Construction and Validation of a Novel Eight-Gene Risk Score to Predict the Malignant Progression and Prognoses in Bladder Cancer Patients

Ruiliang Wang

shanghai tenth people's hospital

Zongtai Zheng

shanghai tenth people's hospital

Shiyu Mao

Shanghai tenth's peoples hospital

Wentao Zhang

Shanghai tenth people's hospital

Ji Liu

Shanghai tenth people's hospital

Cheng Li

Shanghai tenth people's hospital

Shenghua Liu (✉ shenghualiu1990@163.com)

Tongji University School of Medicine <https://orcid.org/0000-0003-0270-3343>

Xudong Yao

Shanghai Tenth People's Hospital <https://orcid.org/0000-0001-7234-3940>

Primary research

Keywords: Bladder cancer, LASSO, WGCNA, risk score, nomogram, malignant progression, prognosis

Posted Date: December 4th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-120263/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

The progression from non-muscle-invasive bladder cancer (NMIBC) to muscle-invasive bladder cancer (MIBC) largely predisposes a life-threatening risk. Owing to this, it is of paramount importance to find new relevant molecular models that will allow for effectively predict the malignant progression of BC.

Based on the RNA-Sequence data of 49 BC patients in our center and weighted gene co-expression network analysis methods, a co-expression network was developed using these genes before selecting the three key modules. Univariate Cox regression was used to select the key module genes with prognostic value in The Cancer Genome Atlas Program (TCGA). Subsequently, we developed an eight-gene risk score using the Least absolute shrinkage and selection operator Cox model. Notably, the eight-gene risk score was observed to be closely related to the malignant clinical features and also showed a favorable predictive power of differentiation between NMIBC and MIBC in the training (TCGA) and the two validation sets (GSE3289 and GSE13507). Furthermore, we generated a nomogram for predicting the overall survival. Both the calibrations and decision curve analysis curves displayed predictive effectiveness of the nomogram. Lastly, the RT-qPCR results revealed that the majority of the eight genes were differentially expressed between BC cell lines and a normal bladder epithelial cell line. Hence, from our study, we established a model of eight-gene risk score, with the potential of predicting malignant progression and determine prognoses of BC patients.

1 Background

Globally, bladder cancer (BC) has been ranked 11th among the cancers and considered as the fourth-most commonly diagnosed cancer in the male[1]. Among the BC patients, 75% have non-muscle invasive bladder cancer (NMIBC) on their first diagnosis whereas the 25% accounts for the muscle-invasive bladder cancer (MIBC). However, > 30% of patients with NMIBC often exhibit recurrence and progresses to MIBC within 5 years. [2]. Moreover, three-quarters of MIBC patients were further progressed to distant metastases with only 15% of long-term survival rate[3].

Cystoscopy had been immensely adopted as the golden standard for the diagnosis of BC[4]. However, its diagnostic ability for progression had yet been uncovered [5]. Despite some predictive models for the recurrence and progression being advanced, the predictive ability remained limited with an accuracy of less than 60%[6, 7]. Although some biomarkers and risk model were employed in predicting clinical outcome of patients with BC by using high-throughput sequencing data, both of their specificity and/or sensitivity remained unsatisfactory [2, 8]. For this reason, it is quite significant to identify valuable model or biomarkers for evaluating prognoses and monitoring malignant progression in BC.

With this article, a new prognostic risk model is presented having eight differently expressed genes between MIBC and NMIBC that predict the progression and malignant progression of BC using RNA-sequence data from our center and that of The Cancer Genome Atlas (TCGA) dataset. Based on the Gene Expression Omnibus (GEO), the gene expression data set was used to further verify the performance of

the risk model. Therefore, this study aimed at determining the expression levels of eight genes between BC cell lines and a normal bladder epithelial cell line. Moreover, we as well purposed to construct and validated an accurate and effective polygenic risk score that will establish a prognostic nomogram with the potential of prognosis of patients.

2 Methods

2.1 Clinical samples and RNA sequencing

A total of 49 BC patients (35 NMIBC and 14 MIBC) in Shanghai Tenth People's Hospital (STPH) who underwent either transurethral resection of bladder tumor or radical cystectomy between November 2019 and April 2020, were included in the training set. Among the criteria followed in the training set were: (1) histologically confirmation of BC; (2) availability of freshly collected tissue from surgery; and (3) availability of clinical data. Informed consent was prior obtained and the ethical approval was granted by the Ethics Committee of Shanghai Tenth People's Hospital. Subsequently, the total RNA of each BC tissue was isolated using Trizol reagent (Invitrogen, Thermo Scientific, Shanghai, China). The cDNA libraries were constructed using a customized protocol following its sequencing using the Illumina Hiseq 2500 sequencer (Sangon Biotech, Shanghai, China).

2.2 Public data processing

Gene expression data from GSE32894 (n=224) and GSE13507 (n=165) datasets were downloaded from GEO (<http://www.ncbi.nlm.nih.gov/geo>) as a series of matrix file format that was earlier processed by the respective study authors through the MAS 5.0 algorithm. The latest TCGA data was downloaded having clinical features and follow-up information using GDC API. The clinical data of the training sets (STPH and TCGA datasets) and validation sets (GSE32894 and GSE13507) were shown in **Table 1**.

2.3 Predictive model establishment and bioinformatics analyses

2.3.1 Weighted gene co-expression network analysis

The “limma” R package was adopted for screening the differential expression genes (DEGs) of between 35 NMIBC and 14 MIBC samples in STPH. A weighted gene co-expression network analysis (WGCNA) was constructed[9] via the ‘WGCNA’ R package v1.68 based on the DEGs (P -value<0.05 and $|\log_2FC| \geq 0.5$). However, the basic flow of WGCNA adhered to the following: (1) Outlier samples were omitted to facilitate the reliability of the weighted gene co-expression network, (2) An appropriate β value was selected using 0.85 as the degree of independence (R^2), (3) A weighted adjacent matrix was transformed into a topological overlap matrix (TOM) to determine the network connectivity of the genes, (4) Genes with similar expression profiles were classified into gene modules based on the average linkage hierarchical clustering following the TOM-based dissimilarity measure, (5) All genes were represented by the expression of module eigengenes (MEs) in a given module. For the modules that were highly

correlated with the NMIBC/MIBC subtype ($|r| \geq 0.3$), they were selected for further analyses, and (6) The genes in the selected modules (brown, turquoise, and yellow) were extracted.

2.3.2 Construction of eight-gene risk score

Based on hub genes extracted from WGCNA, we performed the univariate Cox to select an overall survival (OS) associated genes in the TCGA dataset [10]. Thereafter, the OS associated genes were subject to Least absolute shrinkage and selection operator (LASSO) with ten-fold cross-validation to construct an eight-genes risk score using “glmnet” package. Accordingly, the genes with the minimal influence on the patients overall survival were removed whereas the genes with non-zero coefficients were selected. The risk score for each patient was calculated as follows: risk score = $\text{Coef}_1 \times \text{expression of gene}_1 + \text{Coef}_2 \times \text{expression of gene}_2 + \dots + \text{Coef}_m \times \text{expression of gene}_m$. Coef denotes the corresponding coefficient of the gene. With this data expression, the ‘survminer’ R package v4.6 was used to evaluate the optimal cutoff value of the risk score in each cohort. The determined corresponding optimal cutoff value was used to divide patients into low- and high-risk groups accordingly.

2.3.3 Gene sets enrichment analysis (GSEA)

GSEA (<http://www.broadinstitute.org/gsea/index.jsp>) was conducted to investigate different pathway activities between patients in high- and low-risk groups. Signaling pathway with $P < 0.05$ and a false discovery rate < 0.25 were regarded as statistically significant.

2.3.4 Gene Set Cancer Analysis

Gene Set Cancer Analysis (GSCALite, <http://bioinfo.life.hust.edu.cn/web/GSCALite/>) provides a single nucleotide variation (SNV) module through the maftools[11]. Therefore, in this present study, we employ the SNV module for analyzing and visualizing the SNV of eight genes in BC.

2.4 Eight-Gene Risk score validation

In evaluating the generalizability of the model, similar formula and coefficients to the training sets (STPH and TCGA dataset) in the validation datasets (GSE3289 and GSE13507) were used. The Kaplan-Meier curves of the validation datasets were separately drawn before analyzing their eight-genes risk score prognostic classification efficacy. Furthermore, the receiver operating characteristic (ROC) curve and violin plot were used to verify the relationship between the risk score and the clinical features (MIBC).

2.5 Development of the nomogram

A nomogram was developed from the factors that were significant in the final multivariate Cox regression analyses. The ROC, concordance index (C-index) and calibration plots were conducted purposely to assess the performance of the nomogram. The nomogram, C-index and calibration plots were generated using the ‘Rms’ R package. Furthermore, the decision curve analysis (DCA) was employed to ascertaining the net benefits of nomogram and other crucial prognostic factors.

2.6 Cell culture

The human normal bladder epithelial cell line SV-HUC-1 and BC cell line T24, UMUC3, J82, 5637, EJ obtained from Shanghai Cell Bank of the Chinese Academy of Sciences (Shanghai, China) were used in this study. The SV-HUC-1 cell lines were cultured in F12K having 10% FBS and 1% penicillin/streptomycin (P/S). As of the BC cell lines, they were cultured in RPMI-1640 (Gibco; Thermo Fisher Scientific, Inc.) with 10% FBS (Gibco; Thermo Fisher Scientific, Inc.) and 1% P/S (Gibco; Thermo Fisher Scientific, Inc.). Cell culture reagents were sourced from Gibco and the subjected cells were cultured at 37°C in 5% CO₂.

2.7 Real-time quantitative (RT-qPCR)

TRIzol reagent (Invitrogen, Thermo Scientific, Shanghai, China) was used for extracting RNA of the cell lines. cDNA synthesis was performed using cDNA Synthesis SuperMix Kit (Cat No. 11141ES60; Yeasen, Shanghai, China) and qPCR was performed using qPCR SYBR Green Master Mix KIT (Cat No. 11203ES03; Yeasen, Shanghai, China). The genes expression levels normalized to the level of GAPDH and Primer sequences are presented in **Table S1**.

2.8 Statistical analysis

The Wilcoxon rank-sum test was performed to investigate the differential expression genes between NMIBC and MIBC. The Student's t-test or one-way ANOVA was used to evaluate the risk score in patients that were grouped according to their clinical characteristics. Both the Kaplan-Meier and log-rank tests were employed to compare the OS of the BC patients of between high- and low-risk groups. The performance of the risk score in predicting prognosis and the NMIBC/MIBC subtype was determined using the ROC curves. The heatmap and correlation matrix were generated using 'Pheatmap' R package v1.0.12 and 'corrplot' R package v0.84, respectively. SPSS 22.0 (SPSS, Armonk, NY, USA), R v3.6.1 (<https://www.r-project.org/>) and Graphpad Prism V7 (GraphPad Software, Inc.) were used for data analyses. However, a two-sided P value < 0.05 was considered significant.

3 Results

3.1 Data acquisition and DEGs selection

The flow diagram of the study had been indicated in Fig. 1. A total of 49 patients (35 NMIBC and 14 MIBC) with BC from STPH were recruited for this study. From our findings, a total of 2725 different expression genes were identified between NMIBC and MIBC. The expression of genes of Log₂FC > 4 and P-value < 0.01 was showed in **Figure S1A**.

3.2 Wgcna Analysis Based On Degs Between Nmibc And Mibc

After the hierarchical clustering analysis, one sample was removed as an outlier. A co-expression network was constructed using 48 BC samples with a complete clinical T stage and NMIBC/MIBC subgroups (Fig. 2A). With the chosen power of $\beta = 13$ (scale-free $R^2 = 0.85$) as the soft-thresholding (Fig. 2B), a total of seven modules were obtained (Fig. 2C). Notably, the highest association of the module-trait relationship was found between the three modules and the NMIBC/MIBC subgroups (brown, turquoise, and yellow) (Fig. 2D). However, 1351 genes from three key modules were selected for further analysis.

3.3 Construction And Verification Of The Eight-genes Progression Risk Score

Among the 1315 genes, 159 were statistically significant (P -values < 0.05) in the univariate Cox analysis, hence selected for the subsequent LASSO Cox regression. The Lambda.1se penalty parameter was selected via ten-fold leave-one-out cross-validation based on the minimum criteria. Of the eight genes (CD96, PDCL3, IP6K2, TRIM38, U2AF1L4, DDB1, KCNJ15 and CTU1) obtained with nonzero coefficients (Fig. 3A, B), were subsequently used to calculate the risk scores of each patient. The coefficients of the eight-genes risk score were shown in Fig. 3C. According to the aforementioned formula, risk scores for each patient were calculated. From the Kaplan-Meier analysis of risk, a distinct prognostic ability in the training set ($P < 0.001$) was displayed (Fig. 3D). We further validated the prognostic value of the risk score in two validation sets with prognostic information. Interestingly, patients with high-risk scores had a significantly worse OS than those with low-risk scores in GSE3289 ($P = 0.035$) and GSE13507 ($P = 0.037$) (Fig. 3E, F). According to enrichment map analysis with GSEA, the association between the risk score and malignant behavior was further confirmed. Our results showed that the high score of the model potential activated the 'Focal adhesion', 'pathway in cancer', 'WNT signal', 'bladder tumor', 'ECM receptor' and 'adherens junctions pathway' (**Figure S1B**). Among the 411 BLCA samples with sequencing data in TCGA, mutations of the eight genes were only found in 29 independent samples in TCGA dataset. Moreover, CD96 and DDB1, each recorded the highest rates of mutations of 28% (**Figure S1C**).

3.4 The Association Between Eight-gene Risk Score And Clinicopathology

The patients having high-risk scores demonstrated a relatively higher proportion of non-papillary ($P < 0.001$), age ($P < 0.001$), pathological stage ($P < 0.001$), histological grade ($P < 0.001$), T stage, N stage ($P < 0.001$) and M stage ($P < 0.01$) in TCGA dataset (Fig. 4A). The dot plot shown according to the survival status of the TCGA set (Fig. 4B), indicated that patients with high-risk scores had a higher mortality rate than those with low-risk scores. The ROC revealed that the risk score has a high diagnostic ability to differentiate between NMIBC and MIBC with the AUC of 0.903 (Fig. 4C). The distributions risk scores in both training and validation sets are presented in detail in Fig. 4D that showed a significantly higher score in MIBC than NMIBC (all $P < 0.05$).

3.5 Construction And Calibration Of The Nomogram

The univariate Cox analyses manifested that the T stage, histological grade, age and risk score were significantly related to OS in the TCGA dataset (Fig. 5A). After multivariate Cox analysis, the risk score, age, T stage, and pathological stage still had the prognostic value. A nomogram was constructed for 1-, 3- and 5-year OS prediction based on significant factors in the multivariate Cox analyses (Fig. 5B). The AUC values of ROC curves were 0.764 (1-year ROC), 0.796 (3-year ROC) and 0.805 (5-year ROC), whereas the C-index of the nomogram was 0.75 ± 0.02 (mean $\pm$ SD) for OS (Fig. 5C). Besides, calibration plots proved that the nomogram prediction had a marked agreement with actual 1-, 3- and 5-year OS (Fig. 5D). Furthermore, the comparison between DCA and the clinical variables (age, pathological stage, T stage, and risk score), illustrated a larger eight-genes risk score that enhanced the clinical net benefit. Nevertheless, the nomogram had the best clinical net benefit compared with other factors (Fig. 5E).

3.6 Validation of eight-genes through in vitro experiments.

The expression of the eight selected genes in five human BC cell lines (T24, UMUC3, 5637, J82 and EJ) and one human normal bladder epithelial cell line (SV-HUC-1) were detected using RT-qPCR. The expression levels of five genes (CTU1, IP6K2, KCNJ15, PDCL3 and U2AF1L4) were significantly different between the BC and normal bladder cells (Fig. 6). The expression level of IP6K2 and KCNJ15 was low from all cancer cell lines compared to normal cell lines. The expression of CTU1, PDCL3 and U2AF1L4 were heterogeneous between different BC cell lines.

4 Discussion

With the enormous development in the field of high-throughput sequencing technologies, an increasing number of bioinformatics methods were used to find biomarkers related to the progression of malignant tumors[12, 13]. As a result of these advances, they have positively impacted on the early diagnosis and progression predictions. The recurrence rates of BC are high (range from 30 to 70%) with 20 to 30% of patients showing the progression of the disease[14]. However, the exact biological function changes during the malignant progression of BC remained elusive. Owing to the emergence of the risk assessment model, it is imperative for improving following this current situation. We performed RNA-sequencing on the tumor tissues of 49 patients with BC in Shanghai Tenth People's Hospital and used the prognostic information in the TCGA dataset. Thereafter, we determined the core DEG between NMIBC and MIBC through the WGCNA method where the three modules with significant vital status were identified. Furthermore, univariate Cox regression and LASSO regression analyses constructed eight-genes risk score using the module genes with TCGA prognostic information. Additionally, GSEA confirmed that these eight genes modulated the signaling pathway such as 'WNT signal', 'bladder tumor' and so on. The GSEA results indicated that these genes rarely mutated, thus the potential dysfunction with these genes may not be as a result of genetic mutations, but rather the dysfunction at the transcriptional level[15]. According to the coefficients of genes, genes with risk score were protective genes except for DDB1. These genes had been noticed to play an important role in previous tumor studies. For example, CD96

following its discovery in 1992, it is an IgG superfamily receptor expressed in some hematopoietic stem cells, $\alpha\beta$ and $\gamma\delta$ T cells. Due to more CD96 expression in BC tissue, it is an indication of more immune cells infiltrates with a potential to eradicate cancer cells[16, 17]. Damage-specific DNA binding protein 1 (DDB1), as the only oncogene in our model, was a multifunctional protein. Previous studies had investigated DDB1 interacting with several chaperone proteins in regulating the repair mechanisms, cell cycle and gene transcription [18]. DDB1 can directly interact with Cullin4B (CUL4B) to assemble into CUL4B-DDB1 complex which mediates the occurrence of osteosarcoma through ubiquitination and degradation of a variety of tumor suppressor factors [19]. As for the tripartite motif protein 38 (TRIM38), a member of the TRIM family, was demonstrated having the most protective effect in our model which was involved in various cellular processes such as cell differentiation, proliferation, apoptosis, and antiviral defense. On the other hand, TRIM38 can activate an abnormal NF- κ B pathway, thus exerting an inhibition cancer effect[20, 21]. .

Subsequently, BC patients from the training set were divided into high-and low-risk groups having a markedly different OS. The credibility of our risk score was further verified in the validation datasets individually (GSE3289 and GSE13507), hence displaying an excellent reproducibility. In addition, the results of the multivariate Cox regression model suggested that our risk score had a higher prognostic capability. Up to now, there are several similar BC prediction models[22–24]. However, these models either only used the public datasets without validation set or data in their center, and the genes of their models were not verified by processes such as RT-qPCR, which limits the generalizability and reproducibility of the models. As for our model, we combined the data in our center and public datasets, before verifying the feasibility of the model in the two validation sets thus making them convincing.

A nomogram including the age, T stage, pathological stage and risk score was drawn to predict the OS probability based on the multivariate analysis of OS. The ROC analysis with 1-, 3- and 5-year OS of BC produced a reliable diagnostic. Moreover, the highly fitted calibration plots and C-index revealed that our nomogram was able to provide simple and accurate prognosis predictions for 1-, 3- and 5-year OS. DCA showed the clinical usefulness of the nomogram to perform personalized mortality risk identification and progression in patients with BC.

The results of RT-qPCR in the present study suggested that five among the eight-genes exhibited dysregulated expression between a normal urothelial cell line and BC cell lines, illustrating that dysregulated expression levels of the hub genes served an important role in the malignant progression of BC. On the contrary, the expression of some genes in cell lines does not match our model. Therefore, the possible reason behind the occurrence and development of tumors is as a result of the comprehensive interaction of tumor immune microenvironment[25, 26]. Our RNA-Sequence samples were obtained from fresh tissues of BC patients, which means there are not only the cancer cells but also some other key components such as inflammatory factors, immune cells in the tumor microenvironment. The incompatible genes may be specifically expressed by either immune cells or other components which, despite not being detected in cancer cell lines. The reason behind the three genes being heterogeneous

between different cancer cell lines had been hypothesized as a result of the cell lines from different origins.

5 Conclusion

Herein, the present study demonstrated that the eight-genes risk score confirmed the reliable predictions in malignant progression and prognoses for BC patients. Using nomogram based on the eight-genes risk score for survival prediction can immensely advance clinical decision making. For instance, the biological function analysis of the eight-genes risk score provided new insights into the malignant progression of BC. Generally, the functional studies and mechanistic analyses of the eight hub genes may largely contribute to the development of the treatments for BC.

Declarations

Ethics declarations

The present study had approval from the Ethics Committee of Shanghai Tenth People's Hospital (Shanghai, China). The research has been carried out in accordance with the World Medical Association Declaration of Helsinki. All patients provided written informed consent.

Availability of data and materials

The datasets used during this research are available from the corresponding author upon reasonable request. The public data that support the findings of this study are openly available in TCGA (<http://cancergenome.nih.gov/>) and NCBI GEO dataset (<https://www.ncbi.nlm.nih.gov/>).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the Shanghai Science Committee Foundation (grant number 19411967700) and the Natural Science Foundation of China (grant number 81472389).

Authors' contributions

R.W. and Z.Z. conceived and designed the study. X.Y. and S.L. acquired the funding. R.W., S.M. and W.Z. collected and collated the data. All the authors were involved in the analysis and interpretation of data. R.W. wrote the manuscript, Z.Z., S.M. and W.Z. critically reviewed and revised the manuscript. J.L., C.L.,

and Z.Z. designed the tables and figures. All authors read and approved the manuscript and agreed to be accountable for all aspects of the research in ensuring that the accuracy or integrity of any part of the work were appropriately investigated and resolved.

Acknowledgements

This study was funded by the Shanghai Science Committee Foundation (grant number 19411967700) and the Natural Science Foundation of China (grant number 81472389).

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* 2020;70(1):7–30.
2. Dyrskjot L, Reinert T, Algaba F, Christensen E, Nieboer D, Hermann GG, Mogensen K, Beukers W, Marquez M, Segersten U, et al. Prognostic Impact of a 12-gene Progression Score in Non-muscle-invasive Bladder Cancer: A Prospective Multicentre Validation Study. *Eur Urol.* 2017;72(3):461–9.
3. Abufaraj M, Dalbagni G, Daneshmand S, Horenblas S, Kamat AM, Kanzaki R, Zlotta AR, Shariat SF. The Role of Surgery in Metastatic Bladder Cancer: A Systematic Review. *Eur Urol.* 2018;73(4):543–57.
4. Witjes JA, Bruins HM, Cathomas R, Comperat EM, Cowan NC, Gakis G, Hernandez V, Linares Espinos E, Lorch A, Neuzillet Y, et al: **European Association of Urology Guidelines on Muscle-invasive and Metastatic Bladder Cancer: Summary of the 2020 Guidelines.** *Eur Urol* 2020.
5. Rose TL, Lotan Y. Advancements in optical techniques and imaging in the diagnosis and management of bladder cancer. *Urol Oncol.* 2018;36(3):97–102.
6. Xylinas E, Kent M, Kluth L, Pycha A, Comploj E, Svatek RS, Lotan Y, Trinh QD, Karakiewicz PI, Holmang S, et al. Accuracy of the EORTC risk tables and of the CUETO scoring model to predict outcomes in non-muscle-invasive urothelial carcinoma of the bladder. *Br J Cancer.* 2013;109(6):1460–6.
7. Kluth LA, Black PC, Bochner BH, Catto J, Lerner SP, Stenzl A, Sylvester R, Vickers AJ, Xylinas E, Shariat SF. Prognostic and Prediction Tools in Bladder Cancer: A Comprehensive Review of the Literature. *Eur Urol.* 2015;68(2):238–53.
8. Dyrskjot L, Zieger K, Real FX, Malats N, Carrato A, Hurst C, Kotwal S, Knowles M, Malmstrom PU, de la Torre M, et al. Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: a multicenter validation study. *Clin Cancer Res.* 2007;13(12):3545–51.
9. Shi J, Zhang P, Liu L, Min X, Xiao Y. Weighted gene coexpression network analysis identifies a new biomarker of CENPF for prediction disease prognosis and progression in nonmuscle invasive bladder cancer. *Mol Genet Genomic Med.* 2019;7(11):e982.
10. Li Y, Lu S, Lan M, Peng X, Zhang Z, Lang J. A prognostic nomogram integrating novel biomarkers identified by machine learning for cervical squamous cell carcinoma. *J Transl Med.* 2020;18(1):223.

11. Mayakonda A, Koeffler HP: **Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies.** 2016.
12. Yan X, Fu X, Guo ZX, Liu XP, Liu TZ, Li S. Construction and validation of an eight-gene signature with great prognostic value in bladder cancer. *J Cancer.* 2020;11(7):1768–79.
13. Zheng Z, Mao S, Guo Y, Zhang W, Liu J, Li C, Yao X. N6methyladenosine RNA methylation regulators participate in malignant progression and have prognostic value in clear cell renal cell carcinoma. *Oncol Rep.* 2020;43(5):1591–605.
14. Hoffman KE, Penson DF, Zhao Z, Huang LC, Conwill R, Laviana AA, Joyce DD, Luckenbaugh AN, Goodman M, Hamilton AS, et al. Patient-Reported Outcomes Through 5 Years for Active Surveillance, Surgery, Brachytherapy, or External Beam Radiation With or Without Androgen Deprivation Therapy for Localized Prostate Cancer. *Jama.* 2020;323(2):149–63.
15. Hu MM, Yang Q, Zhang J, Liu SM, Zhang Y, Lin H, Huang ZF, Wang YY, Zhang XD, Zhong B, et al. TRIM38 inhibits TNFalpha- and IL-1beta-triggered NF-kappaB activation by mediating lysosome-dependent degradation of TAB2/3. *Proc Natl Acad Sci U S A.* 2014;111(4):1509–14.
16. Georgiev H, Ravens I, Papadogianni G, Bernhardt G. Coming of Age: CD96 Emerges as Modulator of Immune Responses. *Front Immunol.* 2018;9:1072.
17. Blake SJ, Dougall WC, Miles JJ, Teng MW, Smyth MJ. Molecular Pathways: Targeting CD96 and TIGIT for Cancer Immunotherapy. *Clin Cancer Res.* 2016;22(21):5183–8.
18. Srinivasan S, Chitalia V, Meyer RD, Hartsough E, Mehta M, Harrold I, Anderson N, Feng H, Smith LE, Jiang Y, et al. Hypoxia-induced expression of phosphocyanin-like 3 regulates expression of VEGFR-2 and promotes angiogenesis. *Angiogenesis.* 2015;18(4):449–62.
19. Chen B, Feng Y, Zhang M, Cheng G, Chen B, Wang H. Small molecule TSC01682 inhibits osteosarcoma cell growth by specifically disrupting the CUL4B-DDB1 interaction and decreasing the ubiquitination of CRL4B E3 ligase substrates. *Am J Cancer Res.* 2019;9(9):1857–70.
20. Kim K, Kim JH, Kim I, Seong S, Kim N. TRIM38 regulates NF-kappaB activation through TAB2 degradation in osteoclast and osteoblast differentiation. *Bone.* 2018;113:17–28.
21. Hu MM, Xie XQ, Yang Q, Liao CY, Ye W, Lin H, Shu HB. TRIM38 Negatively Regulates TLR3/4-Mediated Innate Immune and Inflammatory Responses by Two Sequential and Distinct Mechanisms. *J Immunol.* 2015;195(9):4415–25.
22. Wang C, Yang Y, Yin L, Wei N, Hong T, Sun Z, Yao J, Li Z, Liu T. Novel Potential Biomarkers Associated With Epithelial to Mesenchymal Transition and Bladder Cancer Prognosis Identified by Integrated Bioinformatic Analysis. *Front Oncol.* 2020;10:931.
23. Lian P, Wang Q, Zhao Y, Chen C, Sun X, Li H, Deng J, Zhang H, Ji Z, Zhang X, et al. An eight-long non-coding RNA signature as a candidate prognostic biomarker for bladder cancer. *Aging.* 2019;11(17):6930–40.
24. Pan S, Zhan Y, Chen X, Wu B, Liu B. Identification of Biomarkers for Controlling Cancer Stem Cell Characteristics in Bladder Cancer by Network Analysis of Transcriptome Data Stemness Indices. *Front Oncol.* 2019;9:613.

25. Li F, Teng H, Liu M, Liu B, Zhang D, Xu Z, Wang Y, Zhou H. Prognostic Value of Immune-Related Genes in the Tumor Microenvironment of Bladder Cancer. *Front Oncol.* 2020;10:1302.

26. Zhou X, Qiu S, Nie L, Jin D, Jin K, Zheng X, Yang L, Wei Q. Classification of Muscle-Invasive Bladder Cancer Based on Immunogenomic Profiling. *Front Oncol.* 2020;10:1429.

Tables

Table 1: Characteristic of datasets used in this study

Characteristic, n (%)		GSE32894	GSE13507	TCGA	STPH
Subtype	NMIBC	173	103	3	35
	MIBC	51	62	400	14
Age (years)	≥65	70	69	254	35
	<65	154	96	149	14
Gender	Male	163	135	299	42
	Female	61	30	104	7
Stage	I	-	103	2	
	II	-	1	128	
	III	-	0	138	
	IV	-	1	132	
	Unknown	-	60	3	
Grade	Low	-	105	21	5
	High	-	60	378	44
	Grade I	45	-	-	
	Grade II	84	-	-	
	Grade III	93	-	-	
	Unknown	2	-	4	
N stage	N0	26	1	234	
	N1-3	20	104	128	
	Unknown	177	60	41	
M stage	M0	-	1	196	
	M1	-	104	11	
	Unknown	-	60	196	
†NMIBC, non-muscle invasive bladder cancer					
MIBC, muscle invasive bladder cancer					

Figures

Figure 1

Flowchart of the risk score model construction process

Figure 1

Flowchart of the risk score model construction process

Figure 2

The WGCNA network construction and key module identification. (a) Sample dendrogram and trait indicator. The clustering was a visual result of calculations based on Pearson correlation coefficients between samples. The color intensity was proportional to T stage and NMIBC/MIBC subgroup of BC. (b) The analysis of topology for soft threshold powers and $\beta=13$ (scale-free $R^2=0.85$) was set as the soft thresholding for further adjacency calculation. (c) The original modules and merged modules were

displayed at top and bottom under the clustering dendrogram. (d) Module-trait relationships between identified modules and clinical features. The numbers represent Pearson's correlation between the clinical traits and modules. The numbers in the parentheses correspond to the p value. WGCNA: weighted gene correlation network analysis. NMIBC: non-muscle-invasive bladder cancer; MIBC: muscle-invasive bladder cancer. ME: module.

Figure 2

The WGCNA network construction and key module identification. (a) Sample dendrogram and trait indicator. The clustering was a visual result of calculations based on Pearson correlation coefficients between samples. The color intensity was proportional to T stage and NMIBC/MIBC subgroup of BC. (b) The analysis of topology for soft threshold powers and $\beta=13$ (scale-free $R^2=0.85$) was set as the soft thresholding for further adjacency calculation. (c) The original modules and merged modules were displayed at top and bottom under the clustering dendrogram. (d) Module-trait relationships between identified modules and clinical features. The numbers represent Pearson's correlation between the clinical traits and modules. The numbers in the parentheses correspond to the p value. WGCNA: weighted gene correlation network analysis. NMIBC: non-muscle-invasive bladder cancer; MIBC: muscle-invasive bladder cancer. ME: module.

Figure 3

Construction and validation of the risk score. (a) Plots of the cross-validation error rates. (b) Distribution of LASSO coefficients of OS-associated genes. (c) Coefficient values of the eight genes. (d, e, f) Kaplan-Meier curves of OS for BC patients assigned to high- and low-risk groups in TCGA, GSE32894 and GSE13507. LASSO: least absolute shrinkage and selection operator; TCGA: The Cancer Genome Atlas;

BLCA: bladder cancer; NMIBC: non-muscle-invasive bladder cancer; MIBC: muscle-invasive bladder cancer; OS: overall survival.

Figure 3

Construction and validation of the risk score. (a) Plots of the cross-validation error rates. (b) Distribution of LASSO coefficients of OS-associated genes. (c) Coefficient values of the eight genes. (d, e, f) Kaplan-Meier curves of OS for BC patients assigned to high- and low-risk groups in TCGA, GSE32894 and

GSE13507. LASSO: least absolute shrinkage and selection operator; TCGA: The Cancer Genome Atlas; BLCA: bladder cancer; NMIBC: non-muscle-invasive bladder cancer; MIBC: muscle-invasive bladder cancer; OS: overall survival.

Figure 4

Associations between risk score and clinical variables.. (a) Clinical variables and expression levels of eight genes were compared between the low- and high-risk groups (b) Risk scores distribution, OS status of each patient, and heatmaps of eight genes in TCGA dataset. (c) Receiver operating characteristic curve showed the predictive efficiency of discrimination power between NMIBC and MIBC in STPH dataset. (d)

The distribution of risk scores were compared between NMIBC and MIBC in STPH, GSE13507 and GSE32894 datasets.*P<0.05, **P<0.01, ***P<0.001.

Figure 4

Associations between risk score and clinical variables.. (a) Clinical variables and expression levels of eight genes were compared between the low- and high-risk groups (b) Risk scores distribution, OS status of each patient, and heatmaps of eight genes in TCGA dataset. (c) Receiver operating characteristic curve showed the predictive efficiency of discrimination power between NMIBC and MIBC in STPH dataset. (d) The distribution of risk scores were compared between NMIBC and MIBC in STPH, GSE13507 and GSE32894 datasets.*P<0.05, **P<0.01, ***P<0.001.

Figure 5

Nomogram Construction based on the risk score in TCGA dataset. (a) Univariate and multivariate Cox analyses indicated that the risk score was significantly associated with OS. (b) Nomogram for predicting the probability of 1-, 3-, and 5-year OS. (c) ROC curve showed the predictive efficiency of 1-, 3-, and 5-year OS. (d) Calibration plots of the nomogram for predicting the probability of 1-, 3-, and 5-year OS. (e) DCA of

the nomogram predicting 1- 3-, and 5-year OS. OS: overall survival; ROC: Receiver operating characteristic; DCA: decision curves analyses; HR: hazard ratio; CI: confidence interval.

Figure 5

Nomogram Construction based on the risk score in TCGA dataset. (a) Univariate and multivariate Cox analyses indicated that the risk score was significantly associated with OS. (b) Nomogram for predicting the probability of 1-, 3-, and 5-year OS. (c) ROC curve showed the predictive efficiency of 1-, 3-, and 5-year

OS. (d) Calibration plots of the nomogram for predicting the probability of 1-, 3-, and 5-year OS. (e) DCA of the nomogram predicting 1- 3-, and 5-year OS. OS: overall survival; ROC: Receiver operating characteristic; DCA: decision curves analyses; HR: hazard ratio; CI: confidence interval.

Figure 6

RT-qPCR validation of eight genes in bladder cancer cell lines T24, UMUC3, 5637, EJ and J82 and normal bladder epithelial cell line SV-HUC-1

Figure 6

RT-qPCR validation of eight genes in bladder cancer cell lines T24, UMUC3, 5637, EJ and J82 and normal bladder epithelial cell line SV-HUC-1

Supplementary Files

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

- [SuppTable.docx](#)
- [SuppTable.docx](#)
- [newS1.tif](#)
- [newS1.tif](#)