Identification Potential Biomarker for Bladder Cancer using Feature Selection

Qian Yu (yuq@tib.cas.cn)
Tianjin University of Science & Technology
Haofan Dong (donghf@tib.cas.cn)
Tianjin University of Science & Technology
Shufan Liu (liushf@tib.cas.cn)
Tianjin University of Science & Technology
Yu Li (liyu@tust.edu.cn)
Tianjin University of Science & Technology
Junwei Luo (luojunwei@hpu.edu.cn)
Henan Polytechnic University
Xin Wu (wuxin@tib.cas.cn)
Chinese Academy of Sciences

Research Article

Keywords: Bladder cancer, Feature Selection, WGCNA, Immune infiltration, Ferroptosis-related genes, Pan-cancer analysis

DOI: https://doi.org/

Additional Declarations: No competing interests reported.

Identification Potential Biomarker for Bladder Cancer using Feature Selection

Qian Yu [1, 2], Haofan Dong[1, 2], Shufan Liu[1, 2], Yu Li [1], Junwei Luo[3*], Xin Wu [2*]

[1] College of Biotechnology, Tianjin University of Science & Technology, Tianjin, 300222, China;

[2] Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, 300308, China;

[3] School of Software, Henan Polytechnic University, Jiaozuo, 454003, China.

[*]Correspondence authors.

E-mail addresses: wuxin@tib.cas.cn; luojunwei@hpu.edu.cn;

Funding Information

This work was supported by the National Natural Science Foundation of China (62372156).

Declaration of competing interest

The authors have no relevant financial or non-financial interests to disclose.

ORCID

Xin Wu https://orcid.org/0000-0002-9225-5574

ABSTRACT

Background: The aim of this study was to utilize machine learning techniques to identify biomarkers associated with the diagnosis of bladder cancer, providing valuable insights into its early pathogenesis and exploring their potential as prognostic markers and therapeutic targets.

Methods: Initially, we conducted a comparative analysis of the genomes between bladder cancer samples, focusing on identifying the most significant differences between the cancer group and the normal group. Next, we employed machine learning techniques for feature selection and identified a key gene by integrating ferroptosis-related genes into our analysis. Moreover, we integrated transcriptome data, somatic mutation data, and clinical data to perform comprehensive analyses, including functional enrichment analysis, tumor mutation load analysis, immune infiltration analysis, and pan-cancer analysis. These analyses aimed to elucidate the pathological relevance of the candidate genes. Furthermore, we constructed a ceRNA network to identify the genes and regulatory pathways associated with these candidate genes.

Results: We initially conducted screening using the Weighted Gene Co-expression Network Analysis and machine learning techniques, resulting in the identification of six candidate genes: NR4A1, PAMR1, CFD, RAI2, ALG3, and HAAO. Subsequently, by integrating data from the FerrDB database, we identified NR4A1 as a gene associated with ferroptosis. Additionally, our analysis revealed a correlation between the expression of NR4A1 and tumor mutations as well as immune infiltration in patients with bladder cancer.

Conclusion: Our data strongly suggest that NR4A1 could serve as a crucial prognostic biomarker for bladder cancer and may also play a role in the development of various other cancers.

Keywords: Bladder cancer; Feature Selection; WGCNA; Immune infiltration; Ferroptosis-related genes; Pan-cancer analysis

1. Introduction

Bladder cancer (BC) is a significant health concern due to its potential impact on morbidity and mortality. The burden of this disease has remained relatively constant over time, posing a substantial impact on public health[1]. While the incidence of BC has shown a downward trend in recent years, the high recurrence and mortality rates associated with BC remain a significant challenge. The high recurrence rates make BC one of the most difficult and costly diseases to manage effectively[2]. To address the challenges posed by bladder cancer, including its high recurrence and mortality rates, is crucial for improving patient outcomes and reducing the burden of this disease.

In recent years, there has been a growing trend of applying machine learning and bio-inspired computing techniques to the field of medicine, specifically in the areas of diagnosis and prognosis. The utilization of machine learning and deep learning approaches in biology is not new, and the use of prediction methods in medicine has also been prevalent[3,4].Machine learning methods offer powerful statistical techniques for developing classification tools. Unlike traditional approaches based solely on clinical knowledge of diseases and treatments, machine learning methods have the capability to select the best algorithm that minimizes classification errors. These methods are well-suited for handling large volumes of data and numerous prediction variables. They excel in identifying nonlinear relationships, including interactions or Boolean combinations of variables that may have been previously unknown[5]. By utilizing machine learning techniques, researchers can effectively analyze complex datasets and uncover hidden patterns or relationships that may contribute to disease diagnosis and prognosis[6]. These methods provide a valuable tool for improving accuracy and efficiency in medical decision-making by incorporating objective algorithms and data-driven approaches. The integration of machine learning methods in medicine holds great potential for enhancing patient care and advancing medical research.

Bioinformatics analysis technology plays a crucial role in the discovery of potential biomarkers and patterns in various research fields[7]. Among the many available analysis algorithms, the Weighted Gene Co-expression Network Analysis (WGCNA) algorithm has gained popularity among bioinformatics researchers due to its efficiency and accuracy. By leveraging the results of gene co-expression network analysis, researchers have made significant advancements in the study of diseases[8-10], drug research[11,12], and species evolution[13,14]. This approach has been particularly useful in identifying key genes and pathways associated with diseases, including rheumatoid arthritis (RA). In a specific study conducted by Chen Yulan et al., the researchers downloaded a dataset related to rheumatoid arthritis from the GEO database. They obtained differential expression data from this dataset and applied the WGCNA method to elucidate differentially abundant genes. The next step involved identifying candidate biomarkers for RA using the LASSO regression model and SVM-RFE analysis[15]. These methods allowed the researchers to select a subset of genes that showed potential as biomarkers for rheumatoid arthritis.

The aim of this study was to identify potential biomarkers for the diagnosis of bladder cancer by obtaining potential biomarkers using bioinformatics and machine learning methods. In this study, the gene expression data obtained from TCGA were used as the research object, and the NR4A1 gene was obtained by WGCNA analysis and machine learning combined with ferroptosis related genes. A large number of studies have investigated the correlation between ferroptosis related genes and the occurrence, development and prognosis of BC. Certain genes have been identified as inhibitors of ferroptosis in BC cells and are known to promote cancer progression. Combined with transcriptome data, somatic mutation data, clinical data and other 32 cancer datasets in TCGA, enrichment analysis, tumor burden analysis, immune infiltration analysis and pan-cancer analysis were performed to uncover the pathological relevance of NR4A1. A CeRNA network was constructed to identify the regulatory pathways of NR4A1.

2. Materials and Methods

2.1 Datasets

In this study, we selected the bladder cancer dataset from the TCGA (The Cancer Genome Atlas) database as the primary dataset for identifying biomarkers. To further validate and examine the results of biomarker identification, we also utilized a combination of GEO (Gene Expression Omnibus) database. We utilized the TCGAbiolinks package in R to download and organize the gene expression data, clinical data, and somatic mutation data from the TCGA database.

The GEO13507 and GEO37815 datasets were extracted from the GEO database using the GEOquery package in the R. This package allowed us to download and organize these datasets for our study.

2.2 Identification DEGs of BLCA

We utilized the R package DeSeq2, which is a widely used tool in the field of bioinformatics for performing differential gene expression analysis. DeSeq2 provides robust statistical methods for identifying genes that show significant changes in expression levels between different experimental conditions. Specifically, we selected genes with |logFC| > 1 and adj.P.Val < 0.05 as differentially expressed genes. As a result, we identified a total of 4725 differentially expressed genes, including 2024 up-regulated genes and 2701 down-regulated genes.

2.3 WGCNA Analysis

The WGCNA algorithm achieves the goal of quickly locking core genes by grouping modules and associating gene modules with phenotypes.

To construct a weighted co-expression network, a soft threshold (soft threshold powers) as the correlation coefficient needs to be determined. The soft threshold determines the strength of the correlation required for two genes to be considered co-expressed. In this study, we selected the power value when R2 (the squared correlation coefficient) was greater than 0.9 as the threshold, resulting in powers = 6.

The gene tree is constructed using hierarchical clustering based on gene neighbor-joining coefficients. Different colors are used to represent different clustering modules, while gray is used as the default color for genes that cannot be classified into any module. After constructing the WGCNA co-expression modules, these modules were linked to cancer classification metrics to explore the associations between gene synergies and cancer classification. Each row represents a different gene co-expression module, and the values represent the correlation coefficients. Positive and negative correlations are distinguished using red and green colors, respectively. The values in parentheses represent the corresponding significant p-values. Based on the analysis, the yellow module was identified as the module that is positively and strongly correlated with cancer.

2.4 Enrichment Analysis

To explore the underlying mechanism of genes derived from WGCNA analysis and differential analysis, we utilized the R packages clusterProfiler[16] and org.Hs.eg.db. The clusterProfiler package provides a comprehensive set of functions for performing gene ontology (GO) analysis, which includes the investigation of gene molecular function (MF), biological process (BP), and cellular component (CC). Additionally, the package allows for the exploration of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. This analysis helps in unraveling the potential biological mechanisms and pathways underlying the studied phenotypes or conditions. A significance level of $P < 0.05$ was used to determine statistically significant KEGG pathways and GO terms for further investigation.

2.5 Machine Learning

2.5.1 SVM-RFECV

SVM-RFECV ranks the importance of each feature based on its impact on classification performance. This ranking is typically determined by evaluating the decrease in classification performance when a feature is removed from the model. By identifying the most influential features, researchers gain insights into the biological relevance of gene expression data and can better understand the underlying mechanisms and pathways associated with the studied phenotype or condition. And, SVM-RFECV incorporates cross-validation in the feature selection process to enhance the robustness of feature selection. By evaluating performance across multiple iterations of cross-validation, SVM-RFECV provides a more reliable assessment of feature importance and selection. This approach helps to mitigate the potential impact of dataset variations, ensuring that the selected features are more likely to generalize well to unseen data and improving the overall reliability of the feature selection process.

2.5.2 XGBoost

XGBoost is a variant of the Gradient Boosting Machine (GBM), which is a machine learning classifier developed by Chen et al[17]. In cancer research, XGBoost has been shown to consistently outperform other machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), logistic regression (LR), and k-nearest neighbor (KNN) algorithms in terms of accuracy and overall performance[17,18]. Studies have demonstrated that XGBoost can achieve higher accuracy and better predictive capabilities in cancer-related tasks. This advantage makes XGBoost a favorable choice for selecting potential biomarkers in cancer research[19,20].

2.6 Survival Analysis

In clinical research, clinical outcomes can take various statistical forms, including continuous variables or discrete events such as death. Traditional statistical methods like t-tests are not suitable for analyzing clinical outcomes, and instead, survival analysis techniques are employed to assess the impact of specific factors on these outcomes.

Survival analysis encompasses several commonly used methods, including the Kaplan-Meier (KM) estimator, the log-rank test, and the COX proportional hazards model. The KM estimator is particularly useful for analyzing survival data. It generates a step function curve where each vertical drop represents the occurrence of one or more events. By plotting survival probabilities over time, the KM method allows for estimating survival probabilities beyond a certain point and observing changes in survival over time[21].

To analyze the impact of the identified genes on clinical outcomes, we utilized two R packages: survival and Survminer. These packages provide a comprehensive set of tools for conducting survival analysis, including estimating survival probabilities, performing log-rank tests, and visualizing KM curves.

2.7 Receiver Operating Characteristic Curve

The ROC curve (Receiver Operating Characteristic curve) is a graphical representation used to evaluate the performance of binary classification methods. The x-axis of the ROC curve represents 1-specificity, which is the false positive rate. The y-axis represents sensitivity, which is the true positive rate.

The ROC curve is constructed by varying the cut-off value or decision threshold of the binary classification method. By adjusting this threshold, we can observe how the sensitivity and specificity change. The curve illustrates the trade-off between correctly identifying positive cases (sensitivity) and incorrectly classifying negative cases (1-specificity).

The area under the ROC curve (AUC) is a measure of the overall performance of the classification model. AUC values range from 0.5 to 1.0, where a value of 0.5 indicates a random classifier and a value of 1.0 indicates a perfect classifier. The closer the AUC is to 1, the higher the accuracy of the diagnostic model.

To plot the ROC curves in our study, we utilized the pROC package in the R language. This package provides functions and tools specifically designed for ROC analysis, allowing us to generate the ROC curves and calculate the corresponding AUC values. By utilizing the pROC package[22], we were able to assess the diagnostic accuracy of our classification models based on the identified genes.

2.8 Immune infiltration analysis

Immune cell infiltration in tumors plays a crucial role in tumor progression and the effectiveness of anti-cancer therapies. To estimate immune infiltration, we employed three widely used bioinformatics analytical tools: xCell, CIBERSORT, and estimate. These tools provide estimation of cell type enrichment scores or relative levels of distinct cell types from gene expression data. xCell utilizes a gene signature-based approach to infer cell type abundance, CIBERSORT employs a deconvolution algorithm to estimate cell type proportions, and estimate calculates immune cell infiltration scores based on gene expression signatures. By leveraging these tools, we can obtain comprehensive insights into the immune cell composition within the tumor microenvironment and gain a better understanding of the tumor-immune interaction.

3 Results

3.1 Intersection of WGCNA and DGEs

A schematic flow diagram of the performed biomarker identification assay is shown in Fig 1.

In our study, we initially selected 4725 differentially expressed genes based on the criteria of |logFC| > 1 and adj.p.val < 0.05, as depicted in Fig 2A. Subsequently, we conducted WGCNA analysis and determined that the optimal threshold for constructing a scale-free network was 6, as illustrated in Fig 2B. After determining the optimal threshold, we set the merging module threshold to 0.25 and generated the gene clustering diagram, as presented in Fig 2C. Next, we integrated the phenotypic data and calculated the correlation coefficients and module significance p-values between the quantitative module eigenvectors and the phenotypes. These results were visualized as a heatmap representing the module-trait correlation coefficients, as shown in Fig 2D. Based on a significance level of $p < 0.05$ and considering the correlation coefficient, we identified the yellow module as the key gene module most relevant to BLCA tumor tissue ($p < 1e-200$, corr = 0.76), as depicted in Fig 2E. Finally, we obtained a set of 609 genes by intersecting the differentially expressed genes with the key genes from the yellow module, as displayed in Fig 2F.

3.2 Enrichment Analysis

We performed enrichment analysis of GO and KEGG pathways using the clusterProfiler package. To obtain significant GO terms and KEGG pathways, we applied a threshold of qvalueCutoff = 0.05 and pvalueCutoff = 0.05, As shown in Fig 3A, we list the top ten important GO terms for DEGs in biological processes (BP), cellular components (CC), and molecular functions (MF). For example, in BP (Fig 3B), DEGs were significantly enriched in response to ameboidal-type cell migration, wound healing, cell-substrate adhesion, muscle contraction, muscle system process, tissue migration, regulation of cell-substrate adhesion, epithelial cell migration, epithelium migration and muscle tissue development. The GO words of the MF group (Fig 3C), including extracellular matrix structural constituent, actin binding, extracellular matrix binding, DNA-binding transcription activator activity, RNA polymerase II-specific, DNA-binding transcription activator activity, actin filament binding, glycosaminoglycan binding, integrin binding, muscle alpha-actinin binding and transmembrane receptor protein kinase activity, were significantly enriched by DEGs. In the CC group (Fig 3D), DEGs were mainly enriched in collagen-containing extracellular matrix, contractile fiber, myofibril, I band, sarcomere, Z disc, actin filament bundle, focal adhesion, cell-substrate junction and actomyosin.

Fig 3E-G shows the analysis of the KEGG pathway of DEGs. We observed that DEGs are mainly involved in Focal adhesion, MAPK signaling pathway, Proteoglycans in cancer, cGMP-PKG signaling pathway, Vascular smooth muscle contraction, Oxytocin signaling pathway, Cellular senescence, Human T-cell leukemia virus 1 infection, Regulation of actin cytoskeleton and ECM-receptor interaction.

3.3 Feature Selection

We composed a new gene expression dataset using 609 features from WGCNA analysis and differential analysis, dividing the dataset into a training set and a test set, where 75% is the training set and 25% is the testing set. We performed feature selection using SVM-RFECV and XGBoost methods. Using these methods, we selected 28 features with SVM-RFECV and 26 features with XGBoost. These features were chosen based on their importance in predicting the outcome of the cancer dataset. In Fig 4A-C, we present the confusion matrix and classification reports, including Precision, Recall, and F1 score, for the SVM-RFECV model. Similarly, in Fig 4D-F, we show the confusion matrix and classification reports for the XGBoost model. Precision represents the ratio of correctly observed positive results to all observed positive results, while Recall is the ratio of correctly observed positive results to the total results observed in the desired category. F1 score is a performance metric that combines both Precision and Recall, providing a measure of overall model performance. Values greater than 0.5 indicate relatively good categorization, while values less than 0.5 suggest categorization failure. As shown in Fig 4, the models constructed for the cancer dataset all show successful classification results. The accuracy of the test dataset was calculated as 98.15% for the SVM-RFECV model and 100% for the XGBoost model. The accuracy is determined by comparing the predicted labels with the true labels in the test dataset. These specific accuracy values were obtained based on the model's performance in correctly classifying the test samples. Finally, we took the intersection of SVM-RFECV and XGBoost, as shown in Fig 4H, and finally identified six genes, NR4A1, PAMR1, CFD, RAI2, ALG3 and HAAO.

Based on the intersection results mentioned above, we further incorporated ferroptosis-related genes into the analysis. We obtained 567 genes related to ferroptosis from the FerrDB database. Among these genes, NR4A1 was identified as the most relevant ferroptosis-related gene in both the cancer group and the normal group, as depicted in Fig 4I.

3.4 Survival analysis and ROC analysis

We performed Kaplan-Meier survival analysis to assess the survival outcomes of patients based on different gene expression or high/low risk groups. To evaluate the impact of the identified NR4A1, we utilized the TCGAbiolinks package to download clinical data and employed the survminer package for survival analysis. The cut_point function was used to determine the optimal threshold for stratifying patients into high and low gene expression groups. Additionally, we obtained clinical data from the GEO database for further analysis, including TCGA-BLCA, GSE3507, and GSE37815 cohorts. Statistical analysis was performed to compare the overall survival (OS) rates between different expression groups[23]. At the same time, we downloaded the clinical data of GEO data from the GEO database and analyzed TCGA-BLCA, GSE3507 and GSE37815 respectively. Further analysis revealed a significant difference in the OS rates between the high and low expression groups in the TCGA-BLCA cohort ($p = 0.0031$), as shown in Fig 5A. Similarly, in the GSE31507 cohort, there was a significant difference in the OS rates between the low and high expression groups ($p = 0.00095$), as depicted in Fig 5B. Furthermore, in the GSE37815 cohort, the high expression group exhibited a significantly lower OS rate compared to the low expression group ($p = 0.00021$), as shown in Fig 5C.

We evaluated the diagnostic performance of the identified gene by analyzing their AUC values using ROC curve analysis. Firstly, for the TCGA dataset, we compared the expression of NR4A1 between the cancer group and the normal group, as shown in Fig 5D. Secondly, we assessed the sensitivity and specificity of these genes for diagnosing BLCA by generating ROC curves. The AUC value for NR4A1 was calculated as 0.9, indicating a high discriminatory power, as depicted in Fig 5E. Additionally, for the GEO datasets, we processed the batch effect using the sva (R/Bioconductor) package and merged the datasets. Subsequently, we calculated the inter-group differences and generated ROC curves. The AUC value obtained for the GEO dataset was 0.697, as shown in Fig 5(F-G). The AUC value represents the area under the ROC curve and is a measure of the overall diagnostic performance of a test. AUC values range from 0 to 1, where a value of 1 indicates a perfect discriminatory power, and a value of 0.5 suggests no discriminatory power (equivalent to random chance). In our analysis, the AUC value of 0.9 for NR4A1 in the TCGA dataset indicates a high accuracy in distinguishing between BLCA and normal samples. Similarly, the AUC value of 0.697 for the GEO dataset suggests a moderate discriminatory power. These results suggest that NR4A1 has potential as a diagnostic biomarker for BLCA.

Finally, we performed a differential analysis of NR4A1 expression levels in TCGA-BLCA, comparing it with stage, N, M, T, age, and sex. As shown in Fig 6, we observed significant differences in stages, especially in stage I+II compared to stage III and VI respectively. The differences are striking. Furthermore, consistent with Fig 5D, we observed a significant decrease in the expression of NR4A1 in the cancer group.

3.5 CeRNA network analysis

In order to gain insights into the mechanism of “NR4A1” in BLCA, we employed themultiMiR” (R/Bioconductor) package to identify the miRNAs that potentially regulate NR4A1. multiMiR incorporates eight different predicted miRNA-target gene interaction databases (diana_microt, elmmo, microcosm, miranda, mirdb, pictar, pita, and targetscan), which greatly facilitates research on disease pathogenesis, diagnosis, and treatment based on the regulatory relationship between miRNAs and target genes, as depicted in Fig7(A-B).

Based on the results obtained, we focused on the miRNAs that were predicted to target NR4A1 in at least six out of the eight databases and utilized the mirnet website to predict the target genes of these miRNAs. Subsequently, we constructed a ceRNA (competing endogenous RNA) network diagram, as depicted in Fig7C. This network diagram provides a visual representation of the interactions between miRNAs, NR4A1, and other target genes, shedding light on the potential regulatory mechanisms involved in BLCA. This network provides novel insights into the post-transcriptional regulation of NR4A1 and may help to reveal potential therapeutic targets for BLCA. The results are shown in the Supplementary Table.

3.6 Tumor mutation burden estimation

Due to the association between tumor mutation burden (TMB) and the response to immunotherapy and prognosis of cancer, we utilized the maftools (R/Bioconductor) tool to analyze and visualize somatic mutation data in tissues with high and low expression of NR4A1. The results of this analysis are presented in Fig 8.

In Fig 8C, we performed a differential mutation analysis using Fisher’s exact test on all genes present in the maf files of the high expression and low expression groups. Our analysis revealed that genes such as RYR2, POLN, and CNTNAP2 exhibited significant differences in mutation frequency between the two groups.

3.7 Immune analysis

In this study, our specific objective was to investigate the potential association between NR4A1 expression and the infiltration levels of immune cells in bladder cancer. Understanding this association can provide valuable insights into the role of NR4A1 in modulating immune responses within the tumor microenvironment, potentially leading to the development of novel therapeutic strategies for bladder cancer treatment.

We utilized the xCell, cibersort and estimate (R/Bioconductor) to analyze the differences between immune cells with high and low expression levels of NR4A1. As shown in Fig9A, significant differences in StromaScore and MicroenvironmentScore were observed in immune cells including adipocytes, chondrocytes, endothelial cells, fibroblasts, HSCs, endothelial cells, megakaryocytes, mesangial cells, and Pericytes. The stromal score and immune score are shown in Fig 9B.

We employed cibersort to analyze and compare the differences in the abundance of 22 immune cell types between the NR4A1 high and low expression groups (as depicted in Fig 9C). The results indicated that Macrophages M1 and Mast cells activated exhibited higher levels in the NR4A1 high expression group compared to the low expression group, and these differences were found to be statistically significant (P<0.05). Additionally, T cells regulatory (Tregs) showed a significant increase in the NR4A1 low expression group. Fig 9D shows a heat map of high and low expression of NR4A1 in immunoassays. Therefore, we suggest that the NR4A1 may play a crucial role in immune cell regulation in BC.

3.8 pan-cancer analysis

To further analyze NR4A1, we conducted an analysis of NR4A1 expression in 23 different tumor types from the TCGA database, comparing cancer tissues with corresponding normal tissues. Our findings revealed that in 15 cancer types (BLCA, BRCA, GBM, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, STAD, THCA and UCEC), the expression level of NR4A1 was significantly increased in the corresponding normal tissues ($p<0.05$), as depicted in the Fig10A.

Furthermore, when considering the overall significance, we observed that NR4A1 plays an important role in five cancers: ACC (p=0.02), CESC (p=0.0015), KICH (p=0.039), KIRC (p=0.0022), and TGCT (p=0.029). These results indicate significant differences in NR4A1 expression among different pathological stages, as shown in the Fig 10(B-F).

We also investigated the correlation between NR4A1 expression and the prognosis of patients with different cancer. Our analysis revealed significant associations between NR4A1 expression and the prognosis of 14 different cancer types. In the Fig 11, it is evident that high NR4A1 expression is associated with poor prognosis in patients with 9 types of cancer (ACC (p=0.026), COAD (p=0.025), DLBC (p=0.0092), ESCA (p=0.00068), KIRP (p=0.037), LUSC (p=0.041), MESO (p=0.018), OV (p=0.015), THCA (p=0.0027)). Conversely, low NR4A1 expression is associated with poor prognosis in patients with 5 types of cancer (BRCA (p=0.017), KICH (p=0.0059), KIRC (p=0.011), LIHC (p=0.0071), STAD (p=0.032)).

In the immune infiltration analysis, we observed significant findings in the BLCA dataset regarding B cells naive and T cells regulatory (Tregs). B cells naive showed a positive correlation, while Tregs showed a negative correlation. These results are depicted in the Fig 12A.

Regarding the immune infiltration of T cells regulatory (Tregs) and NR4A1 expression, we found a negative correlation in 18 cancer types (BRCA (p = 1.81e-10, r = -0.18), CESC (p = 3.75e-03, r = -

0.16), GBM (p = 0.02, r = -0.17), HNSC (p = 2.39e-03, r = -0.13), KIRC (p = 2.15e-05, r = -0.17), LGG (p = 1.32e-04, r = -0.16), LIHC (p = 4.69e-13, r = -0.34), LUAD (p = 4.14e-06, r = -0.19), LUSC (p = 4.13e-03, r = -0.12), MESO (p = 3.59e-03, r = -0.31), OV (p = 6.42e-04, r = -0.16), PCPG (p = 9.87e-05, r = -0.28), PRAD (p = 2.73e-08, r = -0.23), SARC (p = 0.01, r = -0.15), THCA (p = 7.70e-6, r = -0.33), THYM (p = 2.71e-03, r = -0.27), and UCEC (p = 8.09e-06, r = -0.18)). However, in COAD (p = 0.04, r = 0.09), a positive correlation was observed. These correlations are also depicted in the Fig 12(B-S).

To further investigate immune infiltration, we conducted separate analyses in 32 other cancer datasets. In 17 cancers (BRCA (p = 3.69e-11, r = 0.19), CESC (p = 3.25e-06, r = 0.26), ESCA (p = 4.36e-04, r = 0.25), GBM (p = 3.86e-08, r = 0.40), HNSC (p = 0.03, r = 0.09), KICH (p = 1.99e-03, r = 0.32), KIRC (p = 8.23e-14, r = 0.30), KIRP (p = 1.52e-10, r = 0.35), MESO (p = 1.61e-04, r = 0.39), OV (p = 2.33e-04, r = 0.18), PRAD (p = 1.29e-05, r = 0.18), READ (p = 0.01, r = 0.19), SARC (p = 3.03e-03, r = 0.18), STAD (p = 1.06e-03, r = 0.15), TGCT (p = 4.11e-04, r = 0.28), THYM (p = 5.82e-03, r = 0.25), and UCEC (p = 5.57e-07, r = 0.20)), we found a positive correlation between immune infiltration of B cells naive and NR4A1 expression. Conversely, in LAML (p = 0.02, r = -0.19), a negative correlation was observed. These correlations are shown in the Fig 13.

4. Discussion

NR4A1, also known as TR3, Nur77, or NGF-IB, is a member of the NR subfamily 4 (NR4A) receptor and belongs to the steroid/thyroid hormone receptor superfamily. It functions as a transcription factor and is considered an early response gene that can be induced by various stimuli, such as serum, inflammatory factors, growth factors, and stress, in different cell types and organs. NR4A1 plays a crucial role in regulating diverse biological processes, including cell growth, apoptosis, and metastasis. The expression and function of NR4A1 have been extensively studied in various cancers, including melanoma, colorectal cancer, breast cancer, and hepatocellular carcinoma. In these cancers, NR4A1 has been shown to play a significant role in tumor progression and metastasis. It regulates key cellular processes associated with cancer, such as cell proliferation, survival, angiogenesis, and immune evasion. Additionally, NR4A1 has been implicated in the regulation of metabolic processes in cancer cells, including glycolysis, fatty acid synthesis, and amino acid metabolism[24,25]. Chang[26] isolated NR4A1 from a human prostate lambda gt11 cDNA library. Then it is found in various tissues and cells, including cancer cells.

Identification of genes critical for bladder cancer diagnosis may not only improve our understanding of the mechanisms underlying bladder cancer progression, but also provide molecular targets for novel therapies and drugs. As a key gene, NR4A1 plays an important role in bladder cancer. The NR4A1-centered ceRNA network may provide important targets for future studies of NR4A1 in bladder cancer. In bladder cancer, NR4A1 may interact with other RNA molecules through the ceRNA network, thereby influencing the development and progression of the disease. By identifying lncRNAs and miRNAs that interact with NR4A1, we can gain insights into their functions and regulatory networks in bladder cancer, and explore novel therapeutic targets. These findings will contribute to the advancement of individualized treatment and precision medicine in the field of bladder cancer, offering patients more effective treatment options.

Our study demonstrated that NR4A1 expression was significantly lower in 15 cancer types compared to the normal group, including BC, based on the analysis high and low expression levels of NR4A1 across 23 different cancers. However, the role of NR4A1 in cancer remains controversial.

Studies have demonstrated that NR4A1 can have both pro-tumor and tumor suppressor roles in cancer cells and tumors[27]. Knockdown of NR4A1 in cancer cells has been shown to inhibit cell growth, induce apoptosis, and reduce angiogenesis[28,29]. Conversely, NR4A1 has also been considered a potent tumor suppressor due to its involvement in growth inhibition and induction of apoptosis[30-33]. Thus, NR4A1 has both tumor suppressor and oncogenic roles in cancer development.

Overexpression of NR4A1 in breast cancer has been identified as a poor prognostic factor associated with decreased survival and increased metastasis[34]. miR-506 inhibits the proliferation and migration of colorectal cancer cells by downregulating the expression of NR4A1[35]. In contrast, overexpression of NR4A1 has been shown to activate the Wnt/β-catenin signaling pathway, thereby promoting colon tumor growth, colony formation, and migration[28]. However, studies have also shown that overexpression of NR4A1 inhibits the proliferation, invasion, and migration of endometrioid endometrial cancer cells, while promoting apoptosis[36]. Overexpression of NR4A1 inhibits the growth and invasiveness of triple-negative breast cancer cells[37]. These results suggest that NR4A1 expression may have different roles in different cancers. There is already growing evidence that this receptor can be targeted by anticancer drugs that induce cell death through NR4A1-dependent and independent pathways.

Furthermore, we conducted a prognostic analysis of NR4A1 expression using both TCGA-BLCA and GEO datasets, which revealed that high expression of NR4A1 in bladder cancer was associated with poor prognosis. Additionally, we observed correlations between NR4A1 expression and clinical parameters such as bladder cancer stage, T, N, M, age, and gender. We observed a significant association between NR4A1 expression and cancer stage. Among the other 32 cancers, high expression of NR4A1 in ACC, COAD, DLBC, ESCA, KIRP, LUSC, MESO, OV, and THCA was associated with poor prognosis. Conversely, low NR4A1 expression in BRCA, KICH, KIRC, LIHC, and STAD was associated with poor prognosis. Furthermore, NR4A1 expression was significantly higher in the normal group compared to the cancer group in BLCA, KIRP, LUSC, BRCA, KICH, KIRC, LIHC, and STAD. Additionally, we analyzed the pan-cancer data from TCGA and found significant differences in NR4A1 expression among different stages of ACC, CESC, KICH, KIRC, and TGCT. Therefore, based on the aforementioned analyses, we believe it would be valuable to conduct further molecular and cellular experiments to confirm the molecular function of NR4A1 in KICH, KIRC, and BLCA.

Previous studies have shown that immune cells play a dual role in tumors, with the ability to both promote and inhibit tumor progression[38]. Regulatory T cells (Tregs) play a crucial role in maintaining immune system homeostasis and immune tolerance, making them an important mechanism in the regulation of tumor immunity. Tregs are currently a research hotspot in this field, primarily due to their potential as therapeutic targets. They exert suppressive effects on the activation and differentiation of CD4 helper T cells and CD8 cytotoxic T cells, leading to reduced reactivity to autoantigens and tumor-expressed antigens[39-41]. Our results analyzed the relationship between Nr4a1 expression and immune cell infiltration. Among the 19 cancers (BRCA, CESC, GBM, HNSC, KIRC, LGG, LIHC, LUAD, LUSC, MESO, OV, PCPG, PRAD, SARC, THCA, THYM, UCEC, and COAD), we observed a negative correlation between NR4A1 expression and regulatory T cells (Tregs).

Studies have demonstrated the co-localization and synergistic effects of tumor-infiltrating CD20$^+$ B cells and CD8$^+$ T cells in human cancers, highlighting the significance of T-cell-B cell interactions in promoting effective antitumor immunity. B cells can play a defensive role against tumors under specific conditions, primarily through the production of tumor-specific antibodies and presentation of tumor antigens. However, certain subsets of B cells and specific antibodies can also impede anti-tumor immunity and facilitate tumor growth[42-44]. Among the 18 cancers (BRCA, CESC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, MESO, OV, PRAD, READ, SARC, STAD, TGCT, THYM, UCEC, and LAML), we observed a negative correlation between NR4A1 expression and B cells navie.

The identification of NR4A1 as a key candidate gene suggests its potential involvement in the initiation and progression of bladder cancer, making it a promising molecular target for the diagnosis and treatment of the disease. While our study provides valuable evidence regarding the role of NR4A1 in tumorigenesis and immune regulation within the tumor microenvironment, it is important to acknowledge the limitations of our study. This is based on pure bioinformatics analysis and relies entirely on available open access database information and has not been experimentally validated. However, our bioinformatics analysis has provided initial insights into the involvement of NR4A1 in bladder cancer and pan-cancer mechanisms, highlighting its potential as a biomarker for further investigation. However, additional molecular biology experiments are required to validate its utility as a biomarker in pan-cancer studies. These studies help advance the development of NR4A1 as a valuable new target for cancer.

5. Conclusion

Firstly, we identified NR4A1 as a key gene using the TCGA-BLCA dataset. We then integrated transcriptome data, somatic mutation data, and clinical data to perform functional enrichment analysis, tumor mutation burden analysis, immune infiltration analysis, and pan-cancer analysis, aiming to elucidate the pathological relevance of this candidate gene. We constructed a ceRNA network to identify the genes and regulatory pathways associated with NR4A1 and other candidate genes. However, it is important to note that our findings are based on bioinformatics analysis and rely on data from existing databases. Therefore, experimental validation is required to confirm these results. Furthermore, machine learning encounters challenges such as high dimensionality and small sample sizes. Additionally, gene expression data often exhibit an imbalanced sample distribution, with a significantly higher number of diseased samples compared to normal samples. Addressing these issues constitutes an important research focus in the field of bioinformatics.

Data Availability

The TCGA datasets was obtained from TCGA database (GDC (cancer.gov)). the GSE13507 and GSE37815 datasets were obtained from GEO database (National Center for Biotechnology Information (nih.gov)).

References

1 Lobo, N. *et al.* Epidemiology, Screening, and Prevention of Bladder Cancer. *Eur Urol Oncol* **5**, 628-639 (2022). https://doi.org:10.1016/j.euo.2022.10.003

2 Ge, L. *et al.* Study Progress of Radiomics With Machine Learning for Precision Medicine in Bladder Cancer Management. *Front Oncol* **9**, 1296 (2019). https://doi.org:10.3389/fonc.2019.01296

3 Islam, M. M. *et al.* Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science* **1** (2020). https://doi.org:10.1007/s42979-020-00305-w

4 Chen, J. H. & Asch, S. M. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* **376**, 2507-2509 (2017). https://doi.org:10.1056/NEJMp1702071

5 Noone, A. M. *et al.* Machine Learning Methods to Identify Missed Cases of Bladder Cancer in Population-Based Registries. *JCO Clin Cancer Inform* **5**, 641-653 (2021). https://doi.org:10.1200/CCI.20.00170

6 Wu, J. *et al.* Glycosyltransferase-related prognostic and diagnostic biomarkers of uterine corpus endometrial carcinoma. *Computers in Biology and Medicine* **163**, 107164 (2023). https://doi.org:https://doi.org/10.1016/j.compbiomed.2023.107164

7 Jiang, Y. *et al.* Screening of Biomarkers in Liver Tissue after Bariatric Surgery Based on WGCNA and SVM-RFE Algorithms. *Dis Markers* **2023**, 2970429 (2023). https://doi.org:10.1155/2023/2970429

8 Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-384 (2011). https://doi.org:10.1038/nature10110

9 Hua, Y., He, Z. & Zhang, X. A pan-cancer analysis based on weighted gene co-expression network analysis identifies the biomarker utility of lamin B1 in human tumors. *Cancer Biomark* **34**, 23-39 (2022). https://doi.org:10.3233/CBM-203247

10 Zhang, G. *et al.* Identification and targeting of cancer-associated fibroblast signature genes for prognosis and therapy in Cutaneous melanoma. *Computers in Biology and Medicine* **167**, 107597 (2023). https://doi.org:https://doi.org/10.1016/j.compbiomed.2023.107597

11 Chen, C. *et al.* Two gene co-expression modules differentiate psychotics and controls. *Mol Psychiatry* **18**, 1308-1314 (2013). https://doi.org:10.1038/mp.2012.146

12 Iskar, M. *et al.* Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol Syst Biol* **9**, 662 (2013). https://doi.org:10.1038/msb.2013.20

13 Delahaye-Duriez, A. *et al.* Rare and common epilepsies converge on a shared gene regulatory network providing opportunities for novel antiepileptic drug discovery. *Genome Biol* **17**, 245 (2016). https://doi.org:10.1186/s13059-016-1097-7

14 Filteau, M., Pavey, S. A., St-Cyr, J. & Bernatchez, L. Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish. *Mol*

Biol Evol **30**, 1384-1396 (2013). https://doi.org:10.1093/molbev/mst053

15 Chen, Y., Liao, R., Yao, Y., Wang, Q. & Fu, L. Machine learning to identify immune-related biomarkers of rheumatoid arthritis based on WGCNA network. *Clin Rheumatol* **41**, 1057-1068 (2022). https://doi.org:10.1007/s10067-021-05960-9

16 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012). https://doi.org:10.1089/omi.2011.0118

17 Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785-794 (2016).

18 Huang, Z. *et al.* An Artificial Intelligence Model for Predicting 1-Year Survival of Bone Metastases in Non-Small-Cell Lung Cancer Patients Based on XGBoost Algorithm. *Biomed Res Int* **2020**, 3462363 (2020). https://doi.org:10.1155/2020/3462363

19 Wang, T., Jiao, M. & Wang, X. Link Prediction in Complex Networks Using Recursive Feature Elimination and Stacking Ensemble Learning. *Entropy (Basel)* **24** (2022). https://doi.org:10.3390/e24081124

20 Sung, J. *et al.* Classification of Stroke Severity Using Clinically Relevant Symmetric Gait Features Based on Recursive Feature Elimination With Cross-Validation. *IEEE Access* **10**, 119437-119447 (2022). https://doi.org:10.1109/access.2022.3218118

21 Schober, P. & Vetter, T. R. Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. *Anesth Analg* **127**, 792-798 (2018). https://doi.org:10.1213/ANE.0000000000003653

22 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* **12**, 1-8 (2011).

23 Budczies, J. *et al.* Cutoff Finder: A Comprehensive and Straightforward Web Application Enabling Rapid Biomarker Cutoff Optimization. *PLoS ONE* **7**, e51862 (2012). https://doi.org:10.1371/journal.pone.0051862

24 Winoto, A. & Littman, D. R. Nuclear hormone receptors in T lymphocytes. *Cell* **109 Suppl**, S57-66 (2002). https://doi.org:10.1016/s0092-8674(02)00710-9

25 Deng, S., Chen, B., Huo, J. & Liu, X. Therapeutic potential of NR4A1 in cancer: Focus on metabolism. *Front Oncol* **12**, 972984 (2022). https://doi.org:10.3389/fonc.2022.972984

26 Chang, C., Kokontis, J., Liao, S. S. & Chang, Y. Isolation and characterization of human TR3 receptor: a member of steroid receptor superfamily. *J Steroid Biochem* **34**, 391-395 (1989). https://doi.org:10.1016/0022-4731(89)90114-3

27 Lee, S.-O., Li, X., Khan, S. & Safe, S. Targeting NR4A1 (TR3) in cancer cells and tumors. *Expert Opinion on Therapeutic Targets* **15**, 195-206 (2011). https://doi.org:10.1517/14728222.2011.547481

28 Wu, H. *et al.* Regulation of Nur77 expression by β-catenin and its mitogenic effect in colon cancer cells. *Faseb j* **25**, 192-205 (2011). https://doi.org:10.1096/fj.10-166462

29 Lee, S. O. *et al.* Inactivation of the orphan nuclear receptor TR3/Nur77 inhibits pancreatic cancer cell and tumor growth. *Cancer Res* **70**, 6824-6836 (2010). https://doi.org:10.1158/0008-5472.Can-10-1992

30 Woronicz, J. D., Calnan, B., Ngo, V. & Winoto, A. Requirement for the orphan steroid receptor Nur77 in apoptosis of T-cell hybridomas. *Nature* **367**, 277-281 (1994). https://doi.org:10.1038/367277a0

31 Liu, Z. G., Smith, S. W., McLaughlin, K. A., Schwartz, L. M. & Osborne, B. A. Apoptotic signals delivered through the T-cell receptor of a T-cell hybrid require the immediate-early gene nur77. *Nature* **367**, 281-284 (1994). https://doi.org:10.1038/367281a0

32 Lin, B. *et al.* Conversion of Bcl-2 from protector to killer by interaction with nuclear orphan receptor Nur77/TR3. *Cell* **116**, 527-540 (2004). https://doi.org:10.1016/s0092-8674(04)00162-x

33 Mullican, S. E. *et al.* Abrogation of nuclear receptors Nr4a3 and Nr4a1 leads to development of acute myeloid leukemia. *Nat Med* **13**, 730-735 (2007). https://doi.org:10.1038/nm1579

34 Hedrick, E., Lee, S.-O., Doddapaneni, R., Singh, M. & Safe, S. NR4A1 Antagonists Inhibit β1-Integrin-Dependent Breast Cancer Cell Migration. *Molecular and Cellular Biology* **36**, 1383-1394 (2016). https://doi.org:10.1128/MCB.00912-15

35 Huang, M. *et al.* MiR-506 Suppresses Colorectal Cancer Development by Inhibiting Orphan Nuclear Receptor NR4A1 Expression. *J Cancer* **10**, 3560-3570 (2019). https://doi.org:10.7150/jca.28272

36 Sun, L. *et al.* Lnc-NA inhibits proliferation and metastasis in endometrioid endometrial carcinoma through regulation of NR4A1. *Journal of Cellular and Molecular Medicine* **23**, 4699-4710 (2019). https://doi.org:https://doi.org/10.1111/jcmm.14345

37 Wu, H. *et al.* Nuclear receptor NR4A1 is a tumor suppressor down-regulated in triple-negative breast cancer. *Oncotarget* **8** (2017).

38 Zhang, Y. *et al.* Macrophage-Associated PGK1 Phosphorylation Promotes Aerobic Glycolysis and Tumorigenesis. *Mol Cell* **71**, 201-215.e207 (2018). https://doi.org:10.1016/j.molcel.2018.06.023

39 van der Veeken, J. *et al.* Memory of Inflammation in Regulatory T Cells. *Cell* **166**, 977-990 (2016). https://doi.org:10.1016/j.cell.2016.07.006

40 Newton, R., Priyadharshini, B. & Turka, L. A. Immunometabolism of regulatory T cells. *Nat Immunol* **17**, 618-625 (2016). https://doi.org:10.1038/ni.3466

41 Li, M. O. & Rudensky, A. Y. T cell receptor signalling in the control of regulatory T cell differentiation and function. *Nature Reviews Immunology* **16**, 220-233 (2016). https://doi.org:10.1038/nri.2016.26

42 Budczies, J. *et al.* A gene expression signature associated with B cells predicts benefit from immune checkpoint blockade in lung adenocarcinoma. *Oncoimmunology* **10**, 1860586 (2021). https://doi.org:10.1080/2162402x.2020.1860586

43 Nielsen, J. S. *et al.* CD20+ tumor-infiltrating lymphocytes have an atypical

CD27- memory phenotype and together with CD8+ T cells promote favorable prognosis in ovarian cancer. *Clin Cancer Res* **18**, 3281-3292 (2012). https://doi.org:10.1158/1078-0432.Ccr-12-0234

44 Yang, F. *et al.* Transcriptome Profiling Reveals B-Lineage Cells Contribute to the Poor Prognosis and Metastasis of Clear Cell Renal Cell Carcinoma. *Front Oncol* **11**, 731896 (2021). https://doi.org:10.3389/fonc.2021.731896

Figures

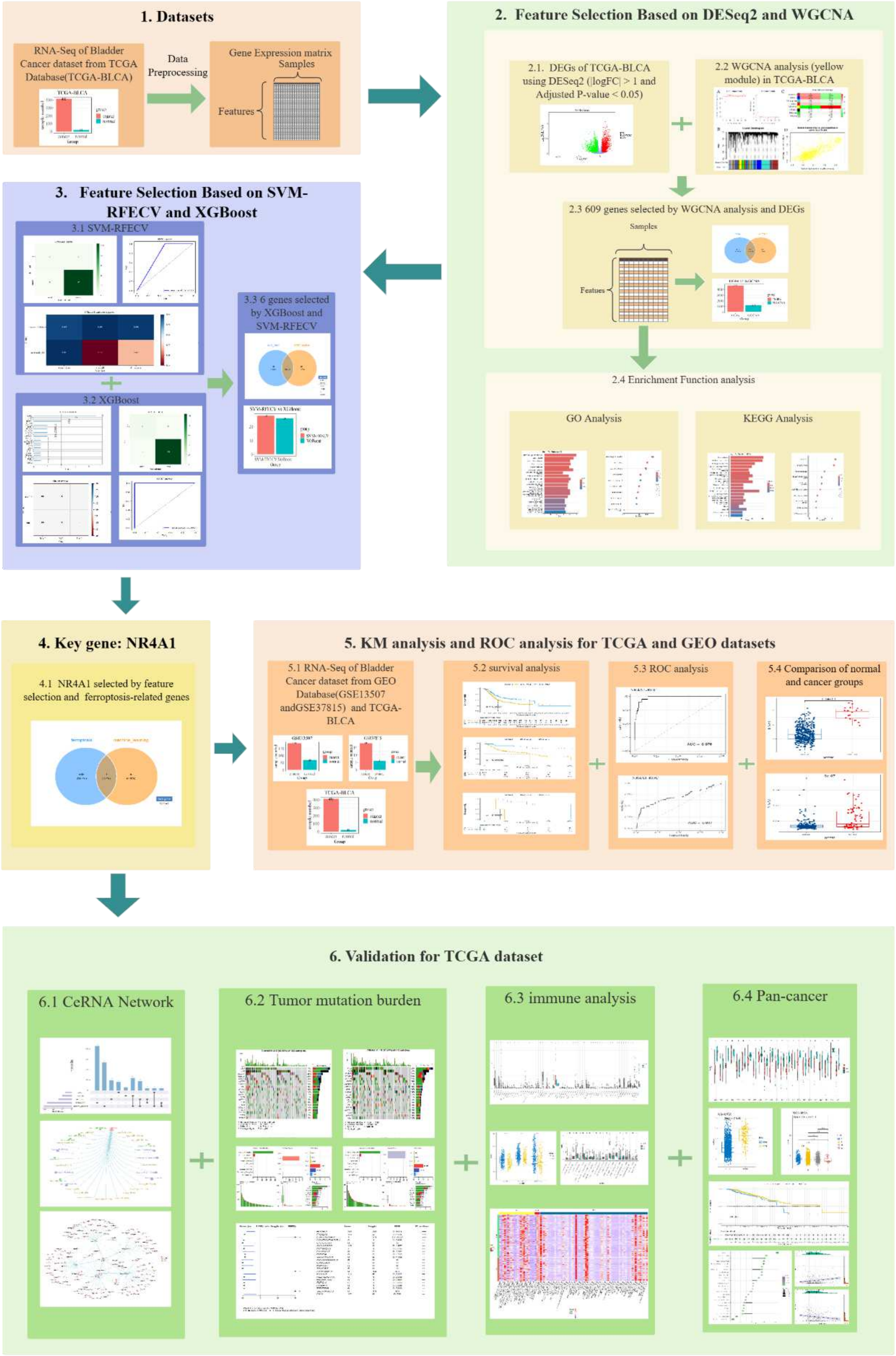

1. Datasets
RNA-Seq of Bladder Cancer dataset from TCGA Database(TCGA-BLCA)
Data Preprocessing
Gene Expression matrix
Samples
Features
2. Feature Selection Based on DESeq2 and WGCNA
2.1. DEGs of TCGA-BLCA using DESeq2 (|logFC| > 1 and Adjusted P-value < 0.05)
2.2 WGCNA analysis (yellow module) in TCGA-BLCA
2.3 609 genes selected by WGCNA analysis and DEGs
Samples
Featues
2.4 Enrichment Function analysis
GO Analysis
KEGG Analysis
3. Feature Selection Based on SVM-RFECV and XGBoost
3.1 SVM-RFECV
3.2 XGBoost
3.3 6 genes selected by XGBoost and SVM-RFECV
4. Key gene: NR4A1
4.1 NR4A1 selected by feature selection and ferroptosis-related genes
5. KM analysis and ROC analysis for TCGA and GEO datasets
5.1 RNA-Seq of Bladder Cancer dataset from GEO Database(GSE13507 andGSE37815) and TCGA-BLCA
5.2 survival analysis
5.3 ROC analysis
5.4 Comparison of normal and cancer groups
6. Validation for TCGA dataset
6.1 CeRNA Network
6.2 Tumor mutation burden
6.3 immune analysis
6.4 Pan-cancer

Fig 1. Schematic workflow of analyses.

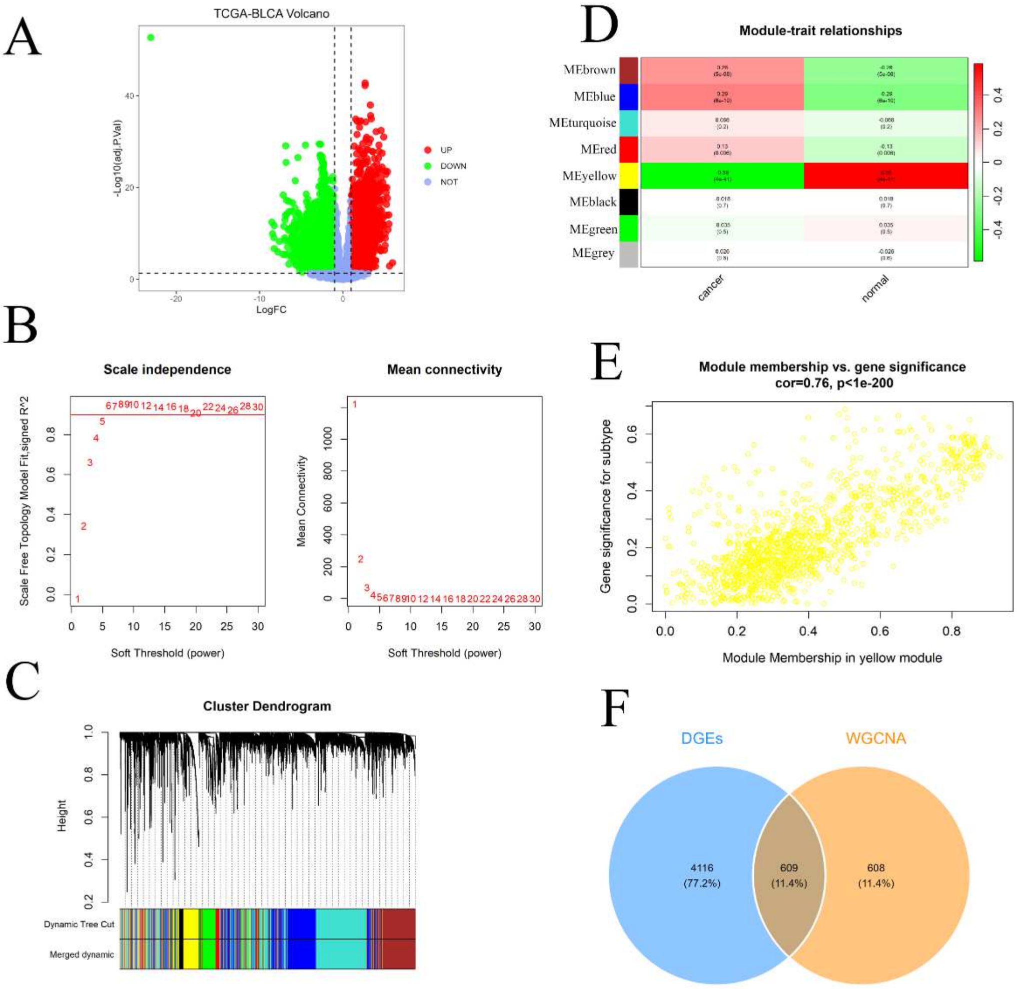
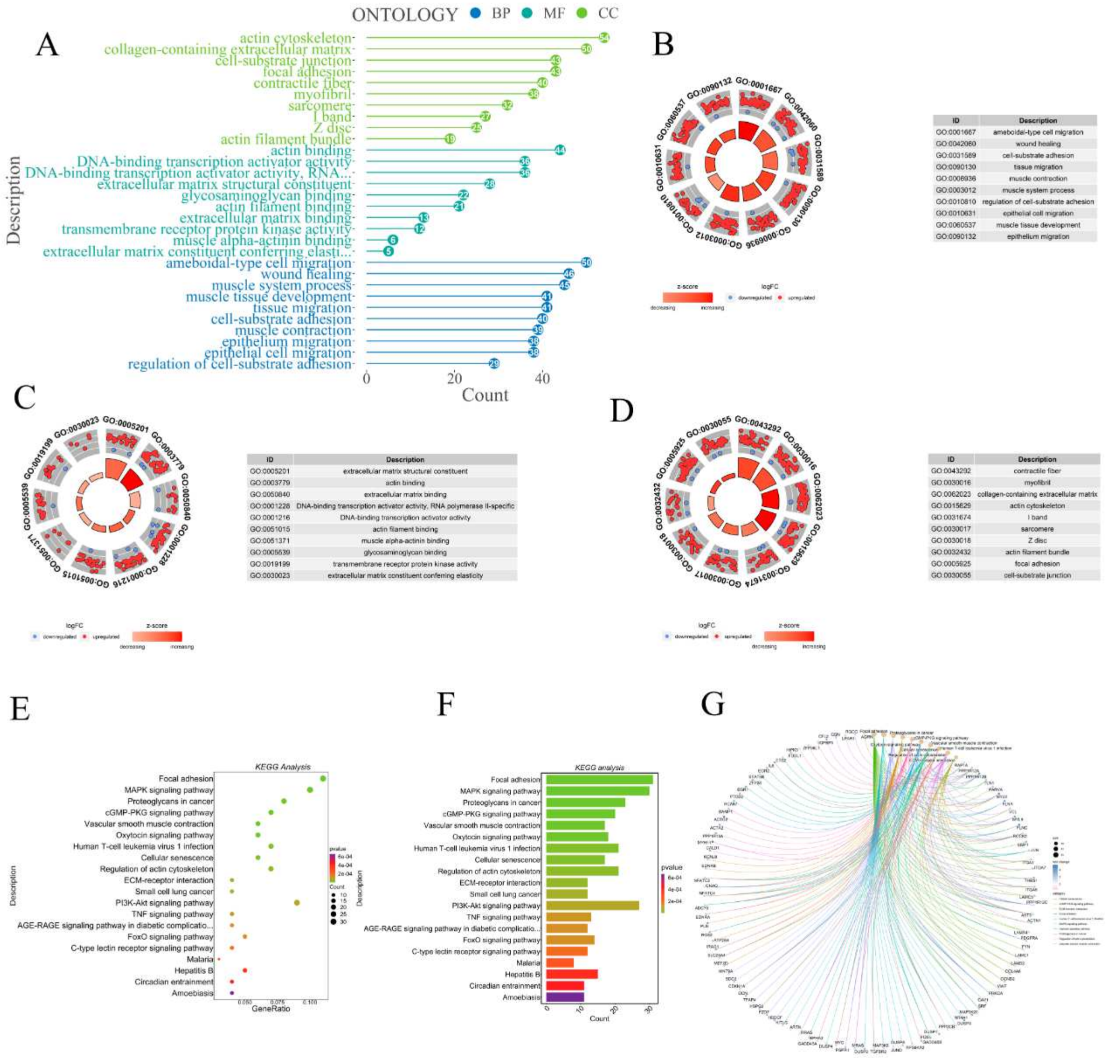


Fig 2. Differential gene analysis and WGCNA. (A) Volcano plots for differential analysis, Red and green indicate DEGs with up-regulated and down-regulated genes, respectively. The x-axis represents logFC, and the y-axis represents log10 (adj.P.Val). (B) Pick soft thresholds based on near scale-free topology criteria. (C) Identification of modules significantly associated with phenotypic data in cancer and normal groups. (D) Hierarchical clustering dendrogram for module identification. (E) Yellow modules with high association with cancer phenotypes. (F) Intersection of DGEs and WGCNA.



ID	Description
GO:0001667	ameboidal-type cell migration
GO:0042060	wound healing
GO:0031589	cell-substrate adhesion
GO:0090130	tissue migration
GO:0006936	muscle contraction
GO:0003012	muscle system process
GO:0010810	regulation of cell-substrate adhesion
GO:0010631	epithelial cell migration
GO:0060537	muscle tissue development
GO:0090132	epithelium migration

ID	Description
GO:0005201	extracellular matrix structural constituent
GO:0003779	actin binding
GO:0050840	extracellular matrix binding
GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-specific
GO:0001216	DNA-binding transcription activator activity
GO:0051015	actin filament binding
GO:0051371	muscle alpha-actinin binding
GO:0005539	glycosaminoglycan binding
GO:0019199	transmembrane receptor protein kinase activity
GO:0030023	extracellular matrix constituent conferring elasticity

ID	Description
GO:0043292	contractile fiber
GO:0030016	myofibril
GO:0062023	collagen-containing extracellular matrix
GO:0015629	actin cytoskeleton
GO:0031674	I band
GO:0030017	sarcomere
GO:0030018	Z disc
GO:0032432	actin filament bundle
GO:0005925	focal adhesion
GO:0030055	cell-substrate junction

Fig 3. Enrichment analysis. GO analysis (Biological Process, Cellular Component, and Molecular Function) of top 10 terms respectively. (A) lollipop chart. Circleplot as (B) BP. (C) MF. (D) CC. KEGG Analysis. (E-G) KEGG enrichment analysis of 607 genes, p<0.05 was considered to be statistically significant; BP: biological process; CC: cell component; MF: molecular function.

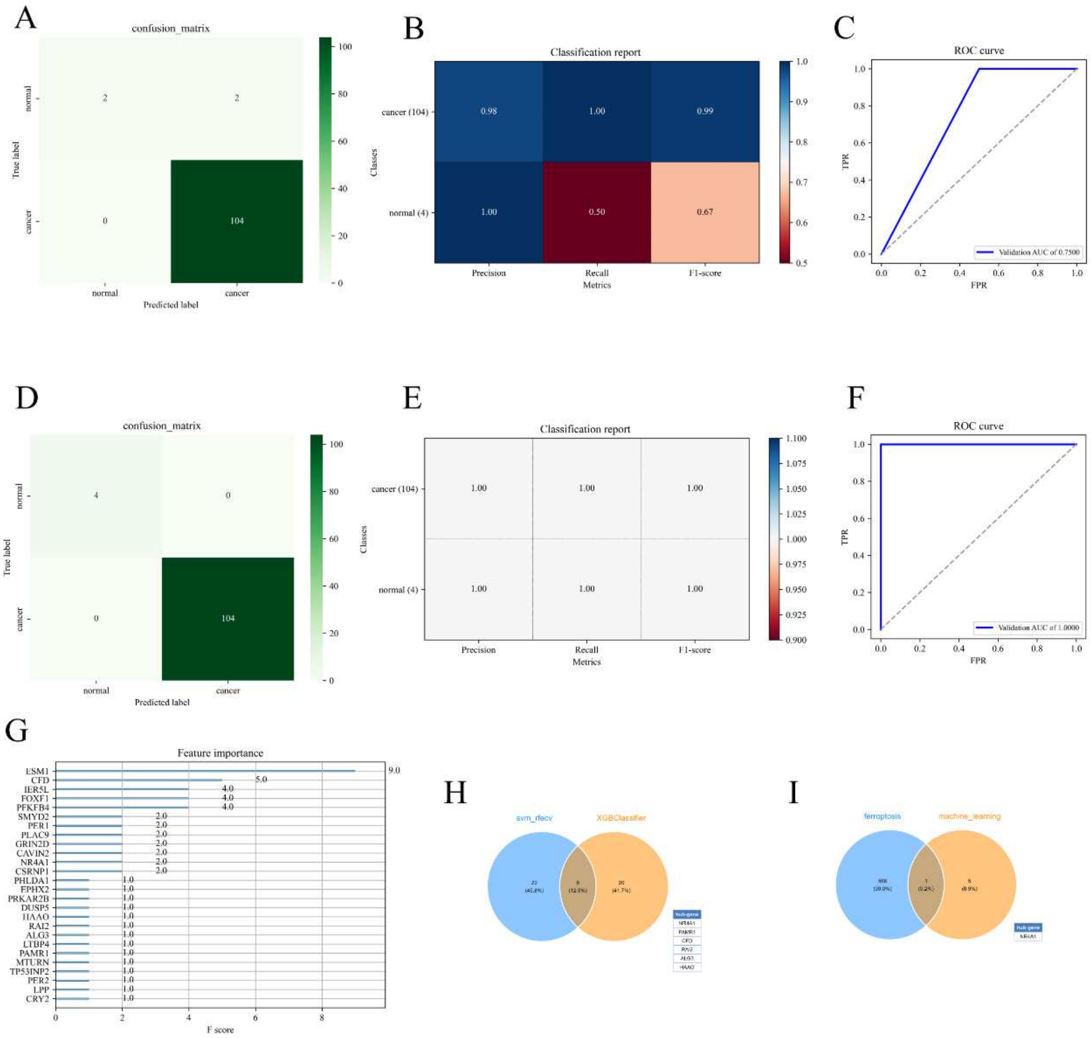


Fig 4. Machine learning. SVM-RFECV Analytics: (A) confusion matrix. (B) ROC curve. (C) classification reports. XGBoost Analytics: (D) confusion matrix. (E) classification reports. (F) ROC curve. (G) XGBoost feature importance graph. Genes with importance scores in the TCGA-BLCA gene expression prediction task and their specific scores. (H) Six key genes for SVM_RFECV and XGBoost. (I) key gene for machine and ferroptosis-related genes.

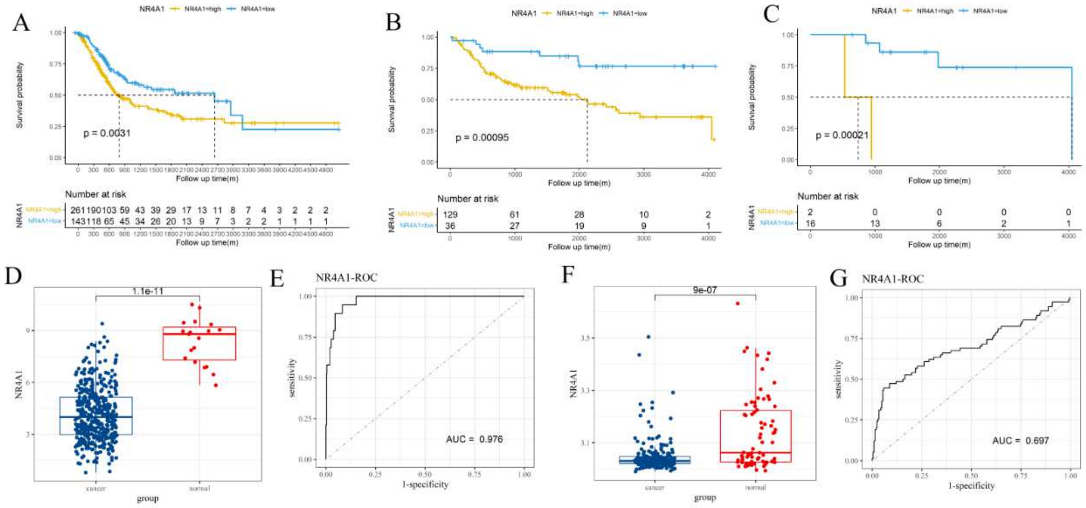


Fig 5. Prognostic value of identified genes for BC in TCGA-BLCA. Kaplan-Meier survival curves for patients of BC with high and low indicated gene expression in TCGA-BLCA, GSE13507 and GSE87315.

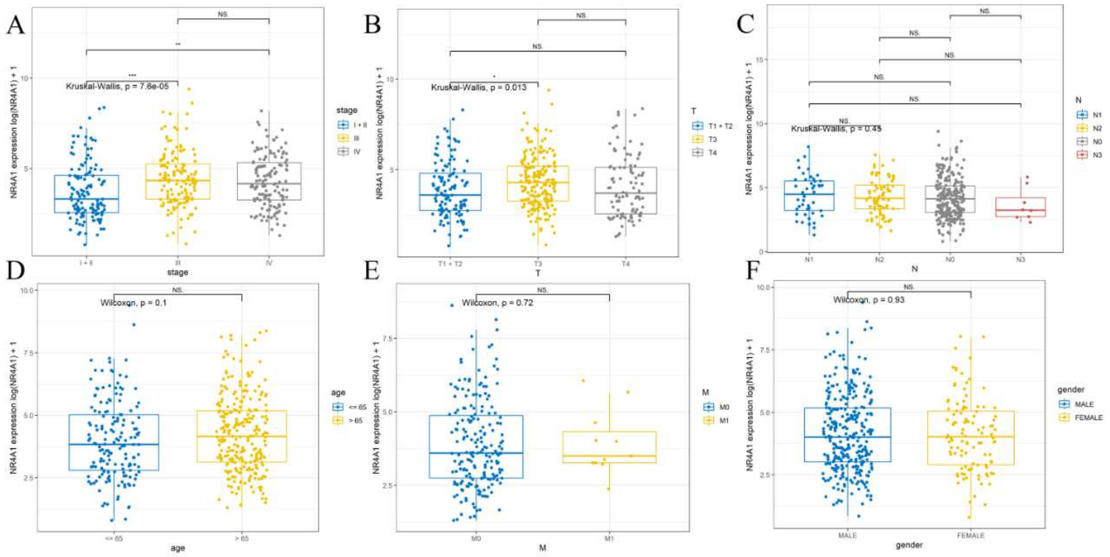


Fig 6. Pathological analysis of TCGA-BLCA. (A) stage. (B) T. (C) N. (D) age. (E) M. (F) gender.

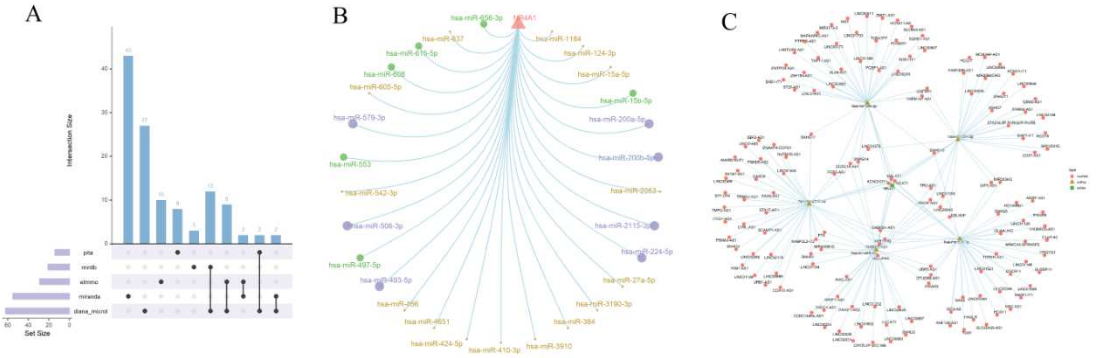


Fig 7. CeRNA network. (A, B) The ceRNA network's target miRNAs were predicted based on the Diana_microt, elmmo, microcosm, mirdb, pictar, pita, targetscan and miranda databases. Purple indicates that the miRNA is present in at least six databases, green suggests that it is present in five databases, and brown indicates that it is present in four databases. (C) Network of ceRNA interactions. Brown represents miRNAs, and red represents lncRNAs.

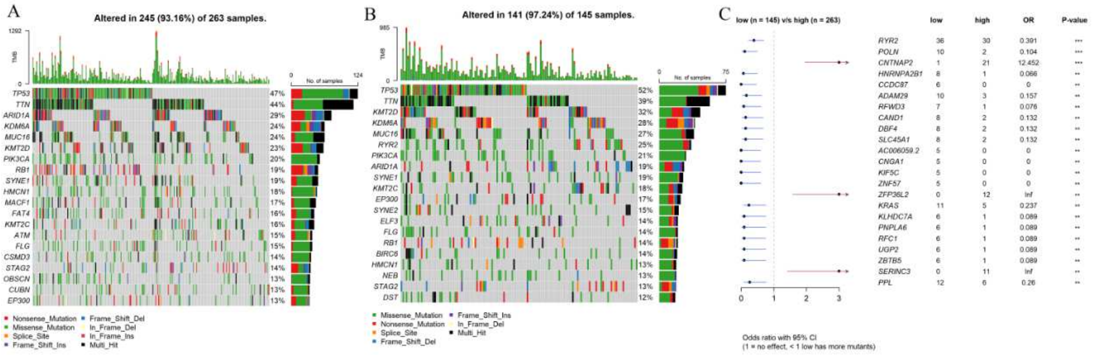


Fig 8. The relationship between TMB and the expression of NR4A1. (A, B) The oncoplots of the mutation genes in for the high and low NR4A1 expression groups. (C) Comparison of low and high expression of NR4A1.

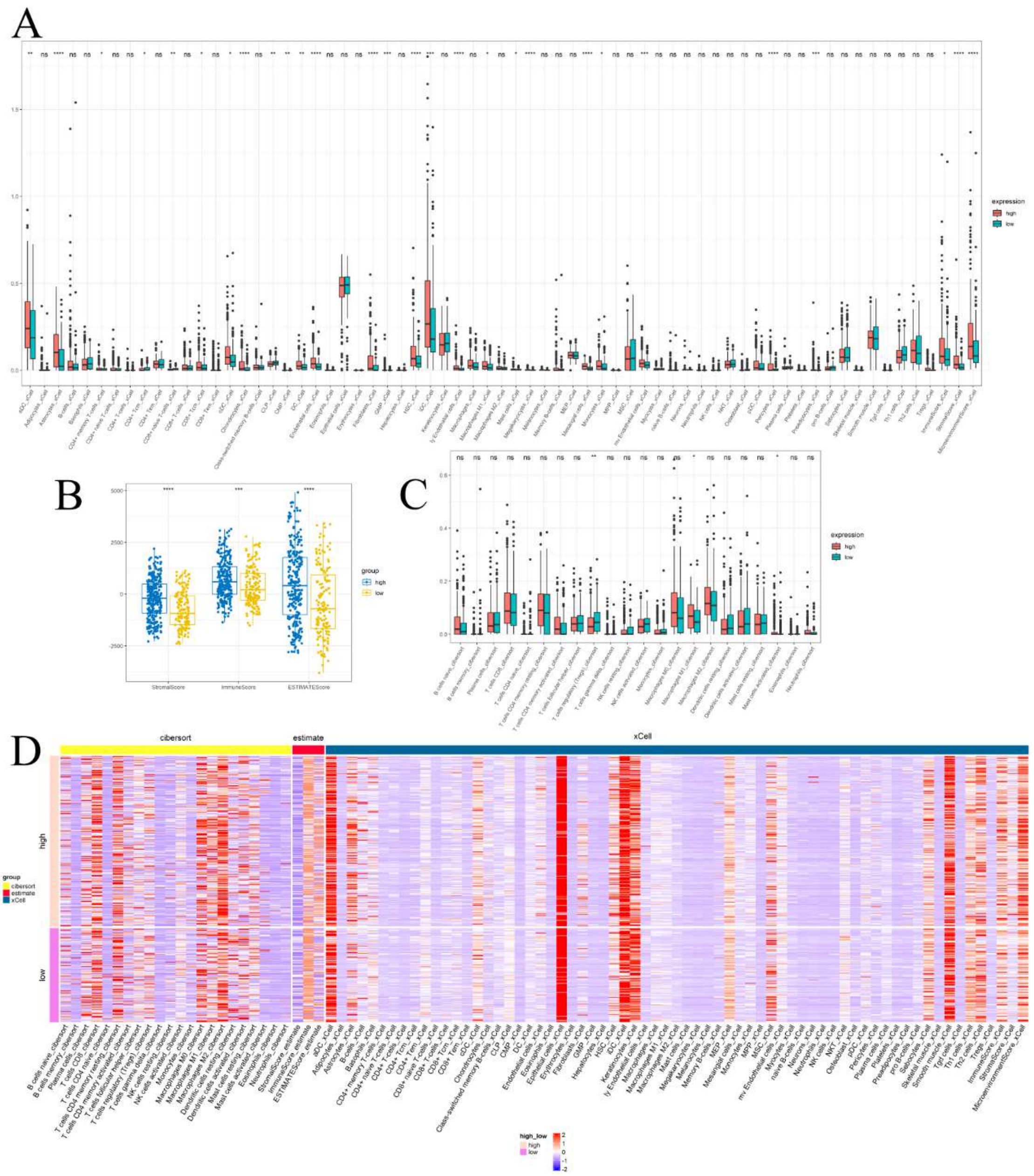


Fig 9. Immune infiltration analysis for the high and low NR4A1 expression groups. (A-D) Violin plot showing differences in immune cell types between the high and low-risk groups in xCell, estimate and CIBERSORT. (A) xCell (B) estimate (C) CIBERSORT.

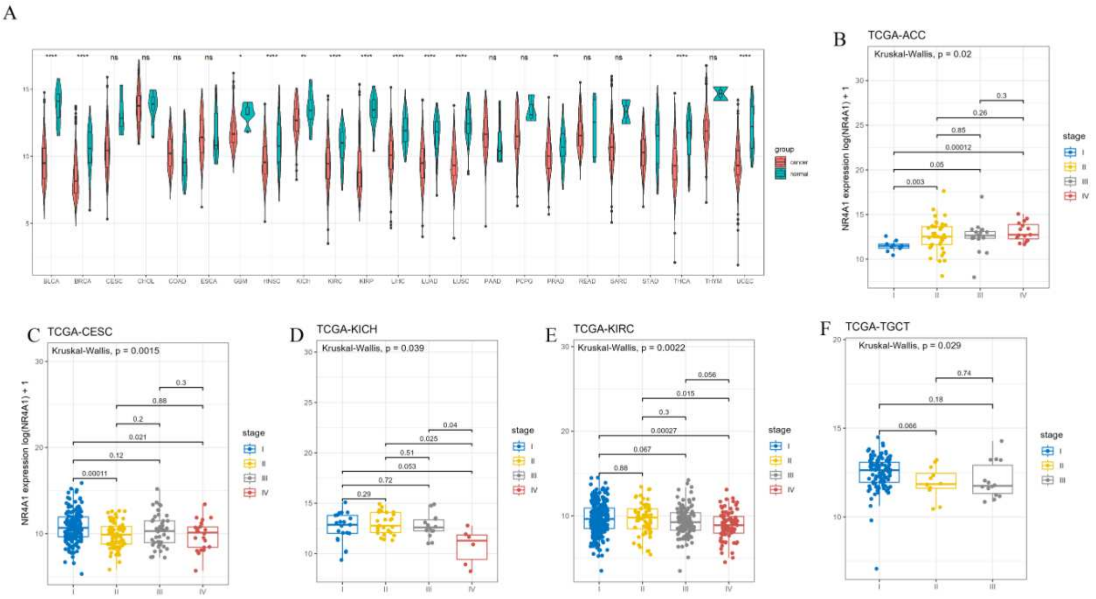


Fig 10. Pan-cancer analysis. (A) Expression levels of NR4A1 in different cancers compared with normal tissues. Red (green) indicates the cancer group (normal group). (B-F) Analysis of NR4A1 expression levels in cancers with stage. Only p<0.05 was shown.

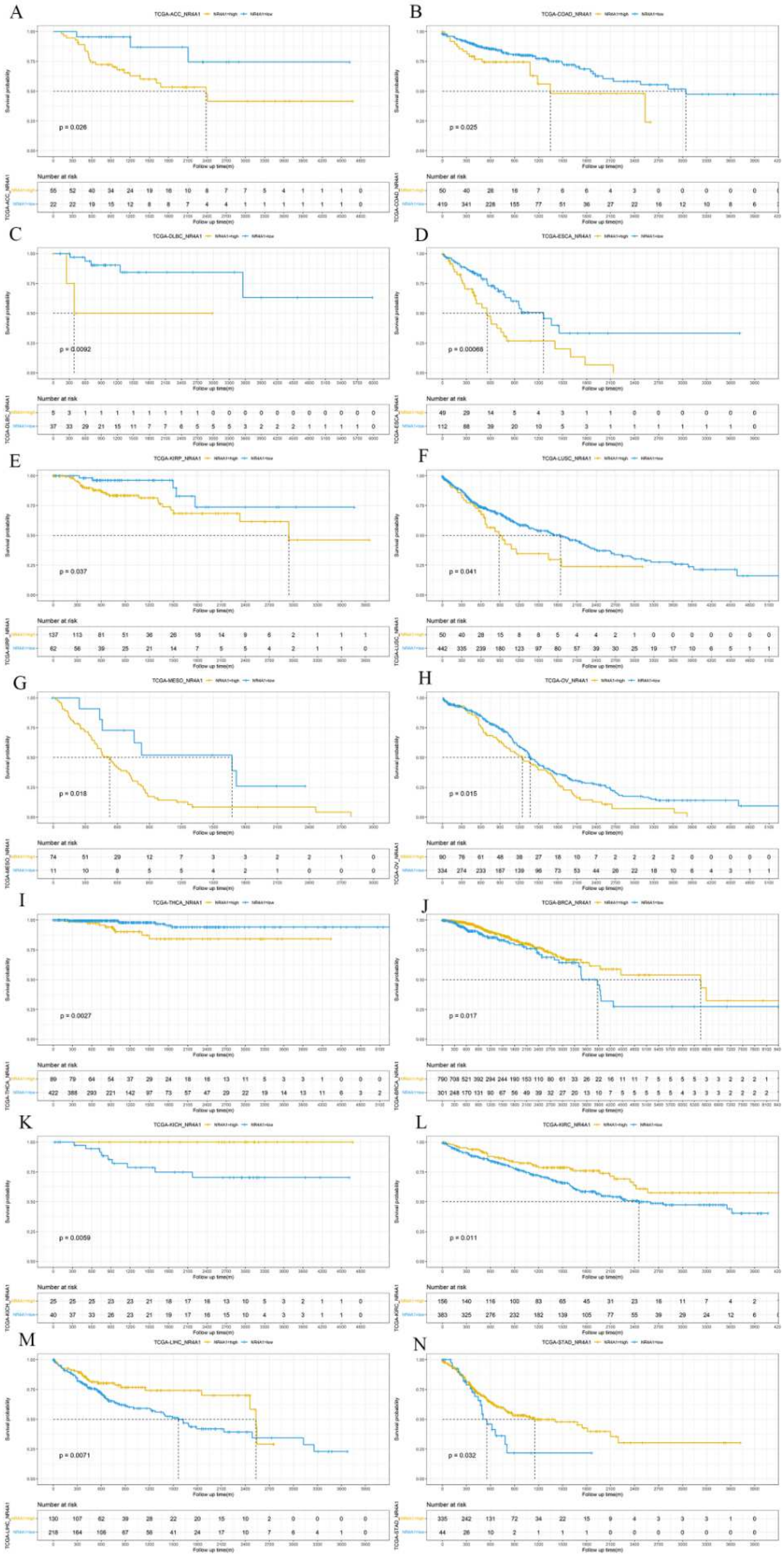


Fig 11. Survival curves in pan-cancer. (A) ACC, (B) COAD, (C) DLBC, (D) ESCA, (E) KIRP, (F) LUSC, (G) MESO, (H) OV, (I) THCA, (J) BRCA, (K) KICH, (L) KIRC, (M) LIHC, (N) STAD.

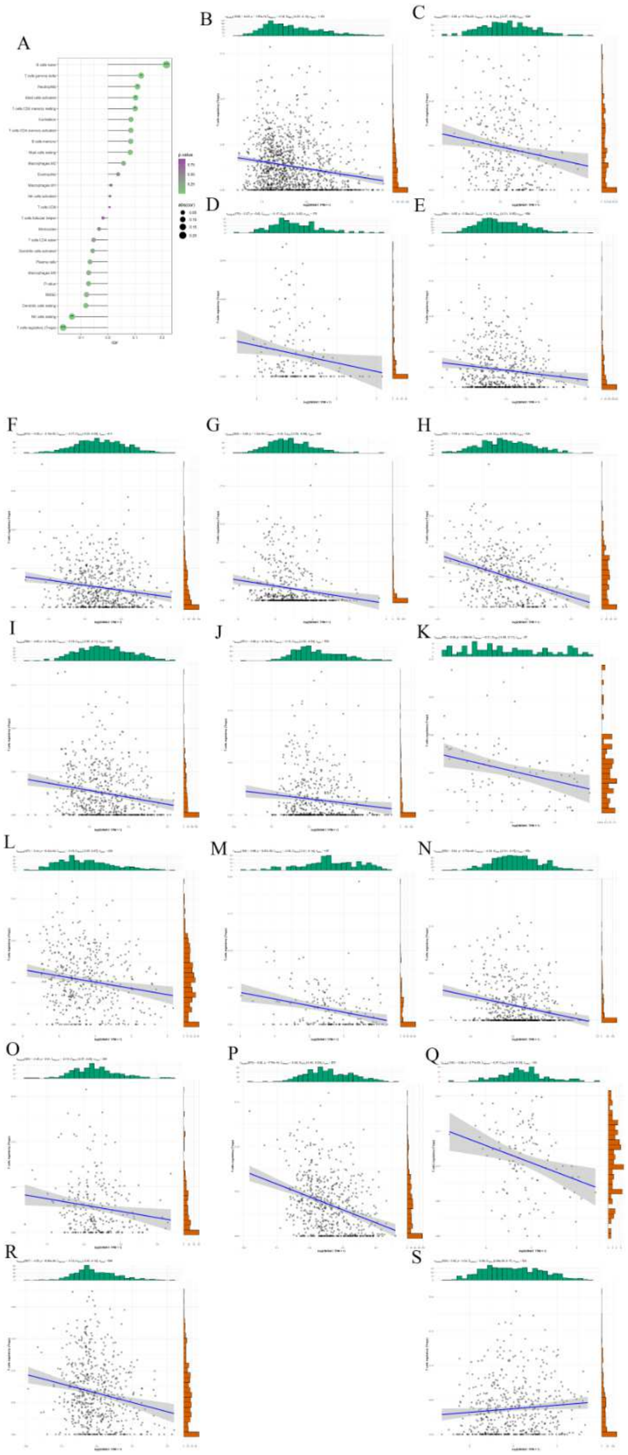


Fig 12. The correlation between NR4A1 expression and 22 kinds of immune cells, and the correlation between T cells regulatory (Tregs) and NR4A1 expression. (A) The correlation between NR4A1 expression and 22 kinds of immune cells, (B) BRCA, (C) CESC, (D) GBM, (E) HNSC, (F) KIRC, (G) LGG, (H) LIHC, (I) LUAD, (J) LUSC, (K) MESO, (L) OV, (M) PCPG, (N) PRAD, (O) SARC, (P) THCA, (Q)

THYM, (R) UCEC, (S) COAD.

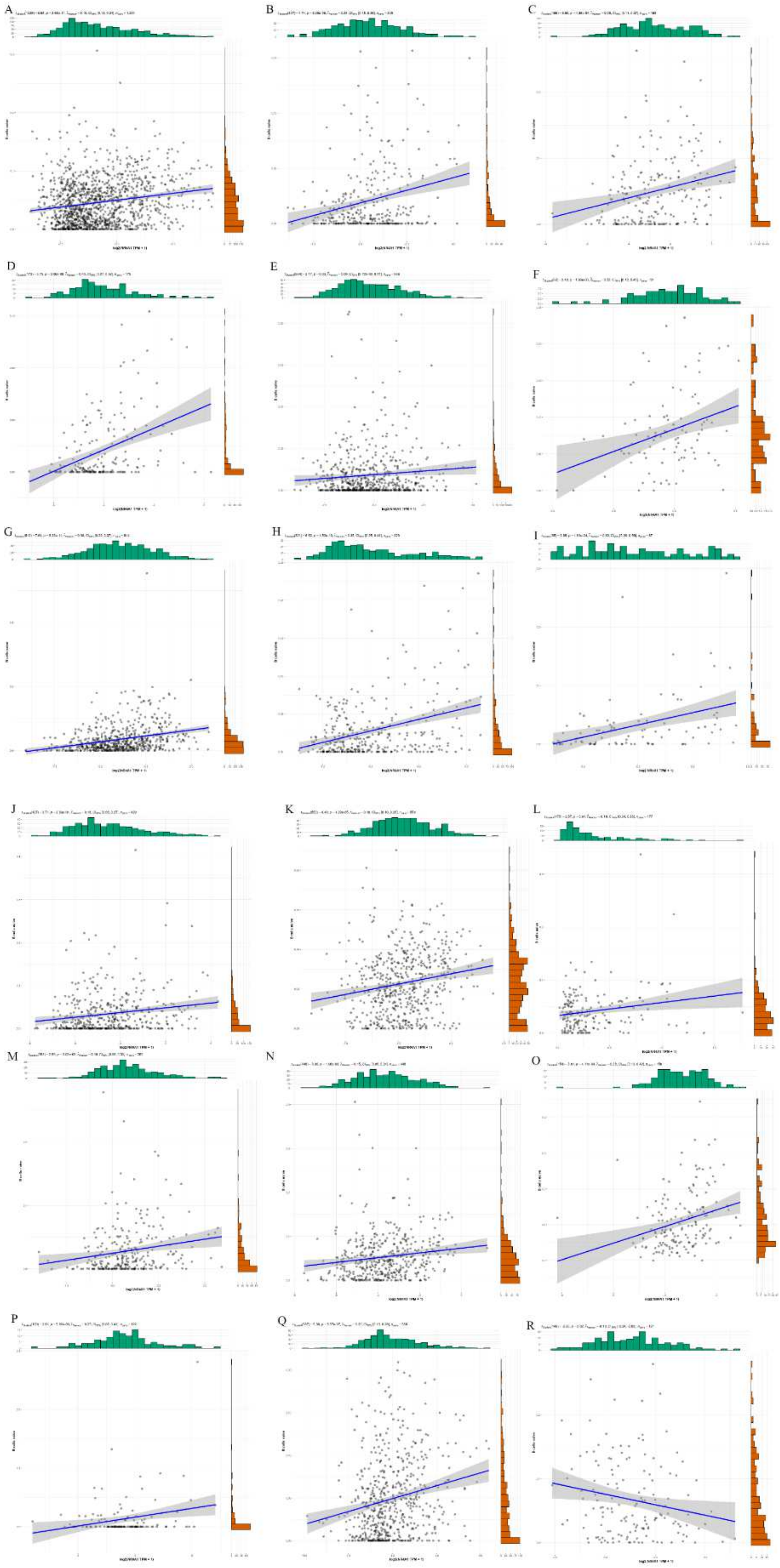


Fig 13. The correlation between B cells naive and NR4A1 expression. (A) BRCA, (B) CESC, (C) ESCA, (D) GBM, (E) HNSC, (F) KICH, (G) KIRC, (H) KIRP, (I) MESO, (J) OV, (K) PRAD, (L) READ, (M) SARC,

(N) STAD, (O) TGCT, (P) THYM, (Q) UCEC, (R) LAML.