scientific reports

Check for updates

OPEN

Construction and validation of a novel gene signature for predicting the prognosis of osteosarcoma

Jinpo Yang[1,4], Anran Zhang[2,4], Huan Luo[3] & Chao Ma[3]

Osteosarcoma (OS) is the most common type of primary malignant bone tumor. The high-throughput sequencing technology has shown potential abilities to illuminate the pathogenic genes in OS. This study was designed to find a powerful gene signature that can predict clinical outcomes. We selected OS cases with gene expression and survival data in the TARGET-OS dataset and GSE21257 datasets as training cohort and validation cohort, respectively. The univariate Cox regression and Kaplan–Meier analysis were conducted to determine potential prognostic genes from the training cohort. These potential prognostic genes underwent a LASSO regression, which then generated a gene signature. The harvested signature's predictive ability was further examined by the Kaplan–Meier analysis, Cox analysis, and receiver operating characteristic (ROC curve). More importantly, we listed similar studies in the most recent year and compared theirs with ours. Finally, we performed functional annotation, immune relevant signature correlation identification, and immune infiltrating analysis to better study he functional mechanism of the signature and the immune cells' roles in the gene signature's prognosis ability. A seventeen-gene signature (*UBE2L3, PLD3, SLC45A4, CLTC, CTNNBIP1, FBXL5, MKL2, SELPLG, C3orf14, WDR53, ZFP90, UHRF2, ARX, CORT, DDX26B, MYC, and SLC16A3*) was generated from the LASSO regression. The signature was then confirmed having strong and stable prognostic capacity in all studied cohorts by several statistical methods. We revealed the superiority of our signature after comparing it to our predecessors, and the GO and KEGG annotations uncovered the specifically mechanism of action related to the gene signature. Six immune signatures, including *PRF1, CD8A, HAVCR2, LAG3, CD274*, and *GZMA* were identified associating with our signature. The immune-infiltrating analysis recognized the vital roles of T cells CD8 and Mast cells activated, which potentially support the seventeen-gene signature's prognosis ability. We identified a robust seventeen-gene signature that can accurately predict OS prognosis. We identified potential immunotherapy targets to the gene signature. The T cells CD8 and Mast cells activated were identified linked with the seventeen-gene signature predictive power.

Osteosarcoma (OS) is a bone tumor that occurs predominantly in adolescents and young adults[1–3]. In the 0–24 age group, the incidence of osteosarcoma among men, women and children is 4.4 per million persons per year[1–3]. The latest advances in molecular genetics of osteosarcoma have changed our views on the cause of the disease and the continued treatment of patients[1–3]. Surgical removal of clinically visible tumors and systemic chemotherapy are currently popular disease management strategies[1–3]. Although the cure rate for patients with local disease is close to 70%, the 5-year overall survival rate for patients with metastatic disease is less than 25%, and most patients die from lung metastases[4]. Unfortunately, the treatment paradigm for OS has remained unchanged for approximately 30 years[1,4]. Therefore, continued efforts are urgently needed for a steady prognostic model for OS patients.

[1]Department of Medical Oncology, The Affiliated Cancer Hospital of Zhengzhou University, Henan Cancer Hospital, Zhengzhou, China. [2]Department of Oncology, Henan Provincial People's Hospital, Zhengzhou University People's Hospital, Henan University People's Hospital, Zhengzhou, China. [3]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and the Berlin Institute of Health, Berlin, Germany. [4]These authors contributed equally: Jinpo Yang and Anran Zhang. email: huan.luo@charite.de; chao.ma@charite.de

The rise of throughput sequencing technology helps clarify disease-causing genes, explore disease pathogenesis, develop biomarkers, and profoundly change our understanding of biology and human diversity[5]. Researchers have developed many statistical models that use genomic data to accurately predict whether the prognostic risk of cancer patients is high or low[6–10]. Many researchers have screened multiple biomarkers related to OS by mining gene expression data[5]. Gene signature can contain more than one single gene with a unique characteristic pattern of gene expression resulting from an altered or unaltered biological process or pathogenic medical condition[11]. Gene signature has a more stable ability and higher fault tolerance for prognostic prediction in cancer studies[7–12].

Finding multiple molecules from the OS gene profile to construct a gene signature can better predict outcome potentially. To fill in the void and find a promising gene signature that targets OS outcomes, this work tried to identify a prognostic gene signature from the TARGET database. More importantly, the signature we found was further tested in an independent dataset for its prognostic ability and was compared to the models built in the most recent year for its superiority. In the end, the functional annotation, immune relevant signature correlation analysis, and 22 tumor-infiltrating immune cells (TICs) analysis were conducted for the full understanding of the gene signature we discovered.

Materials and methods

Database selection. The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) is a dynamically updated database of the National Cancer Institute (NCI) Office of Cancer Genomics (OCG). Its mission is to advance the molecular understanding of cancer to improve patient prognosis[13]. The TARGET Osteosarcoma (TARGET-OS) project has elucidated a comprehensive molecular profile to identify the genetic changes that drive the occurrence and development of high-risk or difficult-to-treat childhood cancers. OS datasets are available without restrictions on their use in publications or presentations and can be obtained from the official web patrol (https://ocg.cancer.gov/programs/target/projects/osteosarcoma) or GDC Xena Hub (https://gdc.xenahubs.net). TARGET-OS was set as training cohort. Eighty-eight OS cases were included, and their gene expression profile, survival time, survival status, and clinical characteristics were obtained. Gene Expression Omnibus (GEO) is an internationally recognized and widely researched public repository for archiving and free distribution of microarray[14]. We searched the GEO using the keyword "Osteosarcoma" and set the filters as follows: (1) organisms: homo sapiens; (2) entry type: series; (3) study type: expression profiling by array; (4) the number of samples with expression data is greater than 50; (5) the number of samples with survival data is greater than 50. One dataset named GSE21257 ($n = 53$) was obtained from GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21257) and treated as validation cohorts to exam the gene signature we constructed. We strictly obeyed the guidelines of the two databases. This study was approved by the Institutional Review Board of Henan Cancer Hospital, which waived the requirement for informed consent due to the use of data obtained from the public databases. All methods were performed in accordance with relevant guidelines and regulations.

Identification of the potential prognostic genes. We implemented a univariate Cox proportional hazard model and Kaplan–Meier estimator to identify genes with potential prognostic ability. In our research, gene expression, survival status, and survival time were input into the R language. With the help of the "survival" and "survminer" R packages, Kaplan–Meier estimator could sort out genes with the ability to distinguish patients' outcomes. The Kaplan–Meier significance threshold was set to $p < 0.05$. Similarly, each gene's univariate Cox model was built using the gene expression data, survival status, survival time, and the adoption of the "survival" R package. The univariate Cox model significance threshold was set to $p < 0.05$. The gene in both tests having a p value < 0.05 was considered the potential prognostic gene.

Gene signature construction and validation. Subsequently, we put the potential prognostic genes identified in the previous step into the LASSO model to detect the best lambda[15–18]. Specifically, we utilized the expression data of the potential prognostic genes, patients' survival data, and the "glmnet" R package to perform the LASSO Cox regression with tenfold cross-validation. Then the R program outputted a list of prognostic genes with coefficients based on the best lambda value selected. According to the instructions and characteristics of the "glmnet" R software package, the selected genes with coefficients would be out putted. The calculation method of the risk score level of each OS is using the following formula:

$$Riskscore = \sum_{i}^{n} Expi * \beta i$$

In the above formula, n represents each hub gene in the gene signature; Expi represents the expression level of each gene; βi represents the coefficient of each gene.

To test our signature's ability in all studied cohorts, the Kaplan–Meier analysis was used to determine the outcome differences between high- and low-risk patients, of which the OS were classified according to the median risk score. In addition, univariate and multivariable Cox analyses further examined the predictive potential of the gene signature. The area under the curve (AUC) is a measure of the classifier's ability to distinguish classes and is used as a summary of the ROC curve[19]. The higher the AUC, the better the model's performance in determining between positive and negative classes.

Comparison of gene signature with previously published models. We searched PubMed (https://pubmed.ncbi.nlm.nih.gov/) using the keyword "gene signature prognosis osteosarcoma" and made the selection based on criteria we set as follows: (1) the impact factor > 4 (Journal Citation Reports Year 2020, Clarivate, https://jcr.clarivate.com/jcr/home); (2) the online publication date of the article is the most recent year (i.e. May

18, 2020, to May 18, 2021); (3) the candidate study contains specifically findings of the signature's composition and coefficients. We extracted the gene signatures and the coefficients from the studies and applied them to the studied cohorts to calculate the risk score of each case. The most important thing was that we used the risk scores to build Kaplan–Meier analysis and Cox model to strictly assess the prognostic ability of our predecessors and ours, thus for horizontal comparison.

Function analysis of the gene signature in OS. Gene ontology (GO), including Biological Process (BP), Cellular Components (CC), and Molecular Functions (MF), and Kyoto Encyclopedia of Genes and Genomes (KEGG) were conducted to find the potential function of genes between high- and low-risk groups[20–22]. Enrich items with p value < 0.05 were considered significant.

Correlations between gene signature and immune relevant signatures. We analyzed the immune activity and tolerance of low- and high-risk groups in the training cohort. Firstly, we picked *CD274, CTLA4, HAVCR2, IDO1, LAG3*, and *PDCD1* as immune-checkpoint-relevant signatures, and *CD8A, CXCL10, CXCL9, GZMA, GZMB, IFNG, PRF1, TBX2,* and *TNF* as immune-activity-relevant signatures. We adopted an integrated analysis including the Pearson correlation coefficient and Wilcoxon rank-sum to determine the interaction between gene signature and immune relevant signatures.

Determine the relationships between our signature and 22 TICs. We applied a comprehensive analysis based on the Pearson coefficient and Wilcoxon's rank-sum test to evaluate the relationship between 22 TICs and the signatures of this study. In the following analysis, in order to determine the prognostic ability of the 22 TICs, we combined two kinds of statistical approaches, including univariate Cox models and Kaplan–Meier analysis. Together with the evidence found in the first half of this section, we could infer the potential TICs that play crucial roles in the signature's prognosis ability.

Statistical analysis. We adopt the "CIBERSORT" R package to estimate the abundance of 22 TICs using the gene expression data of the cohorts. LASSO regression was carried out by the "glmnet" R package. Kaplan–Meier plots were constructed by the integration of the "survival" and "survminer" R packages. Cox models, including univariate and multivariable were built via the "survival" R package. The ROC curves was made possible with the help of the "pROC" R package. R software (version 4.0.4, Windows 64-bit) carried out all the processes in this study.

Results

Cohorts' characteristics. As Fig. 1 demonstrates, 88 OS cases that came from the TARGET-OS cohort were taken for model training. The dataset GSE21257, contained 53 OS cases, were selected for model validating. For patients included in the study, we have collected their clinical characteristics and shown them in Table 1 in detail.

Prognostic gene signature identification. The univariate Cox regression and Kaplan–Meier analysis were conducted to test each gene's prognostic ability. As shown in Table S1, 70 genes were identified by the Kaplan–Meier estimator, while, 80 genes were determined from the univariate Cox regression model, which has the predictive ability. We intersected them, found 57 genes suitable for our study, and included them in our next analyses (Table S2). The LASSO algorithm displayed when 17 genes existed, the model could achieve the optimized ability (Fig. 2A,B). Table 2 shows the coefficients of the 17 genes.

Confirmation of the prognostic capacity of the seventeen-gene signature. In the risk plot in Fig. 3, we displayed the survival time, survival status, and relative expression of the hub genes for each sample, so as to show the distinguishing ability of the signature in a macroscopic view. In the training cohort, *UBE2L3, PLD3, SLC45A4, CLTC, CTNNBIP1, FBXL5, MKL2, SELPLG, C3orf14, WDR53*, and *ZFP90* have protective abilities for OS patients, while *UHRF2, ARX, CORT, DDX26B, MYC,* and *SLC16A3* display unfavored for the OS prognosis (Figure S1A).

After drawing the risk plot, we first chose the Kaplan–Meier estimator to estimate the ability of the model we built. As shown in Fig. 4, the survival probability of the high-risk group in the training cohort is lower than that of the low-risk group (p value = 1.764E−08), the same is happening in the validation cohort (p value = 8.915E−03), which demonstrated significant survival differences occurred in the signature distinguished patients.

The univariate and multivariable Cox regression were established to exam the signature's prognostic capacity (Fig. 5). Analysis in the training cohort showed that the risk score affected the OS patients' outcomes (p value ≤ 5.42E−06). Consistently, the results in the validation cohort proved that risk score was the best one affecting prognosis in either univariate or multivariable examination, furtherly confirmed the powerful predictive capacity of the gene signature (p value ≤ 4.53E−05).

As shown in Fig. 6A, ROC analysis indicated that the area under the curve (AUC) for our seventeen-gene signature risk score reached 0.891 (95% CI 0.780–0.995, best cutoff = − 5.633), which was the best among other clinical factors. In the GSE21257 cohort, the AUC as well arrived at 0.777 (95% CI 0.780–0.995, best cutoff = − 9.553), topping other characteristics (Fig. 6B).

Our gene signature is superior to previous ones. Based on the screening criteria set, we found nine studies that suit for our comparison (Table 3). We applied these discovered signatures and their risk score equa-

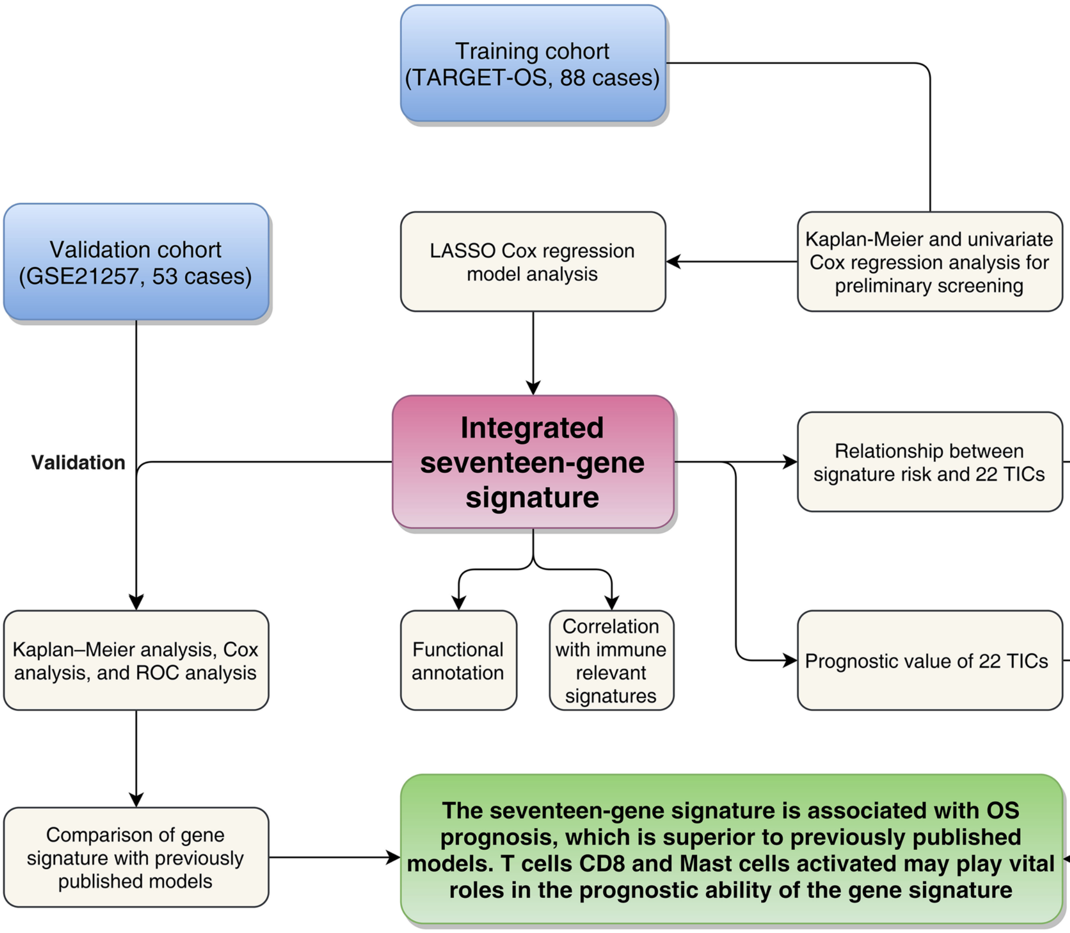

Training cohort (TARGET-OS, 88 cases)
Validation cohort (GSE21257, 53 cases)
LASSO Cox regression model analysis
Kaplan-Meier and univariate Cox regression analysis for preliminary screening
Validation
Integrated seventeen-gene signature
Relationship between signature risk and 22 TICs
Kaplan–Meier analysis, Cox analysis, and ROC analysis
Functional annotation
Correlation with immune relevant signatures
Prognostic value of 22 TICs
Comparison of gene signature with previously published models
The seventeen-gene signature is associated with OS prognosis, which is superior to previously published models. T cells CD8 and Mast cells activated may play vital roles in the prognostic ability of the gene signature

Figure 1. Flow chart of the study. *LASSO* least absolute shrinkage and selection operator Cox regression model, *ROC* receiver operating characteristic, *OS* osteosarcoma, *TICs* tumor-infiltrating immune cells.

tions to our training and validation cohorts to calculate the risk score of each OS patient. Then, the Kaplan–Meier estimators were built against our signature and previous models (Fig. 7), demonstrating Yang et al.'s and ours have the ability predicting the outcomes of OS. However, our gene signature (p value ≤ 8.915E−03) seemed to be stronger than Yang et al.'s (p value ≤ 3.602E−02) in terms of their p values.

Additionally, Cox univariate and multivariable regression were constructed using these selected prognosis models (Fig. 8). The results in the training cohort demonstrated that only our gene signature (p value ≤ 1.33E−06) having the prognosis capabilities in both the univariate and multivariable analyses. The Cox analysis of the validation cohort determined that only our gene signature passed all the univariate and multivariable tests (p value ≤ 4.31E−06).

GO and KEGG enrichment analysis with the seventeen-gene signature.

According to the risk score for each case in the TARGET-OS cohort, we conducted GO and KEGG enrichment analysis between high-risk and low-risk groups. The GO enrichment result showed the differences between the two groups mainly focus on extracellular matrix organization, extracellular structure organization, collagen—containing extracellular matrix, endoplasmic reticulum lumen, and extracellular matrix structural constituent (Figure S2A). KEGG analysis was showed that the enriched items were mainly related to protein digestion and absorption, complement and coagulation cascades, and Wnt signaling pathway (Figure S2B).

Relationships between the seventeen-gene signature and immune relevant signatures.

We observed that 8/15 of the immune relevant signatures in the high-risk group were significantly under expressed, as demonstrated by the Wilcoxon test (Fig. 9A). The Pearson coefficient test discovered 7/15 of the immune relevant signatures correlated with the seventeen-gene signature (Fig. 9B, Table S3). Incorporating the above findings, six genes, including *PRF1, CD8A, HAVCR2, LAG3, CD274,* and *GZMA* were identified associating with the seventeen-gene signature.

The seventeen-gene signature and 22 TICs.

The GO and KEGG analysis suggested that the difference between the two groups was related to the immune response, so we further conducted 22 TICs analysis to better study how the seventeen-gene signature interact with the immune microenvironment. CIBERSORT algorithm

Characteristics	Training cohort (TARGET-OS, n = 88)	Validation cohort (GSE21257, n = 53)
Age		
< 14	39 (44.32%)	15 (28.3%)
≥ 14	45 (51.14%)	38 (71.7%)
Unknown	4 (4.55%)	0
Gender		
Female	37 (42.05%)	19 (35.85%)
Male	47 (53.41%)	34 (64.15%)
Unknown	4 (4.55%)	0
Race		
Non-White	13 (14.77%)	NA
White	51 (57.95%)	NA
Unknown	24 (27.27%)	NA
Ethnicity		
Not Hispanic or Latino	52 (59.09%)	NA
Hispanic or Latino	11 (12.5%)	NA
Unknown	25 (28.41%)	NA
Tumor location		
Femur	NA	27 (50.94%)
Fibula	NA	2 (3.77%)
Humerus	NA	8 (15.09%)
Tibia	NA	15 (28.3%)
Unknown	NA	1 (1.89%)
Histological subtype		
Chondroblastic	NA	6 (11.32%)
Fibroblastic	NA	5 (9.43%)
Osteoblastic	NA	32 (60.38%)
Others	NA	10 (18.87%)
Metastatic status		
Non-metastatic	63 (71.59%)	39 (73.58%)
Metastatic	21 (23.86%)	14 (26.42%)
Unknown	4 (4.55%)	0
Survival status		
Alive	58 (65.91%)	30 (56.6%)
Dead	27 (30.68%)	23 (43.4%)
Unknown	3 (3.41%)	0

Table 1. Clinical characteristics of patients involved in the study.

was used to determine the proportion of the tumor-infiltrating immune subpopulations. We visual outputted the 22 TICs distribution and inner correlation in Figure S3.

Combining the findings from difference analysis (Fig. 10A) and correlation analysis (Fig. 10B, Table S4), three TICs (Fig. 10C), including T cells CD8, Mast cells activated, and T cells CD4 memory activated were identified associating with the seventeen-gene signature. Among them, Mast cells activated were found positively correlated with the gene signature, while the others negatively.

We further tested the 22 TICs prognostic abilities by consulting the Kaplan–Meier estimator and univariate Cox proportional-hazard model. As displayed in Fig. 11, the univariate Cox proportional-hazard model (Fig. 11A) indicated that T cells CD8, T cells CD4 memory activated, T cells CD4 naïve, Dendritic cells resting, and Mast cells activated impacted prognosis. Additionally, Kaplan–Meier estimator (Fig. 11B; Table S5) highlighted that T cells CD8, T cells CD4 naïve, and Mast cells activated can predict the survival rate of OS. From the above survival analysis, it can be determined that T cells CD8, T cells CD4 naïve, and Mast cells activated have potential prognostic ability in OS.

The results of this part found that T cells CD8 and Mast cells activated were significantly related to our signature and closely related to the OS prognosis, potentially implying that T cells CD8 and Mast cells activated infiltrations play vital roles in the discovered signature in OS patients.

Discussion

In this study, we innovatively discovered a robust seventeen-gene prognostic signature for the OS prognosis by mining TARGET and GEO databases. Specifically, our novelty lay in using univariate Cox model, Kaplan–Meier estimator, and LASSO regression in the model training phase. The adoption of an independent

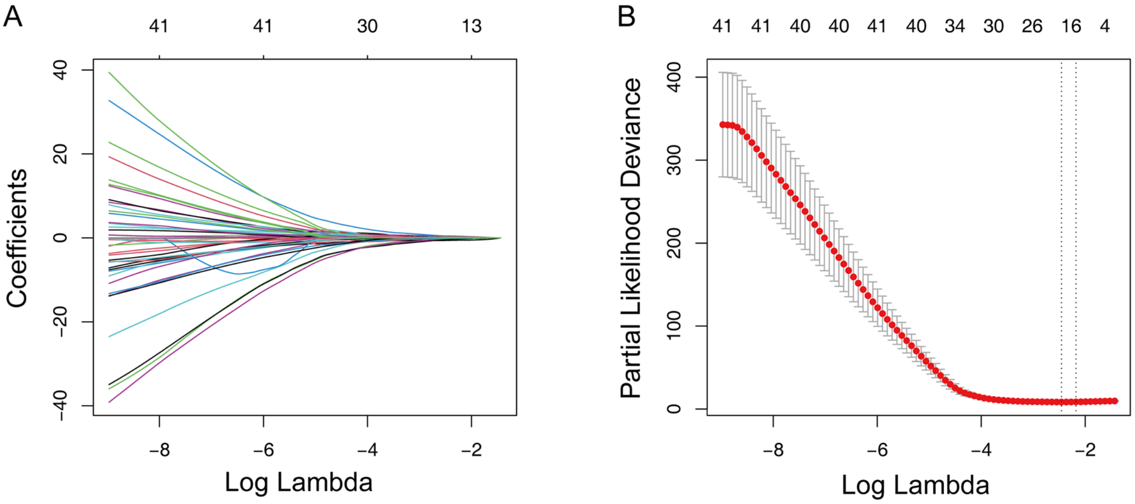


Figure 2. LASSO regression analysis for the construction of prognostic gene signature. (**A**) Cross-validation for tuning parameter screening upon LASSO regression analysis. (**B**) Screening of optimal parameter (lambda) at which the vertical lines were drawn. *LASSO* the least absolute shrinkage and selection operator Cox regression model.

Gene symbol	Description	Risk coefficient
C3orf14	Chromosome 3 Open Reading Frame 14	− 0.106872216
UHRF2	Ubiquitin Like With PHD And Ring Finger Domains 2	0.321564173
DDX26B	Integrator Complex Subunit 6 Like	0.334217953
ZFP90	ZFP90 Zinc Finger Protein	− 0.505473926
FBXL5	F-Box And Leucine Rich Repeat Protein 5	− 0.24934174
UBE2L3	Ubiquitin Conjugating Enzyme E2 L3	− 0.308369838
MYC	MYC Proto-Oncogene, BHLH Transcription Factor	0.250664103
CLTC	Clathrin Heavy Chain	− 0.356348745
ARX	Aristaless Related Homeobox	0.444234486
CTNNBIP1	Catenin Beta Interacting Protein 1	− 0.601488248
CORT	Cortistatin	0.220902785
SELPLG	Selectin P Ligand	− 0.070201073
WDR53	WD Repeat Domain 53	− 0.059234551
SLC16A3	Solute Carrier Family 16 Member 3	0.002731127
MKL2	Myocardin Related Transcription Factor B	− 0.028375916
SLC45A4	Solute Carrier Family 45 Member 4	− 0.156290246
PLD3	Phospholipase D Family Member 3	− 0.128837662

Table 2. Prognostic genes obtained from LASSO Cox regression model.

cohort, Kaplan–Meier analysis, Cox regression, ROC curve in the validation process, moreover, highlighted our innovativeness. Most importantly, we compared our signature with published research to prove ours' superiority. At the end of the study, we discovered important mechanisms related to gene signature through function annotations, immune gene correlation analysis, and immune infiltration analysis and speculated that the T cells CD8 and Mast cells activated might potentially help the predictive ability of the signature. This study we worked on designed to shed light on the development of future OS research.

Our signature consists of seventeen genes (Table 2), which were *UBE2L3, PLD3, SLC45A4, CLTC, CTNNBIP1, FBXL5, MKL2, SELPLG, C3orf14, WDR53, ZFP90, UHRF2, ARX, CORT, DDX26B, MYC,* and *SLC16A3*, respectively. After tested in the two cohorts (Figure S1), *UBE2L3, PLD3, SLC45A4, CTNNBIP1, FBXL5, SELPLG, WDR53,* and *ZFP90*, showed solid protective impacts on OS, while *UHRF2, ARX, CORT, DDX26B, MYC,* and *SLC16A3* witnessed effects on OS prognosis unfavorably. Our findings suggest that *CTNNBIP1* is a suppressor of cancer migration, thus making it a potential prognostic predictor for OS. Rothzerg et al. also demonstrated that high expression of CTNNB1 was associated with a good OS prognosis, which is consistent with our findings[23]. In the *UHRF* family, *UHRF1* and UHRF2 have a multidomain architecture and have similarities in sequence and domain organization[24]. *UHRF1* is a well-known epigenetic regulator. Significant *UHRF1* overexpression has been shown in many kinds of tumors[25]. Liu et al. reported that *UHRF1* promotes the proliferation of human OS cells

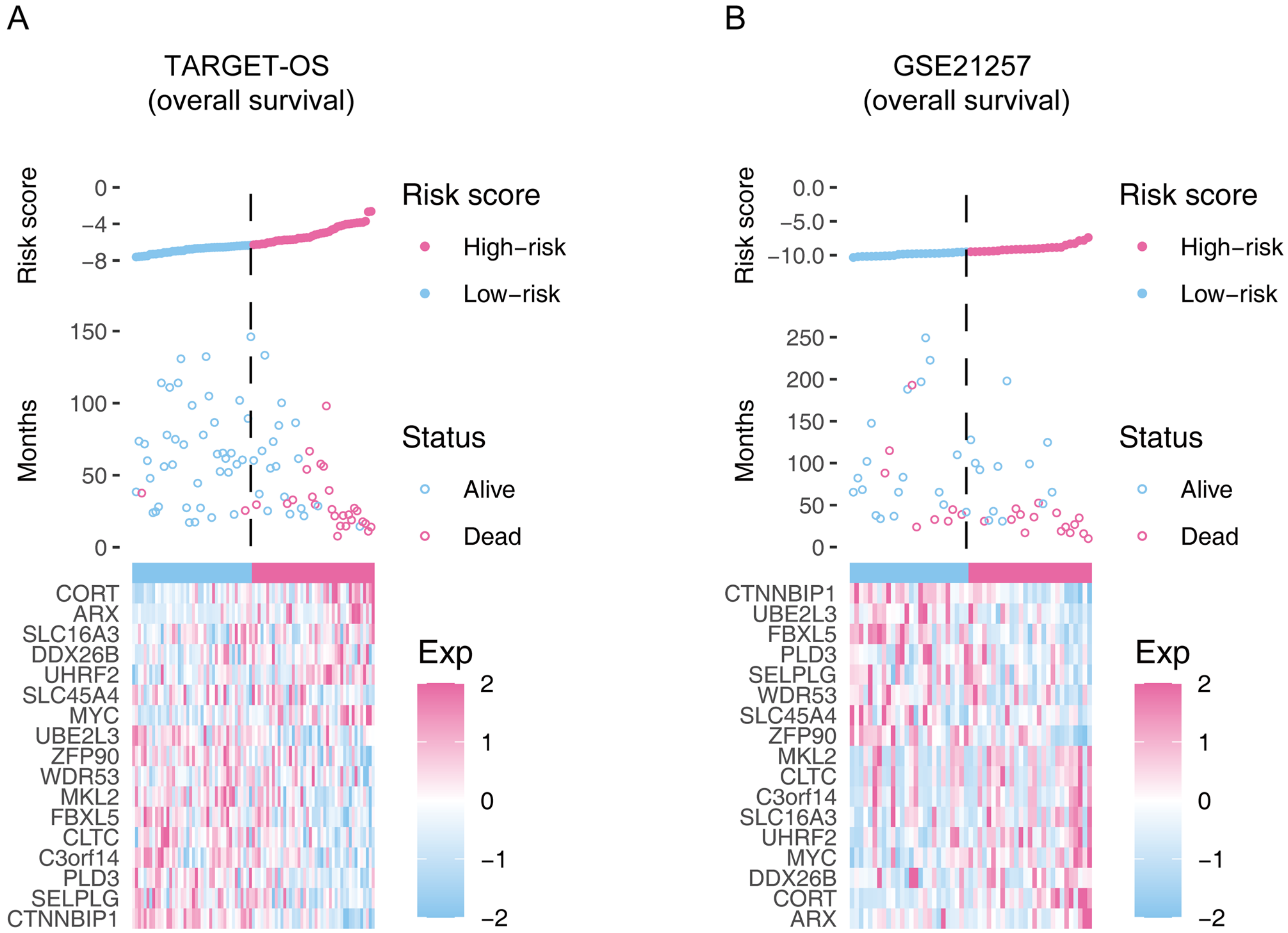


Figure 3. The overall distributions of the risk score (upper), survival status (middle), and gene expression profiles (bottom) of the seventeen-gene signature in the training (**A**) and validation (**B**) cohorts.

and increases the invasiveness of human OS cells by down-regulating the E-cadherin expression and increasing EMT in an Rb1-dependent manner[25]. *CORT* is an endogenous cyclic neuropeptide, which can regulate the growth and metastasis of lung cancer and thyroid cancer, and regulate inflammation by inhibiting immune infiltration[26]. Wu et al. found that the expression of *CORT* was higher in high-risk OS populations, confirmed that the high expression of *CORT* was related to the poor prognosis of OS[27]. According to previous studies, *MYC* is widely involved in many cancers, and its expression is estimated to be elevated or dysregulated in up to 70% of human cancers[28]. *MYC* mediated transcriptional amplification through super enhancers is an important hallmark of cancer[29]. The dysregulated expression of the oncogene *MYC* is usually associated with the oncogenesis and progression of OS[30]. *MYC* proto-oncogene boosts the oncogenic transcription amplification process in cancer and is a crucial target for cancer therapy[30]. It is reported that the *MYC* gene is amplified in OS, and its expression is often up regulated in patients with OS[30]. *MYC* overexpression, coupled with the loss of Ink4a/Arf, can further the transformation of bone marrow stromal cells into OS[30]. Above all, high *MYC* levels are related to low apoptosis and poor outcomes in patients with OS[30]. Chen's team recently demonstrated that *MYC*-driven super-enhancer signaling is essential for OS tumorigenesis, and the *MYC*/super-enhancer axis targeting therapeutic strategy to be a promising perspective for OS patients[30]. The genes *ARX*[31,32], *DDX26B*[33,34], and *SLC16A3*[35,36] have been reported to be involved in the occurrence and development of certain cancers, but whether they play an important role in OS has not yet been revealed, implied more efforts are needed.

Freshly, with the widespread application of bioinformatics, potential gene signatures associated with OS prognosis were generated from the publicly databases, which witnessed by more and more involved research. To judge the pros or cons of our signature, we found nine studies published in the most recent past year and compared them horizontally[37–45]. The comparison results once again confirmed our discovery is superior in predicting the prognosis of OS.

KEGG analysis was showed that the enriched items were mainly related to protein digestion and absorption, complement and coagulation cascades, and Wnt signaling pathway (Figure S3B). Wnt signaling is one of the key cascades regulating development and stemness, and it is also closely related to cancer[46]. The role of Wnt signaling in carcinogenesis has been most prominently described in colorectal cancer, but abnormal Wnt signaling has been observed in more cancer entities[46]. Constitutive Wnt signal activation is common in human OS, while gene mutations that activate components of the Wnt pathway are rare in OS[1]. Wnt signaling may play a key role in OS proliferation, metastasis and OS cancer stem cell maintenance[1].

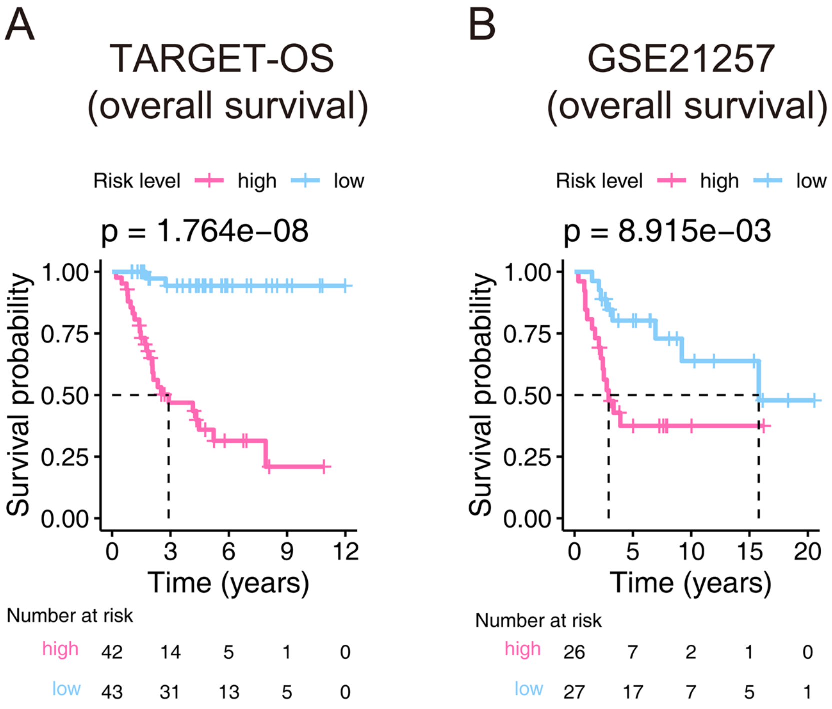
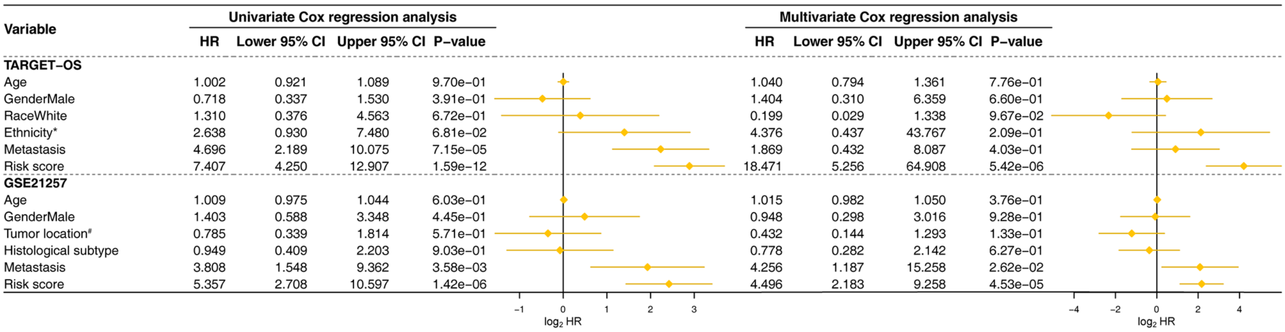


Figure 4. Kaplan–Meier estimator that evaluating the prognosis capacity of the seventeen-gene signature in the training (**A**) and validation (**B**) cohorts. The bottom part indicates the number of patients at risk. The two-sided log-rank test measured the differences between the high- and low-risk groups with a p value < 0.05.



Variable	Univariate Cox regression analysis				Multivariate Cox regression analysis			
	HR	Lower 95% CI	Upper 95% CI	P−value	HR	Lower 95% CI	Upper 95% CI	P−value
TARGET−OS								
Age	1.002	0.921	1.089	9.70e−01	1.040	0.794	1.361	7.76e−01
GenderMale	0.718	0.337	1.530	3.91e−01	1.404	0.310	6.359	6.60e−01
RaceWhite	1.310	0.376	4.563	6.72e−01	0.199	0.029	1.338	9.67e−02
Ethnicity*	2.638	0.930	7.480	6.81e−02	4.376	0.437	43.767	2.09e−01
Metastasis	4.696	2.189	10.075	7.15e−05	1.869	0.432	8.087	4.03e−01
Risk score	7.407	4.250	12.907	1.59e−12	18.471	5.256	64.908	5.42e−06
GSE21257								
Age	1.009	0.975	1.044	6.03e−01	1.015	0.982	1.050	3.76e−01
GenderMale	1.403	0.588	3.348	4.45e−01	0.948	0.298	3.016	9.28e−01
Tumor location#	0.785	0.339	1.814	5.71e−01	0.432	0.144	1.293	1.33e−01
Histological subtype	0.949	0.409	2.203	9.03e−01	0.778	0.282	2.142	6.27e−01
Metastasis	3.808	1.548	9.362	3.58e−03	4.256	1.187	15.258	2.62e−02
Risk score	5.357	2.708	10.597	1.42e−06	4.496	2.183	9.258	4.53e−05

Figure 5. Univariate and multivariate Cox proportional-hazards models that built for testing the predicting ability of the seventeen-gene signature in two cohorts. *HR* hazard ratio, *CI* confidence interval.

Immunotherapy is a type of therapy that helps the individual's immune system eliminate or control cancer[47]. Recently, immunotherapy has begun to show good prospects in various adult cancers, but whether this method is effective in OS is still rarely reported. Several biological characteristics of OS suggest that the regulation of the immune response may bring benefits, and the various immune approaches available now make immunotherapy potential for OS[48]. One of the main challenges of immunotherapy is identifying biomarkers that predict response so that treatments can be tailored to maximize benefits[48]. In the present study, six genes, including *PRF1, CD8A, HAVCR2, LAG3, CD274* (*PD-L1*), and *GZMA*, were identified as closely related to our seventeen-gene signature and might guide future OS immunotherapy.

Combining the findings of immune infiltration analysis and the 22 TICs survival analysis, we speculated that the extensive infiltration of T cells CD8 and Mast cells activated in tumors may help our signature to achieve stable predictive ability. There is evidence that *PD-1* is involved in the progression of OS disease, and the percentage of *PD-1* in peripheral blood CD4 + and CD8 + T lymphocytes in OS patients is significantly up-regulated[49]. More importantly, in vivo and in vitro experiments conducted by researchers have confirmed that *PD-L1* in OS is significantly expressed[49]. Therefore, inhibition of *PD-1/PD-L1* is an interesting therapeutic target that can restore the function of the immune system to OS cells[49]. Mast cells are immune cells that accumulate in tumors and their microenvironment during disease progression[50]. They play a multi-faceted role in the tumor microenvironment by regulating various events in tumor biology, such as angiogenesis, cell proliferation, and survival[50].

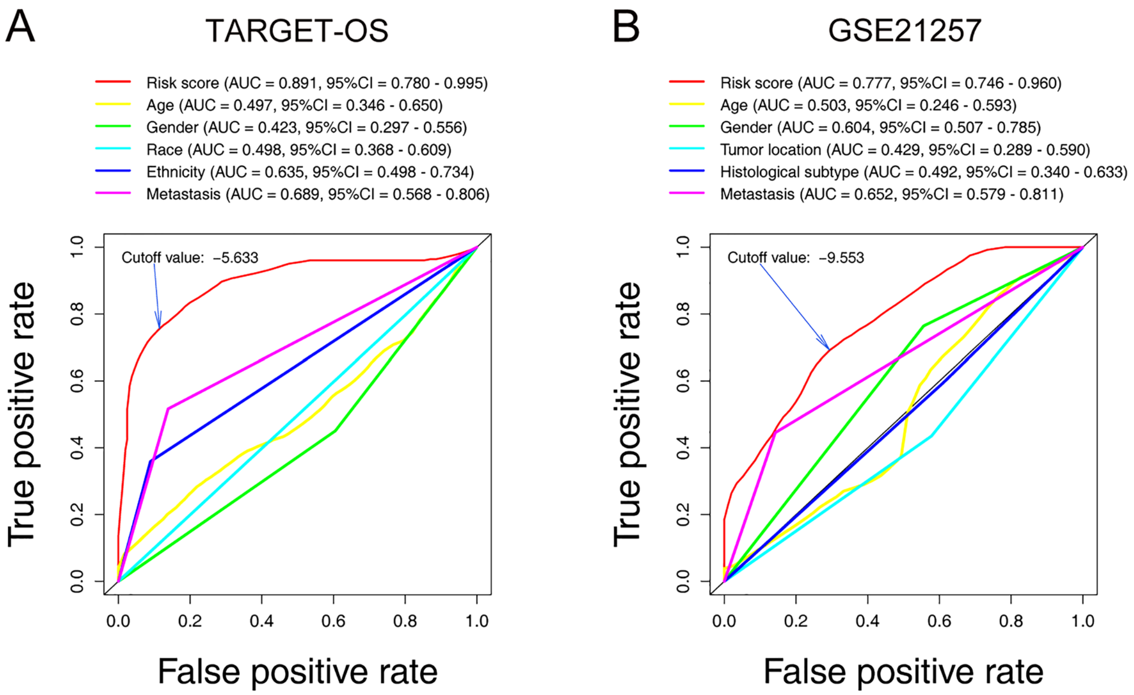


Figure 6. ROC curves that constructed for examining the predictive ability of the seventeen-gene signature in the training (**A**) and validation (**B**) cohorts. *ROC* receiver operating characteristic, *AUC* area under the ROC curve, *95% CI* 95% confidence interval.

Authors	Published online date	PMID	Gene signature composition
Fu et al	2021 Mar 18	33816483	*DCN, P4HA1*
Yang et al	2021 May 5	33952718	*P4HA1, ABCB6, STC2*
Cao et al	2020 Dec 23	33425993	*GJA5, APBB1IP, NPC2, FKBP11*
Xiao et al	2020 Dec 15	33384961	*IFITM3, VAMP8, ACTA2, GZMA, CDCA7, EVI2B, SLC7A7*
Chen et al	2020 Dec 14	33381518	*MSR1, TLR7*
Wen et al	2020 Dec 3	33281116	*COCH, MYOM2, PDE1B*
Yu et al	2020 Aug 21	32820615	*CXCR3, SSTR3, SAA1, CCL4, PYY, CCR9, CXCL9, CXCL11, C3, CXCL2, S1PR4, CXCL10, CXCR6*
Song et al	2020 Jul 24	32850346	*CD4, CD68, CSF1R*
Zhu et al	2020 Jun 22	32581649	*SLC18B1, RBMXL1, DOK3, HS3ST2, ATP6V0D1, CCAR1, C1QTNF1*

Table 3. Candidate research for comparison to our signature. *PMID* PubMed ID.

Invasion and transfer. Mast cells are recruited in the early stages of tumor development and play a key role in angiogenesis and tissue remodeling and promote tumor occurrence and growth[50]. As tumor growth progresses, mast cells recruit immune cells or suppress anti-tumor responses[50]. We know from previous studies that mast cells affect the homeostasis of OS and affect tumor progression, but we have not yet understood its underlying mechanism[50–52]. Interestingly, our research showed that T cells CD8 and Mast cells activated can potentially target the gene signature in OS treatment. Thus, further research should consider closely to the roles that the T cells CD8 and Mast cells activated play in the remodeling of the tumor microenvironment.

In the end, we must clarify the limitations of this research. The seventeen-gene signature we derived was from retrospective data. We believe that more prospective data can make our results more effective and rigorous. In addition, although it has absolute superiority compared with previous studies, its proof results are derived from the analysis results of three public databases. There is still no wet laboratory data to explain and support the prognostic ability of these 17 genes and their role in immune infiltration. Therefore, ongoing research is needed to reveal more evidence to for the seventeen -gene signature's promising future.

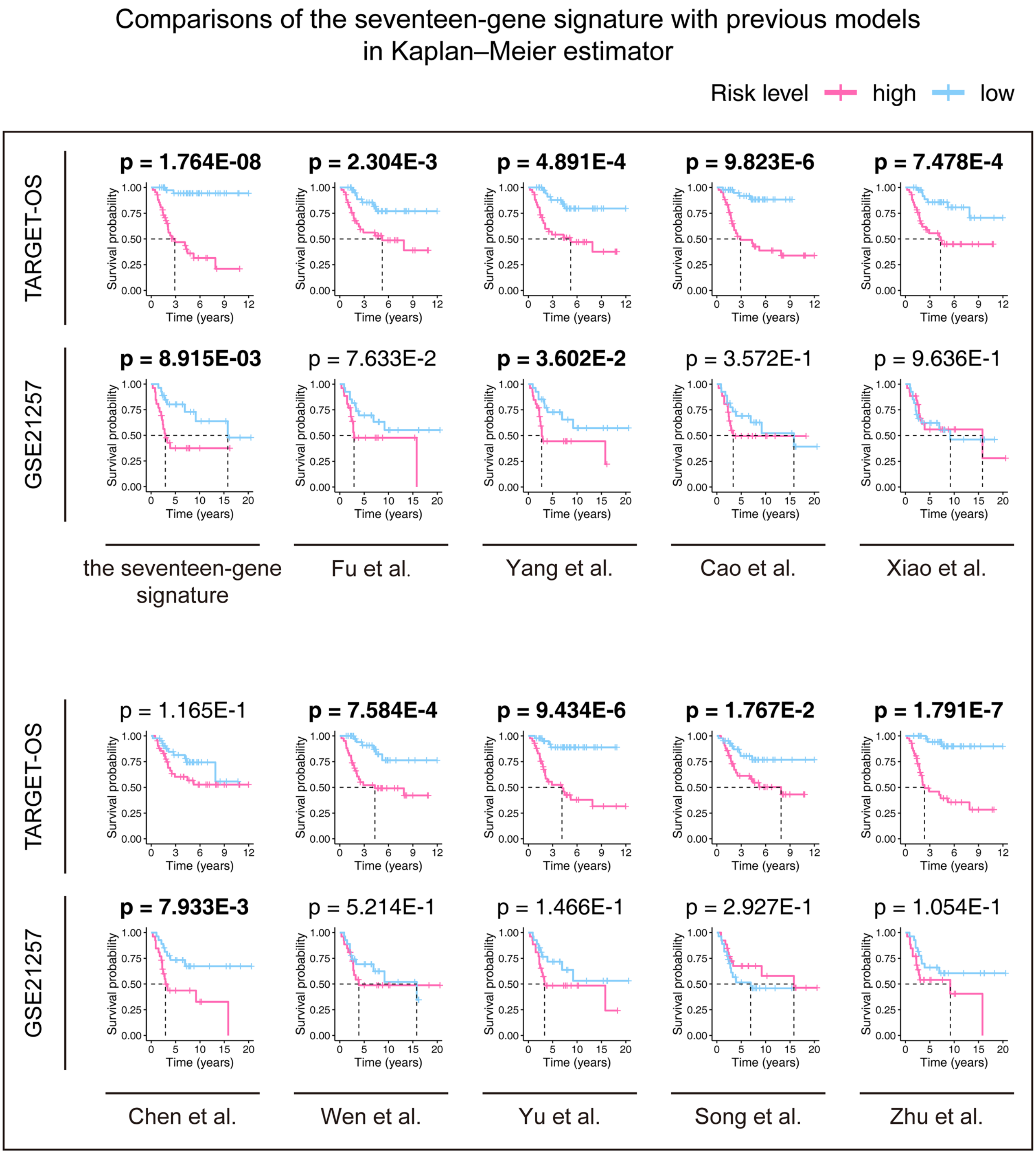
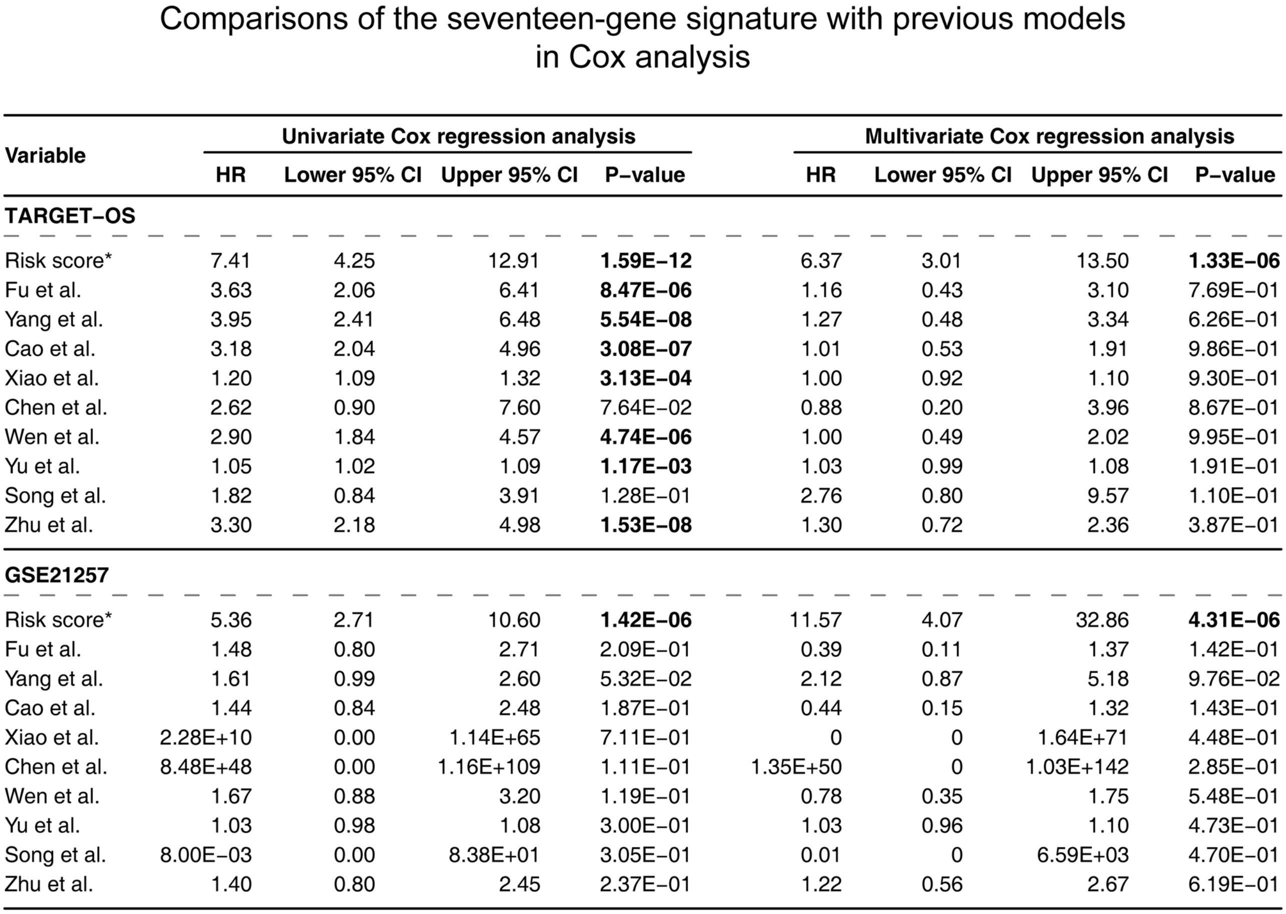


Figure 7. Comparisons between the seventeen-gene signature and previous studies conducted in the training and validation cohorts using Kaplan–Meier estimator. The two-sided log-rank test measured the differences between the high- and low-risk groups. The bold p value indicates that < 0.05, which considers significantly.



Comparisons of the seventeen-gene signature with previous models in Cox analysis

Variable	Univariate Cox regression analysis				Multivariate Cox regression analysis			
	HR	Lower 95% CI	Upper 95% CI	P−value	HR	Lower 95% CI	Upper 95% CI	P−value
TARGET−OS								
Risk score*	7.41	4.25	12.91	**1.59E−12**	6.37	3.01	13.50	**1.33E−06**
Fu et al.	3.63	2.06	6.41	**8.47E−06**	1.16	0.43	3.10	7.69E−01
Yang et al.	3.95	2.41	6.48	**5.54E−08**	1.27	0.48	3.34	6.26E−01
Cao et al.	3.18	2.04	4.96	**3.08E−07**	1.01	0.53	1.91	9.86E−01
Xiao et al.	1.20	1.09	1.32	**3.13E−04**	1.00	0.92	1.10	9.30E−01
Chen et al.	2.62	0.90	7.60	7.64E−02	0.88	0.20	3.96	8.67E−01
Wen et al.	2.90	1.84	4.57	**4.74E−06**	1.00	0.49	2.02	9.95E−01
Yu et al.	1.05	1.02	1.09	**1.17E−03**	1.03	0.99	1.08	1.91E−01
Song et al.	1.82	0.84	3.91	1.28E−01	2.76	0.80	9.57	1.10E−01
Zhu et al.	3.30	2.18	4.98	**1.53E−08**	1.30	0.72	2.36	3.87E−01
GSE21257								
Risk score*	5.36	2.71	10.60	**1.42E−06**	11.57	4.07	32.86	**4.31E−06**
Fu et al.	1.48	0.80	2.71	2.09E−01	0.39	0.11	1.37	1.42E−01
Yang et al.	1.61	0.99	2.60	5.32E−02	2.12	0.87	5.18	9.76E−02
Cao et al.	1.44	0.84	2.48	1.87E−01	0.44	0.15	1.32	1.43E−01
Xiao et al.	2.28E+10	0.00	1.14E+65	7.11E−01	0	0	1.64E+71	4.48E−01
Chen et al.	8.48E+48	0.00	1.16E+109	1.11E−01	1.35E+50	0	1.03E+142	2.85E−01
Wen et al.	1.67	0.88	3.20	1.19E−01	0.78	0.35	1.75	5.48E−01
Yu et al.	1.03	0.98	1.08	3.00E−01	1.03	0.96	1.10	4.73E−01
Song et al.	8.00E−03	0.00	8.38E+01	3.05E−01	0.01	0	6.59E+03	4.70E−01
Zhu et al.	1.40	0.80	2.45	2.37E−01	1.22	0.56	2.67	6.19E−01

Figure 8. Comparisons between the seventeen-gene signature and previous studies conducted in the training and validation cohorts using Cox models. *HR* hazard ratio, *CI* confidence interval. *The seventeen-gene signature that identified in this study; the bold p value indicates that < 0.05, which considers significantly.

Conclusion

The present work identified a novel and robust seventeen-gene signature for the OS prognosis by mining TARGET and GEO databases. In addition, we determined the reliability and applicability of the signature by applying it to an independent cohort. Through comparison, we confirm that our signature is superior to previous research. We identified our signature's potential immunotherapy targets and the important role of T cells CD8 and Mast cells activated in the seventeen-gene signature prognostic capacity. The real-world influence of the seventeen-gene signature and the underlying mechanisms between it and tumor immunity in OS remained a lack of research and warranted further investigation.

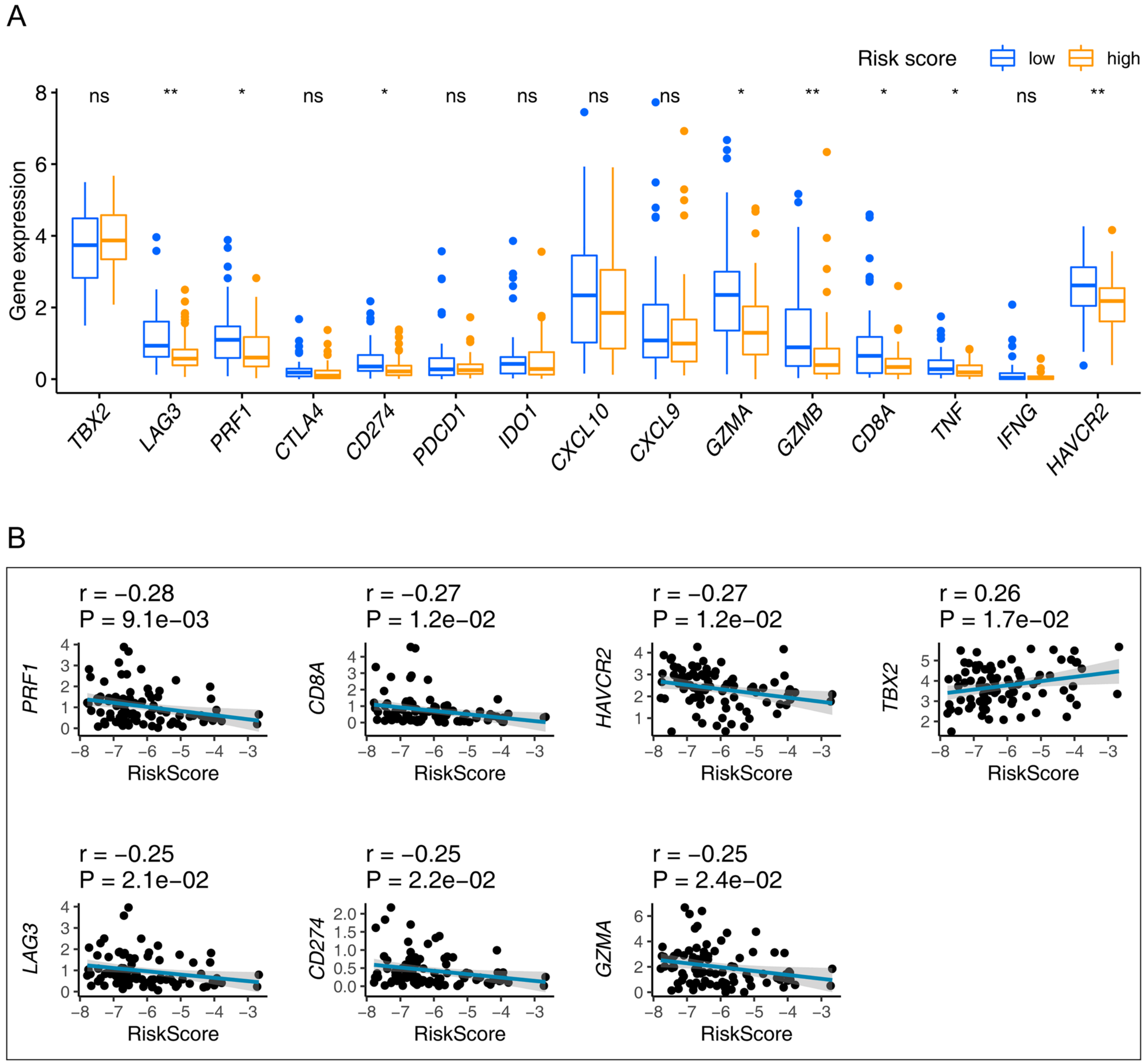


Figure 9. Identification of the relationships between the seventeen-gene signature and immune relevant signatures. (**A**) Wilcoxon rank-sum was adopted to differentiate immune relevant signatures between the high- and low-risk groups. (**B**) The Pearson coefficient was applied for the correlation test between the immune relevant signatures and seventeen-gene signature. Only correlations with p value < 0.05 were plotted. ns: p value > 0.05; *p value < 0.05; **p value < 0.01; ***p value < 0.001; p value < 0.05 was considered statistically significant.

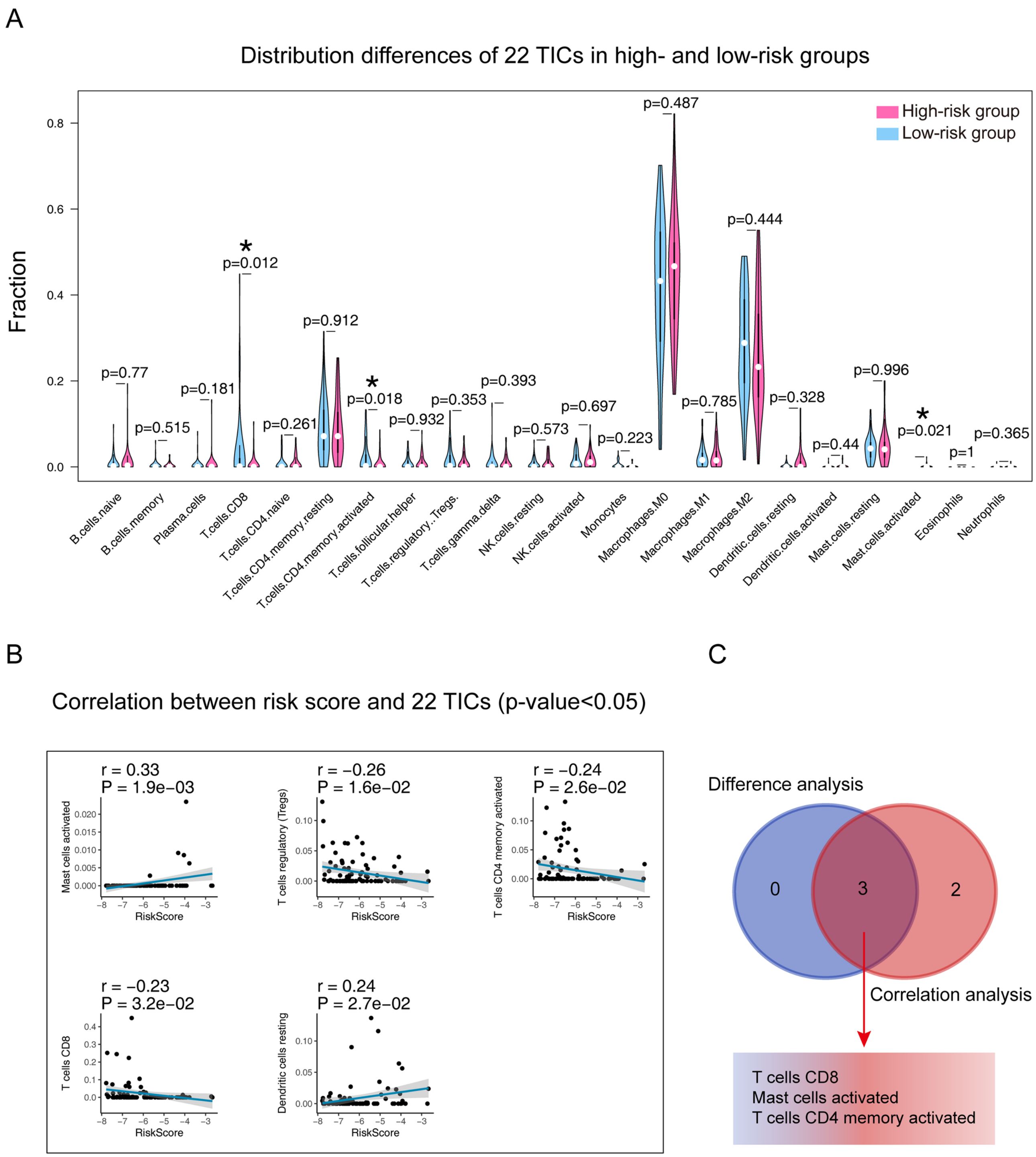
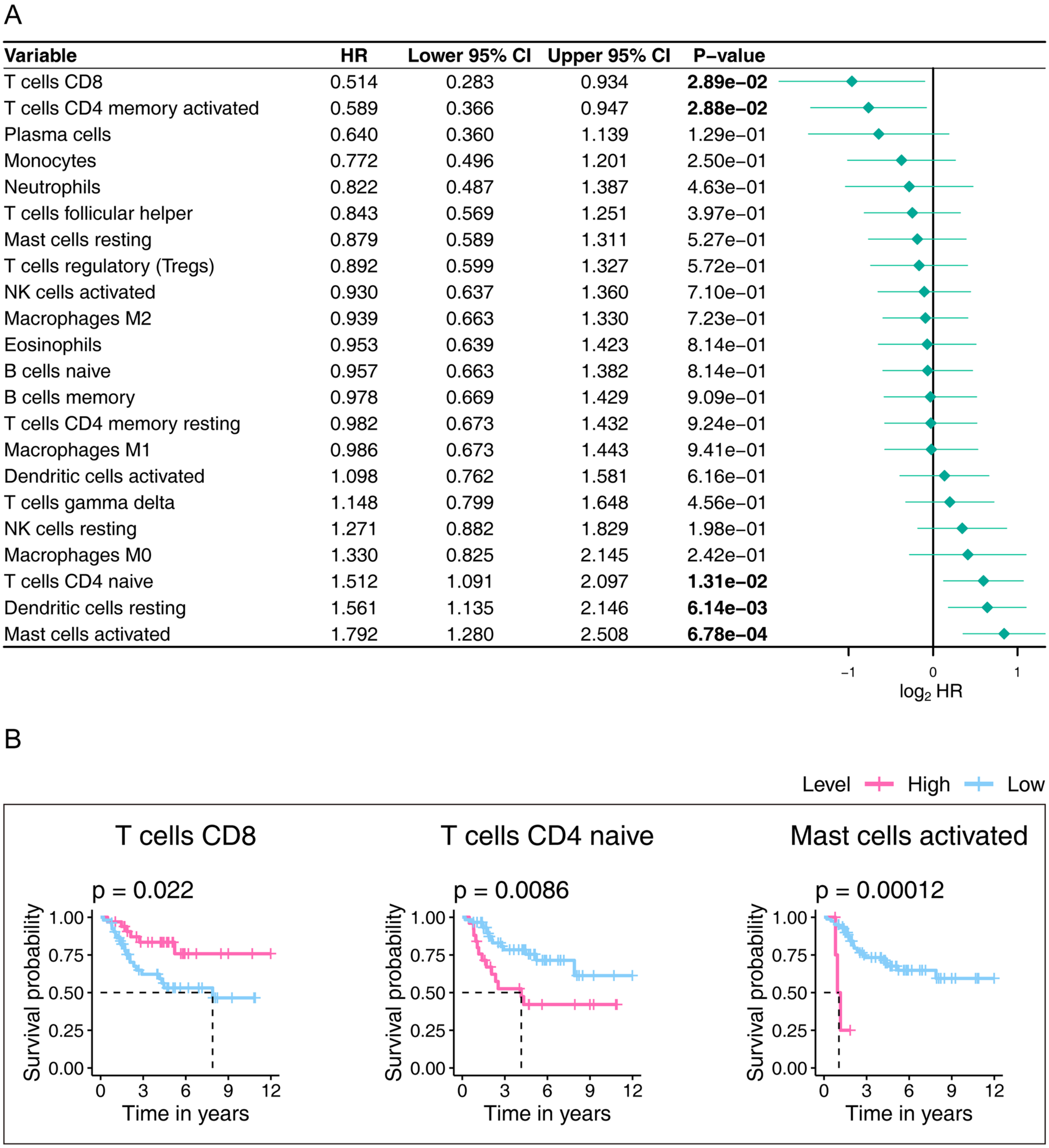


Figure 10. Integrating analysis for the relationship between TICs and the seventeen-gene signature. (**A**) Wilcoxon rank-sum was adopted to differentiate each of 22 TICs between the high- and low-risk groups. (**B**) The Pearson coefficient was applied for the correlation test between the TICs and the seventeen-gene signature. Only correlations with p value < 0.05 were plotted. (**C**) The Venn diagram that integrating the results from (**A**) and (**B**). TIC: tumor-infiltrating immune cell; *p value < 0.05; p value < 0.05 was considered statistically significant.



Variable	HR	Lower 95% CI	Upper 95% CI	P–value
T cells CD8	0.514	0.283	0.934	**2.89e–02**
T cells CD4 memory activated	0.589	0.366	0.947	**2.88e–02**
Plasma cells	0.640	0.360	1.139	1.29e–01
Monocytes	0.772	0.496	1.201	2.50e–01
Neutrophils	0.822	0.487	1.387	4.63e–01
T cells follicular helper	0.843	0.569	1.251	3.97e–01
Mast cells resting	0.879	0.589	1.311	5.27e–01
T cells regulatory (Tregs)	0.892	0.599	1.327	5.72e–01
NK cells activated	0.930	0.637	1.360	7.10e–01
Macrophages M2	0.939	0.663	1.330	7.23e–01
Eosinophils	0.953	0.639	1.423	8.14e–01
B cells naive	0.957	0.663	1.382	8.14e–01
B cells memory	0.978	0.669	1.429	9.09e–01
T cells CD4 memory resting	0.982	0.673	1.432	9.24e–01
Macrophages M1	0.986	0.673	1.443	9.41e–01
Dendritic cells activated	1.098	0.762	1.581	6.16e–01
T cells gamma delta	1.148	0.799	1.648	4.56e–01
NK cells resting	1.271	0.882	1.829	1.98e–01
Macrophages M0	1.330	0.825	2.145	2.42e–01
T cells CD4 naive	1.512	1.091	2.097	**1.31e–02**
Dendritic cells resting	1.561	1.135	2.146	**6.14e–03**
Mast cells activated	1.792	1.280	2.508	**6.78e–04**

Figure 11. Univariate Cox proportional-hazards model (**A**) and Kaplan–Meier estimator (**B**) that built for evaluating the prognostic ability of 22 TICs. Only graphs with a p value < 0.05 in the log-rank test were plotted in (**B**). The bold p value indicates that < 0.05, which considers significant. *TIC* tumor-infiltrating immune cell.

Data availability

Publicly available datasets were used in this study. Data from TARGET-OS (https://ocg.cancer.gov/programs/target/projects/osteosarcoma) and data from GSE21257 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21257) were downloaded and analyzed in this work.

Received: 14 August 2021; Accepted: 5 January 2022
Published online: 24 January 2022

References

1. Fang, F. *et al.* Targeting the Wnt/beta-catenin pathway in human osteosarcoma cells. *Oncotarget* **9**, 36780–36792 (2018).
2. Harrison, D. J., Geller, D. S., Gill, J. D., Lewis, V. O. & Gorlick, R. Current and future therapeutic approaches for osteosarcoma. *Expert Rev. Anticancer Ther.* **18**, 39–50 (2018).
3. Isakoff, M. S., Bielack, S. S., Meltzer, P. & Gorlick, R. Osteosarcoma: Current treatment and a collaborative pathway to success. *J. Clin. Oncol.* **33**, 3029–3035 (2015).
4. Whelan, J. S. & Davis, L. E. Osteosarcoma, chondrosarcoma, and chordoma. *J. Clin. Oncol.* **36**, 188–193 (2018).
5. Jia, Y., Liu, Y., Han, Z. & Tian, R. Identification of potential gene signatures associated with osteosarcoma by integrated bioinformatics analysis. *PeerJ* **9**, e11496 (2021).
6. Zhang, A., Yang, J., Ma, C., Li, F. & Luo, H. Development and validation of a robust ferroptosis-related prognostic signature in lung adenocarcinoma. *Front. Cell Dev. Biol.* **9**, 616271 (2021).
7. Wang, Y. *et al.* Prognostic implications of immune-related eight-gene signature in pediatric brain tumors. *Braz. J. Med. Biol. Res.* **54**, e10612 (2021).
8. Luo, H. & Ma, C. A novel ferroptosis-associated gene signature to predict prognosis in patients with uveal melanoma. *Diagnostics (Basel)* **11**, 20 (2021).
9. Luo, H., Ma, C., Shao, J. & Cao, J. Prognostic implications of novel ten-gene signature in uveal melanoma. *Front. Oncol.* **10**, 567512 (2020).
10. Ma, C. *et al.* Identification of a novel tumor microenvironment-associated eight-gene signature for prognosis prediction in lung adenocarcinoma. *Front. Mol. Biosci.* **7**, 571641 (2020).
11. Cantini, L. *et al.* Classification of gene signatures for their information value and functional redundancy. *NPJ Syst. Biol. Appl.* **4**, 2 (2018).
12. Yu, S., Shao, F., Liu, H. & Liu, Q. A five metastasis-related long noncoding RNA risk signature for osteosarcoma survival prediction. *BMC Med. Genom.* **14**, 124 (2021).
13. Mao, R. *et al.* Prognostic nomogram for childhood acute lymphoblastic leukemia: A comprehensive analysis of 673 patients. *Front. Oncol.* **10**, 1673 (2020).
14. Clough, E. & Barrett, T. The gene expression omnibus database. *Methods Mol. Biol.* **1418**, 93–110 (2016).
15. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
16. Goeman, J. J. L1 penalized estimation in the Cox proportional hazards model. *Biom. J.* **52**, 70–84 (2010).
17. Sauerbrei, W., Royston, P. & Binder, H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat. Med.* **26**, 5512–5528 (2007).
18. Tibshirani, R. The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385–395 (1997).
19. Cao, R. & Lopez-de-Ullibarri, I. ROC curves for the statistical analysis of microarray data. *Methods Mol. Biol.* **1986**, 245–253 (2019).
20. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
21. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).
22. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
23. Rothzerg, E., Xu, J., Wood, D. & Koks, S. 12 Survival-related differentially expressed genes based on the TARGET-osteosarcoma database. *Exp. Biol. Med. (Maywood)* **246**, 2072–2081 (2021).
24. Iguchi, T. *et al.* Identification of UHRF2 as a negative regulator of epithelial-mesenchymal transition and its clinical significance in esophageal squamous cell carcinoma. *Oncology* **95**, 179–187 (2018).
25. Liu, W. *et al.* UHRF1 promotes human osteosarcoma cell invasion by downregulating the expression of Ecadherin in an Rb1dependent manner. *Mol. Med. Rep.* **13**, 315–320 (2016).
26. Delgado-Maroto, V. *et al.* The neuropeptide cortistatin attenuates experimental autoimmune myocarditis via inhibition of cardiomyogenic T cell-driven inflammatory responses. *Br. J. Pharmacol.* **174**, 267–280 (2017).
27. Wu, Z. L. *et al.* Development of a novel immune-related genes prognostic signature for osteosarcoma. *Sci. Rep.* **10**, 18402 (2020).
28. Dang, C. V. MYC on the path to cancer. *Cell* **149**, 22–35 (2012).
29. Kress, T. R., Sabo, A. & Amati, B. MYC: Connecting selective transcriptional control to global RNA production. *Nat. Rev. Cancer* **15**, 593–607 (2015).
30. Chen, D. *et al.* Super enhancer inhibitors suppress MYC driven transcriptional amplification and tumor progression in osteosarcoma. *Bone Res.* **6**, 11 (2018).
31. Hackeng, W. M. *et al.* Assessment of ARX expression, a novel biomarker for metastatic risk in pancreatic neuroendocrine tumors, in endoscopic ultrasound fine-needle aspiration. *Diagn. Cytopathol.* **48**, 308–315 (2020).
32. Liu, W. B. *et al.* Epigenetic silencing of Aristaless-like homeobox-4, a potential tumor suppressor gene associated with lung cancer. *Int. J. Cancer* **134**, 1311–1322 (2014).
33. Chen, H. *et al.* Small RNA-induced INTS6 gene up-regulation suppresses castration-resistant prostate cancer cells by regulating beta-catenin signaling. *Cell Cycle* **17**, 1602–1613 (2018).
34. Lui, K. Y. *et al.* Integrator complex subunit 6 (INTS6) inhibits hepatocellular carcinoma growth by Wnt pathway and serve as a prognostic marker. *BMC Cancer* **17**, 644 (2017).
35. Yu, S. *et al.* Comprehensive analysis of the SLC16A gene family in pancreatic cancer via integrated bioinformatics. *Sci. Rep.* **10**, 7315 (2020).
36. Javaeed, A. & Ghauri, S. K. MCT4 has a potential to be used as a prognostic biomarker—a systematic review and meta-analysis. *Oncol. Rev.* **13**, 403 (2019).
37. Yang, M. *et al.* Identification of a novel glycolysis-related gene signature for predicting the prognosis of osteosarcoma patients. *Aging (Albany NY)* **13**, 12896–12918 (2021).
38. Fu, Y. *et al.* Development and validation of a hypoxia-associated prognostic signature related to osteosarcoma metastasis and immune infiltration. *Front. Cell Dev. Biol.* **9**, 633607 (2021).
39. Cao, M. *et al.* Identification and development of a novel 4-gene immune-related signature to predict osteosarcoma prognosis. *Front. Mol. Biosci.* **7**, 608368 (2020).

40. Xiao, B. *et al.* Identification and verification of immune-related gene prognostic signature based on ssGSEA for osteosarcoma. *Front. Oncol.* **10**, 607622 (2020).
41. Chen, Z., Huang, H., Wang, Y., Zhan, F. & Quan, Z. Identification of immune-related genes MSR1 and TLR7 in relation to macrophage and type-2 T-helper cells in osteosarcoma tumor micro-environments as anti-metastasis signatures. *Front. Mol. Biosci.* **7**, 576298 (2020).
42. Wen, C. *et al.* A three-gene signature based on tumour microenvironment predicts overall survival of osteosarcoma in adolescents and young adults. *Aging (Albany NY)* **13**, 619–645 (2020).
43. Song, Y. J. *et al.* Immune landscape of the tumor microenvironment identifies prognostic gene signature CD4/CD68/CSF1R in osteosarcoma. *Front. Oncol.* **10**, 1198 (2020).
44. Yu, Y. *et al.* Development of a prognostic gene signature based on an immunogenomic infiltration analysis of osteosarcoma. *J. Cell Mol. Med.* **24**, 11230–11242 (2020).
45. Zhu, N. *et al.* Co-expression network analysis identifies a gene signature as a predictive biomarker for energy metabolism in osteosarcoma. *Cancer Cell Int.* **20**, 259 (2020).
46. Zhan, T., Rindtorff, N. & Boutros, M. Wnt signaling in cancer. *Oncogene* **36**, 1461–1473 (2017).
47. Viale, P. H. The American Cancer Society's Facts & Figures: 2020 Edition. *J. Adv. Pract. Oncol.* **11**, 135–136 (2020).
48. Wedekind, M. F., Wagner, L. M. & Cripe, T. P. Immunotherapy for osteosarcoma: Where do we go from here?. *Pediatr. Blood Cancer* **65**, e27227 (2018).
49. Wang, Z., Li, B., Ren, Y. & Ye, Z. T-cell-based immunotherapy for osteosarcoma: Challenges and opportunities. *Front. Immunol.* **7**, 353 (2016).
50. Inagaki, Y. *et al.* Dendritic and mast cell involvement in the inflammatory response to primary malignant bone tumours. *Clin. Sarcoma Res.* **6**, 13 (2016).
51. Maciel, T. T., Moura, I. C. & Hermine, O. The role of mast cells in cancers. *F1000Prime Rep.* **7**, 09 (2015).
52. Campillo-Navarro, M. *et al.* Mast cells in lung homeostasis: Beyond type I hypersensitivity. *Curr. Respir. Med. Rev.* **10**, 115–123 (2014).

Author contributions

J.Y., A.Z., and C.M. drafted the original manuscript. H.L. plotted all figures for data visualization. H.L. and C.M. did the final proofread of the manuscript. All authors reviewed the current form of the manuscript and approved its publication.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-05341-5.

Correspondence and requests for materials should be addressed to H.L. or C.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.