Identification of muscle-invasive related genes in bladder cancer single-cell sequencing data for constructing patient prognostic model

Weizhuo Wang
Suzhou Municipal Hospital, The Affiliated Suzhou Hospital of Nanjing Medical University, Nanjing Medical University

Hengrui Chen
Suzhou Municipal Hospital, The Affiliated Suzhou Hospital of Nanjing Medical University, Nanjing Medical University

Zheng Tang
Suzhou Municipal Hospital, The Affiliated Suzhou Hospital of Nanjing Medical University, Nanjing Medical University

Fei Wang
Suzhou Municipal Hospital, The Affiliated Suzhou Hospital of Nanjing Medical University, Nanjing Medical University

Kai Li
Suzhou Municipal Hospital, The Affiliated Suzhou Hospital of Nanjing Medical University, Nanjing Medical University

Ke Zhang (876274635@qq.com)
Suzhou Municipal Hospital, The Affiliated Suzhou Hospital of Nanjing Medical University, Nanjing Medical University

Article

Keywords: Single-cell sequencing, bladder cancer, muscle-invasive, prognostic model

Posted Date: May 19th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2920456/v1

Abstract

Single-cell sequencing is an emerging sequencing technology that can effectively identify the cell types of tumors. In bladder cancer prognosis, muscular invasion often represents a poor prognosis and affects patients' quality of life. This study aims to extract the expression levels of muscle-invasive related genes(MIRGs) in bladder cancer patients and construct a model of MIRG, which can predict bladder cancer patients' prognosis using bioinformatics methods.

Methods: Single-cell sequencing data of bladder cancer patients were obtained from the GEO database. After conducting quality control and cell type identification, all epithelial cells in the samples were extracted and classified based on their invasive and non-invasive characteristics, followed by a differential analysis. The results were identified as MIRGs. Subsequently, we downloaded and organized gene data of bladder cancer patients from TCGA and determined the intersection of MIRGs and the sequenced gene set of TCGA patients. Clinical information was then associated with the intersection, and the data were divided into training and test sets, with the training set used for model construction and the test set for model verification. Subsequently, the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm and Cox regression were used to construct a prognostic model based on MIRGs. Based on the prognostic features, risk scores were calculated, and patients were classified into high-risk and low-risk groups. We observed the survival information of patients in the high-risk and low-risk groups in both the training and test sets, constructed ROC curves to assess the predictive ability of the model, and subsequently, we generated nomograms.

Results: Three cell types were identified, and epithelial cells were extracted, clustered, and divided into invasive and non-invasive groups based on pathological staging. A total of 411 differentially expressed genes were screened. GO and KEGG analyses revealed that these genes were significantly associated with cellular processes such as apoptosis, cell adhesion, and tumor development and progression.After intersecting the expressed genes, 402 genes were determined for model construction. Following the LASSO algorithm and Cox regression, a risk prediction model consisting of CD74, AKR1B1, EIF3D, EMP1, CRABP2, TRIM31, RPL36A and MRPS6 was established.Survival curves and Receiver Operating Characteristic (ROC) curves demonstrated that the model exhibited good predictive ability. A nomograms was constructed to predict patients' survival rates at 1, 3, and 5 years. The calibration curve of the nomograms indicated that it had a satisfactory prognostic ability for patients.

Conclusion: In this study, based on single-cell sequencing data, TCGA sequencing data and clinical information, the bladder cancer muscle-invasive related gene prognostic model constructed using multi-omics methods demonstrated a certain degree of accuracy and reliability in predicting the survival prognosis of bladder cancer patients. This provides a reference for assessing the prognosis of bladder cancer patients.

Introduction

Single-cell sequencing is a technique that performs genomic, transcriptomic, or epigenomic sequencing analysis on individual cells. Compared to traditional sequencing methods based on whole tissues or cell populations, single-cell sequencing can reveal heterogeneity between cells and provide detailed information about cell states, functions, and subpopulations. This method has broad application prospects in the fields of biology, developmental biology, neuroscience, oncology, and more[1]. In cancer research, the development of single-cell sequencing technology provides powerful tools for studying cell-to-cell differences, developmental processes, disease mechanisms, and potential therapeutic targets[2].

Bladder cancer(BLCA) is a malignant tumor that usually forms cancer cells on the inner wall of the bladder. Hundreds of thousands of people are diagnosed with bladder cancer worldwide each year, and its incidence is increasing[3]. In developed countries, bladder cancer is more common among older populations, while in developing countries, it is often associated with occupational environments and environmental pollution. According to statistics, men are more likely to develop bladder cancer than women, and factors such as smoking, long-term exposure to chemicals, and chronic bladder infections are also associated with the development of bladder cancer[4].

Bladder cancer patients can be divided into non-muscle-invasive and muscle-invasive types. Non-muscle-invasive bladder cancer accounts for about 60%-70% of cases limited to the bladder mucosa (Ta stage), 20%-30% involving the subepithelial connective tissue (T1 stage), and about 10% presenting as carcinoma in situ[5]. The main treatment for non-muscle-invasive bladder cancer is TURBT (transurethral resection of bladder tumor), possibly accompanied by intravesical drug instillation therapy (such as BCG or chemotherapy drugs). The prognosis for non-muscle-invasive bladder cancer is relatively good, with a five-year survival rate of approximately 70–80%. However, it should be noted that the recurrence rate of non-muscle-invasive bladder cancer is high, so regular monitoring is required. Muscle-invasive bladder cancer invades the muscle layer, including invasion of the muscle layer (T2 stage), invasion of surrounding tissues (T3 stage), invasion of surrounding organs such as the prostate, seminal vesicles, uterus, vagina, pelvic wall, and abdominal wall (T4 stage), etc. Given the invasiveness of muscle-invasive bladder cancer, timely diagnosis and treatment are crucial. Current treatments strongly recommend radical cystectomy combined with bilateral pelvic lymph node dissection and cisplatin-based neoadjuvant chemotherapy for all resectable non-metastatic muscle-invasive bladder cancer patients. The five-year survival rate for patients undergoing cystectomy alone is about 50%, and radical cystectomy can improve survival rates, making it the preferred treatment for muscle-invasive bladder cancer. Radiation therapy can be considered as part of a multimodal bladder preservation approach or as palliative treatment for patients who are not suitable for cystectomy. Chemotherapy is the preferred treatment for metastatic bladder cancer or unresectable bladder cancer patients[6]. Whether muscle invasion occurs has a significant impact on the treatment and prognosis of patients, and the treatment methods for the two types of bladder cancer differ significantly. For example, most non-muscle-invasive bladder cancers only require bladder electrocautery and instillation therapy, while muscle-invasive bladder cancers require bladder removal and surrounding tissue dissection, and even bladder reconstruction, although there is no evidence that this method improves long-term outcomes, it has a significant impact

on patients' lives[6]. About 10–20% of non-muscle-invasive bladder cancers progress to muscle-invasive bladder cancers, so people with non-muscle-invasive bladder cancer need continuous follow-up and subsequent treatment[7]. During the development of bladder cancer, existing research has revealed multiple genes, such as PIK3CA, which leads to changes in the invasive growth of bladder cancer, and alterations in the FGFR3 gene, which are effective oncogenic drivers in bladder cancer[8,9]. Therefore, it is necessary to continuously identify new genes that affect the prognosis of bladder cancer patients. Our study is the first to analyze the differential expression of muscle-invasive and non-muscle-invasive bladder cancer in the single-cell sequencing field. Single-cell data can eliminate the confounding effects of gene expression in different cell types, allowing for more accurate identification of differentially expressed genes. Here, we use relevant single-cell sequencing data, extract epithelial cell populations, and identify corresponding differentially expressed genes, and then combine clinical data to construct a prognostic model for bladder cancer patients. We aim to provide new targets and strategies for the treatment of bladder cancer.

Method

1.1 Data source

We downloaded single-cell data (GSE135337) from the GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE135337), and bulk sequencing data along with the corresponding clinical data for bladder cancer from TCGA. Single-cell analysis was performed using the R language. Prior to the analysis, single-cell data underwent quality control and selection, followed by cell type identification.

1.2 Selecting Highly Variable Genes and Determining PCA Dimensions

We utilized the R package 'Seurat' to identify highly variable genes in the single-cell data and applied the 'ElbowPlot' function to determine the number of principal components (PCs) to be used for subsequent dimensionality reduction through principal component analysis (PCA).

1.3 Identification of Epithelial Cell Types and Extraction of Differentially Expressed Genes

After dimensionality reduction of the single-cell data for bladder cancer, we classified the cells into two groups based on their pathological stage: muscle-invasive and non-muscle invasive. We used the SINGER package in R to identify the cell types and subsequently selected the epithelial cells. Genes with $|\log2FC| > 0.5$ and $p < 0.05$ were considered differentially expressed.

1.4 KEGG and GO Analysis of Differentially Expressed Genes

For the obtained muscle-invasive related genes, we used the KEGG rest API (https://www.kegg.jp/kegg/rest/keggapi.html) to obtain the latest KEGG Pathway gene annotations as background, and mapped the genes to the background set. We performed enrichment analysis using the R package clusterProfiler (version 3.14.3) to obtain the gene set enrichment results. We set the minimum gene set to 5 and the maximum gene set to 5000. $P < 0.05$ and $FDR < 0.25$ were considered as significantly different. We further explored the potential molecular functions and cellular components involved in these genes using Gene Ontology (GO) analysis, with the filtering criteria set to $p < 0.01$ and $q < 0.01$.

1.5 Establishing a Risk Prediction Model and Survival Analysis

First, we used univariate Cox regression analysis to evaluate the prognostic value of muscle-invasive related genes and obtained differentially expressed genes related to muscle invasion. Then, we used the least absolute shrinkage and LASSO regression to select predictive variables and avoid overfitting. We then performed multivariate Cox regression analysis to determine the final candidates involved in the risk model. A risk signature was constructed based on muscle-invasive related genes to predict the prognosis of BLCA patients. The calculation method for risk score is as follows:

Riskscore = Σcoef gene_i × gene_i expression level, where the risk value is obtained by weighting the regression coefficients (coef) and expression levels of muscle-invasive genes. Based on the median risk score, BLCA patients in the TCGA dataset were divided into high-risk and low-risk groups.

1.6 Risk Prediction Model Assessment and Nomogram

The correlation analysis between risk score and clinical features was performed by comparing gender, age, American Joint Committee on Cancer (AJCC) stage, and the relationship between risk score, gene expression, and patient subsets. Subsequently, we compared whether there were significant differences in survival time and status between high-risk and low-risk patients with the same clinical pathological features, and used R software to plot Kaplan-Meier (K-M) survival curves. We then drew Receiver Operating Characteristic (ROC) curves to observe the diagnostic value of the model. We also further assessed the prognostic value of the model by grouping patients according to stage (stages - and -) and used R software to plot K-M survival curves. Meanwhile, we downloaded patients' progression-free survival time to observe the accuracy of the model in predicting progression-free survival. Based on the ROC curve evaluation, we incorporated clinical data with a significant impact on patient prognosis, constructed a nomogram together with the risk score to predict patient survival rates at 1, 3, and 5 years, and built calibration curves to evaluate the predictive ability of the nomogram.

Results

2.1 Data source

We analyzed single-cell sequencing data from 7 bladder cancer patients using the R language. The criteria for cell inclusion were: more than 300 genes sequenced, mitochondrial gene percentage less than 10%, and red blood cell gene percentage less than 5%. In total, 36,787 patient cells were included (Fig. 1)[10]. Meanwhile, we organized the expression data of bladder cancer patients in the TCGA dataset, with a total of 392 patients. The detailed information can be found in Table 2.

2.2 Identifying Highly Variable Genes and PCA Dimensionality Reduction

We used the NormalizeData and ScaleData functions to standardize and centralize the cell data. Then, we integrated the 7 single-cell datasets using an anchor-based approach, selecting 2000 genes for integration. After integration, we used the FindVariableFeatures function to search for highly variable genes in the single-cell data, setting the number of highly variable genes to 3000. Next, we performed Principal Component Analysi(PCA)on the highly variable genes and generated an ElbowPlot to determine the number of principal components to be retained in the PCA dimensionality reduction process. Typically, as the number of principal components increases, the proportion of explained variance gradually decreases. When the explained variance ratio has an obvious inflection point in the graph, it can be considered as an appropriate number of principal components. This point is called the "elbow." We selected pca = 20 as the subsequent analysis choice. After selecting the appropriate PC Figure 2a , we then used the FindNeighbors and FindClusters functions to classify the cells into clusters and finally used t-SNE for dimensionality reduction and visualization. Figure 2b is the t-SNE plot grouped by source after integration, and Fig. 2c is the t-SNE plot grouped by cell clusters after integration. After cell cluster identification, we used the HumanPrimaryCellAtlasData dataset from the SingleR package in R[11], which has already identified cell types, to determine the types of our cells. We finally identified 3 cell types (Fig. 2d), which are epithelial cells, endothelial cells, and macrophages[1].

2.3 Cell Clustering and Cell Type Identification

As the histological origin of bladder cancer is predominantly epithelial cells, we subsequently isolated the epithelial cells, resulting in a total of 36,169 cells. We displayed the cell cluster classification before extraction and the epithelial cell extraction results in a t-SNE plot(Figure 3a). Using the pathological information of seven patients(Supplementary Table 1), we divided the epithelial cell clusters into muscle-invasive and non-muscle-invasive groups Figure 3b . We then employed the FindAllMarkers function to extract feature genes, with a selection criteria of p,adj < 0.05 and |log2FC| > 0.5. A total of 411 feature genes were extracted. We have listed the names and expression patterns of the top 20 genes with the highest log2FC differential expression (Table 1).

Table 1
Top 20 differentially expressed genes and their expression levels in muscle-invasive and non-muscle-invasive bladder cancer cells.

Gene	Avg_log2FC	Pct.invasion	Pct.non	P_val_adj
S100A8	4.151861311	0.48	0.055	0
S100A7	3.202385736	0.279	0.006	0
S100A9	2.777596415	0.685	0.336	0
MT1X	2.610142079	0.562	0.35	0
DMKN	2.546192707	0.678	0.535	0
IFI27	2.406189457	0.998	0.758	0
C1orf56	2.244334995	0.482	0.366	0
IGKC	2.201257168	0.603	0.119	0
ZFAND2A	2.195416493	0.678	0.239	0
RAB21	2.159856744	0.506	0.343	0
CCND1	2.048081225	0.911	0.603	0
TMEM19	2.026838153	0.446	0.263	0
DEFB1	2.011937288	0.62	0.13	0
PHLDA2	1.94197265	0.979	0.771	0
MDM2	1.8867406	0.497	0.39	0
HSPB1	1.81725058	1	0.998	0
PI3	1.78192133	0.219	0.005	0
LEAP2	1.760338816	0.445	0.138	0
INSIG1	1.636347151	0.625	0.197	0

2.4 GO analysis and KEGG analysis of genes associated with muscle layer invasion

We performed gene enrichment analysis and KEGG pathway analysis of the muscle-invasive -related genes(MIRGs) using R. Biological processes (BP), cellular components (CC), and molecular functions (MF) were analyzed. In BP analysis(Figure 4a), we found enrichment of processes such as cytoplasmic translation, regulation of apoptotic signaling pathway, and intrinsic apoptotic signaling pathway. In CC analysis Figure 4b , we observed enrichment of components such as cell-substrate junction, focal adhesion, and ribosome. In MF analysis Figure 4c , we found enrichment of functions such as structural

constituent of ribosome, ubiquitin-like protein ligase binding, and ubiquitin protein ligase binding.For KEGG pathway analysis Figure 4d , we found that in addition to Pathways in cancer, the enriched pathways also included Proteoglycans in cancer, IL-17 signaling pathway, p53 signaling pathway, NF-kappa B signaling pathway, and AMPK signaling pathway. These results indicate that the differentially expressed genes are mainly related to cell apoptosis, cell adhesion, and intercellular interactions. In other words, the activity and connectivity of cells and intercellular connections in muscle-invasive bladder cancer cells are different from those in non-muscle-invasive bladder cancer cells. Furthermore, the KEGG analysis showed that these cellular activities are closely related to the occurrence and development of tumors, as IL-17 signaling pathway, p53 signaling pathway, NF-kappa B signaling pathway, and AMPK signaling pathway have been found to play a role in the development of various tumors[12–15].

2.5 Construction of Prognostic Model for Muscle Infiltration-Related Genes Using TCGA Expression Data

Clinical data, including age, gender, pT stage, pN stage, pM stage, AJCC stage, survival status, and survival time of TCGA patients were included (Table 2). Ultimately, a total of 392 patients with both clinical survival information and sample sequencing information were recruited and randomly divided into training set (n = 186,Supplementary table2) and test set (n = 186,Supplementary table3) using the "caret" package in R language. Next, we identified muscle infiltration-related genes and intersected them with the gene expression files in TCGA, obtaining 402 intersection genes. We used these genes to correlate with clinical data and normalized their FPKM values by taking log2 + 1(Supplementary table4). Subsequently, we used univariate screening to identify genes related to patient clinical prognosis, and employed the Least absolute shrinkage and selection operator(LASSO)for variable shrinkage(Fig. 5A, B). Ultimately, we constructed a patient prognosis model composed of eight genes using multivariate Cox regression, including CD74, AKR1B1, EIF3D, EMP1, CRABP2, TRIM31, RPL36A, and MRPS6. The corresponding model coefficients, Hazard ratios, and P values are shown in Fig. 5C,D and Supplementary table5. Patients were further divided into high-risk and low-risk groups according to the median value of their risk scores. Heatmaps were created to show the expression differences of muscle infiltration-related genes between the high-risk and low-risk groups. The results showed that the expression of the eight genes in the high-risk group was different from that in the low-risk group in both the training and tset sets. Subsequently, scatter plots and risk curves were used to display the survival status and risk scores of each bladder cancer patient. Additionally, the K-M survival curves for the test set, test set, and overall population showed significant differences in overall survival rates between high-risk and low-risk patients ($P < 0.05$), with higher mortality rates and hazard ratios in the high-risk group than in the low-risk group (Fig. 6).

Table 2
Clinical information

Covariates	Type	Total	Test	Train	Pvalue
Age	<=65	157(40.05%)	74(37.76%)	83(42.35%)	0.4096
	> 65	235(59.95%)	122(62.24%)	113(57.65%)	
Gender	female	103(26.28%)	58(29.59%)	45(22.96%)	0.1685
	male	289(73.72%)	138(70.41%)	151(77.04%)	
M	M0	188(47.96%)	89(45.41%)	99(50.51%)	0.1151
	M1	11(2.81%)	2(1.02%)	9(4.59%)	
	unknow	193(49.23%)	105(53.57%)	88(44.9%)	
N	N0	227(57.91%)	109(55.61%)	118(60.2%)	0.8918
	N1	46(11.73%)	25(12.76%)	21(10.71%)	
N	N2	74(18.88%)	36(18.37%)	38(19.39%)	
	N3	6(1.53%)	3(1.53%)	3(1.53%)	
	unknow	39(9.95%)	23(11.73%)	16(8.16%)	
Stage	Stage II	124(31.63%)	64(32.65%)	60(30.61%)	0.884
	Stage III	136(34.69%)	66(33.67%)	70(35.71%)	
	Stage IV	132(33.67%)	66(33.67%)	66(33.67%)	
T	T2	115(29.34%)	55(28.06%)	60(30.61%)	0.846
	T3	190(48.47%)	94(47.96%)	96(48.98%)	
	T4	56(14.29%)	30(15.31%)	26(13.27%)	
	unknown	31(7.91%)	17(8.67%)	14(7.14%)	

2.6 Validation of the Risk Model and Nomogram

To evaluate whether the model score, age, gender, and pathological grade are independent prognostic factors for bladder cancer patients, we obtained forest plots through univariate and multivariate Cox regression analyses. The results showed that both AJCC staging and risk scores were independent prognostic factors for bladder cancer patients ($P < 0.05$) (Fig. 7A, B), indicating that our prognostic model is an independent factor for patient prognosis and has diagnostic value, regardless of other clinical characteristics.To assess the accuracy of risk scores and clinical features in predicting the prognosis of bladder cancer patients, we plotted clinical information ROC curves and time-dependent ROC curves for

the training set, test set, and the entire cohort (Fig. 8A,B,C). The training set showed Area Under Curve (AUC) values of 0.780, 0.805, and 0.821 for 1-year, 3-year, and 5-year risk scores, respectively. The test set showed AUC values of 0.735, 0.632, and 0.635 for 1-year, 3-year, and 5-year risk scores, respectively. The overall patient population showed AUC values of 0.746, 0.721, and 0.724 for 1-year, 3-year, and 5-year risk scores, respectively(Fig. 8D,E,F). Additionally, we plotted ROC curves predicting the 1-year survival probability for the training set, test set, and the entire cohort, including all clinical information and the model. The largest AUC value was observed for the model, suggesting that our model has greater diagnostic value for prognosis compared to other clinical characteristics. Moreover, there was no correlation between gender, age, and grade and risk scores of bladder cancer patients ($P < 0.05$). To further study the prognostic value of the model, we grouped patients by stage (I-II and III-IV) (Fig. 9A, B). K-M survival curve analysis showed that patients with higher risk values had shorter survival times and poorer prognosis for patients with the same AJCC stage of bladder cancer ($P < 0.05$). This indicates that our model has diagnostic value for predicting patient prognosis. We also performed an analysis of progression-free survival based on patient data (Fig. 9C). The results showed that the progression-free survival time of high-risk patients was significantly longer than that of low-risk patients, indicating that the expression of model genes may have an impact on patients' progression-free survival period. Finally, we found that age, stage, and risk score had an impact on patient prognosis, so we constructed a nomogram based on these three indicators to score patients and predict their one-year, three-year, and five-year survival probabilities. The calibration curve of the nomogram showed that the predicted values were close to the actual survival probabilities, indicating that the nomogram and overall model have good diagnostic value(Fig. 9D,E).

Discusion

Single-cell sequencing is a high-throughput sequencing technology that allows for the determination of the genome, transcriptome, or epigenome at the single-cell level. Single-cell sequencing can resolve heterogeneity among different cell types within cancer tissue and provide a deeper understanding of the tumor microenvironment and intercellular interactions. It can determine the cell type, such as epithelial cells, endothelial cells, macrophages, T cells, and more, which is important for cancer therapy. Through single-cell sequencing, personalized genomic information can be provided for patients, which helps identify drug targets and achieve precise treatment. Despite the many advantages of single-cell sequencing in cancer research, there are also limitations such as high cost and complex data processing. Therefore, combining multiple techniques for cancer research can help better understand the biological characteristics of tumors[16].

Bladder cancer is a common malignant tumor that can be divided into invasive and non-invasive types. The clinical treatment methods for the two types differ significantly. Because invasive bladder cancer invades deeper tissues, more invasive treatment methods, such as surgery, radiotherapy, and chemotherapy, are usually required. Non-invasive bladder cancer is typically treated with local methods, such as transurethral resection and photodynamic therapy. Although the treatment methods for the two types differ significantly, the cost of treatment is still substantial. In addition, the prognosis of invasive

bladder cancer is usually worse than that of non-invasive bladder cancer[7]. This is because invasive bladder cancer is more likely to spread to lymph nodes and other organs, leading to recurrence and metastasis. The prognosis of non-invasive bladder cancer is usually better, but recurrence and metastasis can also occur. Moreover, the treatment of bladder cancer is often accompanied by a large amount of follow-up and subsequent treatment, which imposes a heavy economic burden on patients. For patients with non-invasive bladder cancer, an effective target that can identify their prognosis is needed, which can roughly predict the progression of their disease and adjust treatment accordingly. For patients with invasive bladder cancer, factors affecting their survival rate need to be observed to determine whether they are suitable for more aggressive surgeries, such as radical cystectomy[6]. In recent years, with the continuous development of single sequencing technology, especially emerging single-cell sequencing, more specific genes can be discovered. In addition, there have been significant advances in the research of immunotherapy and targeted therapy for bladder cancer. These treatment methods can target specific cancer cell molecules or immune cells for targeted therapy, thereby improving treatment effectiveness and prognosis. Bladder cancer tissue usually originates from epithelial cells[3], so this study utilizes single-cell sequencing data to extract differentially expressed genes to eliminate interference from other cell types.

Using multi-omics analysis, this study utilized bladder cancer single-cell sequencing data to perform quality control, high variance gene selection, cell clustering, and cell type identification. All epithelial cells were grouped based on patient T staging, resulting in the identification of 411 differentially expressed genes, which were subjected to KEGG and GO analysis. Subsequently, we took the intersection of these genes and those expressed in TCGA patient data, resulting in 402 differentially expressed genes. These genes were further analyzed in conjunction with clinical information to investigate patient prognosis, ultimately revealing eight key prognostic genes closely related to bladder cancer: CD74, AKR1B1, EIF3D, EMP1, CRABP2, TRIM31, RPL36A and MRPS6. Based on these eight genes, a prognosis model was constructed and a line chart was created, and survival analysis, ROC curves, and the calibration curve of the line chart all indicated the accuracy of the model.

Among the genes involved in the model construction, the protein encoded by CD74 is related to the major histocompatibility complex class II (MHC) and serves as an important partner in regulating antigen presentation in immune response. It also acts as a cell surface receptor for macrophage migration inhibitory factor (MIF), which, when bound to the encoded protein, initiates survival pathways and cell proliferation. CD74 has been found to be significantly correlated with better prognosis in immune-related diseases and hepatocellular carcinoma[17], consistent with the negative coefficient of CD74 in the model indicating that high expression of CD74 would lower patients' risk scores. AKR1B1 encodes a member of the aldo/keto reductase superfamily, encompassing over 40 known enzymes and proteins. In lung cancer, its expression has been found to suppress de novo glutathione synthesis, thereby overcoming acquired resistance to EGFR-targeted therapy[18]. It is also considered a prognostic marker for endometrial cancer and is closely associated with the development of various cancers, such as colorectal and cervical cancer[19,20].The protein encoded by EIF3D is the major RNA binding subunit of the Eukaryotic translation

initiation factor-3 (eIF3) complex. In gallbladder cancer, EIF3D promotes disease progression by stabilizing GRK2 kinase and activating the PI3K-AKT signaling pathway[21]. It is also involved in poor prognosis of various cancers, such as lung adenocarcinoma and gastric cancer[22,23]. EMP1 is involved in bleb assembly and cell death, and is located in the plasma membrane. In colorectal cancer, cells with high EMP1 expression are considered the source of metastatic recurrence[24]. Furthermore, it has been found to play a role in the invasion and metastasis of various tumors, such as promoting the proliferation and invasion of ovarian cancer cells via the activation of the MAPK pathway[25], and regulating the proliferation, migration, and stemness of glioma cells through PI3K-AKT signaling and CD44[26]. This aligns with our study as the HR value indicates EMP1 as a factor for poor prognosis in TCGA patients. CRABP2 encodes a member of the retinoic acid (RA, a form of vitamin A) binding protein family and the lipocalin/cytosolic fatty-acid binding protein family. It has been identified as a novel biomarker for high-risk endometrial cancer[27]. In breast cancer, it is thought to regulate invasion and metastasis via an ER-dependent hippocampal pathway[28]. Overexpression in ovarian cancer has been found to suppress apoptosis, promote cell invasion, and increase the expression of epithelial-mesenchymal transition (EMT) markers, with transfection of si-CRABP2 having the opposite effect[29]. TRIM31 encodes a protein that functions as an E3 ubiquitin-protein ligase. This gene exhibits altered expression in certain tumors and may act as a negative regulator of cell growth. Loss of TRIM31 promotes breast cancer progression by regulating K48- and K63-linked ubiquitination of p53[30]. In this model, the HR value for TRIM31 is negative, which mirrors its role in breast cancer research. In breast cancer, muscle infiltration also indicates a poor prognosis, suggesting that TRIM31 could be a key target in the progression of infiltrating muscle cancer.RPL36A encodes a ribosomal protein that is a component of the 60S subunit. Sharing sequence similarity with yeast ribosomal protein L44, it belongs to the L44E (L36AE) family of ribosomal proteins. Studies suggest that RPL36A could serve as a prognostic marker for hepatocellular and renal cell carcinoma[31,32].The protein expressed by MRPS6 participates in the construction of mammalian mitochondrial ribosomal proteins and has been found to be highly expressed in breast cancer, being linked to poor prognosis in patients[33]. Knockdown of MRPS6 has also been shown to decrease the proliferation of breast cancer cells[34].

In summary, we identified muscle infiltration-related genes by analyzing single-cell sequencing data from bladder cancer patients, and subsequently constructed a prognostic model by linking TCGA bladder cancer second-generation sequencing data and clinical data, providing a promising avenue for personalized survival and clinical outcome prediction for bladder cancer patients. However, this study has certain limitations as it is based on public databases and has not been verified by clinical trials or basic research. The eight genes identified in this study, CD74, AKR1B1, EIF3D, EMP1, CRABP2, TRIM31, RPL36A and MRPS6, have been found to be involved in tumor invasion and metastasis in multiple cancers, often associated with patient prognosis. However, there is still limited research focused specifically on bladder cancer, and our future research will focus on these eight genes to uncover their key biological markers and therapeutic targets for bladder cancer muscle infiltration.

Declarations

DATA AVAILABILITY

All the data used in this study can be retrieved from public databases. GSE135337 is obtained from the Gene Expression Omnibus (GEO) database, accessible at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE135337. TCGA data are obtained from The Cancer Genome Atlas, available at https://portal.gdc.cancer.gov/.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS STATEMENT

WeizhuoWang: Conceptualization, Funding acquisition Investigation, Data curation, Writing – original draft, Formal analysis. **HengruiChen**: Investigation, Data curation, Writing – original draft. **ZhengTang**: Investigation, Data curation, Methodology. **FeiWang**: Investigation, Validation, Visualization. : Investigation, Data curation. **KaiLi**: Investigation. **KeZhang**:Investigation, Methodology.

References

1. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):1–14. doi:10.1038/s12276-018-0071-8
2. Lei Y, Tang R, Xu J, et al. Applications of single-cell sequencing in cancer research: progress and perspectives. J Hematol OncolJ Hematol Oncol. 2021;14(1):91. doi:10.1186/s13045-021-01105-2
3. Dobruch J, Oszczudłowski M. Bladder Cancer: Current Challenges and Future Directions. Med Kaunas Lith. 2021;57(8):749. doi:10.3390/medicina57080749
4. Rozanec JJ, Secin FP. [Epidemiology, etiology and prevention of bladder cancer.]. Arch Esp Urol. 2020;73(10):872–878.
5. Babjuk M, Burger M, Capoun O, et al. European Association of Urology Guidelines on Non-muscle-invasive Bladder Cancer (Ta, T1, and Carcinoma in Situ). Eur Urol. 2022;81(1):75–94. doi:10.1016/j.eururo.2021.08.010
6. Witjes JA, Bruins HM, Cathomas R, et al. European Association of Urology Guidelines on Muscle-invasive and Metastatic Bladder Cancer: Summary of the 2020 Guidelines. Eur Urol. 2021;79(1):82–104. doi:10.1016/j.eururo.2020.03.055
7. Cathomas R, Lorch A, Bruins HM, et al. The 2021 Updated European Association of Urology Guidelines on Metastatic Urothelial Carcinoma. Eur Urol. 2022;81(1):95–103. doi:10.1016/j.eururo.2021.09.026

8. Jing W, Wang G, Cui Z, et al. FGFR3 Destabilizes PD-L1 via NEDD4 to Control T-cell-Mediated Bladder Cancer Immune Surveillance. Cancer Res. 2022;82(1):114–129. doi:10.1158/0008-5472.CAN-21-2362
9. Aleksakhina SN, Imyanitov EN. Cancer Therapy Guided by Mutation Tests: Current Status and Perspectives. Int J Mol Sci. 2021;22(20):10931. doi:10.3390/ijms222010931
10. Balzer MS, Ma Z, Zhou J, Abedini A, Susztak K. How to Get Started with Single Cell RNA Sequencing Data Analysis. J Am Soc Nephrol JASN. 2021;32(6):1279–1292. doi:10.1681/ASN.2020121742
11. Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–172. doi:10.1038/s41590-018-0276-y
12. Li X, Bechara R, Zhao J, McGeachy MJ, Gaffen SL. IL-17 receptor-based signaling and implications for disease. Nat Immunol. 2019;20(12):1594–1602. doi:10.1038/s41590-019-0514-y
13. Hernández Borrero LJ, El-Deiry WS. Tumor suppressor p53: Biology, signaling pathways, and therapeutic targeting. Biochim Biophys Acta Rev Cancer. 2021;1876(1):188556. doi:10.1016/j.bbcan.2021.188556
14. Peng C, Ouyang Y, Lu N, Li N. The NF-κB Signaling Pathway, the Microbiota, and Gastrointestinal Tumorigenesis: Recent Advances. Front Immunol. 2020;11:1387. doi:10.3389/fimmu.2020.01387
15. Ciccarese F, Zulato E, Indraccolo S. LKB1/AMPK Pathway and Drug Response in Cancer: A Therapeutic Perspective. Oxid Med Cell Longev. 2019;2019:8730816. doi:10.1155/2019/8730816
16. Suvà ML, Tirosh I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. Mol Cell. 2019;75(1):7–12. doi:10.1016/j.molcel.2019.05.003
17. Xiao N, Li K, Zhu X, et al. CD74 + macrophages are associated with favorable prognosis and immune contexture in hepatocellular carcinoma. Cancer Immunol Immunother CII. 2022;71(1):57–69. doi:10.1007/s00262-021-02962-z
18. Zhang KR, Zhang YF, Lei HM, et al. Targeting AKR1B1 inhibits glutathione de novo synthesis to overcome acquired resistance to EGFR-targeted therapy in lung cancer. Sci Transl Med. 2021;13(614):eabg6428. doi:10.1126/scitranslmed.abg6428
19. Hojnik M, Šuster NK, Smrkolj Š, et al. AKR1B1 as a Prognostic Biomarker of High-Grade Serous Ovarian Cancer. Cancers. 2022;14(3):809. doi:10.3390/cancers14030809
20. Syamprasad NP, Rajdev B, Jain S, et al. Pivotal role of AKR1B1 in pathogenesis of colitis associated colorectal carcinogenesis. Int Immunopharmacol. 2023;119:110145. doi:10.1016/j.intimp.2023.110145
21. Zhang F, Xiang S, Cao Y, et al. EIF3D promotes gallbladder cancer development by stabilizing GRK2 kinase and activating PI3K-AKT signaling pathway. Cell Death Dis. 2017;8(6):e2868. doi:10.1038/cddis.2017.263
22. He J, Wang X, Cai J, Wang W, Qin X. High expression of eIF3d is associated with poor prognosis in patients with gastric cancer. Cancer Manag Res. 2017;9:539–544. doi:10.2147/CMAR.S142324

23. Wang D, Jia Y, Zheng W, Li C, Cui W. Overexpression of eIF3D in Lung Adenocarcinoma Is a New Independent Prognostic Marker of Poor Survival. Dis Markers. 2019;2019:6019637. doi:10.1155/2019/6019637
24. Cañellas-Socias A, Cortina C, Hernando-Momblona X, et al. Metastatic recurrence in colorectal cancer arises from residual EMP1 + cells. Nature. 2022;611(7936):603–613. doi:10.1038/s41586-022-05402-9
25. Liu Y, Ding Y, Nie Y, Yang M. EMP1 Promotes the Proliferation and Invasion of Ovarian Cancer Cells Through Activating the MAPK Pathway. OncoTargets Ther. 2020;13:2047–2055. doi:10.2147/OTT.S240028
26. Wang J, Li X, Wu H, et al. EMP1 regulates cell proliferation, migration, and stemness in gliomas through PI3K-AKT signaling and CD44. J Cell Biochem. 2019;120(10):17142–17150. doi:10.1002/jcb.28974
27. Egan D, Moran B, Wilkinson M, et al. CRABP2 - A novel biomarker for high-risk endometrial cancer. *Gynecol Oncol*. Published online September 23, 2022:S0090-8258(22)01840-6. doi:10.1016/j.ygyno.2022.09.020
28. Feng X, Zhang M, Wang B, et al. CRABP2 regulates invasion and metastasis of breast cancer through hippo pathway dependent on ER status. J Exp Clin Cancer Res CR. 2019;38(1):361. doi:10.1186/s13046-019-1345-2
29. Xie T, Tan M, Gao Y, Yang H. CRABP2 accelerates epithelial mesenchymal transition in serous ovarian cancer cells by promoting TRIM16 methylation via upregulating EZH2 expression. Environ Toxicol. 2022;37(8):1957–1967. doi:10.1002/tox.23542
30. Guo Y, Li Q, Zhao G, et al. Loss of TRIM31 promotes breast cancer progression through regulating K48- and K63-linked ubiquitination of p53. Cell Death Dis. 2021;12(10):945. doi:10.1038/s41419-021-04208-3
31. Song MJ, Jung CK, Park CH, et al. RPL36 as a prognostic marker in hepatocellular carcinoma. Pathol Int. 2011;61(11):638–644. doi:10.1111/j.1440-1827.2011.02716.x
32. Zhong W, Huang C, Lin J, et al. Development and Validation of Nine-RNA Binding Protein Signature Predicting Overall Survival for Kidney Renal Clear Cell Carcinoma. Front Genet. 2020;11:568192. doi:10.3389/fgene.2020.568192
33. Lin X, Guo L, Lin X, Wang Y, Zhang G. Expression and prognosis analysis of mitochondrial ribosomal protein family in breast cancer. Sci Rep. 2022;12(1):10658. doi:10.1038/s41598-022-14724-7
34. Oviya RP, Gopal G, Shirley SS, Sridevi V, Jayavelu S, Rajkumar T. Mitochondrial ribosomal small subunit proteins (MRPS) MRPS6 and MRPS23 show dysregulation in breast cancer affecting tumorigenic cellular processes. Gene. 2021;790:145697. doi:10.1016/j.gene.2021.145697

Figures

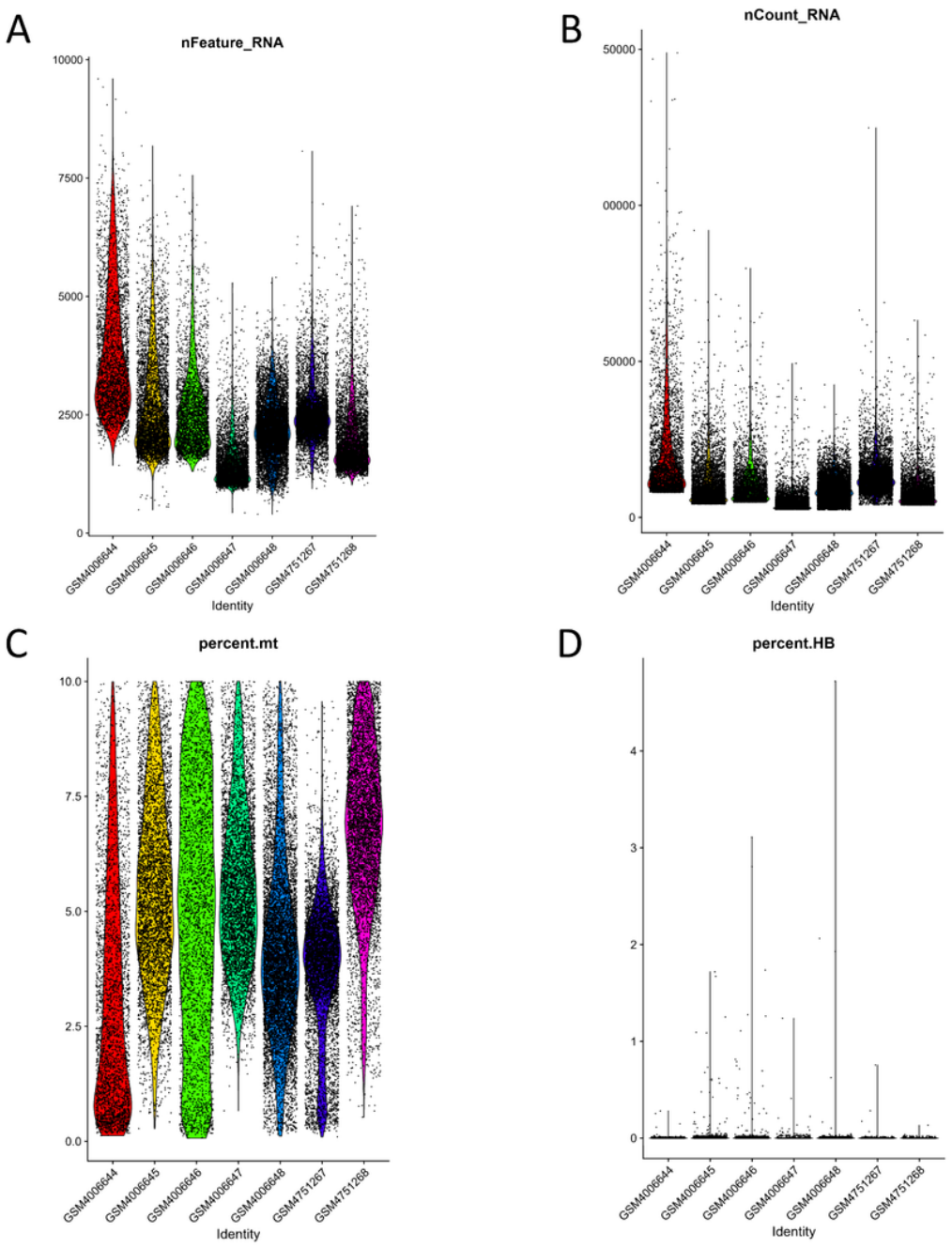

Figure 1

(A) Number of genes detected in the 7 samples; (B) Number of sequencing counts obtained in the 7 samples; (C) Mitochondrial gene ratio in the cells of the 7 samples; (D) Red blood cell gene ratio in the 7 samples.

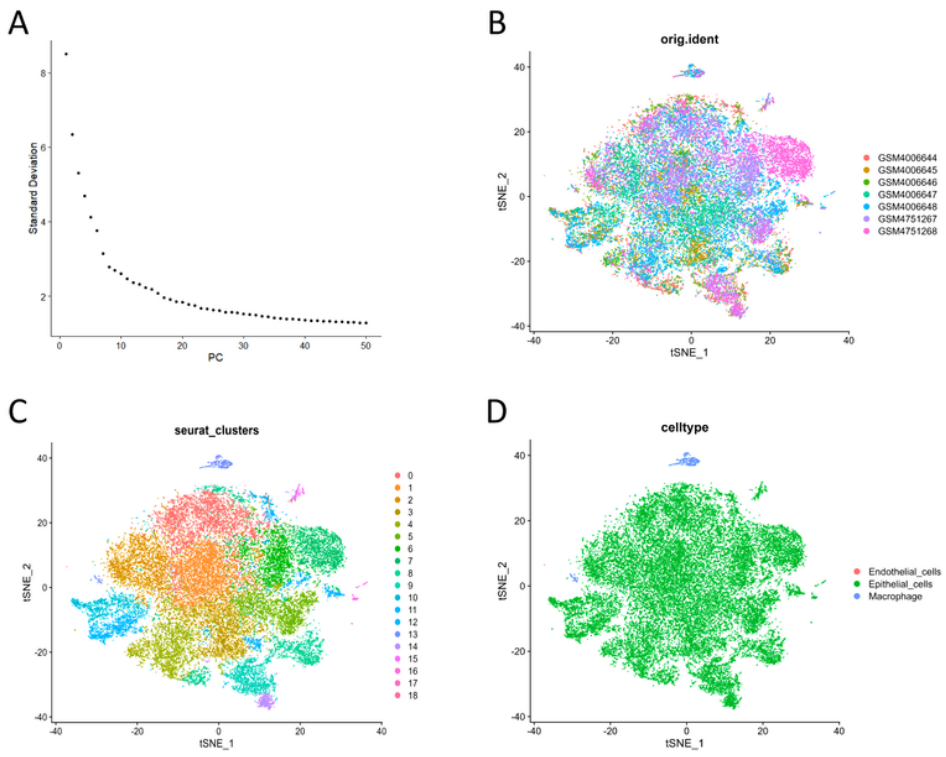

Figure 2

(A) Elbow plot of variance explained vectors.(B) t-SNE visualization of single cells colored by the origin of the samples.(C) t-SNE visualization of single cells colored by cell clusters.(D) t-SNE visualization of single cells colored by cell type identification.

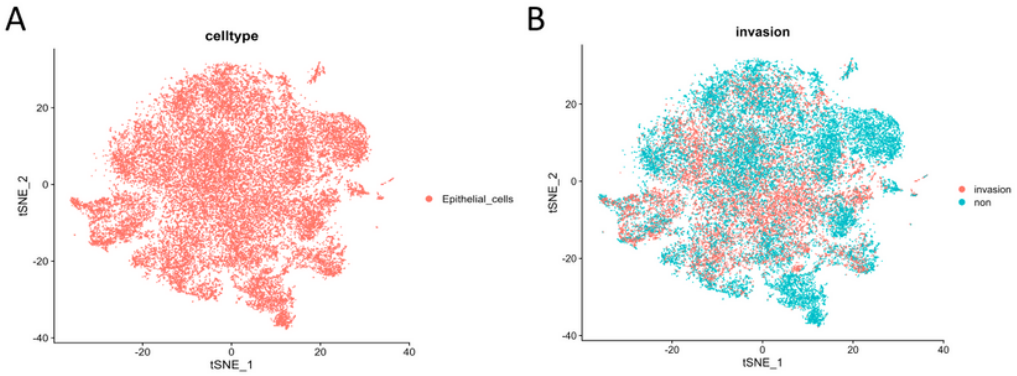

Figure 3

(A) Epithelial cells individually extracted in t-SNE dimensionality reduction; (B) t-SNE dimensionality reduction displayed according to the presence or absence of muscle invasion.

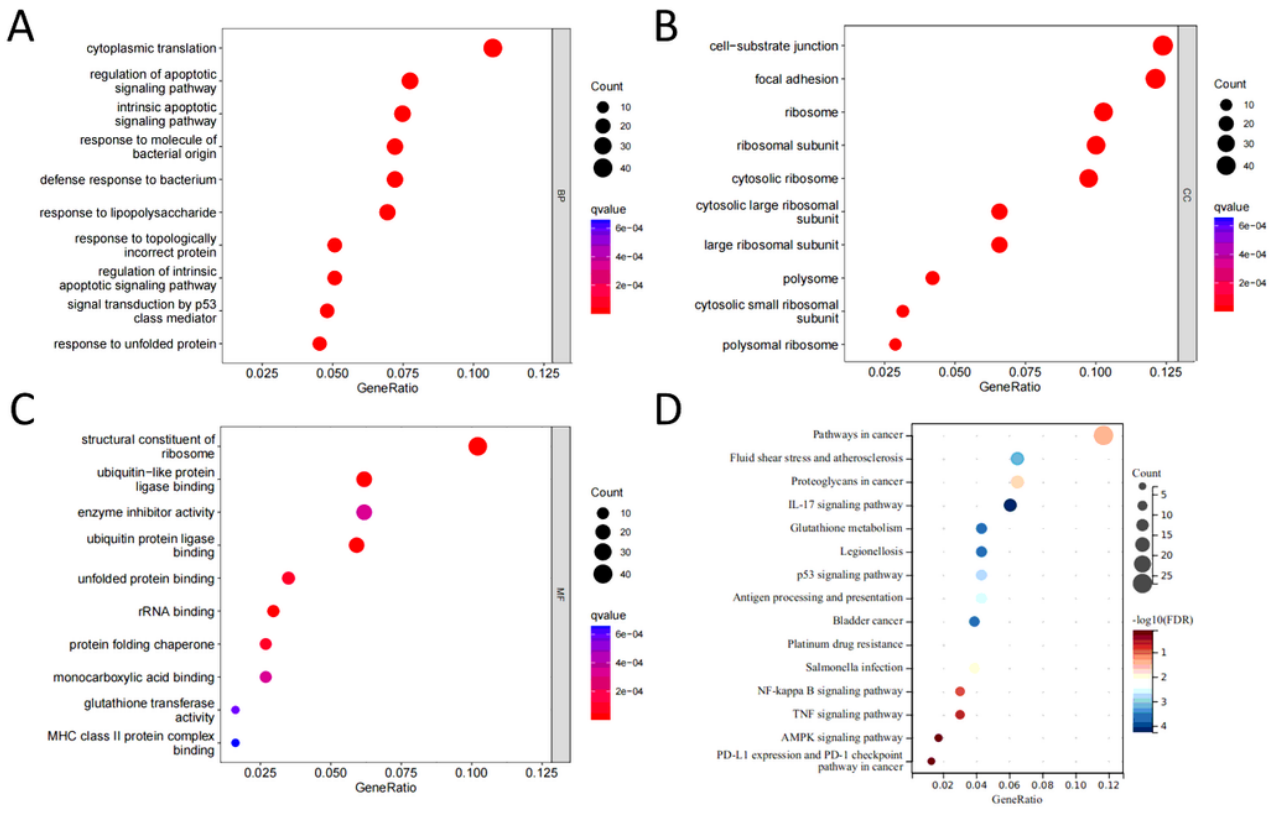

Figure 4

(A) identifies the BP in GO analysis, (B) identifies the CC in GO analysis, (C) identifies the MF in GO analysis, (D) KEGG analysis results.

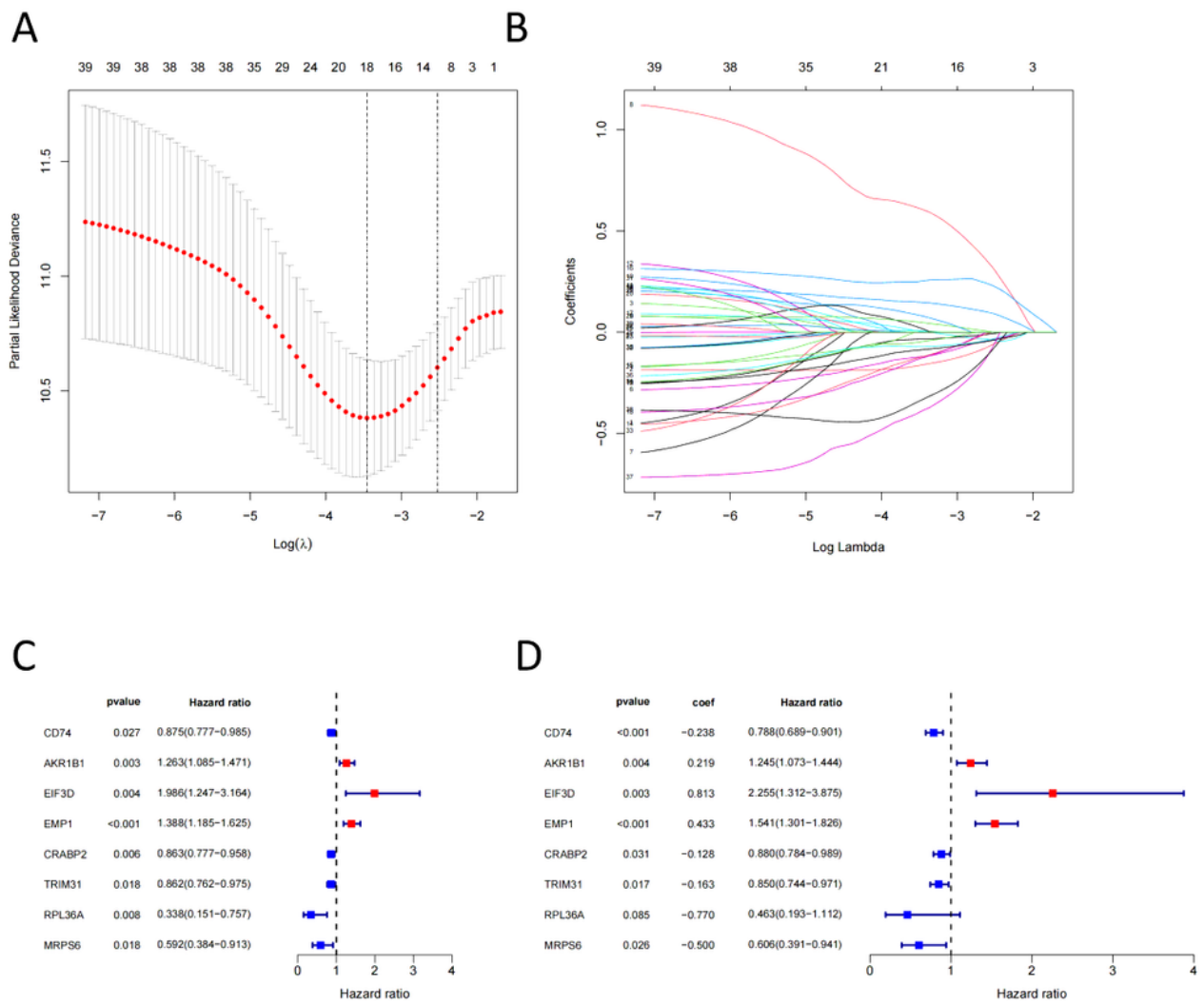

Figure 5

(A,B) The number of genes selected by LASSO regression; (C) Single-factor Cox regression; (D) Multi-factor Cox regression and model coefficients.

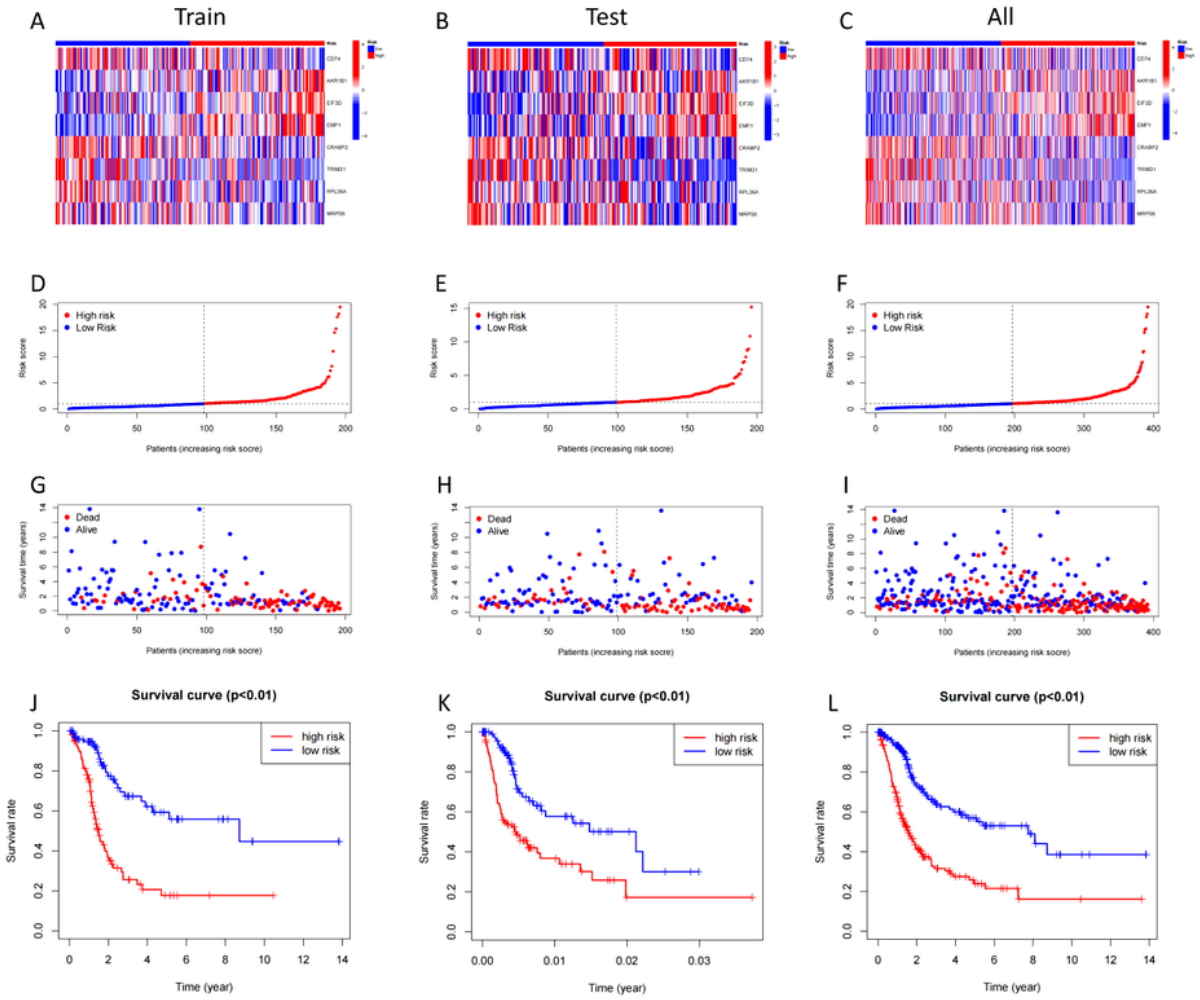

Figure 6

(A, D, G, J) Heatmaps of gene expression levels in the training set across risk score, Distribution of risk scores, Scatter plot of risk scores and survival status, Kaplan-Meier curves of high- and low-risk groups. (B, E, H, K) Heatmaps of gene expression levels in the training set across risk score, Distribution of risk scores, Scatter plot of risk scores and survival status, Kaplan-Meier curves of high- and low-risk groups. (C, F, I, L) Heatmaps of gene expression levels in the entire cohort across risk score, Distribution of risk

scores in the training set, Scatter plot of risk scores and survival status, Kaplan-Meier curves of high- and low-risk groups.

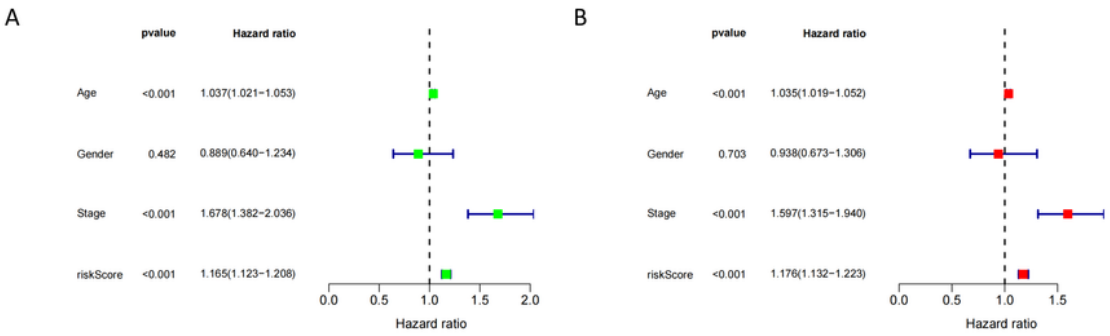

Figure 7

(A) Univariate analysis of independent prognostic factors, (B) Multivariate analysis of independent prognostic factors

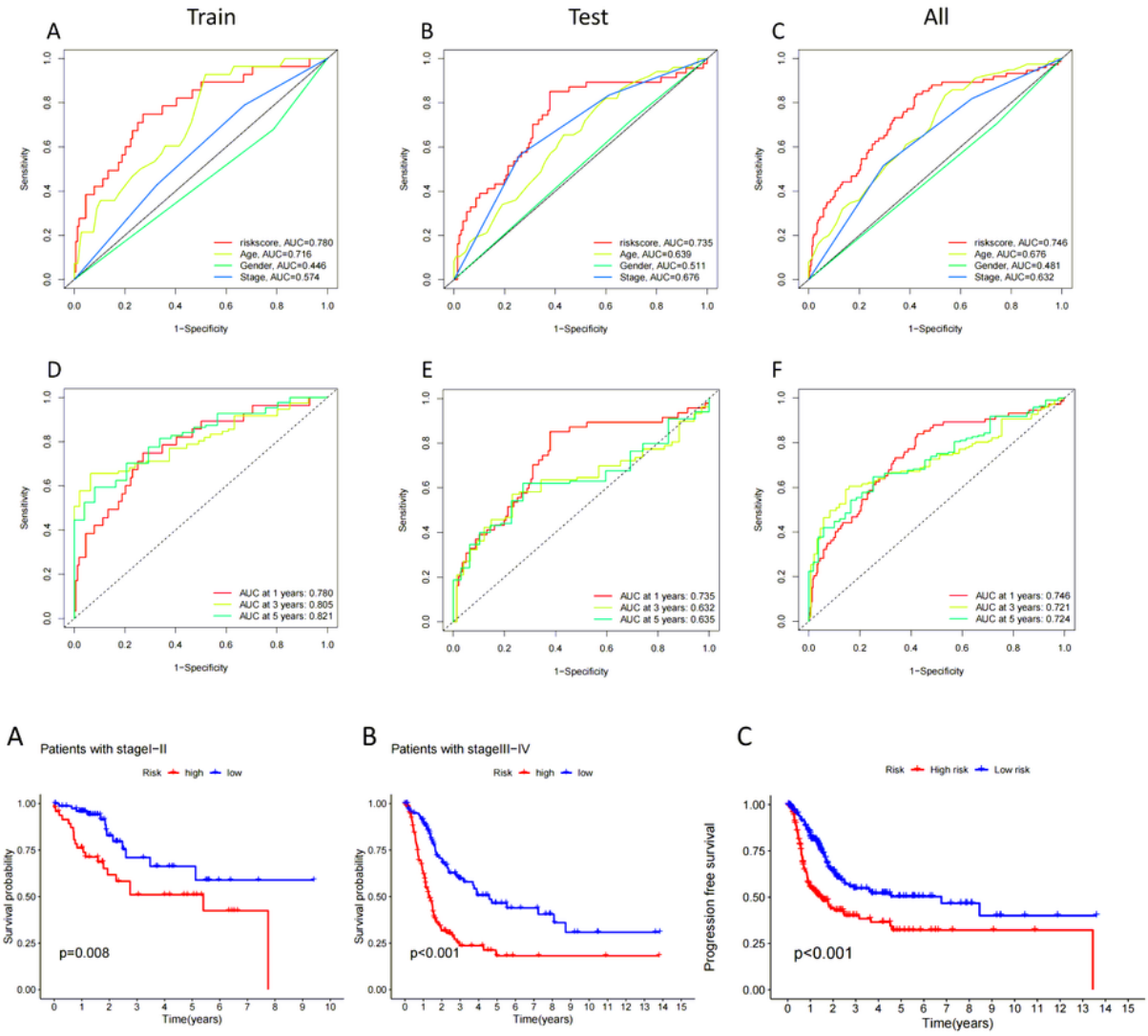

Figure 8

(A, D) Training set clinical characteristics ROC curves, 1-, 3-, and 5-year ROC curves for the training set; (B, E) Test set clinical characteristics ROC curves, 1-, 3-, and 5-year ROC curves for the test set; (C, F) Overall clinical characteristics ROC curves, 1-, 3-, and 5-year ROC curves for the entire cohort.

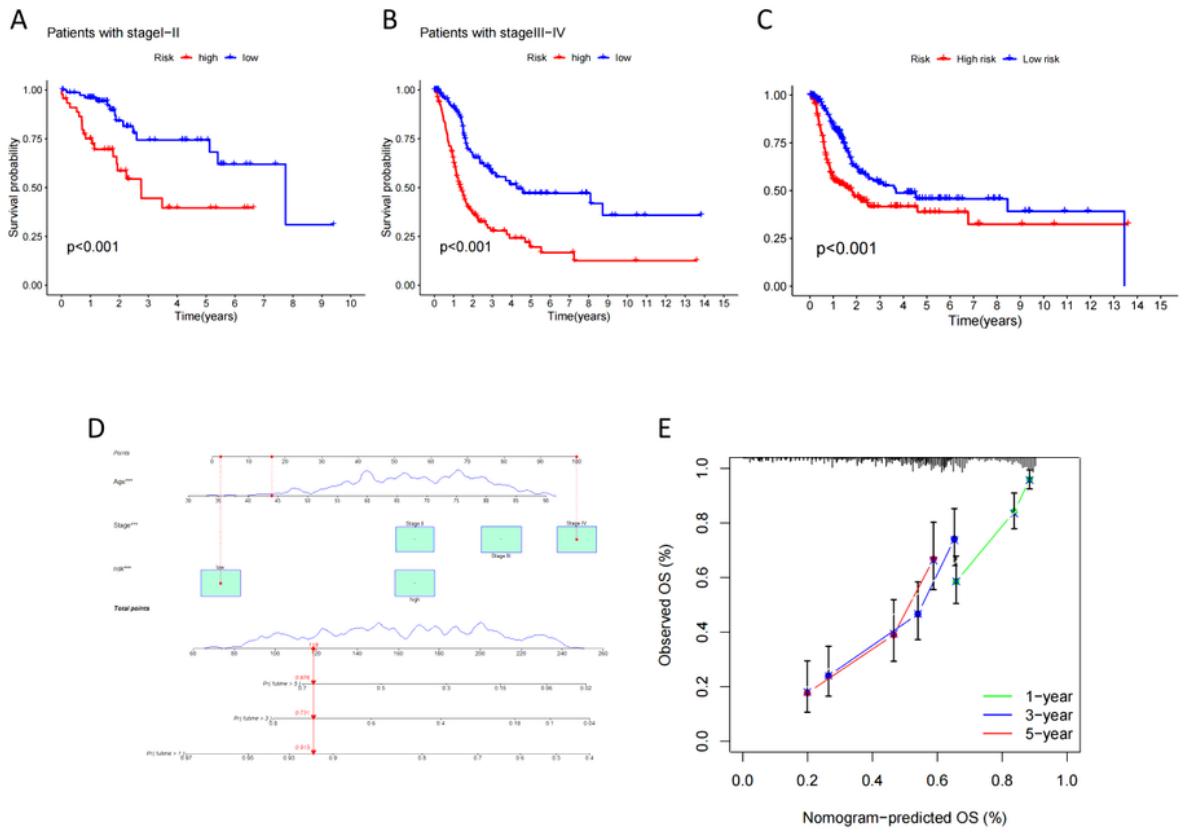

Figure 9

(A) Survival curve of patients with stage I-II; (B) Survival curve of patients with stage III-IV; (C) Progression-free survival curve; (D) Prognostic nomogram of patients; (F) Calibration curve of the nomogram.

Supplementary Files

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

- SupplementaryTable1.xlsx
- Supplementarytable2.xlsx
- Supplementarytable3.xlsx
- Supplementarytable4.xlsx
- Supplementarytable5.xlsx