Journal of Cancer Research and Clinical Oncology (2021) 147:153–165
https://doi.org/10.1007/s00432-020-03396-3

ORIGINAL ARTICLE – CANCER RESEARCH

An immune-related gene signature for determining Ewing sarcoma prognosis based on machine learning

En-hui Ren[1,2] · Ya-jun Deng[2] · Wen-hua Yuan[2] · Zuo-long Wu[2] · Guang-zhi Zhang[2] · Qi-qi Xie[2,3]

Received: 26 May 2020 / Accepted: 16 September 2020 / Published online: 23 September 2020

Abstract

Purpose Ewing sarcoma (ES) is one of the most common malignant bone tumors in children and adolescents. The immune microenvironment plays an important role in the development of ES. Here, we developed an optimal signature for determining ES patient prognosis based on immune-related genes (IRGs).

Methods We analyzed the ES gene expression profile dataset, GSE17679, from the GEO database and extracted differential expressed IRGs (DEIRGs). Then, we conducted functional correlation and protein–protein interaction (PPI) analyses of the DEIRGs and used the machine learning algorithm-iterative Lasso Cox regression analysis to build an optimal DEIRG signature. In addition, we applied ES samples from the ICGC database to test the optimal gene signature. We performed univariate and multivariate Cox regressions on clinicopathological characteristics and optimal gene signature to evaluate whether signature is an important prognostic factor. Finally, we calculated the infiltration of 24 immune cells in ES using the ssGSEA algorithm, and analyzed the correlation between the DEIRGs in the optimal gene signature and immune cells.

Results A total of 249 DEIRGs were screened and an 11-gene signature with the strongest correlation with patient prognoses was analyzed using a machine learning algorithm. The 11-gene signature also had a high prognostic value in the ES external verification set. Univariate and multivariate Cox regression analyses showed that 11-gene signature is an independent prognostic factor. We found that macrophages and cytotoxic, CD8 T, NK, mast, B, NK CD56bright, TEM, TCM, and Th2 cells were significantly related to patient prognoses; the infiltration of cytotoxic and CD8 T cells in ES was significantly different. By correlating prognostic biomarkers with immune cell infiltration, we found that FABP4 and macrophages, and NDRG1 and Th2 cells had the strongest correlation.

Conclusion Overall, the IRG-related 11-gene signature can be used as a reliable ES prognostic biomarker and can provide guidance for personalized ES therapy.

Keywords Ewing sarcoma · Prognosis analysis · Machine learning · Iterative Lasso regression · Immune cell infiltration

En-hui Ren, Ya-jun Deng, and Wen-hua Yuan have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00432-020-03396-3) contains supplementary material, which is available to authorized users.

✉ Qi-qi Xie
jieqq16@lzu.edu.cn

En-hui Ren
reneh17@lzu.edu.cn

Ya-jun Deng
dengyajun205@163.com

Wen-hua Yuan
yuanwh17@lzu.edu.cn

Zuo-long Wu
wuzl18@lzu.edu.cn

Guang-zhi Zhang
zhanggzh18@lzu.edu.cn

[1] Department of Orthopaedics, Xining First People's Hospital, Xining, Qinghai 810000, People's Republic of China

[2] Lanzhou University, Lanzhou, Gansu 730000, People's Republic of China

[3] Breast Disease Diagnosis and Treatment Center, Affiliated Hospital of Qinghai University & Affiliated Cancer Hospital of Qinghai University, Xining 810000, People's Republic of China

Introduction

Ewing sarcoma (ES) is the second most common bone tumor in children and adolescents (Kaatsch et al. 2016). Patients' responses to traditional surgical resection combined with radiotherapy and chemotherapy vary greatly (Andreou et al. 2020), posing several problems to ES treatment. Therefore, the early identification of high-risk patients is crucial for the development of personalized treatment plans. Immune cell infiltration plays an important role in the development of many tumors, including ES. For example, Hingorani et al. (2015) systematically evaluated more than 70 different immunophenotypes in ES via flow cytometry, and found that CD14 + HLA-DRlo/neg cells and CTLA-4 + T-cell infiltration were more than CD4 + T-cell infiltration. Additionally, the targeted killing effect of NK cells and T lymphocytes on ES cells has been long proven (Cho et al. 2010; Meyer-Wentrup et al. 2005), lymphocytes can be used as a prognostic indicator for early patient recovery after chemotherapy (De Angulo et al. 2007). NK cells have also been shown to play a direct killing effect on tumor cells by activating NKG2D and DNAM-1 receptors (Verhoeven et al. 2008). Tumor-associated macrophages (TAM) play an important role in promoting tumor cell invasion and vascular proliferation by recruiting chemokines to tumor sites and secreting cytokines and growth factors (Wyckoff et al. 2007). Also, previous studies have shown that CD68 + TAM infiltration in ES is related to poor overall survival (Fujiwara et al. 2011). Although the role of immune cells in ES has been explored to some extent, the specific mechanism of immune regulation in ES remain unclear. However, only a few studies exist on immunotherapy targets and prognosis biomarkers, and even fewer immunotherapy schemes apply to ES.

Therefore, exploring new ES immunotherapy targets and prognostic biomarkers is essential for the development of immunotherapy strategies and personalized treatment.

Machine learning algorithm–iterative Lasso Cox regression can cross-validate the relationship between genes and prognosis multiple times, and fit the prediction performance of genes according to a specific algorithm, which not only contains a lot of prognostic information, but also improves the accuracy of prediction (Eschrich et al. 2005; Sveen et al. 2012; Wang et al. 2004). However, there is no research on the construction of ES prognostic biomarkers based on immune-related genes (IRGs) using machine learning algorithms. In this study, we constructed an optimal 11-gene signature for determining ES prognosis, thereby significantly improving ES prognostic biomarkers.

Materials and methods

Data download

Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) (Barrett et al. 2013) is an international public gene repository that stores a large number of disease-related gene data. International Cancer Genome Consortium (ICGC; https://icgc.org) (Hudson et al. 2010) is a database used to describe the genomic changes in a variety of tumors. In this study, the ES expression profiling dataset, GSE17679, was downloaded and analyzed using the R language GEOquery package (Davis and Meltzer 2007), which was used as the training set to screen the prognosis gene signature. The samples included 88 ES samples and 18 matched normal skeletal muscle samples from human beings. The platform was based on GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. We used 57 ES samples from ICGC for external verification set of prognostic gene signature. The baseline information of all patients is shown in Table 1.

Data preprocessing and differential expressed immune-related gene (DEIRG) screening

The affy package (R version 3.6.3) (Gautier et al. 2004) was used to perform background correction and data normalization on the original GSE17679 expression profile data. The effects, before and after normalization, were visualized via box chart and principal component analysis (PCA). Differential expressed genes (DEGs) were obtained by analyzing the gene expression matrix using the limma package (Li et al. 2013). The screening criteria were: $|\log_2$ fold change $(\log_2 FC)| > 1$, adjust $p < 0.05$. Then, we downloaded and organized the IRG list from the ImmPort database (Bhattacharya et al. 2018), and extracted DEIRGs from DEGs according to the IRG list. A volcano map was made using

Table 1 Baseline data for all patients

Variables	Training set ($n = 88$)	External verification set ($n = 57$)
Age, years, no. (%)		
≤16	44 (50.0)	36 (63.2)
>16	44 (50.0)	21 (36.8)
Gender, no. (%)		
Male	56 (63.6)	31 (54.4)
Female	32 (36.4)	26 (45.6)
Vital status		
Alive	37 (42.0)	28 (49.1)
Dead	51 (58.0)	29 (50.1)

the ggplot2 package (Maag 2018) to show the distribution of DEIRGs.

Gene ontology (GO) analysis of DEIRGs

GO is commonly used to annotate gene function (Ashburner et al. 2000), mainly including the molecular function (MF), biological process (BP), and cellular component (CC). To explore the role of DEIRGs on the immune mechanism of ES, we used the R language clusterProfiler package (Yu et al. 2012a) to perform GO analysis of DEIRGs.

DEIRG PPI network construction, screening of key IRGs, and definition of key IRG similarity

Search tool for the retrieval of interacting genes (STRING) (https://string-db.org; version: 11.0) (Szklarczyk et al. 2019) is an online tool for analyzing PPI networks. We used STRING to construct PPI network data of DEIRGs and imported Cytoscape for visualization (Doncheva et al. 2019). Next, we used the plug-in cytoHubba (Chin et al. 2014) to identify key IRGs in the PPI network. To explore the similarity of key IRGs, we used the R language GOSemSim (Yu et al. 2010) and clusterProfiler packages to calculate similarities between the MF and CC of key IRGs, and applied ggplot2 for visualization.

Construction of the optimal prognosis gene signature

To obtain the optimal gene signature combination capable of determining ES prognosis, we used a single-factor Cox regression to screen DEIRGs related to ES prognosis as candidate genes, with a screening criterion of $p < 0.05$. Lasso regression is an algorithm based on the L_1 norm penalty term for filtering variables (Frost and Amos 2017); it is mainly used for supervised learning of high-dimensional data. Using Lasso regression, each iteration produced a new gene combination, and so, we conducted 1000 Lasso regression models for candidate genes and screened gene combinations with frequencies > 50 (Sveen et al. 2012). To ensure the reliability of the risk prediction of gene combinations, we used Cox regression to verify the optimal prognostic gene signature.

Verification of optimal gene signature

To evaluate the clinical application value of the optimal prognostic signature, we applied the RNA-seq data and clinical prognostic data of 58 ES samples in ICGC to verify the optimal gene signature. Cox regression model was used to detect the prognostic value of ES in the independent external cohort. In addition, we conducted an ROC comparison analysis among the optimal gene signature and the prognostic biomarkers (SOX2, STAG2, TP53, and CXCR4) (Berghuis et al. 2012; Sannino et al. 2019; Tirode et al. 2014) in the external verification set.

Univariate and multivariate Cox regression analyses of optimal prognostic signature and its correlation analysis with immune cell infiltration

We performed univariate and multivariate Cox regression analyses on clinicopathological characteristics (gender, age, and metastatic) and 11-gene characteristics in the training set and external validation set to evaluate whether signature is an important prognostic factor. Single-sample gene set enrichment analysis (ssGSEA) calculates gene enrichment by calculating the degree of gene regulation in a single sample (Krug et al. 2018). In this study, we used human tumor immune cell markers defined by Bindea et al. (2013). The gene expression matrix was analyzed by ssGSEA using the GSVA package (Hänzelmann et al. 2013) in R version 3.6.3, and the infiltration of 24 kinds of immune cells was observed. We used PCA, the most widely used data dimensionality reduction method in machine learning (Ringner 2008), to cluster the filtered data and to detect differences in immune cell infiltration between ES and the normal group; we used ggplot2 to draw the PCA cluster map. The correlation, infiltration difference, and interaction of 24 kinds of immune cells were visualized using corrplot (Friendly 2002), ggplot2, and igraph (Adai et al. 2004) packages. The interaction criteria were $p < 0.05$ and | Correlation coefficient |> 0.15. To explore the correlation between optimal gene signature and immune cell infiltration, we analyzed the correlation and visualized it using ggplot2.

Results

Screening of DEIRGs

The box chart results before and after correction (Supplementary Fig. 1a, b) showed that the mean value of gene expression after pretreatment was consistent, and the homogenization result was reliable. PCA showed a significant difference between the groups before and after homogenization, with a reliable sample data source, which could be used for subsequent analysis (Supplementary Fig. 1c, d). Through the differential expression analysis of the gene expression profile data, a total of 249 DEIRGs were screened and their distribution was displayed via a volcano map (Fig. 1a).

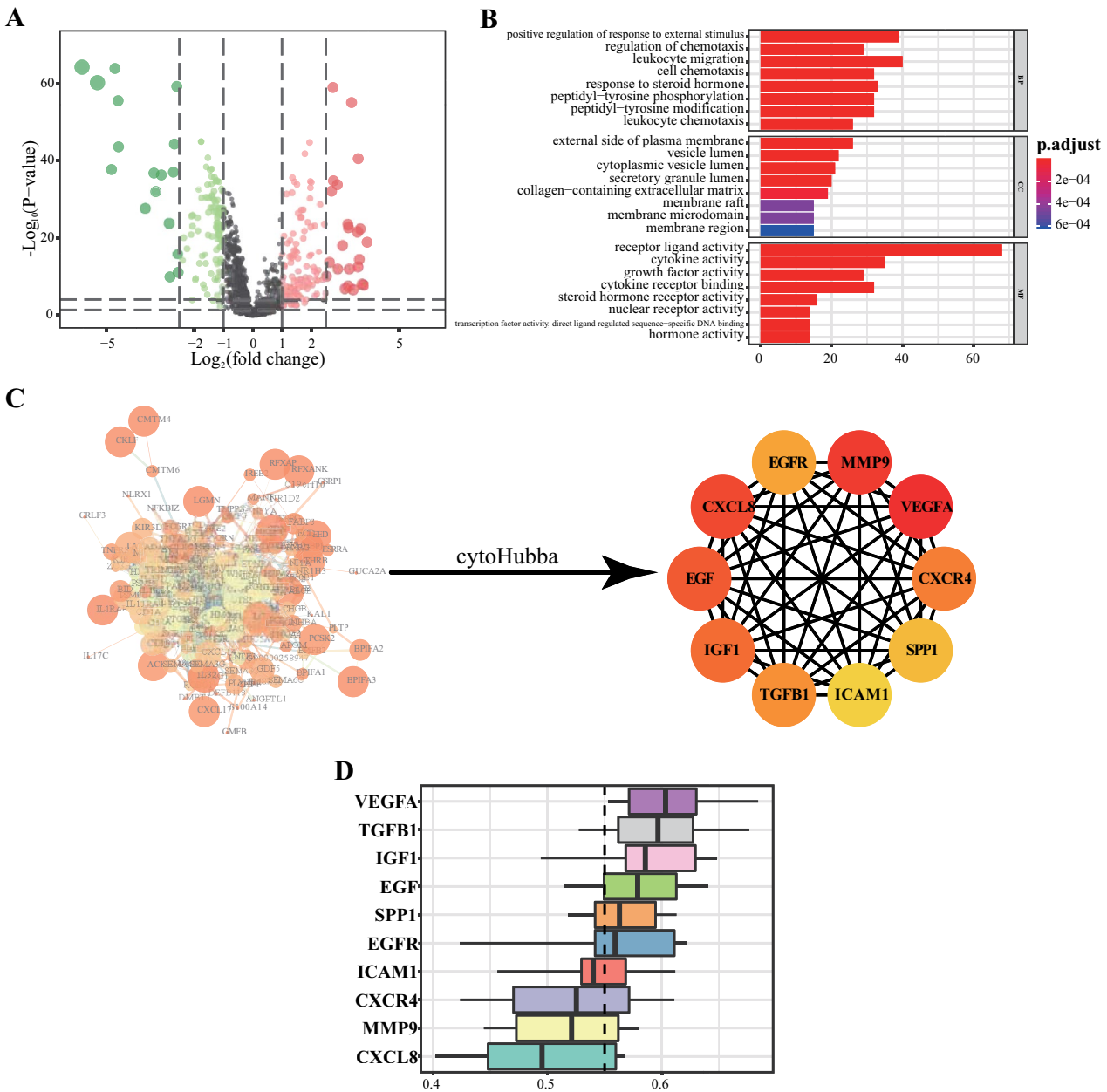

Fig. 1 Correlation analysis of DEIRGs. **a** Volcanic map of DEIRGs; red represents upregulated genes, green represents downregulated genes. **b** GO analysis of DEIRGs; the ordinate represents the name of gene, the abscissa represents the number of genes, and the color represents the adjusted *p* value. **c** PPI network of DEIRGs; node size represents clustering coefficient, node color represents degree, connection thickness represents comprehensive score, and connection color represents co-expression. Interaction network of key IRGs; the darker the color, the higher the MCC algorithm score. **d** The best friends graph of key IRGs' similarity; the vertical coordinate represents the name of key IRGs, and the horizontal coordinate represents the average value of gene similarity with other genes. *DEIRGs* differential expressed IRGs, *GO* gene ontology, *PPI* protein–protein interaction, *IRGs* immune-related genes, *MCC* Matthews correlation coefficient

GO and PPI network analyses of DEIRGs, and key IRG screening and similarities

To further study DEIRG function, we conducted GO and PPI network analyses, screened key IRGs, and determined the similarities between key IRGs. GO analysis (Fig. 1b) showed that the main changes in BP were related to processes involving position regulation and responses to external stimulus, regulation of chemotaxis, leukocyte migration, cell chemotaxis, response to steroid hormones, peptidyl-tyrosine phosphorylation, peptidyl-tyrosine modification, and leukocyte chemotaxis. Changes in CC were mainly concentrated on the external side of the plasma membrane, vesicle lumen,

cytoplasmic vesicle lumen, secretory granule lumen, collagen-containing extracellular matrix, membrane raft, membrane microdomain, and membrane region. Changes in MF mainly focused on cytokine activity, growth factor activity, cytokine receptor binding, steroid hormone receptor activity, nuclear receptor activity, transcription factor activity, direct ligand-regulated sequence-specific DNA binding, and hormone activity. PPI network analysis via Cytoscape and its plug-in, cytoHubba, revealed ten key IRGs with the highest correlation scores: VEGFA, MMP9, CXCL8, EGF, IGF1, CXCR4, TGFB1, EGFR, SPP1, and ICAM1 (Fig. 1c). Analysis of the semantic similarity of GO annotations of key IRGs showed a strong similarity in biological effects between EGFR and SPP1, and CXCR4 and MMP9 (Fig. 1d).

Optimal prognostic gene signature

The optimal prognosis gene signature constructed using 1000 iterations of the Lasso–Cox regression prediction model (Fig. 2a) yielded an 11-gene signature (CRLF3, ECD, FABP4, FGF6, GNRH2, NDRG1, PAK2, PLTP, PTGDS, RBP1, and ZC3HAV1), which had the largest area under the ROC curve and were most closely related to prognosis. An evaluation of the 5-year prognostic value of the 11-gene signature using the ROC curve showed that the prognostic value was significant (AUC = 0.993; Fig. 2b). Furthermore, the 11-gene signature was found to be significantly correlated with poor patient prognosis (Log-rank $p < 0.001$; Fig. 2c).

To evaluate the predictive power of the 11-gene signature in ES, we used Cox regression model to calculate

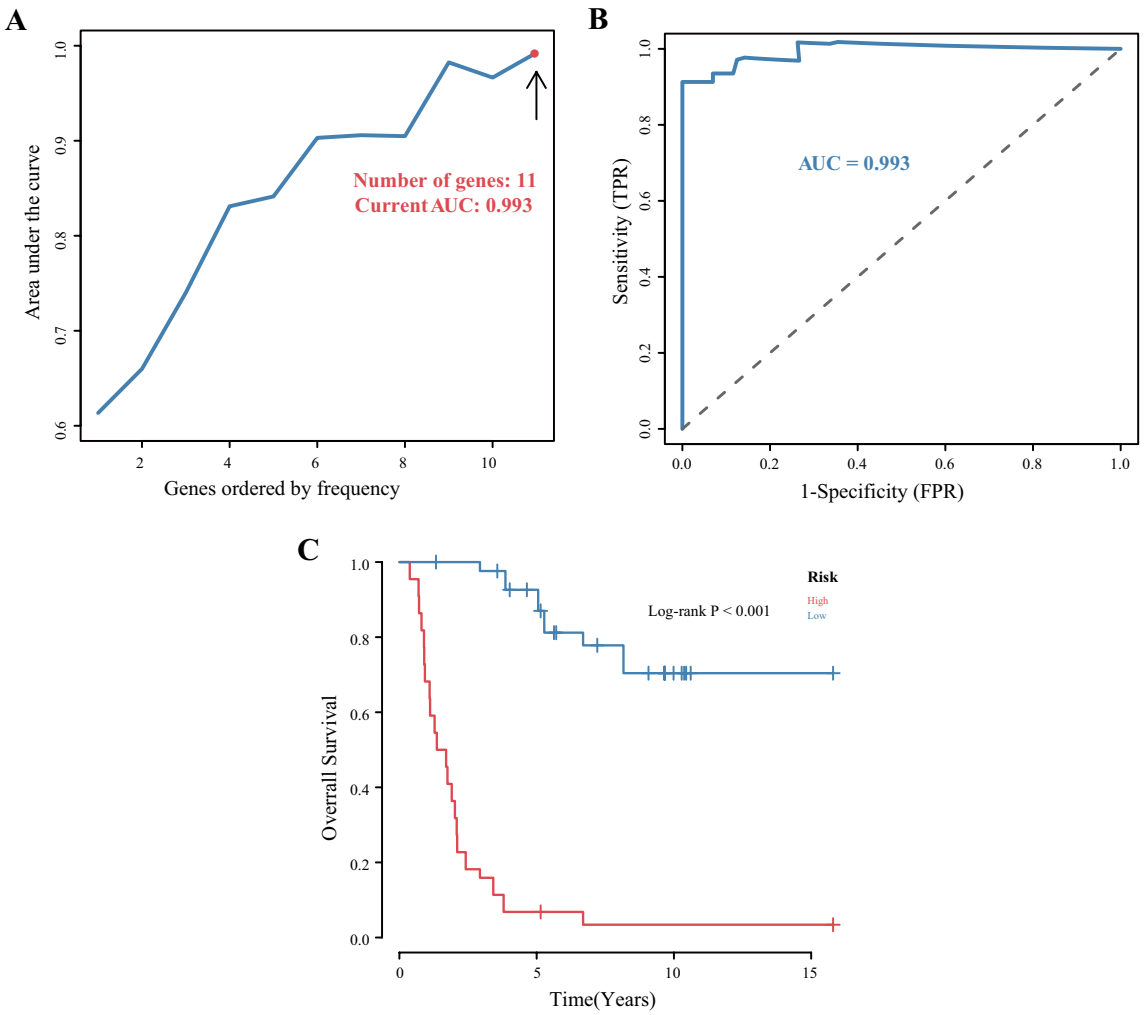

Fig. 2 Determination of genes in prognostic expression characteristics. **a** The relationship between gene signature frequency and area under the ROC curve shows that the 11-gene signature has the highest AUC. **b** ROC curve of 11-gene signature. **c** Stratified survival curves of patients with ES according to 11-gene signature: all survival curves were calculated by Kaplan–Meier analysis and compared by log-rank tests. *ROC* receiver-operating characteristic, *AUC* area under the curve, *ES* Ewing sarcoma

patient risk factors, relapse-free survival (RFS), and the expression of each gene in the gene combination (Fig. 3). The risk factor was significantly higher for patients in the high-risk group than in the low-risk group, while the RFS was significantly lower for high-risk patients than low-risk patients. The expression of each gene was significantly different between both groups.

Verification of optimal gene signature in external verification set

We conducted an ROC analysis of the optimal gene signature through an external validation set, and the results showed (Fig. 4a) that the AUC for 1, 3, and 5 years were 1, 0.92, and 0.97, respectively. It is suggested that the optimal gene signature has higher prognostic value in the external verification set of ES. The survival analysis of the optimal gene signature showed that there was a significant difference between the high-risk group and the low-risk group ($p < 0.0001$) (Fig. 4b). The results of Cox regression model validation (Fig. 4c) showed that there were significant differences in RFS and gene expression between high-risk and low-risk groups according to the optimal gene signature risk score.

We conducted ROC analysis of the optimal gene signature and other known biomarkers (SOX2, STAG2, TP53, and CXCR4) in the external verification set. The results showed (Fig. 4d) that the AUC of the optimal gene signature at different times was higher than the other prognostic

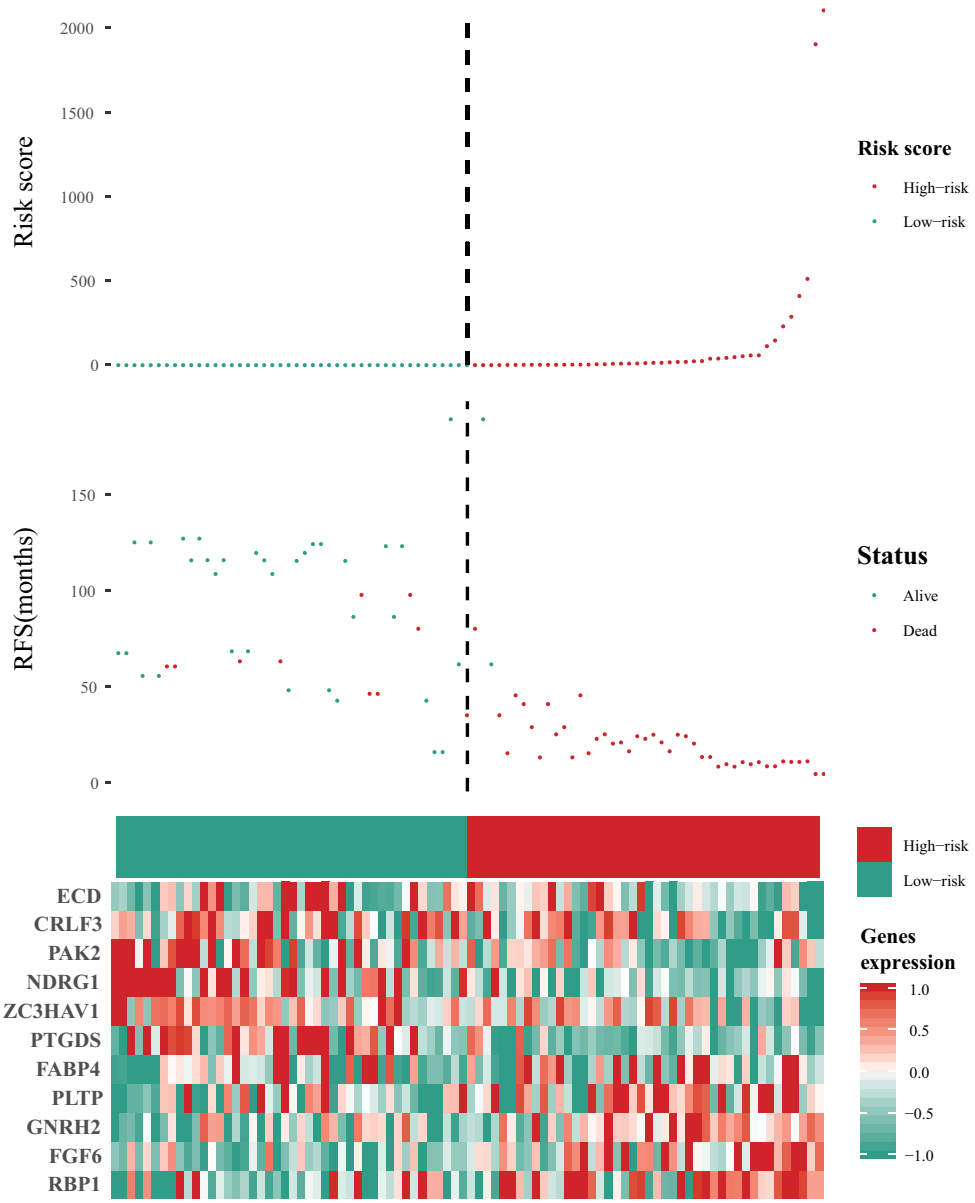

Fig. 3 Evaluation of the relationship between the prognostic gene signature and patient prognosis via a COX regression model. The risk factors in the high- and the low-risk groups, gene expression, relapse-free survival (RFS), and the gene signature are shown for each case. The risk scores are arranged in ascending order from left to right

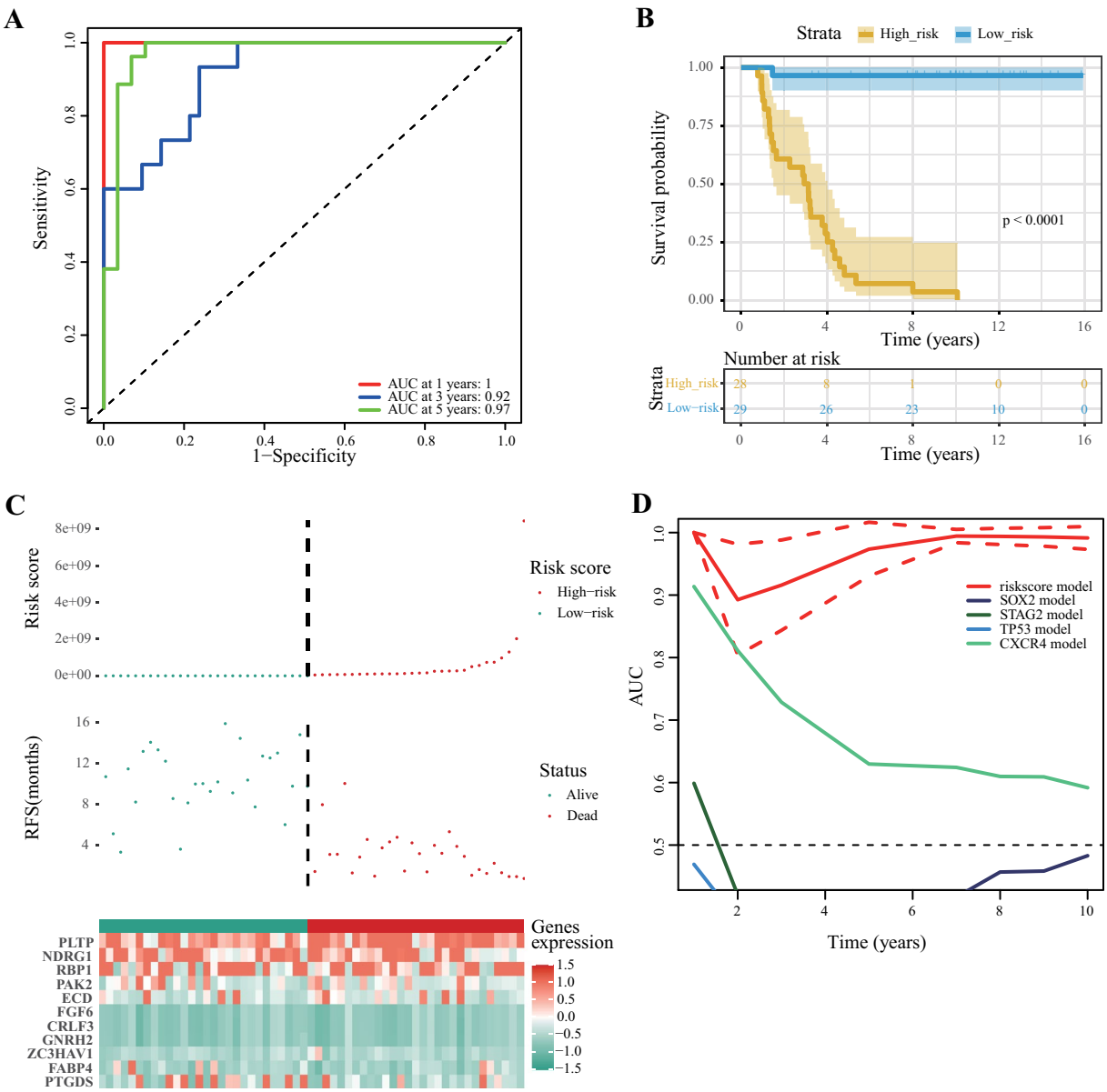

Fig. 4 Verification of 11-gene signature in external verification set. **a** ROC curves of 11-gene signature at different time stages. **B** The Kaplan–Meier curve based on the 11-gene signature. **c** Cox regression model evaluates the prognostic value of 11-gene signature. **d** ROC analysis results of 11-gene signature and known prognostic biomarkers

biomarkers, indicating that the gene signature has a high clinical prognostic value.

The results of univariate and multivariate Cox regressions on clinicopathological characteristics (gender, age, and metastatic) and 11-gene signature

Univariate Cox regression showed that metastasis (HR = 1.479, $p = 0.016$) was significantly correlated with the prognosis of Ewing's sarcoma patients, while 11-gene signature was significantly correlated in both training set (HR = 1.185, $p < 0.001$) and external validation set (HR = 2.105, $p < 0.001$) (Table 2). In multivariate Cox regression analysis, 11-gene signature was significantly correlated with prognosis in both training set (HR = 1.185, $p < 0.001$) and external validation set (HR = 2.098, $p = 0.003$) (Table 2). The above results suggested that the 11-gene signature was an important factor affecting prognosis.

Table 2 Univariate and multivariate Cox regressions on clinicopathological characteristics and optimal gene signature

Variables	Univariate Cox		Multivariate Cox	
	HR (95% CI)	*p* value	HR (95% CI)	*p* value
Training set				
Gender	0.902 (0.508–1.601)	0.724	1.101 (0.606 – 1.999)	0.753
Age	0.892 (0.559–1.660)	0.892	0.934 (0.540 – 1.613)	0.805
Metastatic	1.479 (0.845–2.591)	0.016	1.356 (0.771 – 2.384)	0.291
Risk score	1.185 (1.081 – 1.299)	< 0.001	1.185 (1.075 – 1.306)	< 0.001
External verification set				
Gender	1.039 (0.500 – 2.160)	0.919	1.044 (0.473 – 2.306)	0.915
Age	1.553 (0.743 – 3.244)	0.242	0.930 (0.404 – 2.138)	0.864
Metastatic	1.695 (0.797 – 3.602)	0.170	1.477 (0.656 – 3.325)	0.346
Risk score	2.105 (1.367 – 3.242)	< 0.001	2.098 (1.288 – 3.418)	0.003

Infiltration of immune cells in ES

PCA clustering graphs related to immune cell infiltration showed a significant difference in immune cell infiltration between ES and normal patients (Fig. 5a). The correlation thermogram for 24 immune cells showed that there was a significant positive correlation between B and cytotoxic T cells, negative correlations between CD8 T, NK, and NK CD56bright cells, and between NK CD56bright and Th2 cells (Fig. 5b). The interaction network of 24 kinds of immune cells (Fig. 5c) showed that NK and NK CD56bright cells had the strongest interaction with other immune cells, while Th2 and TEM cells had the weakest interaction with other immune cells. The differences in the infiltration of 24 immune cells in ES (Fig. 5d) showed that, compared with the normal control group, more cytotoxic T cells ($p = 0.022$), but fewer CD8 T cells infiltrated ES ($p = 0.025$).

Relationship between immune cell infiltration and ES prognosis

Analysis of the relationship between immune cell infiltration and patient prognosis (Fig. 6) showed that the infiltration of B (HR: 2.27, 95% CI: 1.31–3.94, $p = 0.003$), CD8 T (HR: 2.15, 95% CI: 1.24–3.7, $p = 0.01$), NK CD56bright (HR: 2.02, 95% CI: 1.17–3.51, $p = 0.01$), NK (HR: 2.33, 95% CI: 1.25–4.35, $p = 0.001$), and Th2 (HR: 3.67, 95% CI: 2.03–6.63, $p < 0.001$) cells indicated a poor prognosis, whereas the infiltration of cytotoxic T cells (HR: 0.45, 95% CI: 0.22–0.89, $p = 0.004$), macrophages (HR: 0.43, 95% CI: 0.24–0.76, $p = 0.001$), mast cells (HR: 0.49, 95% CI: 0.28–0.84, $p = 0.01$), TCM (HR: 0.38, 95% CI: 0.22–0.65, $p < 0.001$), and TEM (HR: 0.45, 95% CI: 0.24–0.85, $p = 0.003$) suggested that the patient had a good prognosis.

Correlation between optimal gene signature and immune cells

To explore the relationship between prognostic biomarkers and immune cells, we analyzed the correlation between the prognostic biomarkers and immune cells. The positive correlation of FABP4 with macrophages, and negative correlation of NDRG1 with Th2 cells were found to be the strongest (Fig. 7).

Discussion

Immunotherapy can activate patient autoimmune systems to kill tumor cells, completely reforming cancer treatment; currently, this is the most anticipated treatment method for tumors. In ES, although EWS-FLI1, FOXO1, Gli, and other immunotherapeutic targets have been discovered (Guo et al. 2006; Sankar et al. 2014; Yang et al. 2010), none of these have been applied clinically. Therefore, the immunotherapeutic target (11-gene signature) identified in this study is of great significance for the development of new immunotherapeutic strategies.

GO analysis showed that DEIRGs are mainly enriched in biological processes, such as leukocyte migration, leukocyte chemotaxis, and cytokine receptor binding, with roles in ES that have long been confirmed by various studies (Abadie et al. 2004; McCoy et al. 1977). This shows that our analysis is true and reliable. By constructing a PPI network, we found that VEGFA, MMP9, CXCL8, EGF, IGF1, CXCR4, TGFB1, EGFR, SPP1, and ICAM1 may play an important role in the development of ES. In addition, using GO annotation semantics to investigate the functional similarity of key IRGs, a strong biological functional similarity was found between EGFR and SPP1, and CXCR4 and MMP9. VEGFA and IGF1 are not only related to tumor cell proliferation, but also play an important role in tumor angiogenesis (Amaral

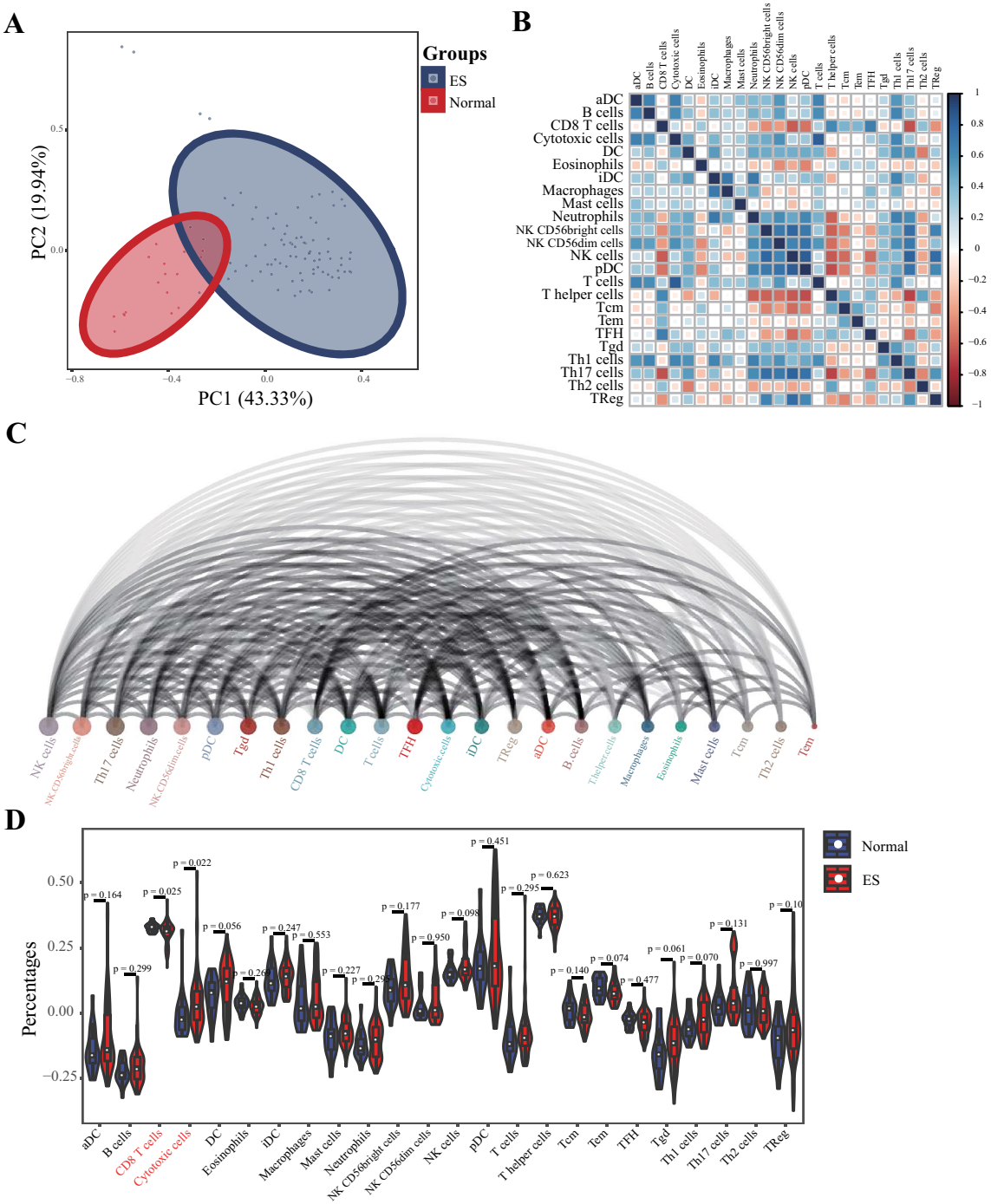

Fig. 5 Infiltration of immune cells in ES. **a** PCA chart of immune cell infiltration between ES samples and normal samples. **b** Correlation heat map of 24 immune cells; the size of the colored square represents the intensity of correlation; the color represents the intensity of correlation—the darker the color, the stronger the correlation. Blue represents a positive correlation, and red represents a negative correlation. **c** Network diagram of 24 immune cell interactions; the size of the circle represents the intensity of the interaction. **d** Violin diagram of the differences in infiltration of 24 immune cells in the ES and normal group, marked with different immune cells in red. *ES* Ewing sarcoma, *PCA* principal component analysis

et al. 2015; Zhou et al. 2018). Both CXCR4 and MMP9 have been reported to promote ES invasion and metastasis through angiogenesis and tumor cell proliferation in ES (Hamdan et al. 2014; Sanceau et al. 2003) which is similar to our results, indicating the reliability of our analysis. EGF–EGFR is associated with the molecular mechanisms of

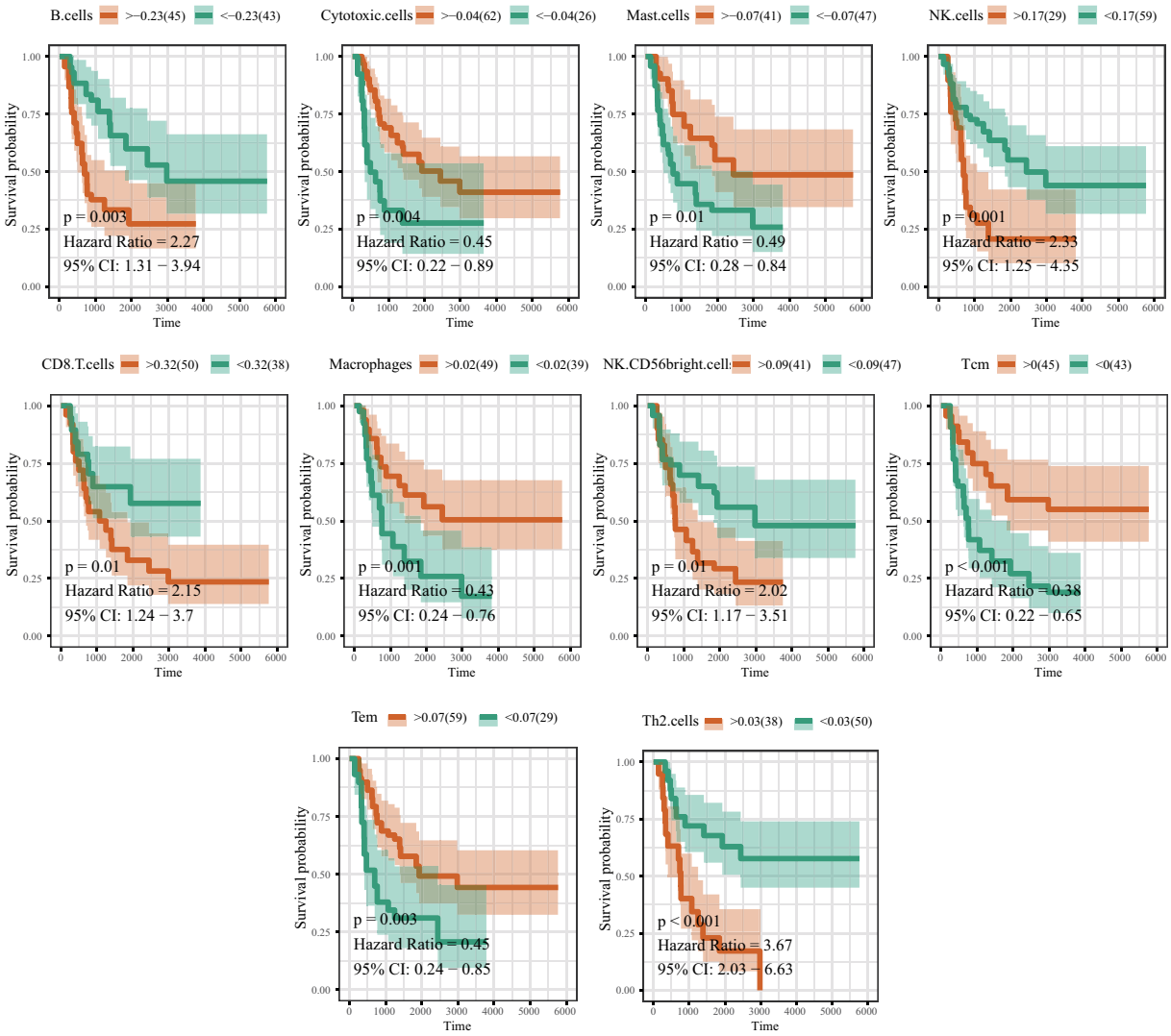

Fig. 6 Kaplan–Meier survival curve of immune cell infiltration and prognosis of patients with Ewing sarcoma (ES). The horizontal axis represents the total survival time, and the vertical axis represents the survival rate

multiple tumors that stimulate ES cell proliferation mainly by regulating EWS-FLI1 and neurotrophin signaling (Kersting et al. 2018). Our results indicate that EGFR and SPP1 have structural and functional similarity, so we speculate that SSP1 may be closely associated with ES cell proliferation, which we intend to study further. Currently, we have not found any report on the effects of TGFB1, SPP1, ICAM1, and CXCL8 in ES, which may help us to explore new immunotherapy targets in ES, yet it is worth further discussion.

Prognostic biomarkers are beneficial for the individualized treatment of ES patients (Walther et al. 2014). Therefore, researchers are paying more and more attention to the predictive function of characteristic molecular markers for ES patient prognosis while focusing on ES immunotherapy targets. Existing studies have shown that the expression of genes such as SOX2, MTAP, Connexin, and 43 Nucleophosmin is related to ES patient prognosis (Abrahao-Machado et al. 2018; Bui et al. 2011; Kikuta et al. 2009; Sannino et al. 2019). However, the use of a single biomarker for evaluating prognosis will undoubtedly omit a lot of prognostic evaluation information, making it difficult to provide guidance for the clinic. We obtained the 11-gene signature with the strongest correlation with ES prognosis via machine learning algorithm–iterative Lasso Cox regression. Compared with the single ES prognostic biomarkers found so far, the 11-gene signature provides a more reliable prognosis assessment, as it avoids the omission of prognostic information and also eliminates the redundancy of prognostic information through 1000 Lasso regressions. This multi-gene fitting method for evaluating patient prognosis has been applied to prognosis evaluations for various tumors, such as colon,

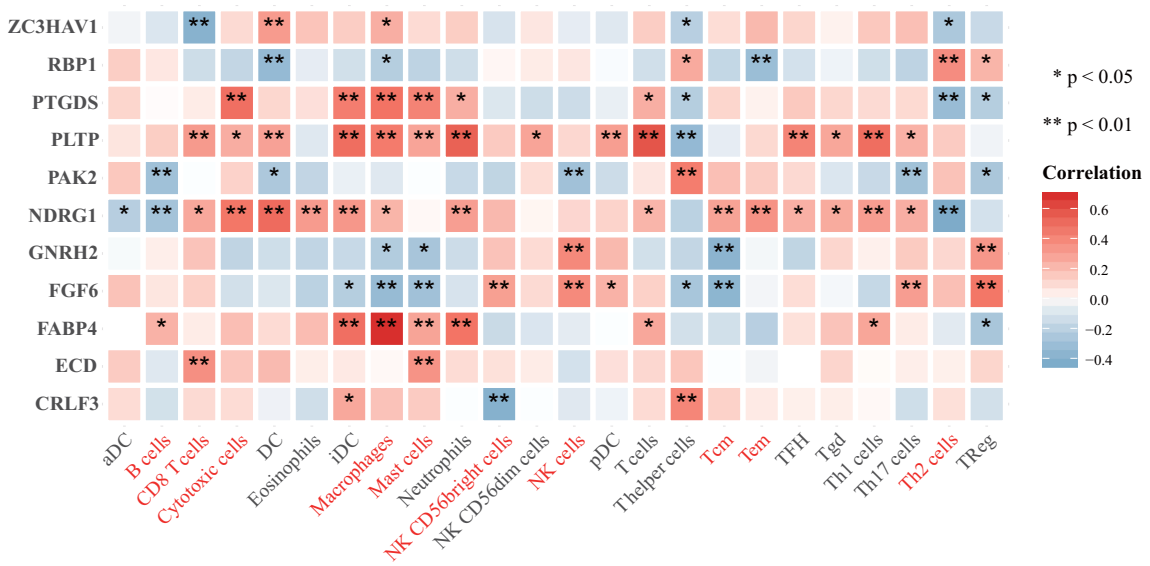

Fig. 7 Correlation analysis of genes in gene signature and immune cells. The ordinate is the gene name, the abscissa is the immune cell name, and the color represents the correlation coefficient, $*p < 0.05$, $**p < 0.01$

gastric, and breast cancers (Cui et al. 2019; Jiang et al. 2019; Salazar et al. 2011). However, the evaluation of ES prognosis is yet to be reported. In order to verify the accuracy of the 11-gene signature in ES prognosis, we applied a Cox regression model. The results showed that the risk factors were significantly higher for patients in the high-risk group than those in the low-risk group. The prognosis evaluation system further ensured the reliability of our research results. To verify the accuracy of 11-gene signature in ES evaluation, we also verified 11-gene signature through external verification set. The results show that 11-gene signature has a high predictive value of prognosis, and the prognostic value of 11-gene signature is higher than that of ES known prognostic biomarkers. Moreover, Li et al. (2017) identified six biomarkers of inflammatory prognosis in the blood of ES patients: CRP (AUC = 0.676), GPS (AUC = 0.634), NLR (AUC = 0.618), PLR (AUC = 0.613), LMR (AUC = 0.574), and NPC (AUC = 0.571). Yu et al. (2012b) showed that circulating cell-free mitochondrial DNA (AUC = 0.708) can be used to identify non-invasive ES patients via real-time fluorescence quantitative PCR. Compared with the previous prognostic biomarkers, the 11-gene signature (AUC = 0.993) has a significantly higher ability to predict ES prognosis, but still needs to be verified using a large clinical research sample size, such that the 11-gene signature can be accurately and reliably applied to the clinic.

We analyzed the infiltration of immune cells in ES by ssGSEA and evaluated the relationship between the 11-gene signature and ES cell infiltration. The results showed that macrophages and cytotoxic, CD8 T, NK, mast, B, NK CD56bright, TEM, TCM, and Th2 cells are obviously related to patient prognosis, but cytotoxic and CD8 T-cell infiltration were different. By exploring the correlation between the 11-gene signature and immune cell infiltration, we found that FABP4 was positively correlated with macrophages, while NDRG1 was negatively correlated with Th2 cells. Mature TAM can promote tumor cell invasion and vascular proliferation by influencing cytokine and growth factor secretion in a variety of tumors (Wyckoff et al. 2007). Clinical studies also show that high TAM expression suggests poor prognosis (Balkwill 2004; Pollard 2004). In ES, TAM mainly promotes osteoclast differentiation through RANKL- and TNF-α-dependent mechanisms, and produces osteoclasts, thus participating in osteolysis (Lau et al. 2007). Additionally, the relationship between TAM infiltration and prognosis has been established (Fujiwara et al. 2011). NK cells can kill tumor cells by activating NKG2D and DNAM-1 receptors (Verhoeven et al. 2008), because NK cells cannot recognize antigens, which leads to allergic reactions; studies have shown that allograft NK cells are more effective than autologous NK cells. The killing effect of tumors is more obvious (Ljunggren and Malmberg 2007), so NK cells may be beneficial for ES immunotherapy. Our analysis yielded similar results as the study on macrophages and NK cells, indicating that our study on the infiltration of immune cells in ES is accurate and reliable. Currently, the relationship between ES and cytotoxic, CD8 T, mast, B, NK CD56bright, TEM, TCM, and Th2 cells has not been reported and requires further study. We found a strong correlation between FABP4 and macrophages, and NDRG1 and Th2 cells, which provides ideas for further exploration

of the immune regulation mechanism of ES, for which no study currently exists.

Conclusion

In summary, we used machine learning algorithm–iterative Lasso Cox regression to cross check the gene–ES prognosis relationship, identified an 11-gene signature that is strongly related to ES prognosis, and verified it via the external validation set. We also found that FABP4 has the strongest correlation with macrophases, also NDRG1 and Th2 cells in 11-gene signature, which is helpful to further study the molecular mechanism of gene signature participating in ES. However, the 11-gene signature still needs high-quality clinical samples for verification and in-depth discussion of molecular mechanisms to be accurately and reliably applied in the clinic.

Acknowledgements This work has benefited from GEO and ICGC. We thank the GEO and ICGC network for its generous sharing large amounts of data. We would like to thank FigureYa (Xiao Ya Hua Tu) for the figure technology support. We would like to thank Editage (www.editage.cn) for English language editing.

Author contributions Study conception and design: ER and QX; Data acquisition and analysis: ER, YD, and QX; Data visualization: WY, ZW, and GZ; manuscript writing: ER, YD, and WY; Manuscript editing: ER and QX; manuscript reviewing: ZW and GZ; All authors approved the final version of the paper.

Data availability The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Compliance with ethical standards

Conflict of interest All authors declared no conflict of interests.

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

References

Abadie A, Besancon F, Wietzerbin J (2004) Type I interferon and TNF alpha cooperate with type II interferon for TRAIL induction and triggering of apoptosis in SK-N-MC EWING tumor cells. Oncogene 23:4911–4920. https://doi.org/10.1038/sj.onc.1207614

Abrahao-Machado LF et al (2018) Loss of MTAP expression is a negative prognostic marker in Ewing sarcoma family of tumors. Biomark Med 12:35–44. https://doi.org/10.2217/bmm-2017-0152

Adai AT, Date SV, Wieland S, Marcotte EM (2004) LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. J Mol Biol 340:179–190. https://doi.org/10.1016/j.jmb.2004.04.047

Amaral AT et al (2015) Trabectedin efficacy in Ewing sarcoma is greatly increased by combination with anti-IGF signaling agents. Clin Cancer Res 21:1373–1382. https://doi.org/10.1158/1078-0432.Ccr-14-1688

Andreou D et al (2020) Which factors are associated with local control and survival of patients with localized Pelvic Ewing's sarcoma? A retrospective analysis of data from the Euro-EWING99 trial. Clin Orthopaed Relat Res 478:290–302. https://doi.org/10.1097/corr.0000000000000962

Ashburner M et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29. https://doi.org/10.1038/75556

Balkwill F (2004) Cancer and the chemokine network. Nat Rev Cancer 4:540–550. https://doi.org/10.1038/nrc1388

Barrett T et al (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–995. https://doi.org/10.1093/nar/gks1193

Berghuis D et al (2012) The CXCR4-CXCL12 axis in Ewing sarcoma: promotion of tumor growth rather than metastatic disease. Clin Sarcoma Res 2:24. https://doi.org/10.1186/2045-3329-2-24

Bhattacharya S et al (2018) ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. Sci Data 5:180015. https://doi.org/10.1038/sdata.2018.15

Bindea G et al (2013) Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity 39:782–795. https://doi.org/10.1016/j.immuni.2013.10.003

Bui MM, Han G, Acs G, Reed D, Gonzalez RJ, Pasha TL, Zhang PJ (2011) Connexin 43 is a potential prognostic biomarker for ewing sarcoma/primitive neuroectodermal tumor. Sarcoma 2011:971050. https://doi.org/10.1155/2011/971050

Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY (2014) cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 8(Suppl 4):S11. https://doi.org/10.1186/1752-0509-8-s4-s11

Cho D, Shook DR, Shimasaki N, Chang YH, Fujisaki H, Campana D (2010) Cytotoxicity of activated natural killer cells against pediatric solid tumors. Clin Cancer Res 16:3901–3909. https://doi.org/10.1158/1078-0432.Ccr-10-0735

Cui Q et al (2019) A prognostic eight-gene expression signature for patients with breast cancer receiving adjuvant chemotherapy. J Cell Biochem. https://doi.org/10.1002/jcb.29550

Davis S, Meltzer PS (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics (Oxford, England) 23:1846–1847. https://doi.org/10.1093/bioinformatics/btm254

De Angulo G, Hernandez M, Morales-Arias J, Herzog CE, Anderson P, Wolff J, Kleinerman ES (2007) Early lymphocyte recovery as a prognostic indicator for high-risk Ewing sarcoma. J Pediatr Hematol/Oncol 29:48–52. https://doi.org/10.1097/MPH.0b013e31802d3e3e

Doncheva NT, Morris JH, Gorodkin J, Jensen LJ (2019) Cytoscape StringApp: network analysis and visualization of proteomics data. J Proteome Res 18:623–632. https://doi.org/10.1021/acs.jproteome.8b00702

Eschrich S et al (2005) Molecular staging for survival prediction of colorectal cancer patients. J Clin Oncol 23:3526–3535. https://doi.org/10.1200/jco.2005.00.695

Friendly M (2002) Corrgrams: exploratory displays for correlation. Matrices 56:316–324

Frost HR, Amos CI (2017) Gene set selection via LASSO penalized regression (SLPR). Nucleic Acids Res 45:114. https://doi.org/10.1093/nar/gkx291

Fujiwara T et al (2011) Macrophage infiltration predicts a poor prognosis for human ewing sarcoma. Am J Pathol 179:1157–1170. https://doi.org/10.1016/j.ajpath.2011.05.034

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy–analysis of Affymetrix GeneChip data at the probe level. Bioinformatics (Oxford, England) 20:307–315. https://doi.org/10.1093/bioinformatics/btg405

Guo W, Tang XD, Tang S, Yang Y (2006) Preliminary report of combination chemotherapy including Arsenic trioxide for stage III osteosarcoma and Ewing sarcoma. Zhonghua wai ke za zhi [Chin J Surg] 44:805–808

Hamdan R, Zhou Z, Kleinerman ES (2014) Blocking SDF-1alpha/CXCR4 downregulates PDGF-B and inhibits bone marrow-derived pericyte differentiation and tumor vascular expansion in Ewing tumors. Mol Cancer Ther 13:483–491. https://doi.org/10.1158/1535-7163.Mct-13-0447

Hänzelmann S, Castelo R, Guinney J (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinf 14:7. https://doi.org/10.1186/1471-2105-14-7

Hingorani P et al (2015) Increased CTLA-4(+) T cells and an increased ratio of monocytes with loss of class II (CD14(+) HLA-DR (lo/neg)) found in aggressive pediatric sarcoma patients. J Immunother Cancer 3:35. https://doi.org/10.1186/s40425-015-0082-0

Hudson TJ et al (2010) International network of cancer genome projects. Nature 464:993–998. https://doi.org/10.1038/nature08987

Jiang Y et al (2019) Tumor immune microenvironment and chemosensitivity signature for predicting response to chemotherapy in gastric cancer. Cancer Immunol Res 7:2065–2073. https://doi.org/10.1158/2326-6066.Cir-19-0311

Kaatsch P, Strothotte J, Becker C, Bielack S, Dirksen U, Blettner M (2016) Pediatric bone tumors in Germany from 1987 to 2011: incidence rates, time trends and survival. Acta Oncologica (Stockholm, Sweden) 55:1145–1151. https://doi.org/10.1080/0284186x.2016.1195509

Kersting N et al (2018) Epidermal growth factor receptor regulation of ewing sarcoma cell function. Oncology 94:383–393. https://doi.org/10.1159/000487143

Kikuta K et al (2009) Nucleophosmin as a candidate prognostic biomarker of Ewing's sarcoma revealed by proteomics. Clin Cancer Res 15:2885–2894. https://doi.org/10.1158/1078-0432.Ccr-08-1913

Krug K et al (2018) A curated resource for phosphosite-specific signature analysis. Mol Cell Proteom. https://doi.org/10.1074/mcp.TIR118.000943

Lau YS, Adamopoulos IE, Sabokbar A, Giele H, Gibbons CL, Athanasou NA (2007) Cellular and humoral mechanisms of osteoclast formation in Ewing's sarcoma. Br J Cancer 96:1716–1722. https://doi.org/10.1038/sj.bjc.6603774

Li J et al (2013) Identifying differentially expressed genes and small molecule drugs for prostate cancer by a bioinformatics strategy. Asian Pac J Cancer Prevent 14:5281–5286

Li YJ, Yang X, Zhang WB, Yi C, Wang F, Li P (2017) Clinical implications of six inflammatory biomarkers as prognostic indicators in Ewing sarcoma. Cancer Manag Res 9:443–451. https://doi.org/10.2147/cmar.S146827

Ljunggren HG, Malmberg KJ (2007) Prospects for the use of NK cells in immunotherapy of human cancer. Nat Rev Immunol 7:329–339. https://doi.org/10.1038/nri2073

Maag JLV (2018) gganatogram: an R package for modular visualisation of anatograms and tissues based on ggplot2. F1000Research 7:1576. https://doi.org/10.12688/f1000research.16409.2

McCoy JL et al (1977) Leukocyte migration inhibition in patients with Ewing's sarcoma by 3-M potassium chloride extracts of fresh and tissue-cultured Ewing's sarcomas. J Natl Cancer Inst 59:1119–1125. https://doi.org/10.1093/jnci/59.4.1119

Meyer-Wentrup F, Richter G, Burdach S (2005) Identification of an immunogenic EWS-FLI1-derived HLA-DR-restricted T helper cell epitope. Pediatr Hematol Oncol 22:297–308. https://doi.org/10.1080/08880010590935194

Pollard JW (2004) Tumour-educated macrophages promote tumour progression and metastasis. Nat Rev Cancer 4:71–78. https://doi.org/10.1038/nrc1256

Ringner M (2008) What is principal component analysis? Nat Biotechnol 26:303–304. https://doi.org/10.1038/nbt0308-303

Salazar R et al (2011) Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol 29:17–24. https://doi.org/10.1200/jco.2010.30.1077

Sanceau J, Truchet S, Bauvois B (2003) Matrix metalloproteinase-9 silencing by RNA interference triggers the migratory-adhesive switch in Ewing's sarcoma cells. J Biol Chem 278:36537–36546. https://doi.org/10.1074/jbc.M304300200

Sankar S et al (2014) Reversible LSD1 inhibition interferes with global EWS/ETS transcriptional activity and impedes Ewing sarcoma tumor growth. Clin Cancer Res 20:4584–4597. https://doi.org/10.1158/1078-0432.Ccr-14-0072

Sannino G et al (2019) Gene expression and immunohistochemical analyses identify SOX2 as major risk factor for overall survival and relapse in Ewing sarcoma patients. EBioMedicine 47:156–162. https://doi.org/10.1016/j.ebiom.2019.08.002

Sveen A et al (2012) ColoGuidePro: a prognostic 7-gene expression signature for stage III colorectal cancer patients. Clin Cancer Res 18:6001–6010. https://doi.org/10.1158/1078-0432.Ccr-11-3302

Szklarczyk D et al (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47:D607–d613. https://doi.org/10.1093/nar/gky1131

Tirode F et al (2014) Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. Cancer Discov 4:1342–1353. https://doi.org/10.1158/2159-8290.Cd-14-0622

Verhoeven DH et al (2008) NK cells recognize and lyse Ewing sarcoma cells through NKG2D and DNAM-1 receptor dependent pathways. Mol Immunol 45:3917–3925. https://doi.org/10.1016/j.molimm.2008.06.016

Walther C et al (2014) A novel SERPINE1-FOSB fusion gene results in transcriptional up-regulation of FOSB in pseudomyogenic haemangioendothelioma. J Pathol 232:534–540. https://doi.org/10.1002/path.4322

Wang Y et al (2004) Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J Clin Oncol 22:1564–1571. https://doi.org/10.1200/jco.2004.08.186

Wyckoff JB et al (2007) Direct visualization of macrophage-assisted tumor cell intravasation in mammary tumors. Cancer Res 67:2649–2656. https://doi.org/10.1158/0008-5472.Can-06-1823

Yang L, Hu HM, Zielinska-Kwiatkowska A, Chansky HA (2010) FOXO1 is a direct target of EWS-Fli1 oncogenic fusion protein in Ewing's sarcoma cells. Biochem Biophys Res Commun 402:129–134. https://doi.org/10.1016/j.bbrc.2010.09.129

Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics (Oxford, England) 26:976–978. https://doi.org/10.1093/bioinformatics/btq064

Yu G, Wang LG, Han Y, He QY (2012a) clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS J Integrative Biol 16:284–287. https://doi.org/10.1089/omi.2011.0118

Yu M, Wan YF, Zou QH (2012b) Cell-free circulating mitochondrial DNA in the serum: a potential non-invasive biomarker for Ewing's sarcoma. Arch Med Res 43:389–394. https://doi.org/10.1016/j.arcmed.2012.06.007

Zhou X et al (2018) MicroRNA-638 inhibits cell growth and tubule formation by suppressing VEGFA expression in human Ewing sarcoma cells. Biosci Rep. https://doi.org/10.1042/bsr20171017