch effect sensitivity of the models. In Figure 3 demonstrate the difference in the mean score value for each group. Although the CPTAC3-ccRCC difference between the Deceased and Tumor-free groups is significant ( $P$ -value = .01), it is less significant than the TCGA-KIRP group ( $P$ -value =  $2.23 \times 10^{-54}$ ). Calibration should be done with at least 10 to 20 samples to get maximum accuracy. The app automatically normalizes the scores to 1 by dividing them using the highest result.

SurviveAI webapp can be accessed at <https://tinyurl.com/surviveai>

## Discussion

Following the significant price decrease of high-throughput sequencing, projects like TCGA have generated vast amounts of data that enable machine learning. Usually, only specific types of cancer cohorts are used to create prediction models, combining multiple sources of OMICs-data to enhance AUC-ROC-based predictions. A multi-OMICs prediction model is more costly and less useful for routine clinical use, due to the increased number of methodologies needed. In order for a model to be user friendly and readily applicable, we based our model on RNA-seq data only, which is affordable and accessible, in clinical and research facilities. We have used 70% of the samples in each TCGA project to train the prediction models, and in order to validate the prediction, we tested them against the rest 30% of the samples from the cohort that was not used for training (test data). In addition, the models were tested against external datasets, CPTAC3-ccRCC and CPTAC3-UCEC. As expected, the models provided low prediction scores for the CPTAC3-UCEC samples, as none of the models were related to uterus cancer. Although KIRP (Kidney Renal Papillary Cell Carcinoma) and ccRCC (Clear cell renal cell carcinoma) are different subtypes of kidney cancer, the TCGA-KIRP model provides excellent predictions (AUC-ROC = 0.86) for the ccRCC dataset samples. Interestingly, the TCGA-SARC model also provides about the same accuracy



(AUC-ROC=0.85) for this dataset, even though the 2 models (KIRP and SARC) do not share any features at all (see Table 2)

We highly recommend that before using the models to calibrate with a truth set that contains at least 10 to 20 samples, as RNA expression level tends to be sensitive to batch effect.

For example, the CPTAC3-ccRCC Tumor-Free samples produce average score results of 0.73 for the Tumor-Free samples while the TCGA-KIRP survived results were between 0.9 and 1 (Table 3).

In this study, we show a novel method of machine learning driven pathways discovery using the simple and robust technique of reverse feature elimination. Also, the decision to use 2 distinct groups (Deceased and Tumor free), allowed us to decipher critical genes and features that are important for progression prediction in some of the projects.

We checked all possible projects available for analysis on the TCGA datasets and used only RNA-seq data for predictions. The reason for this is the relatively low cost and simplicity to produce such data for clinical and research purposes. This allows other researchers to use the models available free online. The Random Forest model is simple and allows us to easily extract the most important features from the data.

In 4 out of 5 models, a significant portion of the models' genes were part of cancer-related pathways. The molecules which were not included might be extensions of those networks or create another unknown network themselves. From a clinical perspective those genes might serve as new drug targets or biomarkers.

### Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>."

### Author Contributions

ON designed the project, collected and analysed the data; NK and EJ contributed to data analysis; NA contributed to the manuscript design and editing; GR supervised and mentored the project.

### ORCID iD

Omri Nayshool  <https://orcid.org/0000-0001-5060-891X>

### Supplemental material

Supplemental material for this article is available online.

### REFERENCES

1. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer*. 2019;144:1941-1953.
2. Kalia M. Biomarkers for personalized oncology: recent advances and future challenges. *Metabolism*. 2015;64:S16-S21.
3. Schwartz M, Parkl M, Phanl JH, Wang MD. Integration of multimodal RNA-seq data for prediction of kidney cancer survival. *Proceedings (IEEE Int Conf Bioinformatics Biomed)*. 2015;2015:1591-1595.
4. Wang J, Chen X, Tian Y, et al. Six-gene signature for predicting survival in patients with head and neck squamous cell carcinoma. *Aging*. 2020;12:767-783.
5. Huang Z, Johnson TS, Han Z, et al. Deep learning-based cancer survival prognosis from RNA-seq data: approaches and evaluations. *BMC Med Genomics*. 2020;13:41.
6. Ma B, Geng Y, Meng F, Yan G, Song F. Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method. *J Cancer*. 2020;11:1288-1298.
7. Gao W-Z, Guo L-M, Xu T-Q, Yin Y-H, Jia F. Identification of a multidimensional transcriptome signature for survival prediction of postoperative glioblastoma multiforme patients. *J Transl Med*. 2018;16:368.
8. Milanez-Almeida P, Martins AJ, Germain RN, Tsang JS. Cancer prognosis with shallow tumor RNA sequencing. *Nat Med*. 2020;26:188-192.
9. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
10. Herrmann M, Probst P, Hornung R, Jurinovic V, Boulesteix A-L. Large-scale benchmark study of survival prediction methods using multi-omics data. *Brief Bioinform*. 2021;22:bbaa167. doi:10.1093/bib/bbaa167
11. Valk PJM, Verhaak RGW, Beijin MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *New Engl J Med*. 2004;350:1617-1628.
12. Kela I, Ein-Dor L, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Breast Cancer Res*. 2005;7:4.38.
13. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46:389-422.
14. Adefuye A, Sales K. Regulation of inflammatory pathways in cancer and infectious disease of the cervix. *Scientifica*. 2012;2012. doi:0
15. Michelson N, Rincon-Torroella J, Quiñones-Hinojosa A, Greenfield JP. Exploring the role of inflammation in the malignant transformation of low-grade gliomas. *J Neuroimmunol*. 2016;297:132-140.
16. Mollenhauer J, Wiemann S, Scheurlen W, et al. DMBT1, a new member of the SRCR superfamily, on chromosome 10q25.3-26.1 is deleted in malignant brain tumours. *Nat Genet*. 1997;17:32-39.
17. Tan H-S, Jiang WH, He Y, et al. KRT8 upregulation promotes tumor metastasis and is predictive of a poor prognosis in clear cell renal cell carcinoma. *Oncotarget*. 2017;8:76189-76203.
18. Xu DH, Zhu Z, Wakefield MR, Xiao H, Bai Q, Fang Y. The role of IL-11 in immunity and cancer. *Cancer Lett*. 2016;373:156-163.
19. Vider BZ, Zimmer A, Chastre E, et al. Deregulated expression of homeobox-containing genes, HOXB6, B8, C8, C9, and Cdx-1, in human colon cancer cell lines. *Biochem Biophys Res Commun*. 2000;272:513-518.
20. Li Y, Jiang A. ST8SIA6-AS1 promotes hepatocellular carcinoma by absorbing miR-5195-3p to regulate HOXB6. *Cancer Biol Ther*. 2020;21:647-655.
21. Wang J, Zhang Q, Zhu Q, et al. Identification of methylation-driven genes related to prognosis in clear-cell renal cell carcinoma. *J Cell Physiol*. 2020;235:1296-1308.
22. Hong B, Zhou J, Ma K, et al. TRIB3 promotes the proliferation and invasion of renal cell carcinoma cells via activating MAPK signaling pathway. *Int J Biol Sci*. 2019;15:587-597.
23. Wang X-J, Li F-F, Zhang Y-J, Jiang M, Ren W-H. TRIB3 promotes hepatocellular carcinoma growth and predicts poor prognosis. *Cancer Biomark*. 2020;29:307-315.
24. Zhang X, Zhong N, Li X, Chen MB. TRIB3 promotes lung cancer progression by activating  $\beta$ -catenin signaling. *Eur J Pharmacol*. 2019;863:172697.
25. Zhao B, Liu L, Mao J, Zhang Z, Wang Q, Li Q. PIM1 mediates epithelial-mesenchymal transition by targeting Smads and c-Myc in the nucleus and potentiates clear-cell renal-cell carcinoma oncogenesis. *Cell Death Dis*. 2018;9:307.
26. Wang J, Kim J, Roh M, et al. Pim1 kinase synergizes with c-MYC to induce advanced prostate carcinoma. *Oncogene*. 2010;29:2477-2487.