www.impactjournals.com/oncotarget/ Oncotarget, 2017, Vol. 8, (No. 37), pp: 61583-61591

Research Paper

Risk score based on three mRNA expression predicts the survival of bladder cancer

Qingzuo Liu[1,*], Ruigang Diao[1,*], Guoyan Feng[1], Xiaodong Mu[1] and Aiqun Li[2]

[1]Yantai Yuhuangding Hospital, Zhifu District, Yantai 264000, China

[2]Yantai Affiliated Hospital of Binzhou Medical University, Muping District, Yantai 264003, China

[*]These authors have contributed equally to this work

Correspondence to: *Xiaodong Mu,* ***email:*** *muxiaodong2017@163.com*
Aiqun Li, ***email:*** *liaiqun2017@163.com*

Keywords: *bladder cancer, prognosis, expression*

Received: *February 27, 2017* ***Accepted:*** *May 23, 2017* ***Published:*** *June 27, 2017*

ABSTRACT

Bladder cancer (BLCA) is one of the most malignant cancers worldwide, and its prognosis varies. 1214 BLCA samples in five different datasets and 2 platforms were enrolled in this study. By utilizing the gene expression in The Cancer Genome Atlas (TCGA) dataset, and another two datasets, in GSE13507 and GSE31684, we constructed a risk score staging system with Cox multivariate regression to evaluate predict the outcome of BLCA patients. Three genes consist of RCOR1, ST3GAL5, and COL10A1 were used to predict the survival of BLCA patients. The patients with low risk score have a better survival rate than those with high risk score, significantly. The survival profiles of another two datasets (GSE13507 and GSE31684), which were used for candidate gene selection, were similar as the training dataset (TCGA). Furthermore, survival prediction effect of risk score staging system in another 2 independent datasets, GSE40875 and E-TABM-4321, were also validated. Compared with other clinical observations, and the risk score performs better in evaluating the survival of BLCA patients. Moreover, the correlation between radiation were also evaluated, and we found that patients have a poor survival in high risk group, regardless of radiation. Gene Set Enrichment Analysis was also implemented to find the difference between high-risk and low-risk groups on biological pathways, and focal adhesion and JAK signaling pathway were significantly enriched. In summary, we developed a risk staging model for BLCA patients with three gene expression. The model is independent from and performs better than other clinical information.

INTRODUCTION

Bladder cancer (BLCA) is one of the most malignant diseases worldwide, with 73,510 new cases and 14,880 deaths [1] in the United States, 2012. According to most recent statistic report in China, there were 80,500 new BLCA cases and 32,900 deaths occurred due to BLCA [2]. Although the improvement of the therapy methods and drugs, a large proportion of patients died within 3 years after diagnosis, which, makes the prognosis of BLCA important [3]. However, as most frequently used prognostic indicators, clinical observations often fail to predict the survival of bladder cancer patients. Thus, the molecular biomarkers are now urgently needed.

Based on previous studies, the performance of single biomarkers in predicting the survival of BLCA patients across datasets are unstable, while combination of biomarkers enhances the performance [4]. In this vein, we implemented Cox multivariate regression model on gene expression of BLCA samples in TCGA dataset. The patients with high risk score had a significantly shorter survival time than those with low risk score, and this finding was further validated in other two cohorts used for candidate gene selection (GSE13507 and GSE31684) and another two totally independent datasets (GSE40875

and E-TABM-4321). Furthermore, according to cox multivariate hazard analyses, the risk score performs better than the other clinical information in prognosis of BLCA patients. The risk score is also effective in estimating the survival of patients whether they underwent radiation or not. Gene Set Enrichment Analysis (GSEA) showed that focal adhesion pathway was significantly altered between high and low risk group, suggesting that the risk score reflects the cell adhesion status of BLCA.

RESULTS

Identification of survival-related genes

With univariate Cox regression model, genes were used to evaluate the correlation between gene expression and overall survival in three independent datasets (TCGA-BLCA, GSE13507 and GSE31684). In order to improve the robustness of the candidate gens, mRNA levels significantly correlated with overall survival in all these three datasets ($p<0.05$) were selected for further analysis, and three genes, RCOR1, ST3GAL5 and COL10A1, were identified. Multivariate cox regression analyses were performed and the risk scores were calculated as the following:

Risk score= 0.14738*RCOR1 + (-0.17272) *ST3GAL5 + 0.18195*COL10A1.

It is noticed that the coefficient of ST3GAL5 is negative, indicating that the expression of this gene is positively related the survival time/rate of BLCA patients while the expression of RCOR1 and COL10A1 are negatively related. Detailed correlation information between overall survival information and the three gene expression was listed in Supplementary Table 1.

Performance of risk score in training dataset

To measure the performance of risk score in predicting the outcome of BLCA patients, the survival of patients with high/low risk score were compared using the median risk score value as cutoff. The overall survival (OS) of patients with high risk score is significantly longer than those with low risk score (Figure 1A, $p=0.00054$). The median survival time of high risk group was 24 months and the median survival time of low risk group was 67.3 months. Re-sampling was also implemented by randomly retrieving 80% of all samples, and the possibility that risk score not significant associated with overall survival ($p>0.05$) was 0.0083 (Supplementary Figure 1). In addition, recurrence free survival (RFS) difference was also calculated between the high and low risk groups, and the result is consistent with the OS profile (Figure 1B, $p<0.001$). As shown in Figure 1C, patients in high risk score were characterized as early relapse, low expression of ST3GAL5, and high expression of RCOR1 and COL10A1. The Receiving operating characteristic curve (ROC) of three-year survival was also plotted according to age, gender, and risk score (Figure 1D), and the area under curve (AUC) was 0.608, 0.5002, and 0.647, respectively, indicating that the risk score performs better in predicting the survival of BLCA patients than other clinical information.

Validation of performance of risk score in test datasets

To evaluate the robustness of our model, after locking the coefficients of each gene, the risk scores in another two independent datasets (GSE13507 and GSE31684) were evaluated. In consistent with the survival profile of the training dataset, the overall survival rate of high risk group is significantly lower than the low risk group in both datasets ($p=0.00064$ and 0.0014 for GSE13507 and GSE31684, respectively, Figure 2A). Since these two datasets were also used in candidate gene selection and over-fit may have brought, we used another two totally independent datasets (GSE48075 and E-TABM-4321) for further validation. Consistent with the observation in training datasets, early recurrence rate in E-TABM-4321 high risk samples was significantly higher than low risk samples. Similar trend was also observed in GSE48075 of overall survival rate (Figure 2B). In addition, the overall expression profiles of candidate genes used for risk score evaluation were also similar, compared to the training datasets (Figures 2C-2D). These results suggest that the risk score was robust in predicting the survival of BLCA patients.

Relationship between risk score, other clinical information, and radiation

In order to compare clinical significance of clinical observations and risk score, multivariate Cox hazard analysis was performed to evaluate the importance of these indicators. As shown in Figure 3A, the most important hazard factor for BLCA is risk score, while gender and histologic grade was not statistically significant. The correlation analyses between risk score and clinicopathological indicators showed that the risk score is significantly associated with age, primary tumor stage and BMI (body mass index), while independent from gender (Figure 3B).

Radiation is one of the most common therapy methods for BLCA. To evaluate whether risk score is also suitable for patients underwent radiation therapy, we artificially divided the patients underwent radiation therapy into high risk group and low risk group using median risk score as cutoff, as usual. The overall survival rate of patients underwent radiation therapy (Figure 3C) with high risk score had a significantly shorter survival rate than these with low risk score. And this trend was also repeated in the patients without radiation therapy (Figure

3D). To facilitate the utilization of risk score, a nomogram was plotted (Figure 3E). All these results above suggest that the prognostic performance of risk score is effective for both patients with and without radiation therapy.

Altered pathways in the high risk score patients

The significantly altered signaling pathways between high risk group and the low risk group were assessed with Gene Set Enrichment Analysis (GSEA) to investigate why the risk score predicts the survival of BLCA patients. The significantly altered pathways in high risk group include "vascular SNARE transport", "ECM receptor interaction", "JAK-STAT signaling pathway", and "focal adhesion" (Figure 4A, Supplementary Table 2). Among these KEGG pathways, "JAK-STAT signaling pathway" (Figure 4B) and "focal adhesion" (Figure 4C) associated genes were noted, suggesting that our risk score reflected the alteration of JAK-STAT signaling pathway and focal adhesion status, and thus predicting the prognosis of BLCA patients.

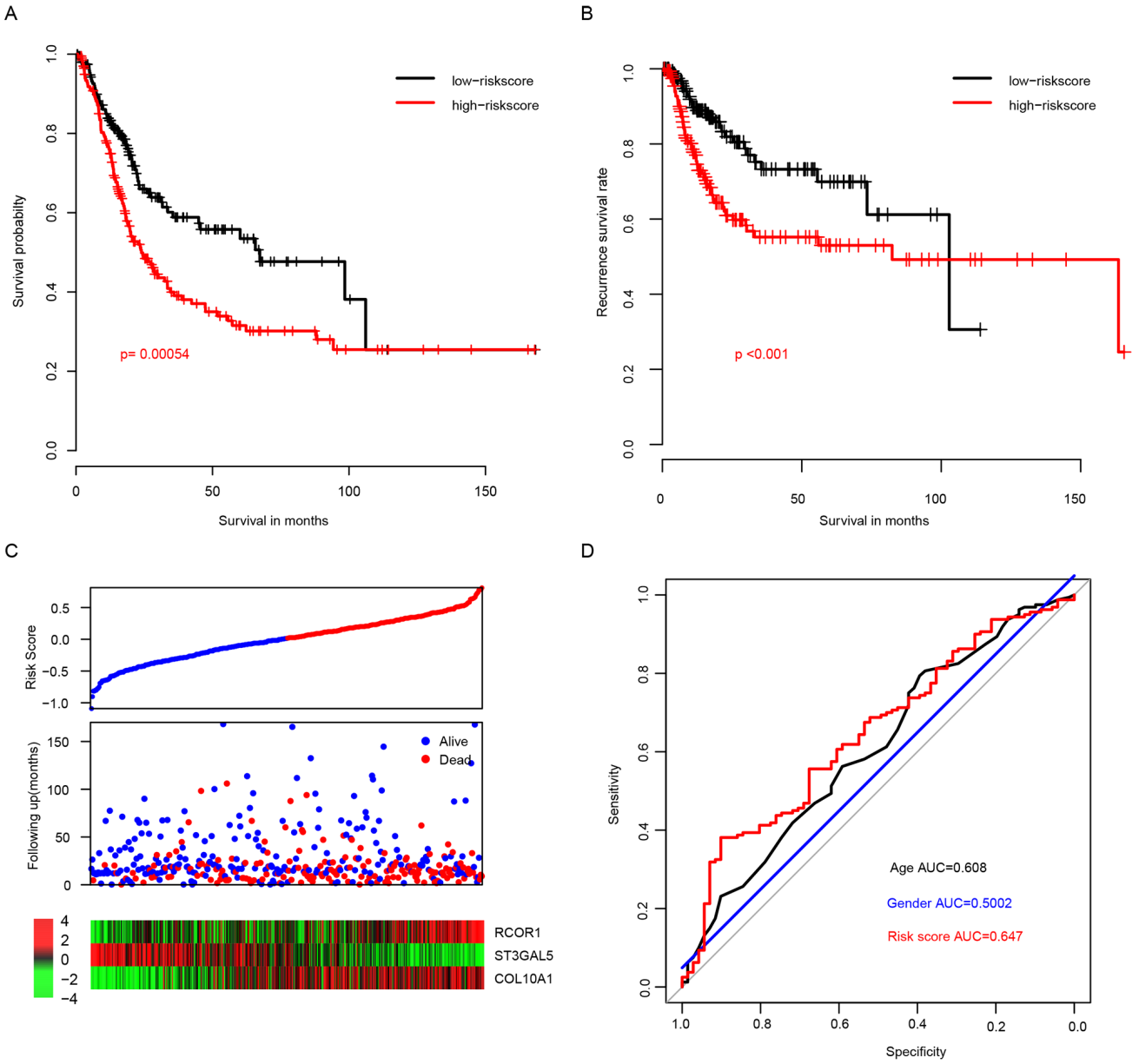

Figure 1: Performance of risk score in the training dataset (TCGA). The overall survival **(A)** and recurrence-free survival **(B)** rate in low-risk group is significantly higher than high-risk group, and the survival details were shown in **(C)** the three-year survival receiving operating characteristic curve (ROC) plotted and area under curves (AUC) were calculated **(D)**.

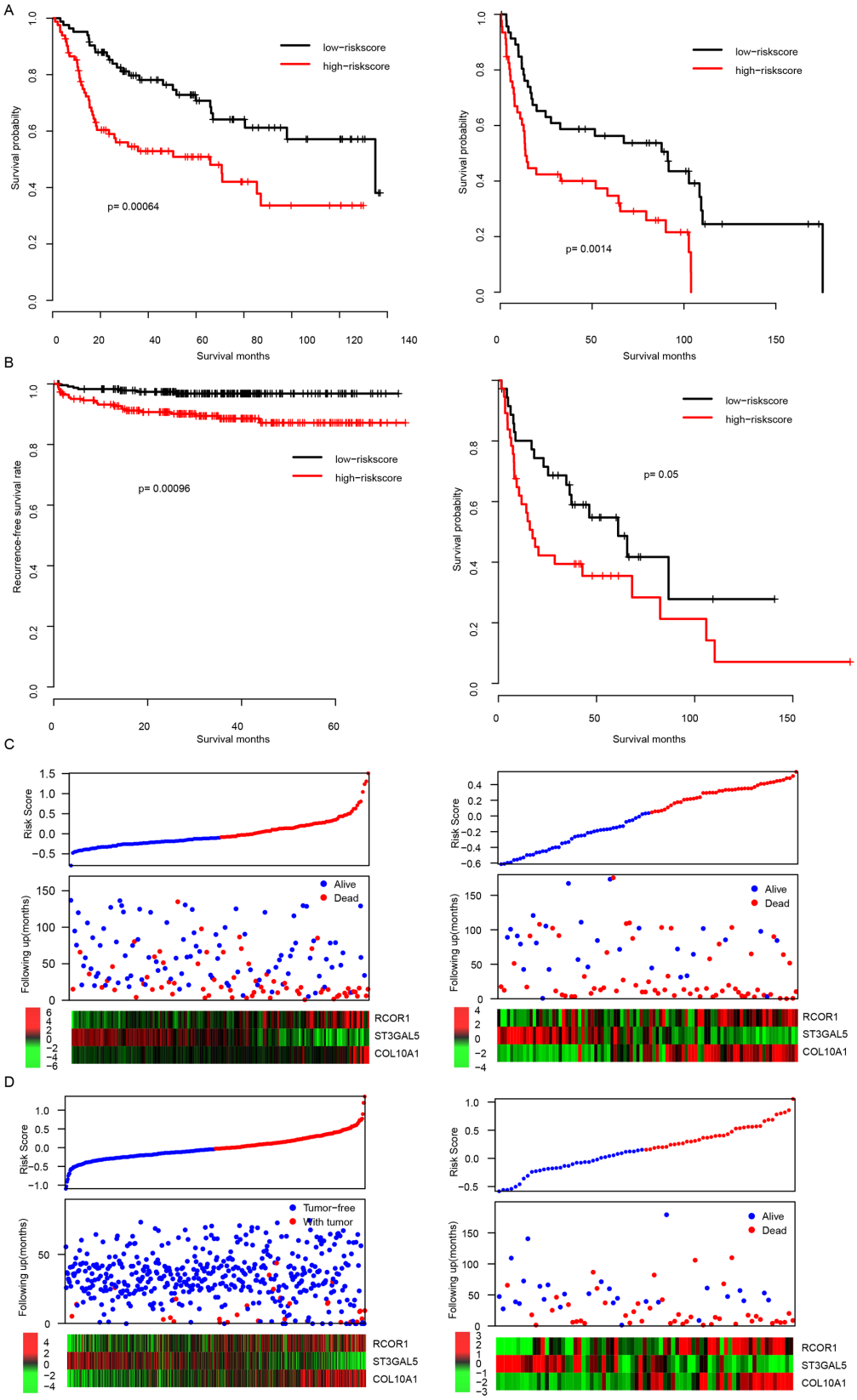

Figure 2: The performance of risk score in validation dataset. The overall survival difference of high/low-risk group were shown in GSE13507 and GSE31684 datasets (**A**, left and right, respectively). Profiles of Recurrence-free survival and overall survival rate of another two totally independent datasets (E-TABM-4321 and GSE40875) were similar **(B)**. Detailed survival information was shown (**C** left for GSE13507, right for GSE31684, **D** left for E-TABM-4321 right for GSE40875).

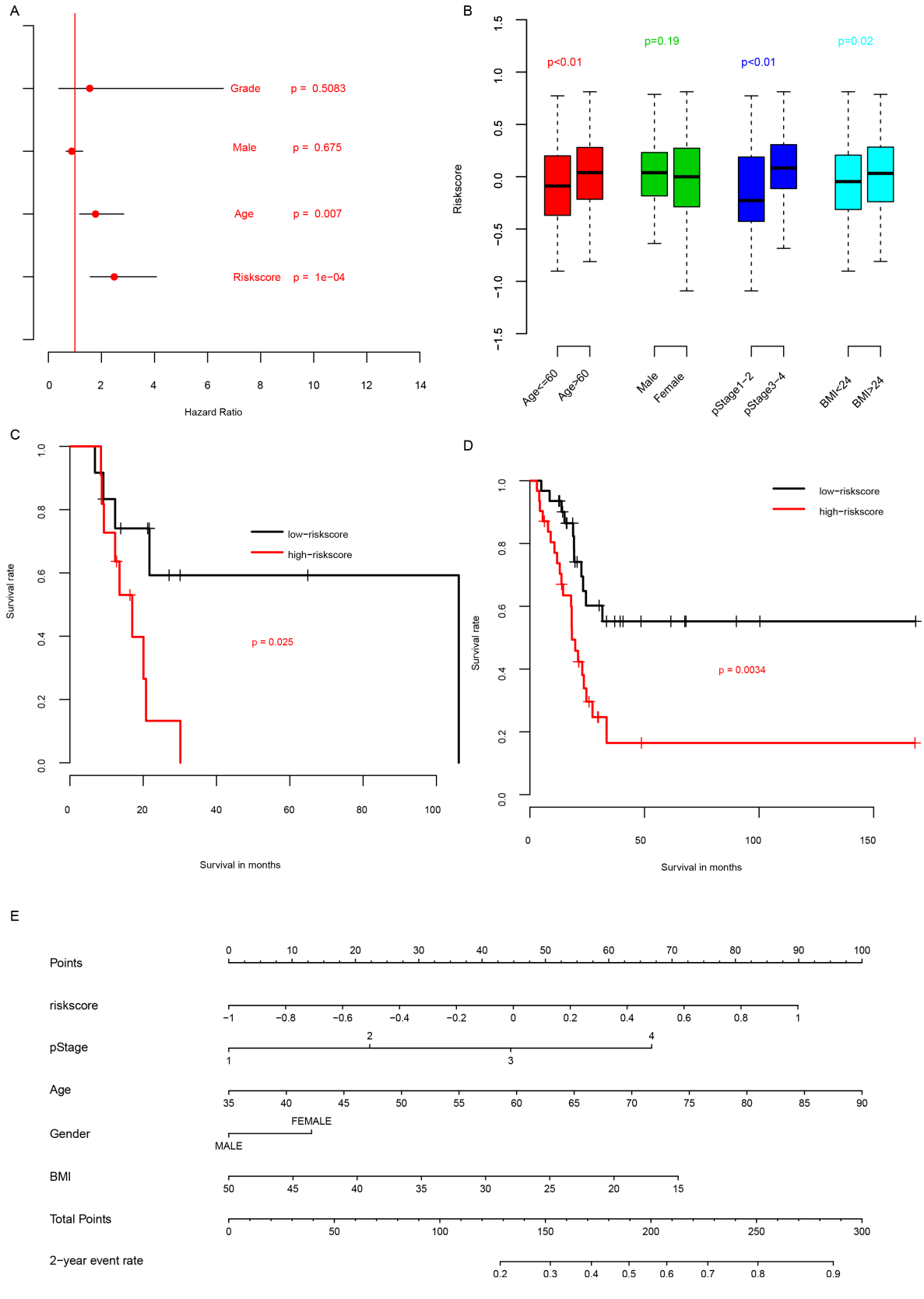

Figure 3: Clinical information and risk score. The clinical significance of clinical information and risk score **(A)**, and association between them **(B)**. The performance of risk score on patients underwent radiation **(C)** and without radiation **(D)** was also plotted. A nomogram containing clinical information and risk score was plotted **(E)**.

DISCUSSION

The prognosis of BLCA patients is still difficult by clinical information, including TNM staging, age, etc. [5, 6] Thus, the molecular biomarker for prognosis is critically needed for treatment. In the past decades, although a lot of single molecular markers for prognosis have been reported [7–9], the clinical performance across datasets is not very satisfactory. On the other hand, multiple genes' predicting effect has been highlighted [10–14]. In our current work, by using Cox multivariate regression on TCGA datasets, we report that the expression of three genes based risk score successfully predicted the survival of BLCA, and this finding is validated in four independent datasets. Totally, 1214 BLCA samples in 5 distinct datasets and two platforms involved in this study, and our model is effective in all of these datasets. Compared to other clinical information, the risk score contributed more, and performs better for survival predicting. In addition, the risk score predicting ability is robust for patients underwent radiation therapy or not.

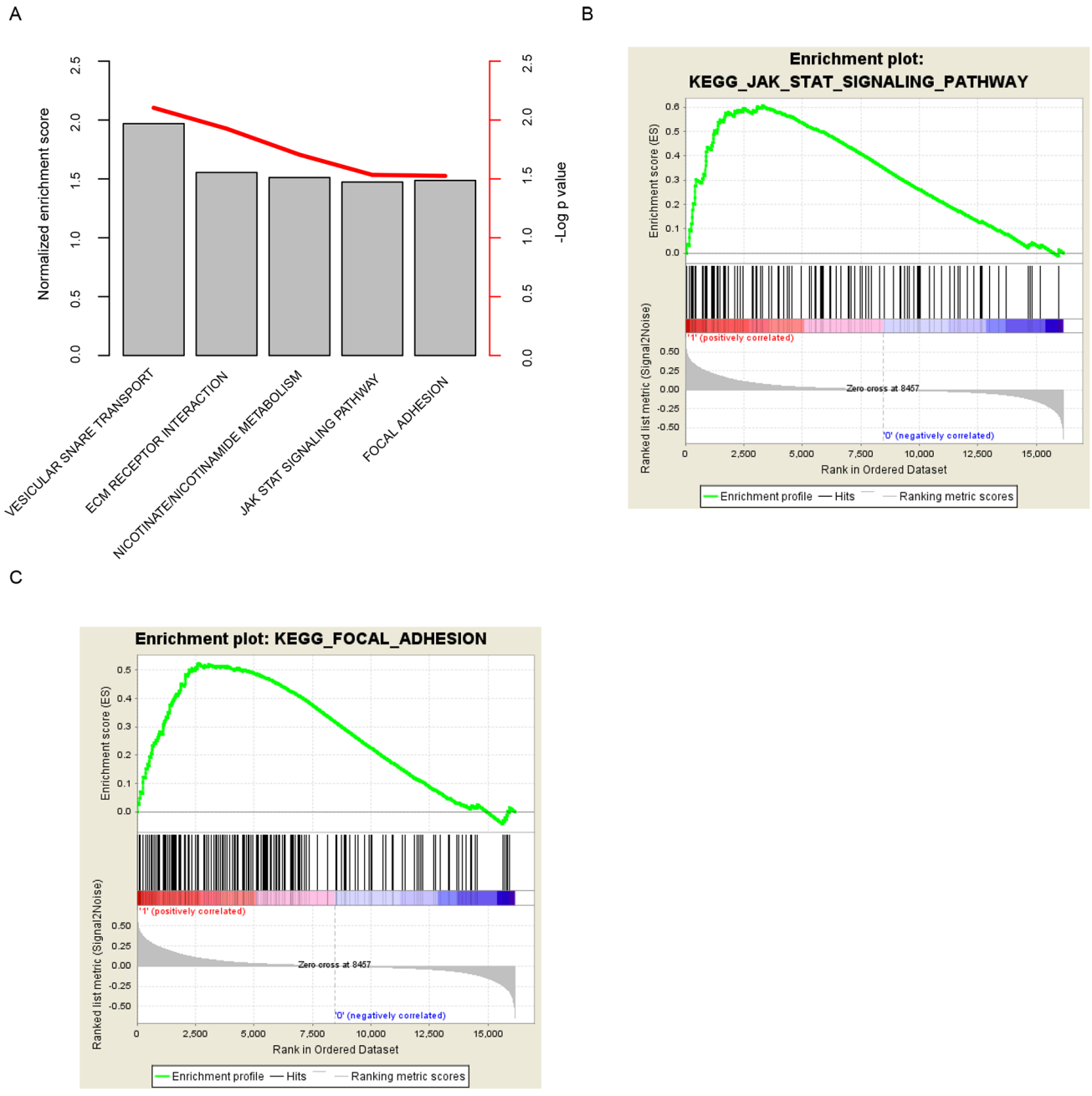

Figure 4: KEGG pathways associated with risk score. The high-risk score associated pathways were calculated **(A)** with GSEA, and JAK-STAT signaling pathway **(B)** and focal adhesion **(C)** were noted.

During the last years, lots of single prognostic biomarkers for bladder cancer have been reported, including mRNAs, lncRNAs, and miRNAs, and clinical significantly associated genes were identified. For example, HMGA2 was found to be associated with epithelial-to-mesenchymal transition in bladder cancer [15], and SKIP was reported to be associated with histological grades and poor prognosis [16]. Up-regulation of miRNAs including miR-141 [17] and miR-34a [18] often indicates a favorable survival. LncRNAs were also reported to predict survival of bladder cancer [19, 20]. However, the clinical utilization of these single biomarkers still need more investigation. One of the evidences is that, none the aforementioned mRNAs are not significantly associated with survival in our datasets. For example, HMGA2 was statistically significantly associated with overall survival only in TCGA dataset (p=0.022), but not GSE13507, GSE31684, GSE48075 and E-TABM-4321 (p=0.0898, 0.476, 0.0553 and 0.183 respectively, data not shown). On the other hand, multiple gene expression considered more information and thus more robust in prognosis, as previous reports [21–26]. Our results showed that the risks score performs better in 1214 samples consist of five independent datasets, consistent with these reports.

Of these three genes, RCOR1 interacts REST (RE1-silencing transcription factor) and modulate chromatin structure together with REST [27]. The prognostic effect of RCOR1 has been elucidated across cancer types, including glioma [28], and diffuse large B-cell lymphoma [29], although the prognostic effect of RCOR1 is still vague in bladder cancer. Another gene, ST3GAL5, has been reported to be positively correlates with the high risk of pediatric acute leukemia [30, 31] and associated with multidrug resistance in human acute myeloid leukemia, indicating the function of ST3GAL5 in carcinogenesis and development. The third gene, COL10A1, has been widely reported to be a oncogene in multiple cancer types [32], contribute to vasculature, and also used as biomarker for neo-adjuvant therapy effect prediction indicator in ER+/HER2+ breast cancer [33]. All the genes except for ST3GAL5 used were oncogenes according to the risk score formula. The positive coefficients of these genes in the risk score formula suggest that these genes contribute to the risk score and thus predict poor survival of bladder cancer, which is consistent with the aforementioned reports, except for ST3GAL5. It is considered that this may result from the heterogeneity among cancers. We also noticed that the functions of genes involved in this study were different, which may explain why the robustness of risk score is better than single gene biomarkers.

The significantly altered pathways include focal adhesion, and other related KEGG pathways, suggesting that the risk score reflected the cell-cell interaction status of BLCAs.

MATERIALS AND METHODS

Data pre-processing

The TCGA dataset were downloaded from UCSC Xena website (http://xena.ucsc.edu/), the expression value were converted to normalized RSEM values, the detailed pre-processing steps, including mapping and normalization, were described on UCSC Xena website. Genes expressed in less than 80% samples were discarded, and for 0 values were replaced with 1/2 of the minimum RSEM value except for 0 values of the corresponding gene. The expression matrix was then transformed with log 2.

Raw data of GSE13507, GSE31684, GSE48075, and E-TABM-4321 was downloaded in. CEL format from GEO (https://www.ncbi.nlm.nih.gov/geo) and array expression (www.ebi.ac.uk/arrayexpress/). After background correction, and normalization with Robust Multiarray Averaging (RMA) using R package "affy" function rma(), probes was mapped to gene name based on the manufacture provided annotation file. Genes matching more than one probe were merged and average values were calculated as their expression values.

Prediction gene selection and cox multivariate regression model

Cox univariate regression were performed in TCGA, GSE13507, and GSE31684 datasets to select the survival-related genes. Gene significantly associated with overall survival ($p<0.05$) in all of these datasets were retained for further analyses, and three genes were selected. Multivariate Cox Regression was performed to develop the risk score staging model with the candidate genes using R package "survival" function coxph(), and coefficients were locked for other datasets. The formula of risk score is described as the following,

$$Risk\ score = \sum_{i}^{n} \beta_i * x_i$$

Where β_i indicates the coefficients of genes and x_i refers to the relative expression of corresponding gene. The coefficients of β_i was calculated in the TCGA datasets and locked for assessing the risk score of samples in the other four independent datasets. Median risk score was used as cutoff values in discriminating the high and low risk group, and the survival difference was compared with Kaplan survival plot.

Statistical analysis

All statistical analysis was carried out with R (https://www.r-project.org/, v3.0.1) and R packages. Normalization of affymetrix raw data was performed

with R package "affy". The survival analysis and cox probability hazard model development was implemented with R package "survival". The ROC curves were plotted with R package "pROC" [13], and nomogram was drawn with R package "rms". The Gene Set Enrichment Analysis was performed with java software GSEA (http://software.broadinstitute.org/gsea/index.jsp) [34].

COMPLIANCE WITH ETHICAL STANDARDS

This article does not contain any studies with human participants or animals performed by any of the authors.

CONFLICTS OF INTEREST

The authors declare no (potential) conflicts of interest.

FUNDING

The authors declare no funding received.

REFERENCES

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015; 65:87-108.
2. Siegel R, Miller K, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015; 65:5-29.
3. Funt SA, Rosenberg JE. Systemic, perioperative management of muscle-invasive bladder cancer and future horizons. Nat Rev Clin Oncol. 2017; 14:221-234.
4. Salomaa V, Havulinna A, Saarela O, Zeller T, Jousilahti P, Jula A, Muenzel T, Aromaa A, Evans A, Kuulasmaa K, Blankenberg S. Thirty-one novel biomarkers as predictors for clinically incident diabetes. PLoS One. 2010; 5:e10100.
5. Zhao M, He XL, Teng XD. Understanding the molecular pathogenesis and prognostics of bladder cancer: an overview. Chin J Cancer Res. 2016; 28:92-98.
6. Dadhania V, Czerniak B, Guo CC. Adenocarcinoma of the urinary bladder. Am J Clin Exp Urol. 2015; 3:51-63.
7. Hong Z, Li H, Li L, Wang W, Xu T. Different expression patterns of histone H3K27 demethylases in renal cell carcinoma and bladder cancer. Cancer Biomark. 2017; 18:125-131.
8. Nandagopal L, Sonpavde G. Circulating biomarkers in bladder cancer. Bladder Cancer. 2016; 2:369-379.
9. Yang J, Platt LT, Maity B, Ahlers KE, Luo Z, Lin Z, Chakravarti B, Ibeawuchi SR, Askeland RW, Bondaruk J, Czerniak BA, Fisher RA. RGS6 is an essential tumor suppressor that prevents bladder carcinogenesis by promoting p53 activation and DNMT1 downregulation. Oncotarget. 2016; 7:69159-69172. doi: 10.18632/oncotarget.12473.
10. Gogalic S, Sauer U, Doppler S, Heinzel A, Perco P, Lukas A, Simpson G, Pandha H, Horvath A, Preininger C. Validation of a protein panel for the non-invasive detection of recurrent non-muscle invasive bladder cancer. Biomarkers. 2017; 19:1-8.
11. Urquidi V, Netherton M, Gomes-Giacoia E, Serie DJ, Eckel-Passow J, Rosser CJ, Goodison S. A microRNA biomarker panel for the non-invasive detection of bladder cancer. Oncotarget. 2016; 7:86290-86299. doi: 10.18632/oncotarget.13382.
12. Li Y, Huang J, Sun J, Xiang S, Yang D, Ying X, Lu M, Li H, Ren G. The transcription levels and prognostic values of seven proteasome alpha subunits in human cancers. Oncotarget. 2017; 8:4501-4519. doi: 10.18632/oncotarget.13885.
13. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77.
14. Kavalieris L, O'Sullivan P, Frampton C, Guilford P, Darling D, Jacobson E, Suttie J, Raman JD, Shariat SF, Lotan Y. Performance characteristics of a multigene urine biomarker test for monitoring for recurrent urothelial carcinoma in a multicenter study. J Urol. 2017; 197:1419-1426.
15. Ding X, Wang Y, Ma X, Guo H, Yan X, Chi Q, Li J, Hou Y, Wang C. Expression of HMGA2 in bladder cancer and its association with epithelial-to-mesenchymal transition. Cell Prolif. 2014; 47:146-151.
16. Wang L, Zhang M, Wu Y, Cheng C, Huang Y, Shi Z, Huang H. SKIP expression is correlated with clinical prognosis in patients with bladder cancer. Int J Clin Exp Pathol. 2014; 7:1695-1701.
17. Wang XL, Xie HY, Zhu CD, Zhu XF, Cao GX, Chen XH, Xu HF. Increased miR-141 expression is associated with diagnosis and favorable prognosis of patients with bladder cancer. Tumour Biol. 2015; 36:877-883.
18. Wang W, Li T, Han G, Li Y, Shi LH, Li H. Expression and role of miR-34a in bladder cancer. Indian J Biochem Biophys. 2013; 50:87-92.
19. Chen T, Xie W, Xie L, Sun Y, Zhang Y, Shen Z, Sha N, Xu H, Wu Z, Hu H, Wu C. Expression of long noncoding RNA lncRNA-n336928 is correlated with tumor stage and grade and overall survival in bladder cancer. Biochem Biophys Res Commun. 2015; 468:666-670.
20. Zhao XL, Zhao ZH, Xu WC, Hou JQ, Du XY. Increased expression of SPRY4-IT1 predicts poor prognosis and promotes tumor growth and metastasis in bladder cancer. Int J Clin Exp Pathol. 2015; 8:1954-1960.
21. Bou Samra E, Klein B, Commes T, Moreaux J. Development of gene expression-based risk score in cytogenetically normal acute myeloid leukemia patients. Oncotarget. 2012; 3:824-832. doi: 10.18632/oncotarget.571.

22. Ito H, Mo Q, Qin LX, Viale A, Maithel SK, Maker AV, Shia J, Kingham P, Allen P, DeMatteo RP, Fong Y, Jarnagin WR, D'Angelica M. Gene expression profiles accurately predict outcome following liver resection in patients with metastatic colorectal cancer. PLoS One. 2013; 8:e81680.

23. Chang W, Gao X, Han Y, Du Y, Liu Q, Wang L, Tan X, Zhang Q, Liu Y, Zhu Y, Yu Y, Fan X, Zhang H, et al. Gene expression profiling-derived immunohistochemistry signature with high prognostic value in colorectal carcinoma. Gut. 2014; 63:1457-1467.

24. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J, Bruin S, Kerr D, Kuppen P, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol. 2011; 29:17-24.

25. Bou Samra E, Klein B, Commes T, Moreaux J. Identification of a 20-gene expression-based risk score as a predictor of clinical outcome in chronic lymphocytic leukemia patients. Biomed Res Int. 2014; 2014:423174.

26. Kim SK, Kim SY, Kim JH, Roh SA, Cho DH, Kim YS, Kim JC. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. Mol Oncol. 2014; 8:1653-1666.

27. Meier K, Brehm A. Chromatin regulation: how complex does it get? Epigenetics. 2014; 9:1485-1495.

28. Yucebas M, Yilmaz Susluer S, Onur Caglar H, Balci T, Dogan Sigva ZO, Akalin T, Oktar N, Dalbasti T, Biray Avci C, Gunduz C. Expression profiling of RE1-silencing transcription factor (REST), REST corepressor 1 (RCOR1), and Synapsin 1 (SYN1) genes in human gliomas. J BUON. 2016; 21:964-972.

29. Chan FC, Telenius A, Healy S, Ben-Neriah S, Mottok A, Lim R. An RCOR1 loss-associated gene expression signature identifies a prognostically significant DLBCL subgroup. Blood. 2015; 125:959-966.

30. Mondal S, Chandra S, Mandal C. Elevated mRNA level of hST6Gal I and hST3Gal V positively correlates with the high risk of pediatric acute leukemia. Leuk Res. 2010; 34:463-470.

31. Ma H, Zhou H, Song X, Shi S, Zhang J, Jia L. Modification of sialylation is associated with multidrug resistance in human acute myeloid leukemia. Oncogene. 2015; 34:726-740.

32. Chapman KB, Prendes MJ, Sternberg H, Kidd JL, Funk WD, Wagner J, West MD. COL10A1 expression is elevated in diverse solid tumor types and is associated with tumor vasculature. Future Oncol. 2012; 8:1031-1040.

33. Brodsky AS, Xiong J, Yang D, Schorl C, Fenton MA, Graves TA, Sikov WM, Resnick MB, Wang Y. Identification of stromal ColXalpha1 and tumor-infiltrating lymphocytes as putative predictive markers of neoadjuvant therapy in estrogen receptor-positive/HER2-positive breast cancer. BMC Cancer. 2016; 16:274.

34. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102:15545-15550.