Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Development and validation of 10-gene signature for predicting survival in patients with prostate cancer

Nan Li
Liuzhou Maternity and Child Health Care Hospital
Kai Yu
GuangXi University
Ling Zhong
Liuzhou Maternity and Child Health Care Hospital
Dingyuan Zeng (zdy13607723199@163.com)
Liuzhou Maternal and Children Healthcare Hospital https://orcid.org/0000-0002-4271-5632

Research

Keywords: Prostate cancer, 10-gene signature, TCGA, Bioinformatics, Risk score

Posted Date: March 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17658/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background. The prognosis for prostate cancer patients remains poor. High-throughput sequencing data provide a solid basis for identifying genes associated with cancer prognosis, but genetic markers are needed to predict the clinical outcome of prostate cancer.

Methods. The Cancer Genome Atlas (TCGA) database (N = 551) was adopted to estimate the prognostic value of immune genes. RNA-seq and clinical follow-up data were downloaded from TCGA. The samples were randomly divided into training and test. Cox regression analyses and least absolute shrinkage and selection operator (LASSO) were conducted to develop an immune risk score. Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and single sample Gene Set Enrichment Analysis (ssGSEA) were used for functional Analysis. Tumor Immune Estimation Resource (TIMER) is used to analyze the immune score, and RMS curve and clinical decision curve analysis is used to analyze the superiority of the comparison with published models.

Results. Survival analyses revealed that 19 genes significantly associated with the overall survival (OS). 10-genes signature was ultimately obtained through random forest feature selection. Riskscore effectively stratified samples in the training, test, and external verification sets and all TCGA sets. The 5-year survival AUC in the training, verification sets and all TCGA sets were around 0.7. Univariate and multivariate analysis showed that 10-genes signature has good predictive performance in clinical. TIMER analysis shows that immunosuppression may reduce the chances of survival for patients with prostate cancer. Compared with published models, our model has a higher C-index.

Conclusion. We constructed a 10-gene signature as a new prognostic marker for predicting survival of prostate cancer patients.

Background

Prostate cancer is the biggest threat to men's health in the world after lung cancer. According to statistics, 164,690 new cases of prostate cancer were diagnosed in 2018, and 29,430 patients died of prostate cancer, making prostate cancer the third leading cause of cancer-related deaths among men in the United States [1]. In China, the incidence of prostate cancer is also increasing year by year, and it is expected to grow faster in the next few years [2]. Globally, there were about 1.3 million new cases of prostate cancer in 2018 and 359,000 deaths from the disease, making it the second leading cause of cancer among men and the fifth leading cause of cancer deaths among men [3].

In recent years, more and more studies have reported methods to predict and stratify survival and prognosis of prostate cancer patients based on gene expression [4–6]. Unfortunately, such studies have not yet been translated into routine clinical practice, which may be due to a small sample size, high data fit or lack of evidence. Currently open and available large-scale databases containing gene expression data, such as The Cancer Genome Atlas (TCGA) and ImmPort, make it possible to mine more reliable biomarkers for prostate cancer to predict and categorize patient outcomes [7, 8]. Every part of the

immune system has been shown to participate in, accelerate and even identify different stages of cancer development and progression. In addition, immune escape has been shown to be a new marker for cancer. Recently, the immunotherapy drug PD-1/PD-L1, based on specific immune checkpoints, has shown surprising results in treating cancer patients [9]. However, in the prostate cancer microenvironment, the molecular events of the interaction between tumor cells and immune cells need to be further explored and summarized, which will ultimately determine their potential to predict the prognosis of prostate cancer patients.

In this study, the TCGA and ImmPort databases were analyzed and patient clinical characteristics were considered to develop and validate prognostic models for prostate cancer based on immune-related genes. This could eventually be used to help clinicians assess the prognosis and treatment options of prostate cancer patients.

Materials And Methods

Data acquisition

TCGA GDC API was used to download the latest clinical follow-up information, which was downloaded on December 12, 2018. A total of 551 RNA-seq data samples were included, including 499 tumor tissue data and 52 normal tissue samples. The immune-related gene set was downloaded from the ImmPort database. After removing the genes with duplicate names, there were 1,811 genes in total. The workflow is shown in Fig. 1.

Data preprocessing

The following steps were used to preprocess the RNA-seq data of 499 samples:

1. samples without clinical data were removed
2. normal tissue sample data were removed
3. the gene with Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) 0 in half of the samples were removed
4. Since the samples whose event was dead were much smaller than those who were alive (10:484), in order to avoid the abnormal model construction, the samples with recurrence prognosis were combined with dead samples as the same status, with a total of 74 and the remaining 420 samples as the status of another group.
5. only the expression profiles of immune-related genes were retained, and a total of 1353 genes were used for subsequent modeling analysis.

Sample grouping

First, 494 samples were divided into training and test sets. To avoid random allocation bias affecting the stability of subsequent modeling, 1000 samples of all samples were put back into random groups in advance, and the most suitable training and test sets were selected according to the following conditions:

1. The two groups are similar in terms of age distribution, clinical stage, follow-up time, and proportion of patient deaths.
2. The number of samples in the two classifications is similar after clustering the gene expression profiles of the two data sets randomly grouped.

The final training set has a total of 246 samples and the test set has a total of 248 samples.

The final training and test set sample information is shown in Table 1.

Table 1
Sample information of training and test set

Features	Overall Set	Training Set	Testing Set	Pvalue
OS	494	246	248	0.433515
T	494	244	247	0.661289
T1	177	92	85	
T2	202	93	109	
T3	109	57	52	
T4	3	2	1	
TX	3	2	1	
N	494	208	213	0.700088
N0	343	167	176	
N1	78	41	37	
NX	73	38	35	
Age	494	246	248	0.493976
0 ~ 50	27	10	17	
50 ~ 60	173	90	83	
60 ~ 70	242	124	118	
70 ~ 100	52	22	30	
RadiationTherapy	494	139	146	0.850781
NO	245	119	126	
YES	40	20	20	

Univariate survival risk analysis

Unilateral Cox proportional risk regression model was developed for each immune-related gene and survival data using training set data. Survival coxph function was used, $p < 0.05$ was selected as the threshold for filtering.

Construction of prognostic immune gene signatures

Least absolute shrinkage and selection operator (LASSO) is a popular method for regression modeling with a large number of potential prognostic features, because it can perform automatic feature selection in a manner that results in signatures with generally good prognostic performance [10]. The LASSO

method has been extended to the Cox model for survival analysis and has been successfully applied for the purpose of building sparse signatures for survival prognosis in many application areas including oncology[11–13], We first use the training set samples to perform univariate Cox proportional hazards regression analysis on each gene, with log rank $p < 0.05$ as a threshold to identify genes with significant prognosis, and Then, the R software package glmnet [14] was used to screen the genes with robust prognostic characteristics. The multivariate Cox regression analysis was further conducted by using the stepwise regression method, and the model was constructed by using 10-fold cross validation. Stepwise regression USES the AIC red pool information criterion, which takes into account the statistical fitting degree of the model and the number of parameters used for fitting. StepAIC method in the MASS package starts from the most complex model and deletes a variable successively to reduce AIC. The smaller the value, the better the model, indicating that the model has obtained sufficient fitting degree with fewer parameters. Risk scoring model:

$$RiskScore = \sum_{k=1}^{n} Exp_k * e^{HR}{}_k$$

where N is the number of prognostic genes, Exp_k is the genes value of prognostic genes, and $e^{HR}{}_k$ is the estimated regression coefficient of genes in the multivariate Cox regression analysis.

Functional enrichment analyze

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed using the R package clusterprofiler[15] for genes, to identify over-represented GO terms in three categories (biological processes, molecular function and cellular component), and KEGG pathway. For this analysis, a FDR < .05 was considered to denote statistical significance.

To estimate KEGG functional enrichment scores, we used a single-sample gene-set enrichment analysis (ssGSEA), which defines enrichment scores to represent the absolute enrichment of gene sets in each sample in a given dataset. Normalized enrichment scores can be calculated for each immune category. SsGSEA analysis was performed in R packaging GSVA.

Differences in tumor infiltrating immune cells

Six types of tumor-infiltrating immune cells were retrieved from the tumor immune estimation resource (TIMER) (https://cistrome.shinyapps.io/timer/), and the immune scores of the two groups were calculated for comparison.

Comparison with other models

Four prognostic risk models: 4-gene signature (Wang) [16], 4-gene signature (Komisarof) [17], 24-gene signature (Long) [18] and 22-gene signature (Erho) [19] were selected for comparison with our 10 genes

model. Four models were applied to TCGA data set to analyze Kaplan-Meier (KM) survival curve, Receiver Operating Characteristic (ROC) curve, Restricted Mean Survival (RMS) curve and clinical decision curve analysis.

Statistical analysis

The KM curve was plotted when the median risk score in each dataset was used as a cutoff to compare the survival risk between the high-risk and low-risk groups. Multivariate Cox regression analysis was performed to test whether immunity related genes (IRGs) markers were independent prognostic factors. Significance was defined as $P < 0.05$, both were bilateral tests. ROC analysis was performed using the R package pROC, c-index calculation was performed using the R package RMS, and all of these analyses, if not specified, were performed using default parameters in R 3.4.3.

Results

Identification of immune-related genes with differential prognosis in patients with prostate cancer and construction of prognostic gene signatures for 10 genes

In the TCGA training set data, we used a univariate Cox proportional hazard regression model to establish the relationship between overall patient survival and immune-related gene expression, found that there were 78 genes with differences. In order to screen for robust immune-related prognostic signature genes, we used the R package glmnet to perform lasso cox regression on these 78 genes for dimensionality reduction analysis. The trajectory of each independent variable shows that as the lambda gradually increases, the number of independent variable coefficients tending to increase gradually. (Fig. 2A). The results of the confidence interval analysis show that the model is optimal when lambda = 0.039. We choose the model when lambda = 0.039 as the final model, which contains 19 genes (Fig. 2B). Furthermore, the 19 genes obtained in the previous step were reduced to 10 genes by multivariate cox survival analysis using AIC red pool information criterion (Table 2). The prognostic KM curve results of 10 genes show that 5 genes (ie, RXRB, KAL1, EED SORT1, and GPI) could significantly reduce the risk of TCGA training set samples in two groups ($p < 0.05$) (Figure S2). The resulting 10-gene signature formula is as follows:

Table 2
Basic information of this 10 Gene.

Gene Symbol	Coef	Pvalue	HR	Low.95.CI.	High.95.CI.
KCNH2	0.2018	0.0001	1.2236	1.1072	1.3522
GREM1	0.1629	0.0001	1.1770	1.0833	1.2787
PSMD14	0.4970	0.0011	1.6438	1.2187	2.2171
RXRB	0.2324	0.0055	1.2616	1.0707	1.4866
BCL10	-0.3032	0.0336	0.7385	0.5584	0.9767
KAL1	-0.3677	0.0641	0.6924	0.4691	1.0218
EED	0.4985	0.0680	1.6462	0.9637	2.8120
SORT1	-0.1302	0.0766	0.8779	0.7602	1.0140
LTB4R	0.2973	0.1104	1.3462	0.9345	1.9391
GPI	0.0282	0.1187	1.0286	0.9928	1.0656

$RiskScore_{10}$ = KCNH2*0.2018 + GREM1*0.1629 + PSMD14*0.4970 + RXRB*0.232 + BCL10*-0.30312 + KAL1*-0.3676 + EED*0.4985 + SORT1*-0.1302 + LTB4R*0.2972 + GPI*0.0282.

Based on the Risk model, the Risk Score of each sample in the TCGA training set was calculated. The training set samples were divided into the high-risk group (risk-H) and the low-risk group (risk-L) as the median Risk Score was taken as the threshold. As patients' risk scores increased, the number of dead samples increased. Heat map analysis showed that high expression of BCL10, KAL1 and SORT1 were correlated with low risk, which were protective factor. High expression of KCNH2, GREM1, PSMD14, RXRB, EED, LTB4R and GPI were correlated with high risk, which were risk factors associated. (Fig. 2C). ROC curve analysis suggests that the Area Under Curve (AUC) at one, three, and five years are greater than 0.70 (Fig. 2D). A risk model based on 10 genes is used to predict the KM survival curve of the risk-H and the risk-L on the training set. There is a significant difference between the risk-H and the risk-L (Fig. 3E).

Robustness of the 10 gene model

In order to verify the stability and reliability of the model, we applied 10 gene signatures to the validation set and TCGA all data, calculated the risk score of each sample. The median risk score as the threshold, the validation set data and all TCGA data set samples were divided into the risk-H and the risk-L. As patients' risk scores increased, the number of deaths increased, and the high-risk group had more deaths than the low-risk group. Heat map analysis showed that high expression of BCL10, KAL1 and SORT1 are correlated with low risk, which are protective factor. High expression of KCNH2, GREM1, PSMD14, RXRB, EED, LTB4R and GPI are risk factors associated with high risk (Fig. 3A, D). The prediction effect of the

model at 1,3 and 5 years was evaluated in the validation set data and all TCGA data sets, and the AUC was greater than 0.70 (Fig. 3B, E). The risk model constructed based on 10 genes was used to predict the k-m survival curve of the high-risk group and the low-risk group on the validation set data and all TCGA data sets, and there was a significant difference between the high-risk group and the low-risk group (Fig. 3C, F).

Analysis between risk models and clinical characteristics

The comprehensive analysis of Riskscore calculated by 10-gene signature and clinical information revealed that The 10-mRNA signature could distinguish T2, T1 + T3 + T4, N0, N1, M0, patients who have not received radiotherapy, young group, and old group significantly different from high-risk group ($p < 0.05$) (Figure S2). This results indicated that our model still has good predictive power in different clinical signs. To identify the independence of the 10-gene signature model in clinical applications, Univariate and multivariate COX regression analysis were used in the clinical information carried throughout the TCGA data. Univariate COX regression analysis found that only T-stage and Riskscore were significantly related to survival (Fig. 4A), However, the corresponding multivariate COX regression analysis only found that the T stage and Risk score (HR = 2.691, 95% CI = 1.200-6.039, p = 0.016) were significantly related to survival (Fig. 4B). The analysis results of the nomogram model show that the RiskScore has the greatest impact on the survival prediction, indicating that the 10 gene-based risk model can better predict the prognosis (Fig. 4C). In addition, we counted 1-, 3-, and 5-year data used to visualize nomogram performance (Fig. 4D).

Riskscore potentially relevant regulatory pathways

In order to observe the relationship between risk score and biological function of different samples, the R software package GSVA carried out ssGSEA analysis and functions with a correlation greater than 0.4 was selected. The results showed that most of the samples were negatively correlated with Riskscore, and a few were positively correlated with Riskscore (Fig. 5A). Furthermore, GSEA was used in the data set of the training set to analyze the significantly enriched pathways in the high-risk group and the low-risk group. The threshold value was $p < 0.05$, and the significantly enriched pathways were obtained (Fig. 5B), among them, PROPANOATE_METABOLISM, PPAR_SIGNALING_PATHWAY were significantly enriched to the high-risk group, which were closely related to the development of cancer.

Analysis of risk models and immune scores

Z score normalization was used for Risk score of each sample in the training set, the samples were divided into high risk group and low risk group. The TIMER tool was used to calculate the immune score of each sample in the training set, and it is found that other than the CD8 Tcell, the other immune scores show significant differences in the high and low risk groups ($p < 0.05$) (Fig. 6).

Comparison with other models

As described in the method, four prognostic risk models were selected: 4-gene signature (Wang), 4-gene signature (Komisarof), 24-gene signature (Long), and 22-gene signature (Erho). In order to make the models comparable, the risk score of each prostate adenocarcinoma (PRAD) sample in the TCGA using the same method based on the corresponding genes in the 4 models were calculated. The OS-KM curve shows that the OS prognosis of the Risk-H and Risk-L in the four models are significantly different except for the Wang model (logrank $p < 0.05$) (Fig. 6A-D). ROC curve reveals that the prediction effect of four models is worse than that of 10 gene signature (Fig. 6E-H). In order to compare the prediction performance of these models on PRAD samples, the RMS curve was drawn using the R language RMS package. The 10 gene signature is more accurate for long-term follow-up (Fig. 6C). Decision Curve Analysis (DCA) results showed that Riskscore had the highest degree of benefit, far higher than the other four subtypes (Fig. 6D).

Discussion

In terms of prognosis, Prostate cancer is a highly heterogeneous disease in that survival times vary substantially among patients with similar tumor node metastasis (TNM) stages. With the diagnosis and treatment of Prostate cancer at earlier stages, traditional clinicopathological indicators such as tumor size, vascular invasion, portal vein thrombus and TNM stage have proven inadequate for predicting individual outcomes, especially risk stratification, as no one-size-fits-all treatment strategy appears to be effective. Consequently, screening prognostic molecular markers that adequately reflect the biological characteristics of tumors would be of great significance for individualized prevention and treatment of Prostate cancer patients. In the present study, we analyzed the expression profiles of 499 Prostate cancer samples from TCGA and identified 10 genes robustly associated with OS. This signature is independent of other clinical factors.

Gene signatures are currently being used in clinical practice. Two examples are Oncotype DX [20, 21], which provides a breast cancer recurrence score based on expression of 21 genes, and Coloprint, which provides a colon cancer recurrence score based on expression of 18 genes [22, 23]. Results obtained with these assays have shown that screening new prognostic cancer markers based on gene expression profiles is a promising high-throughput molecular identification method. In that regard, Shao N et al. [24] developed a seven lncRNAs signature that could predict PCa Rapid biochemical recurrence, but the AUC was only about 0.68, and Abou-Ouf H et al. [25] build a 10-gene model, based on high-dimensional discriminant analysis, but the AUC was only about 0.65. In addition, Lee JY et al.[26] used clustering score (CS) and predictive score (PS) to identify 29 PCa genes (called PCa29) as early biomarkers from two data sets in Gene Expression Omnibus (GEO). Although PCa29 can distinguish between normal and tumor tissues and is specific for prostate cancer, the large number of genes that need to be detected makes this analysis clinically impractical. By contrast, our 6-gene signature has a high AUC using only 6 genes, which makes it conducive to clinical application.

The 10 genes in our signature include KCNH2, GREM1, PSMD14, RXRB, BCL10, KAL1, EED, SORT1, LTB4R and GPI. It has been reported that in the context of androgen receptor-positive prostate cancer, EED

regulate androgen receptor expression levels and androgen receptor downstream targets [27]. The expression of SORT1 was altered in malignant patient tissue, when compared to indolent and normal prostate tissue [28]. Although others have not been previously reported to be related to Prostate cancer, many genes have been reported in relation to other cancers. Ours is the first study to suggest that they can be used as new prognostic markers of Prostate cancer. At the same time, our GSEA results show that the 10-gene signature enrichment significantly correlates with pathways and biological processes associated with the occurrence and development of cancer. This indicates that our model has potential clinical application value and could provide a potential target for diagnosis and for development of new targeted therapies.

Although we have identified potential candidate genes affecting tumor prognosis using bioinformatics technology with large samples, our study has limitations. First, the sample lacks some clinical follow-up information, so we did not consider factors such as the presence of other health conditions to differentiate prognostic biomarkers. Second, the results obtained using bioinformatics analysis alone are insufficient and need to be confirmed through experimental verification. Therefore, further genetic and experimental studies with larger samples and experimental validation are needed.

Conclusions

In this study, univariate and multivariate analysis showed that 10-genes signature has good predictive performance in clinical. TIMER analysis showed that immunosuppression may reduce the chances of survival for patients with prostate cancer. Compared with published models, our model has a higher C-index. In conclusion, our integrated analyses revealed that 19 genes significantly associated with the overall survival (OS) and constructed a 10-gene signature as a new prognostic marker for predicting survival of prostate cancer patients.

Abbreviations

AUC: Area Under Curve; CS: Clustering score; DAC: Decision Curve Analysis; FPKM: Fragments Per Kilobase of transcript per Million fragments mapped; GEO: Gene Expression Omnibus; GO: Gene Ontology; IRGs: Immunity Related Genes; KEGG: Kyoto Encyclopedia of Genes and Genomes; KM: Kaplan-Meier; KMS: Restricted Mean Survival; LASSO: Least absolute shrinkage and selection operator; OS: overall survival; OS: Overall survival; PRAD: prostate adenocarcinoma; PS: predictive score; risk-H: high-risk group; risk-L: low-risk group; ROC: Receiver Operating Characteristic; ssGSEA:single sample Gene Set Enrichment Analysis; TCGA: The Cancer Genome Atlas; TIMER: Tumor Immune Estimation Resource; TNM：Tumor node metastasis

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated during and/or analyses during the current study are available in the TCGA-PRAD database.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by Guangxi Natural Science Foundation under Grant No. (No.2017GXNSFBA198199, No. 2017GXNSFBA198193), the Key Research and Development Program of Guangxi (No. GuikeAB18126056) and Liuzhou Science and Technology Bureau Project (No.2016G020217).

Authors' contributions

NL and KY acquired, analyzed, interpreted the data and drafted the manuscript. LZ and DYZ conceived the research and revised the paper for important content. All authors read and approved the final manuscript.

Acknowledgements

None.

References

1. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2018.** *CA Cancer J Clin* 2018, **68:**7-30.
2. Liu S, Yang L, Yuan Y, Li H, Tian J, Lu S, Wang N, Ji J: **Cancer incidence in Beijing, 2014.** *Chin J Cancer Res* 2018, **30:**13-20.
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: **Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.** *CA Cancer J Clin* 2018, **68:**394-424.
4. Pang X, Xie R, Zhang Z, Liu Q, Wu S, Cui Y: **Identification of SPP1 as an Extracellular Matrix Signature for Metastatic Castration-Resistant Prostate Cancer.** *Front Oncol* 2019, **9:**924.

5. Bakht MK, Lovnicki JM, Tubman J, Stringer KF, Chiaramonte J, Reynolds MR, Derecichei I, Ferraiuolo RM, Fifield BA, Lubanska D, et al: **Differential expression of glucose transporters and hexokinases in prostate cancer with a neuroendocrine gene signature: a mechanistic perspective for FDG imaging of PSMA-suppressed tumors.** *J Nucl Med* 2019.
6. Shi R, Bao X, Weischenfeldt J, Schaefer C, Rogowski P, Schmidt-Hegemann NS, Unger K, Lauber K, Wang X, Buchner A, et al: **A Novel Gene Signature-Based Model Predicts Biochemical Recurrence-Free Survival in Prostate Cancer Patients after Radical Prostatectomy.** *Cancers (Basel)* 2019, **12**.
7. Fraser M, Rouette A: **Prostate Cancer Genomic Subtypes.** *Adv Exp Med Biol* 2019, **1210:**87-110.
8. Sun J, Li S, Wang F, Fan C, Wang J: **Identification of key pathways and genes in PTEN mutation prostate cancer by bioinformatics analysis.** *BMC Med Genet* 2019, **20:**191.
9. Schott DS, Pizon M, Pachmann U, Pachmann K: **Sensitive detection of PD-L1 expression on circulating epithelial tumor cells (CETCs) could be a potential biomarker to select patients for treatment with PD-1/PD-L1 inhibitors in early and metastatic solid tumors.** *Oncotarget* 2017, **8:**72755-72772.
10. Kostareli E, Hielscher T, Zucknick M, Baboci L, Wichmann G, Holzinger D, Mucke O, Pawlita M, Del Mistro A, Boscolo-Rizzo P, et al: **Gene promoter methylation signature predicts survival of head and neck squamous cell carcinoma patients.** *Epigenetics* 2016, **11:**61-73.
11. Zhang JX, Song W, Chen ZH, Wei JH, Liao YJ, Lei J, Hu M, Chen GZ, Liao B, Lu J, et al: **Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis.** *Lancet Oncol* 2013, **14:**1295-1306.
12. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, Yoon CJ, Ellis P, Wedge DC, Pellagatti A, et al: **Clinical and biological implications of driver mutations in myelodysplastic syndromes.** *Blood* 2013, **122:**3616-3627; quiz 3699.
13. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR, Diao L, et al: **Assessing the clinical utility of cancer genomic and proteomic data across tumor types.** *Nat Biotechnol* 2014, **32:**644-652.
14. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *J Stat Softw* 2010, **33:**1-22.
15. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *OMICS* 2012, **16:**284-287.
16. Wang L, Gong Y, Chippada-Venkata U, Heck MM, Retz M, Nawroth R, Galsky M, Tsao CK, Schadt E, de Bono J, et al: **A robust blood gene expression-based prognostic model for castration-resistant prostate cancer.** *BMC Med* 2015, **13:**201.
17. Komisarof J, McCall M, Newman L, Bshara W, Mohler JL, Morrison C, Land H: **A four gene signature predictive of recurrent prostate cancer.** *Oncotarget* 2017, **8:**3430-3440.
18. Long Q, Xu J, Osunkoya AO, Sannigrahi S, Johnson BA, Zhou W, Gillespie T, Park JY, Nam RK, Sugar L, et al: **Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence.** *Cancer Res* 2014, **74:**3228-3237.

19. Erho N, Crisan A, Vergara IA, Mitra AP, Ghadessi M, Buerki C, Bergstralh EJ, Kollmeyer T, Fink S, Haddad Z, et al: **Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy.** *PLoS One* 2013, **8:**e66855.
20. Schildgen V, Warm M, Brockmann M, Schildgen O: **Oncotype DX Breast Cancer recurrence score resists inter-assay reproducibility with RT(2)-Profiler Multiplex RT-PCR.** *Sci Rep* 2019, **9:**20266.
21. Wang F, Reid S, Zheng W, Pal T, Meszoely I, Mayer IA, Bailey CE, Park BH, Shu XO: **Sex Disparity Observed for Oncotype DX Breast Recurrence Score in Predicting Mortality Among Patients with Early Stage ER-Positive Breast Cancer.** *Clin Cancer Res* 2020, **26:**101-109.
22. Maak M, Simon I, Nitsche U, Roepman P, Snel M, Glas AM, Schuster T, Keller G, Zeestraten E, Goossens I, et al: **Independent validation of a prognostic genomic signature (ColoPrint) for patients with stage II colon cancer.** *Ann Surg* 2013, **257:**1053-1058.
23. Tan IB, Tan P: **Genetics: an 18-gene signature (ColoPrint(R)) for colon cancer prognosis.** *Nat Rev Clin Oncol* 2011, **8:**131-133.
24. Shao N, Zhu Y, Wan FN, Ye DW: **Identification of seven long noncoding RNAs signature for prediction of biochemical recurrence in prostate cancer.** *Asian J Androl* 2019, **21:**618-622.
25. Abou-Ouf H, Alshalalfa M, Takhar M, Erho N, Donnelly B, Davicioni E, Karnes RJ, Bismar TA: **Validation of a 10-gene molecular signature for predicting biochemical recurrence and clinical metastasis in localized prostate cancer.** *J Cancer Res Clin Oncol* 2018, **144:**883-891.
26. Lee JY, Lin SY, Lin CY, Chuang YH, Huang SH, Tseng YY, Wang HJ, Yang JM: **Identification of the PCA29 gene signature as a predictor in prostate cancer.** *J Bioinform Comput Biol* 2019, **17:**1940006.
27. Liu Q, Wang G, Li Q, Jiang W, Kim JS, Wang R, Zhu S, Wang X, Yan L, Yi Y, et al: **Polycomb group proteins EZH2 and EED directly regulate androgen receptor in advanced prostate cancer.** *Int J Cancer* 2019, **145:**415-426.
28. Johnson IR, Parkinson-Lawrence EJ, Keegan H, Spillane CD, Barry-O'Crowley J, Watson WR, Selemidis S, Butler LM, O'Leary JJ, Brooks DA: **Endosomal gene expression: a new indicator for prostate cancer patient prognosis?** *Oncotarget* 2015, **6:**37919-37929.

Additional Files

Figure S1

KM survival curve of 10 gene.

Figure S2

KM survival curve of the 10-mRNA signature based on multiple clinical information distinguished high-risk group and low-risk group. A-C: T1 group, T2 group, T3 + T4 group, respectively. D-E: N0 group and N1 group, respectively. F: M0 group. G: patients who have not received radiotherapy, H-I: young group and old group, respectively.

Figures

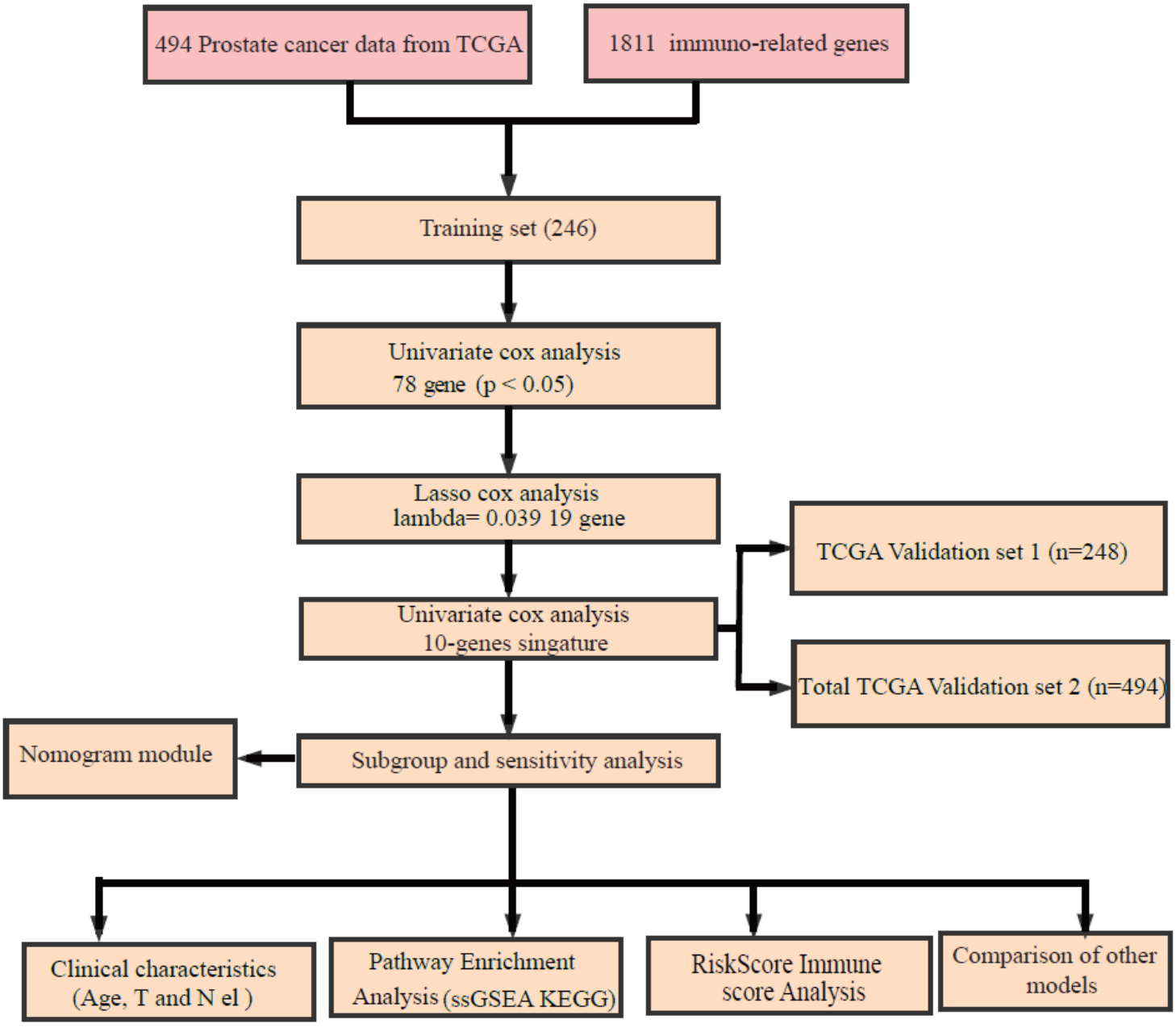


Figure 1

Workflow.

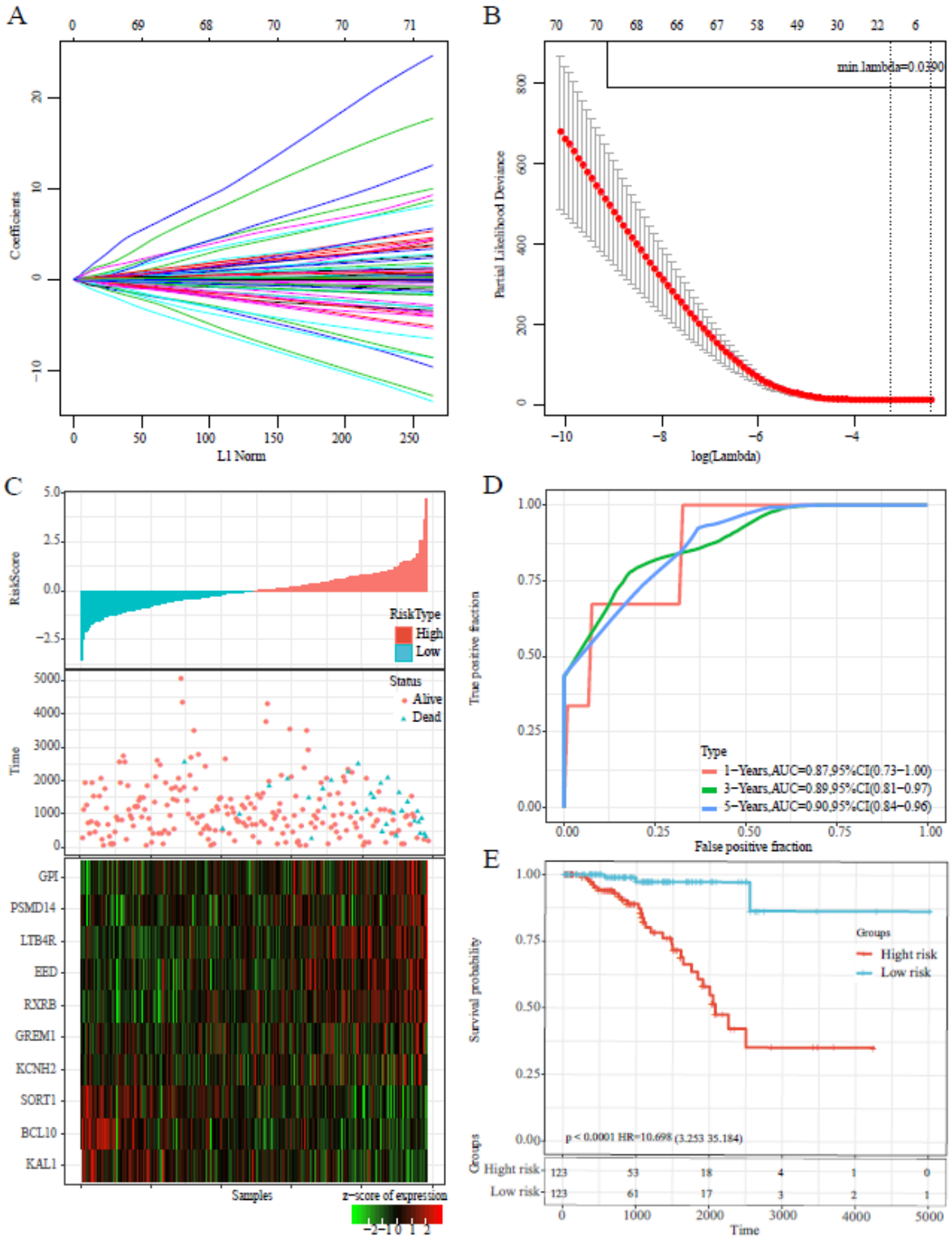


Figure 3

Construction of prognostic gene signatures for 10 genes. A: The trajectory of each independent variable, the horizontal axis represents the log value of the independent variable lambda, and the vertical axis represents the coefficient of the independent variable. B: Confidence interval under each lambda. C: risk score, survival time and survival status, and 10-gene expression in TCGA training set. D: ROC curve and

AUC of 10-gene signature classification. E: KM survival curve distribution of 10-gene signature in the training set.

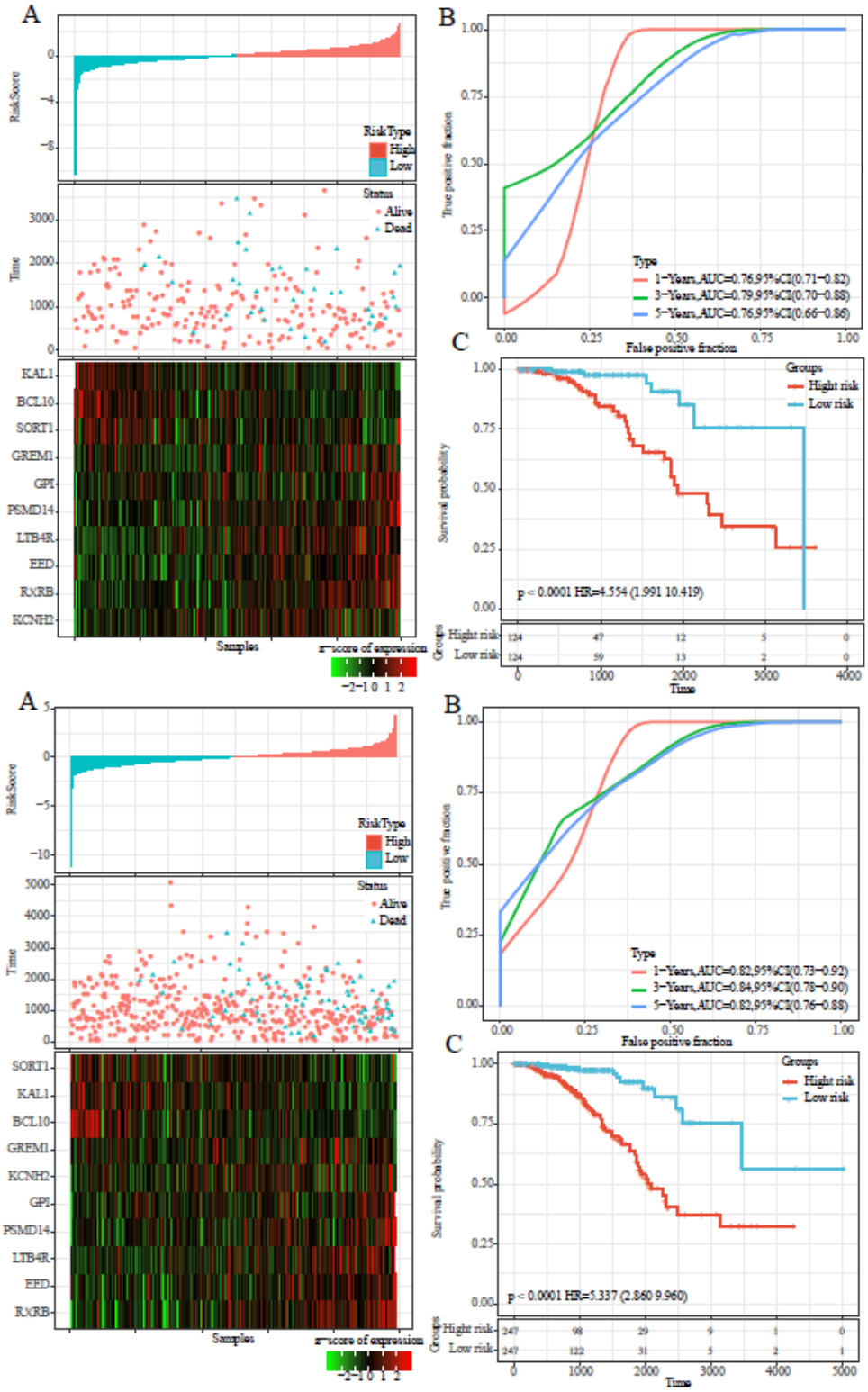


Figure 5

The robustness of 10 gene signatures was verified. A: risk score, survival time and status, and 10-gene expression in the validation set data. B: ROC curve and AUC for 10 gene signature classification in the validation set data. C: KM survival curve distribution of 10-gene signature in the verification set. D: 10- KM

survival curve distribution of gene signature in the TCGA all data set. E: risk score, survival time and status, and 10-gene expression in the validation set data. E: ROC curve and AUC for 10-gene signature classification in the validation set data. D: KM survival curve distribution of 10-gene signature in TCGA all data set.

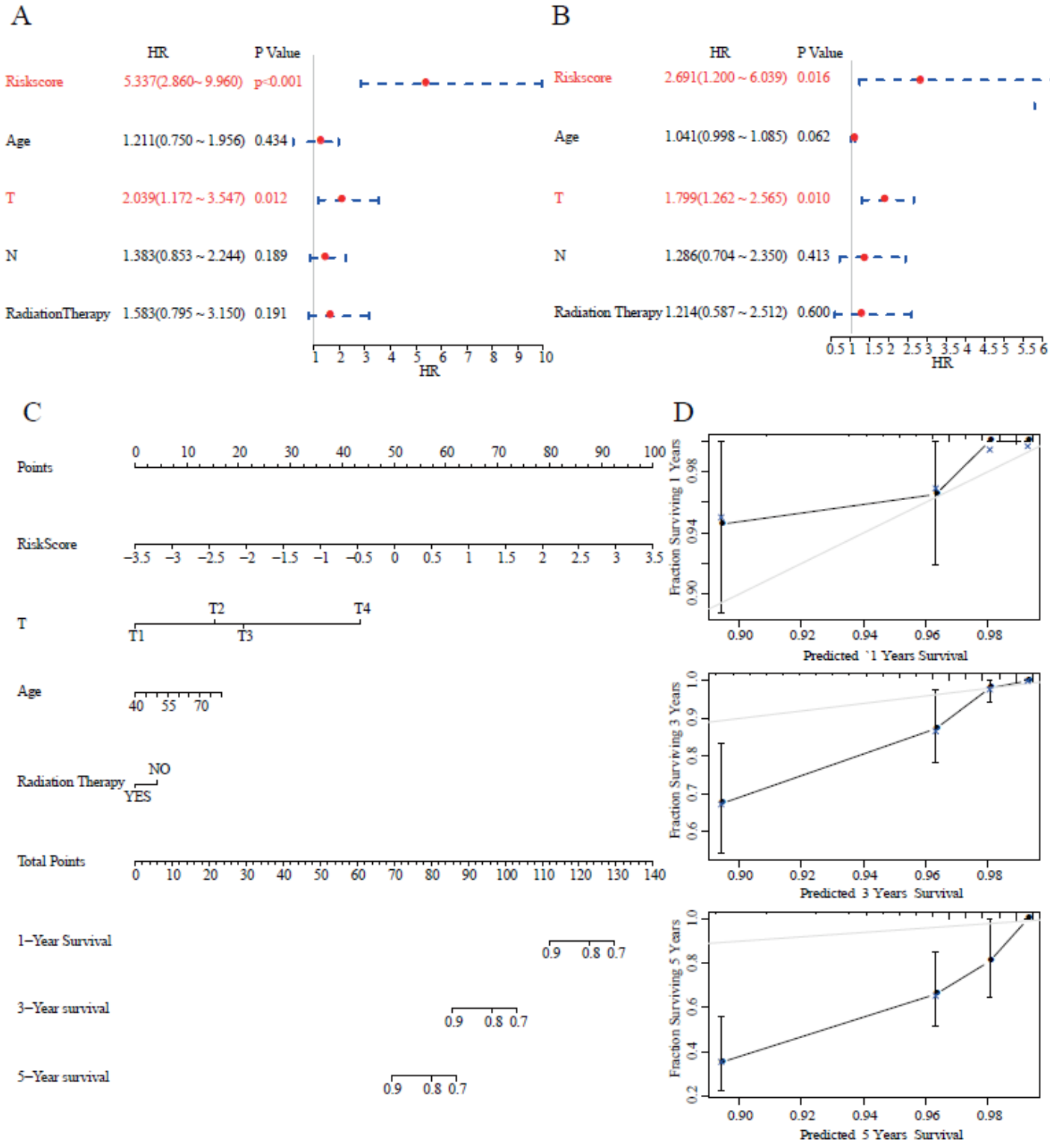


Figure 7

Analysis between risk models and clinical characteristics. A: Forest plot of clinical factors and Riskscore for univariate survival analysis. B: Forest plot of clinical factors and Riskscore for multivariate survival analysis. Among them, red represents significant OS correlation. C: Nomogram of risk score compared with clinical factors. D: A statistical chart of 1 -, 3 -, and 5-year survival in nomograms.

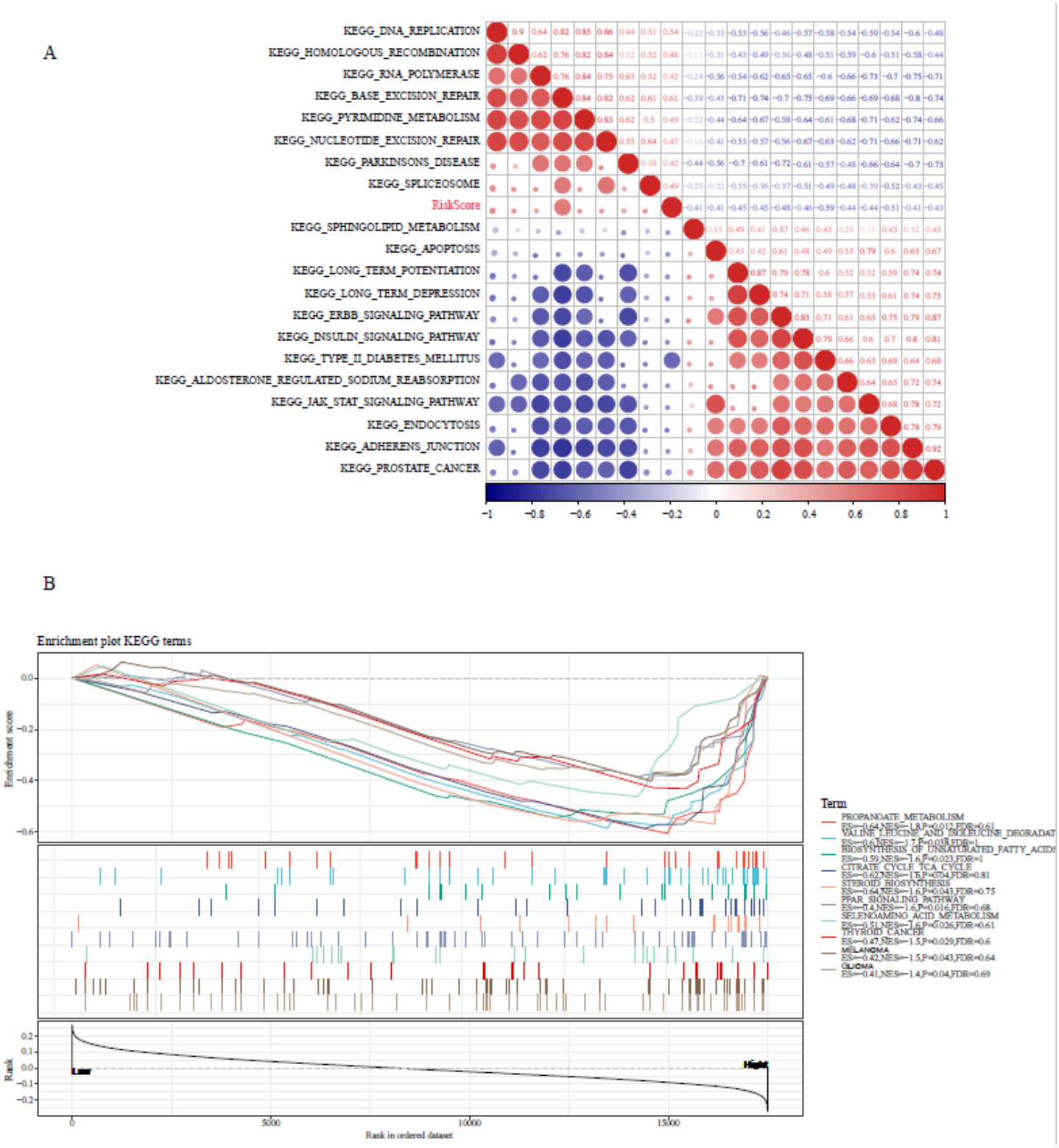


Figure 9

Riskscore potentially relevant regulatory pathways. A: KEGG pathways correlated with risk score greater than 0.4, and the clustering of correlation coefficients between RiskScore and KEGG. B: Pathways enriched in the high-risk group and the low-risk group of the training set by 10-gene signature, in which different colors represent different pathways.

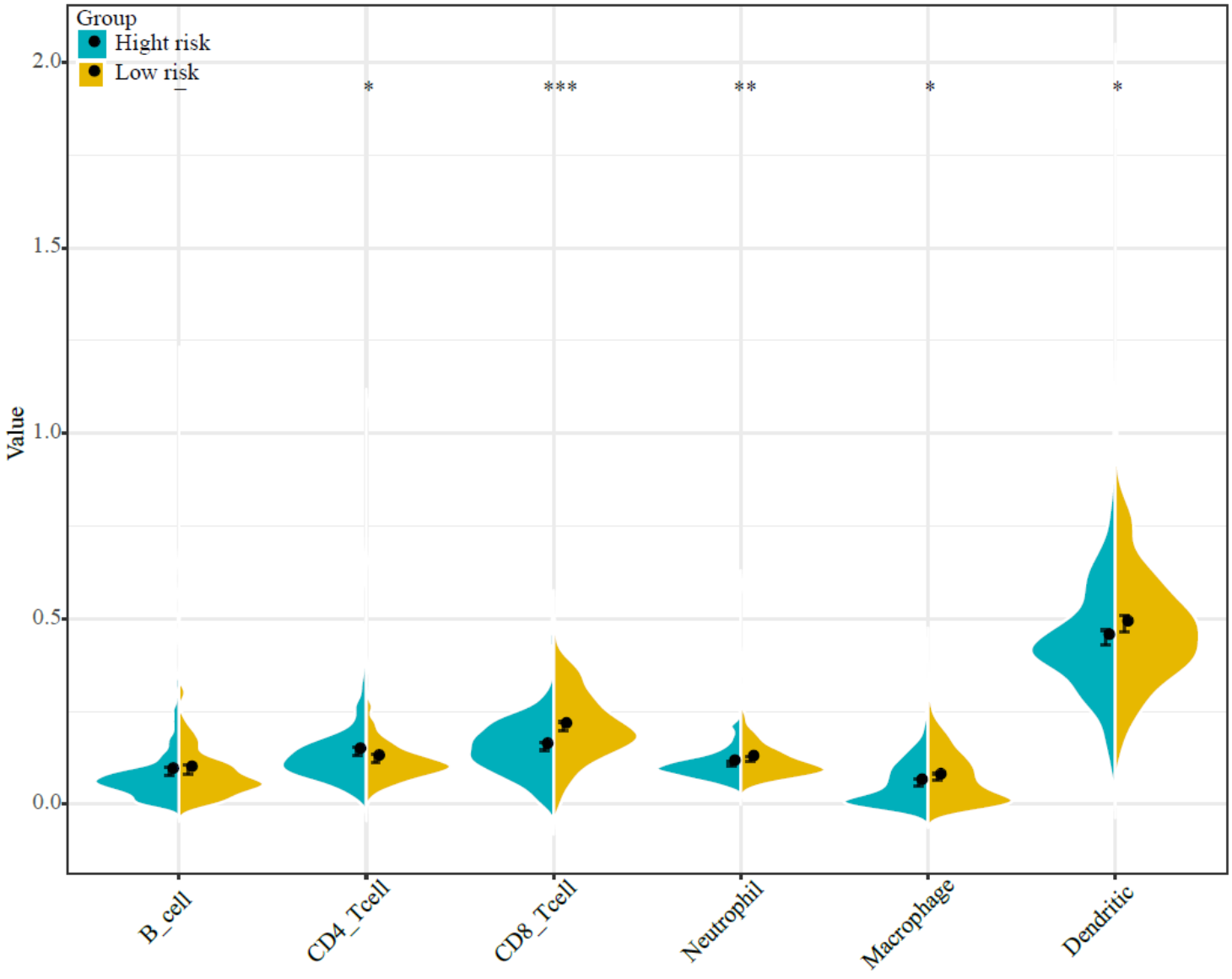


Figure 11

Immune score was detected by Tumor immune estimation resource in high and low risk groups.

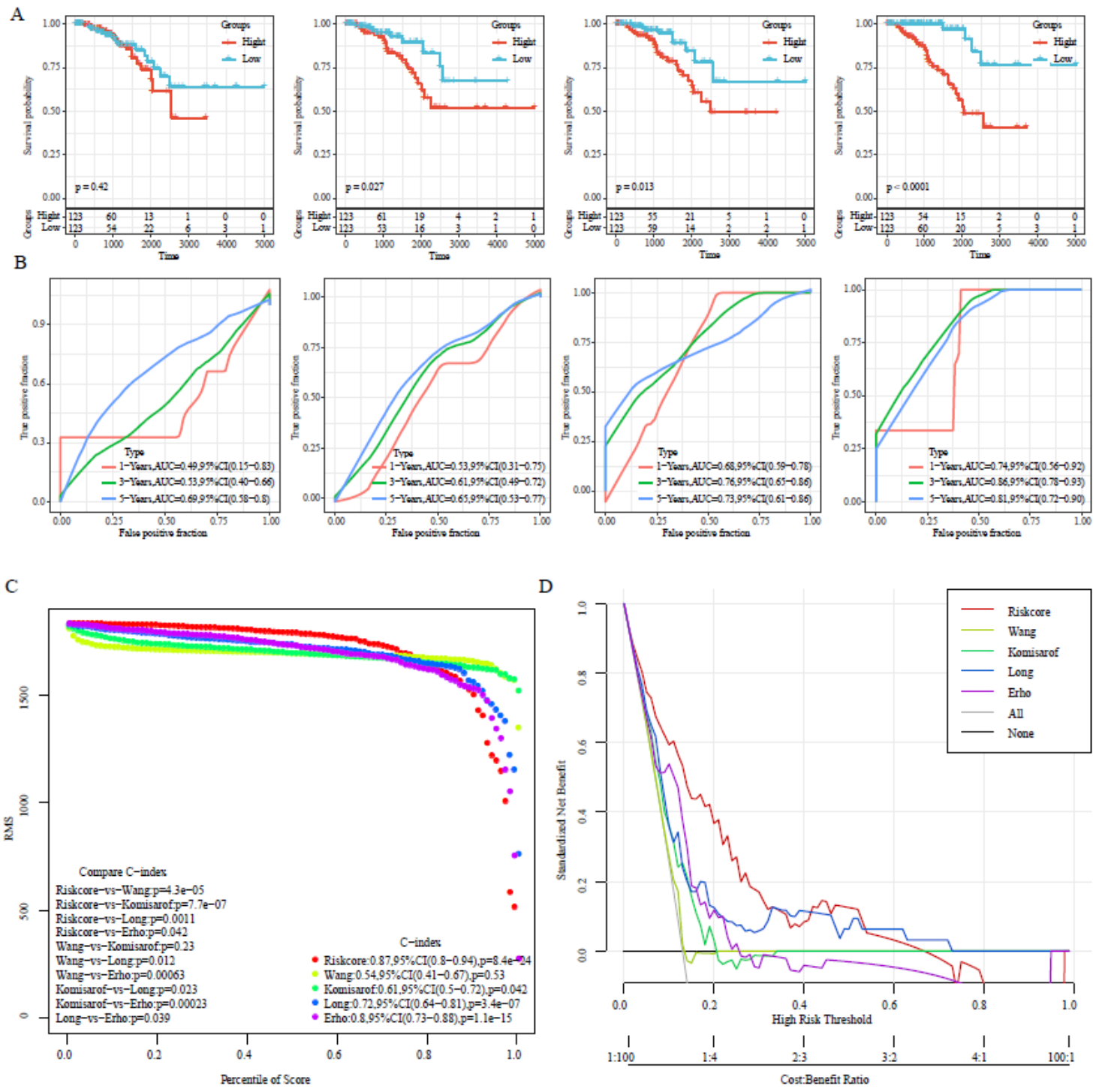


Figure 13

Riskscore versus published models. A-D: KM curves of Wang model, Komisarof model, Long model and Erho model, respectively. E-F: AUC curves of Wang model, Komisarof model, Long model and Erho model, respectively. C: RMS curves of five models. D: DCA curves of five models.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS2.pdf
- FigureS1.pdf
- FigureS1.pdf
- FigureS2.pdf