Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Two-step consensus clustering approach to immune cell infiltration: An integrated exploration and validation of prognostic and immune implications in sarcomas

Ao-Yu Li [a,b], Jie Bu [c], Hui-Ni Xiao [d], Zi-Yue Zhao [a,b], Jia-Lin Zhang [a,b], Bin Yu [a,b], Hui Li [a,b,**], Jin-Ping Li [e,***], Tao Xiao [a,b,*]

[a] *Department of Orthopedics, The Second Xiangya Hospital, Central South University, Changsha, China*
[b] *Orthopedic Biomedical Materials Engineering Laboratory of Hunan Province, Changsha, China*
[c] *Department of Orthopedics, Hunan Cancer Hospital, The Affiliated Cancer Hospital of Xiangya School of Medicine, Central South University, Changsha, China*
[d] *Department of Gastroenterology, The Second Affiliated Hospital, University of South China, Hengyang, China*
[e] *Department of Orthopedics, Changsha Central Hospital, The Affiliated Changsha Central Hospital, Hengyang Medical School, University of South China, Changsha, China*

ARTICLE INFO

Keywords:
Sarcomas

ABSTRACT

To conduct a comprehensive investigation of the sarcoma immune cell infiltration (ImmCI) patterns and tumoral microenvironment (TME). We utilized transcriptomic, clinical, and

Abbreviations: ATRX, Alpha Thalassemia/Mental Retardation Syndrome X-Linked; CD274, PDCD1 Ligand 1; CD276, CD276 Molecule; CD27, T-Cell Activation Antigen CD27; CD28, T-Cell-Specific Surface Glycoprotein CD28; CD40, B-Cell Surface Antigen CD40; CD80, Costimulatory Molecule Variant IgV-CD80; CD8A, T-Cell Surface Glycoprotein CD8 Alpha Chain; CTLA4, Cytotoxic T-Lymphocyte-Associated Antigen 4; CXCL10, Chemokine (C-X-C Motif) Ligand 10; CXCL9, Chemokine (C-X-C Motif) Ligand 9; GITRL, Tumor Necrosis Factor Superfamily Member 18; GZMA, Granzyme A; GZMB, Granzyme B; HAVCR2, Hepatitis A Virus Cellular Receptor 2; HPD, 4-Hydroxyphenylpyruvate Dioxygenase; ICOS, Inducible T Cell Costimulator; IDO1, Indoleamine 2,3-Dioxygenase 1; IFNG, Interferon Gamma; LAG3, Lymphocyte Activating 3; LIGHT, TNFSF14; MUC16, Mucin 16; Cell Surface Associated, PD-1; Programmed Cell Death 1, PD-L2; Programmed Cell Death 1 Ligand 2, PDCD1; Programmed Cell Death 1, PDCD1LG2; Programmed Cell Death 1 Ligand 2, PRF1; Perforin 1, RB1; RB Transcriptional Corepressor 1, TBX2; T-Box Transcription Factor 2, TNF; Tumor Necrosis Factor, TNFSF14; TNF Superfamily Member 14, TNFSF18; TNF Superfamily Member 18, TP53; Tumor Protein P53, TTN; Titin. Other abbreviations: CDF, cumulative distribution function; CIBERSORT, Cell-type Identification by Estimating Relative Subsets of RNA Transcripts; CTLA4, cytotoxic T lymphocyte-associated antigen-4; DEGs, Differentially Expressed Genes; EMA, European Medicines Agency; ESTIMATE, Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data; FDA, Food and Drug Administration; GEO, Gene Expression Omnibus; GO, Gene Ontology; GSEA, Gene Set Enrichment Analysis; ICB, immune checkpoint blockade; ImmCI, immune cell infiltration; ICGC, International Cancer Genome Consortium; KEGG, Kyoto Encyclopedia of Genes and Genomes; MPNST, Malignant Peripheral Nerve Sheath Tumors; OS, overall survival; PCA, principal component analysis; PD1, Programmed Cell Death 1; TARGET, Therapeutically Applicable Research to Generate Effective Treatments; TAMs, tumor-associated macrophages; TCGA, The Cancer Genome Atlas; TIDE, tumor immune dysfunction evaluation; TMB, Tumor Mutation Burden; TME, tumor microenvironment; TPM, Transcripts per kilobase million.

* Corresponding author. Department of Orthopedics, the Second Xiangya Hospital, Central South University, Changsha, China.

** Corresponding author. Department of Orthopedics, the Second Xiangya Hospital, Central South University, Changsha, China.

*** Corresponding author. Department of Orthopedics, Changsha Central Hospital, the Affiliated Changsha Central Hospital, Hengyang Medical School, University of South China, Changsha, China.

E-mail addresses: 208202106@csu.edu.cn (A.-Y. Li), bujie@hnca.org.cn (J. Bu), 2023020092@usc.edu.cn (H.-N. Xiao), 2204140429@csu.edu.cn (Z.-Y. Zhao), 228202130@csu.edu.cn (J.-L. Zhang), yubin@csu.edu.cn (B. Yu), lihuix@csu.edu.cn (H. Li), 2018050734@usc.edu.cn (J.-P. Li), xiaotaoxyl@csu.edu.cn (T. Xiao).

https://doi.org/10.1016/j.heliyon.2024.e38253
Received 29 March 2024; Received in revised form 27 August 2024; Accepted 20 September 2024
Available online 1 October 2024

Tumor microenvironment
Immune cell infiltration
Prognosis
Immune checkpoint blade

mutation data of sarcoma patients (training cohort) obtained from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) server. Cell-type Identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT) and Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data (ESTIMATE) algorithms were applied to decipher the immune cell infiltration landscape and TME profiles of sarcomas. An unsupervised clustering method was utilized for classifying ImmCI clusters (initial clustering) and ImmCI-based differentially expressed gene-driven clusters (secondary clustering). Mortality rates and immune checkpoint gene levels was analyzed among the identified clusters. We calculated the ImmCI score through principal component analysis. The tumor immune dysfunction evaluation (TIDE) score was also employed to quantify immunotherapy efficacy between two ImmCI score groups. We further validated the biomarkers for ImmCI and gene-driven clusters via experimental verification and the accuracy of the ImmCI score in predicting survival outcomes and immunotherapy efficacy by external validation cohorts (testing cohort). We demonstrated that ImmCI cluster A and gene-driven cluster A, were beneficial prognostic biomarkers and indicators of immune checkpoint blockade response in sarcomas via *in-silico* and laboratory experiments. Additionally, the ImmCI score exhibited independent prognostic significance and was predictive of immunotherapy response. Our research underscores the clinical significance of ImmCI scores in identifying sarcoma patients likely to respond to immunotherapy.

1. Introduction

Sarcomas are rare and heterogeneous malignancies, accounting for only 1 % of all adult cancers [1]. Approximately 13,500 new sarcoma cases were diagnosed in the United States in 2019 [2]. Originating from the mesenchymal layer, sarcomas encompass at least 100 histologic subtypes, including fibromatous neoplasms, lipomatous neoplasms, myomatous neoplasms, nerve sheath tumors, and more [1]. Each subtype exhibits distinct clinical characteristics, and within the same histologic subtype, tumoral heterogeneity may still be present [3,4]. Although current sarcoma treatments involve multiple approaches, including adjuvant radiotherapy and chemotherapy following surgical treatment, individualized therapies have not been emphasized, resulting in a dismal prognosis with a 5-year survival rate estimated at 60 % [1,5]. Therefore, there is an urgent need for new targeted therapies. Immunotherapy has emerged as a key treatment option for sarcomas, but its application requires a patient-selective approach [6–8]. Consequently, a comprehensive analysis of the tumor microenvironment (TME) and immune cell infiltration (ImmCI) patterns in sarcomas is crucial to developing an ImmCI prediction model for prognosis and immunotherapy response to design more effective immunotherapies for this patient population.

It is now understood that the TME takes pivotal characters in the occurrence, development, deterioration, metastasis, and recurrence of sarcomas, albeit exhibiting a context-dependent manner [9,10]. The TME in sarcomas comprises the extracellular matrix and various non-tumor cells, including immune cells, vascular endothelial cells, and fibroblasts [9]. Given the heterogeneous nature of sarcomas, the function of the TME is multifaceted. For instance, studies investigating the role of tumor-associated macrophages (TAMs) in sarcomas have yielded inconsistent results. However, in leiomyosarcoma and synovial sarcomas, the number of TAMs has been negatively correlated with the overall survival (OS) of patients [11,12], while a positive correlation has been observed in rhabdomyosarcoma and bone sarcomas [13,14]. Similarly, the presence of CD8$^+$ T lymphocytes has been associated with improved prognosis in angiosarcoma, osteosarcoma, and Ewing sarcoma [15,16]. However, the increased survival rates warrant further multidimensional validation due to the heterogeneous nature of sarcomas. Thus, a comprehensive analysis of TME components in different sarcoma subtypes holds significant value.

Besides, the relationship between TME components and immunotherapy efficacy has been understudied. It is noteworthy that the function of immune cells and the efficacy of immunotherapy may vary among patients, yielding a robust immune response in only some patients [17]. Understanding ImmCI is a critical aspect of the TME, aiding researchers in improving the immunotherapy efficacy by enhancing the immune infiltration level. However, the therapeutic efficacy of Immune Checkpoint Blockades (ICBs) remains relatively uncertain, emphasizing the need for predictive models to determine therapeutic response. For instance, Kim et al. demonstrated that although the *PD-L1* patterns could predict the efficacy of pazopanib in sarcoma treatment, prediction signatures are still needed to identify patients who will benefit from pazopanib [18]. A study that conducted the immunotyping of sarcomas based on ImmCI features revealed a better *PD-1* blockade response with pembrolizumab in subtypes with high ImmCI [19]. Qi et al. investigated the impact of NETosis and anoikis on sarcoma prognosis and the immune microenvironment [20,21], while Li et al. focused on the prognostic significance of cuproptosis in Ewing's sarcoma [22]. Weng et al. conducted cluster analysis on sarcoma patients based on immune infiltration levels, comparing the immune cell characteristics, tumor mutation burden (TMB), gene mutations, and clinical outcomes among the 3 identified immune subtypes. Additionally, they established an immune gene signature with promising prognostic capabilities [23]. Shi et al. identified a system of long non-coding RNAs (lncRNAs) associated with epithelial-mesenchymal transition (EMT) and the immune microenvironment. They also investigated the role of these lncRNAs in predicting sarcoma prognosis and response to tumor immunotherapy [24]. Feng et al. conducted a comparative analysis of tumor immune microenvironment-related scores, including the 90-gene signature and immune infiltration cluster analysis, in leiomyosarcoma samples. Their findings revealed similarities between the scores but also highlighted inconsistencies. The study suggests the necessity for further comprehensive research in this area [25]. Therefore, our research on the ImmCI predictive signature for sarcoma prognosis

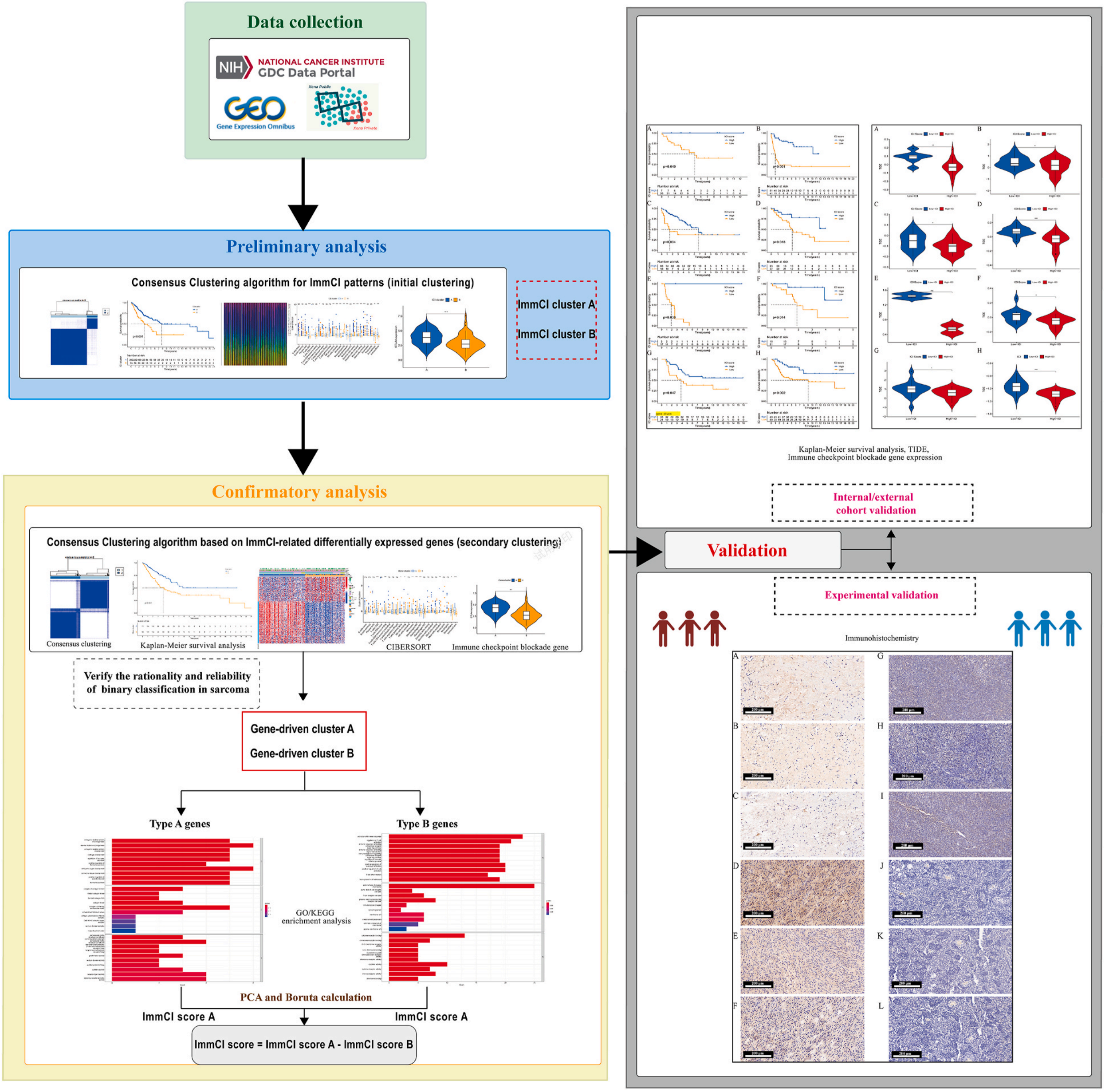

Data collection
NATIONAL CANCER INSTITUTE
GDC Data Portal
Gene Expression Omnibus
Xena Public
Xena Private
Preliminary analysis
Consensus Clustering algorithm for ImmCI patterns (initial clustering)
ImmCI cluster A
ImmCI cluster B
Confirmatory analysis
Consensus Clustering algorithm based on ImmCI-related differentially expressed genes (secondary clustering)
Consensus clustering
Kaplan-Meier survival analysis
CIBERSORT
Immune checkpoint blockade gene
Verify the rationality and reliability of binary classification in sarcoma
Gene-driven cluster A
Gene-driven cluster B
Type A genes
Type B genes
GO/KEGG enrichment analysis
PCA and Boruta calculation
ImmCI score A
ImmCI score A
ImmCI score = ImmCI score A - ImmCI score B
Validation
Kaplan-Meier survival analysis, TIDE, Immune checkpoint blockade gene expression
Internal/external cohort validation
Experimental validation
Immunohistochemistry
A
B
C
D
E
F
G
H
I
J
K
L

Fig. 1. Workflow of the research: 1) Data collection from TCGA, GEO, TARGET and ICGC databases, 2) The CIBERSORT algorithm, along with Consensus Clustering and Kaplan-Meier survival analysis, was employed to achieve ImmC cluster. 3) DEGs between the ImmC cluster were utilized for secondary clustering analysis, pathway enrichment analysis, and the calculation of ImmC scores. 4) Samples from the TARGET database, ICGC database, and clinical patients were utilized for the external validation of secondary cluster analysis, immune microenvironment analysis results, and immunohistochemistry experimental validation.

and immunotherapy response holds clinical significance in optimizing immunotherapy approaches.

This study investigated intra-tumoral immune infiltration and the TME landscape in sarcomas based on three independent cohorts. We further validated the rationality and reliability of binary classification (initial and secondary clustering) in sarcoma samples through *in silico* analysis and experimental validation. Subsequently, we calculated ImmCI scores for each sample and validated their accuracy in predicting survival outcomes and immunotherapy responses in training cohorts and another independent validation cohort. Moreover, we estimated the tumor mutation burden (TMB) of different ImmCI score subgroups and confirmed prognostic performance of ImmCI scores in both training and testing cohorts. The research process is displayed in Fig. 1.

2. Materials and methods

2.1. Ethical statement and clinical samples

The study was supervised by the Declaration of Helsinki and was approved by the Ethics Evaluation Committee of the Changsha Central Hospital (2021-S0100, 2021-08-20). All participants provided written informed consent. A total of 12 sarcoma patients and matching clinical information were obtained from the Changsha Central Hospital. This sarcoma cohort contains 8 subtypes: fibrosarcoma, synovial sarcoma, liposarcoma, rhabdomyosarcoma, nerve sheath tumor, osteosarcoma, Ewing sarcoma, chondrosarcoma. All patients were histologically and clinically diagnosed with sarcoma and had not undergone induction therapies before surgery. All tumors were diagnosed by two experienced senior pathologists.

2.2. Sarcoma data aggregation

Transcriptomic and clinical information from three independent sarcoma cohorts were obtained from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases. The gene expression RNAseq data for TCGA-SARC were retrieved from UCSC Xena, comprising 265 tissue samples representing various subtypes, including fibromatous neoplasms, lipomatous neoplasms, myomatous neoplasms, nerve sheath tumors, soft tissue tumors, sarcomas (NOS), and synovial-like neoplasms. Mutation data in VarScan2 format for 237 sets were downloaded from GDC TCGA-SARC. To ensure data compatibility with GEO, transcripts per kilobase million (TPM) values were derived from TCGA-SARC expression profiles (FPKM values). Selection criteria for GEO datasets included: (1) inclusion of more than 30 patients, (2) presence of at least one sarcoma subtype matching TCGA-SARC, (3) availability of RNA expression profiles, and (4) inclusion of clinical data and prognostic information. Based on these criteria, datasets GSE72545 and GSE17118 were selected. GSE72545 consists of 64 primary high-grade myxofibrosarcomas, and GSE17118 comprises 24 liposarcomas and 16 malignant peripheral nerve sheath tumors (MPNST). The comBat function in the 'sva' package was used to remove the batch effects between RNAseq data of different sources.

2.3. Immune cell infiltration analysis and tumor microenvironment analysis

To quantify immune cell infiltration and the infiltration degree of 22 different immune cell subsets, the Cell-type Identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT) algorithm was applied to the gene expression profiles of the three sarcoma cohorts (TCGA-SARC and GSE72545/17118) [26]. It is well-established that CIBERSORT utilizes a deconvolution method with supporting vector regression (using LM22 datasets) to analyze the portion of designated cell types from mixed cell types. The LM22 dataset contains information on 22 immune cell types. To ensure stability of results, this step was repeated 1,000 times. For tumor microenvironment analysis, immune and stromal cells were calculated using the "Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data (ESTIMATE)" algorithm in each sarcoma sample. The “ImmuneScore” and “StromalScore” were calculated to predict immune infiltration degree and tumor purity, respectively [27].

2.4. Consensus clustering for immune cell infiltration (initial clustering)

The consensus clustering approach was used to classify unsupervised groups in the dataset. For Consensus Clustering, PAM arithmetic and "euclidean" distance were utilized to complete 1000 bootstraps with every bootstrap containing $\geq$80 % of samples. Cluster number k was between 2 and 9. The optimum k was identified as per consensus matrix plots, cumulative distribution function (CDF), delta area plot and clinical meaning of clustering results. The "ConsensusClusterPlus" R package was utilized to stratify sarcoma samples into various ImmCI clusters [28].

2.5. Identify differentially expressed genes (DEGs) based on ImmCI clusters and construct gene-driven clusters (secondary clustering) with verification

To further explore the patterns of ImmCI, ImmCI-related DEGs were catalogized by the “limma” algorithms [29]. The significant filter criteria for DEGs were adjusted P-value$<$0.05 and $|$fold-change$|>$1. Then, Consensus Clustering was utilized to classify sarcoma patients according to the expression matrix of DEGs. The specific parameters for the relevant Consensus Clustering are the same as for "initial clustering" and the "ConsensusClusterPlus" R package was used. We named it “Gene-driven clusters” for this type of classification (secondary clustering). We then checked the expression status of ICB genes (*PD-1, CTLA4, PD-L2, ICOS, GITRL, LIGHT, CD27, CD28, CD40, CD80, IDO1, LAG3*) between ImmCI clusters A/B (initial clustering) and gene-driven clusters A/B (secondary clustering) to further verify the rationality of binary classification for sarcoma samples by a confirmed common tendency phenomenon.

2.6. Calculation of the ImmCI score

The ImmCI score was calculated as follows. Initially, gene-driven cluster A and B were identified by the secondary clustering as described above. Then, ImmCI-related DEGs was named based on their high or low expression in gene-driven cluster. First the "Boruta" algorithm in "Boruta" package was used for further screening of the ImmCI-related DEGs [30]. Next, we used gene-driven cluster A-related genes and gene-driven cluster B-related genes as independent variables, respectively, and make twice PCA analysis on the samples, and use the predict function to get the predicted value of each sample, which we call ImmCI score A/B. Subsequently, the ImmCI score A (corresponding to type A genes signature) and the ImmCI score B (corresponding to type B genes signature) were combined to form a prognostic score. The ImmCI score of each sarcoma sample was calculated using the following formula: *ImmCI score = ΣImmCI score A – ΣImmCI score B*. The ImmCI score was calculated for sarcoma patients, categorized into the high-ImmCI score or low-ImmCI score subgroups based on the optimal cut-off value for the survival analysis. The "Boruta" and "stats" R package, was used for the analysis described above.

2.7. Gene enrichment analysis, gene set enrichment analysis (GSEA) and tumor immune dysfunction evaluation

We annotated gene-driven cluster A/B-related genes with gene ontology by R package "clusterProfiler" [31]. To figure out the function of ImmCI score in sarcomas, GSEA was carried out [32,33]. GSEA software was used for computing and visualizing GSEA results based on different ImmCI subgroups. The GSEA procedure was repeated for 1,000 times to achieve a normalized enrichment score. The standard for statistically significant are $p < 0.05$. Take each top 5 KEGG pathways in high-ImmCI score group and low-ImmCI score group for visualization. Tumor Immune Dysfunction and Exclusion (TIDE) score, calculated for individual tumor samples, can serve as a surrogate biomarker for anticipating responses to immune checkpoint blockade [34]. We obtained the TIDE score of TCGA-SARC, TARGET-osteosarcoma and ICGC-Ewing sarcoma through the TIDE online according to tutorial instruction (http://tide.dfci.harvard.edu/). We demonstrated and visualized the differences of TIDE score between two ImmCI score groups by R-packages “ggpubr”.

2.8. Somatic mutation analysis and verification of ImmCI score based on TMB score

The “maftool” R package was employed for exploring the distribution of mutated genes, calculating the TMB of TCGA-SARC and estimating TMB status of different ImmCI score subgroups. For the definition of high and low TMB groups, we used the ’surv_cutpoint’ function of the survminer package to obtain the best cutoff of TMB for grouping in the survival analysis. After the difference analysis of sarcoma samples with TMB as the standard, the stratified test was carried out with ImmCI score to verify the potential value of ImmCI score for predicting immunotherapy response and prognosis.

2.9. Validation

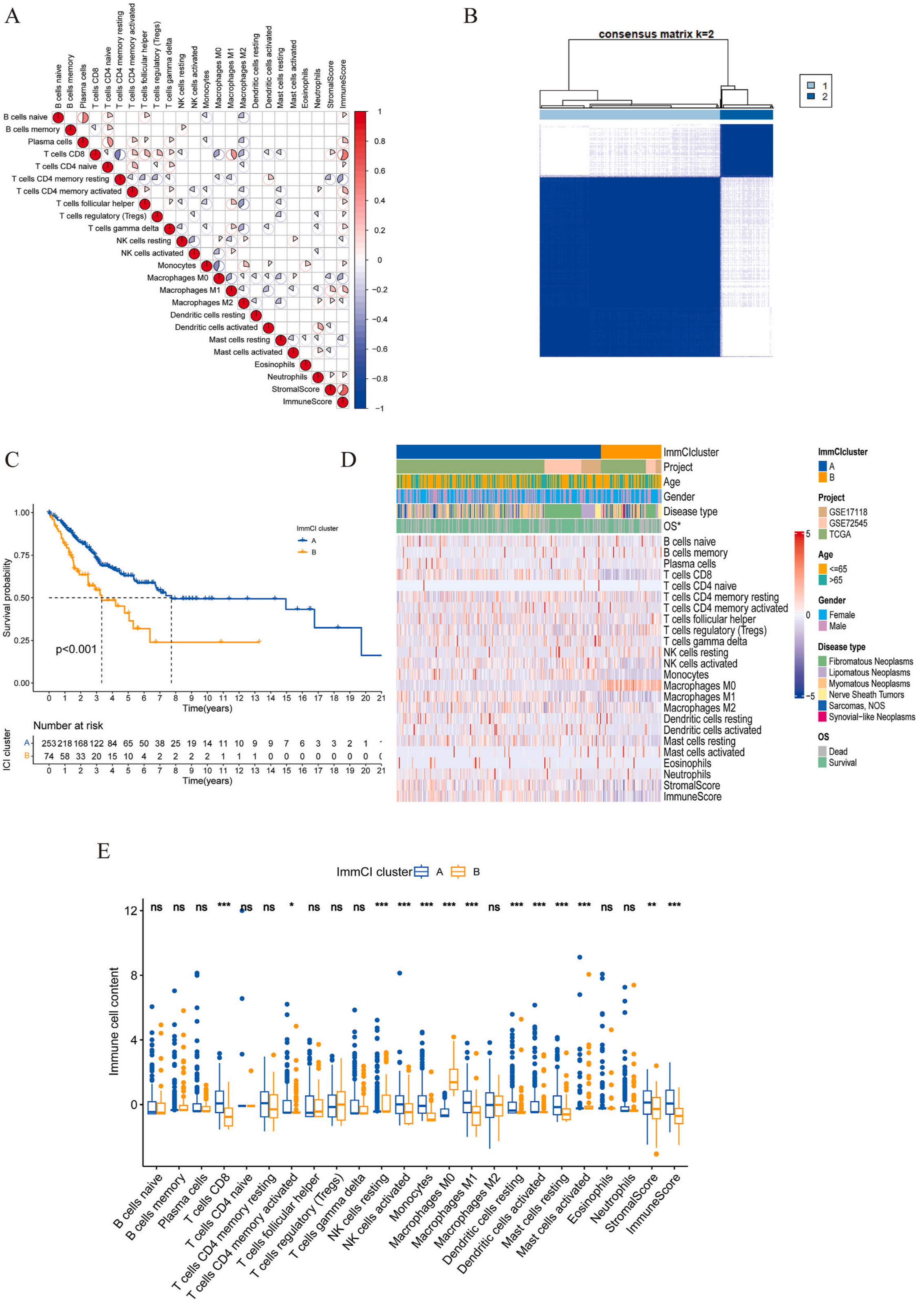
Validation cohorts were obtained from the TARGET-osteosarcoma project (Therapeutically Applicable Research to Generate Effective Treatments, TARGET), ICGC-Ewing sarcoma (International Cancer Genome Consortium, ICGC), and GSE21257. The process of downloading and organizing these validation cohorts was identical to that used for the training cohort, TARGET-osteosarcoma clinical and transcriptomic data, downloaded from UCSC Xena, included 88 samples. Dataset GSE21257 contained 53 osteosarcoma patients. The ICGC-Ewing sarcoma cohort included 108 patients. All samples in the validation cohorts met our quality inspection standards established in the previous training steps.

2.10. Immunohistochemical (IHC) staining

IHC staining was implemented with *CTLA4* antibody (1:50; CST, USA), *PD-1* (1:50; CST, USA), *CD80* antibody (1:50; huabio, China) based on the manufacturer’s protocols. Specific processes can be found in previous studies [35]. Following removal of paraffin by washing successively with xylene, 100 % ethanol, and 95 % (v/v) ethanol, slides were rinsed with deionized water and antigen retrieved in 0.01 mol/L sodium citrate pH 6.0 for 10 min. After cooling, slides were blocked in PBT 5 % (v/v) donkey serum, and an anti-PAM antibody was applied at 1:200 (v:v) in PBT, and the slide incubated 12–16 h in a humidified chamber at 4 °C. Secondary antibody was applied at 1:300 (v:v). Following incubation for 2 h, 3,3′-Diaminobenzidine (DAB) Substrate was prepared and applied for 3 min. The samples were counterstained with hematoxylin for 15 s and imaged. Determination of the degree of staining is performed by the IHC Profiler [36]. Quantitative methods for immunohistochemistry are derived from the IHC Toolbox [37].

2.11. Statistical analysis and software usage

We conducted all analyses using R software (version 4.1.1) and GraphPad Prism (version 8.0). We employed the Wilcoxon test for comparing two groups and the Kruskal-Wallis test for comparing more than two groups. To assess prognostic significance, we utilized kaplan-meier survival plots. The String database (https://cn.string-db.org/) and the cytoscape software (https://cytoscape.org) were used to perform the protein-protein interaction analysis (PPI). The STRING database provides 8 ’interaction’ methods: 1) light blue: from curated database, 2) purple: experimentally determined, 3) green: gene neighborhood, 4) red: gene fusions, 5) blue: gene co-occurrence, 6) yellow: textmining, 7) black: co-expression, 8) light purple: protein homology. The Friends analysis approach assesses the functional correlation between different genes in a pathway, suggesting that a gene is more likely to be expressed if it interacts with other genes in the same pathway, and it is widely used to identify critical genes. The mgeneSim function in the “GOSemSim” R package was utilized to calculate the relationship between a specific gene and other genes in Gene Ontology data. This calculation includes pairwise semantic similarities in categories such as ’Biological Process’, ’Molecular Function’, and ’Cellular Component’. The overall correlation between the two genes was determined by calculating the geometric mean of the pairwise semantic similarities in the three categories of BP, MF, and CC. The box plots and rain cloud plots were created using the “ggplot2”



(caption on next page)

Fig. 2. The ImmCI landscape and two independent ImmCI subtypes (initial clustering) in the training sarcoma samples. (A) The relationship among immune infiltrating characteristics (immune infiltrating cells and Stromal/Immune Score). Red color stands for positive correlation, and blue stands for negative correlation. (B) The optimal ImmCI consensus number is 2. (C) OS curve in different ImmCI subgroups of sarcomas (Kaplan–Meier, $p < 0.001$). (D) Unsupervised clustering of project, genders, ages, ImmCI cluster, and survival status in sarcoma samples. The abscissa represents independent samples, and the ordinate represents independent immune infiltrating characteristics. (E) The degrees of immune cell infiltration and Stromal/Immune Scores in different ImmCI clusters. $*p < 0.05$; $**p < 0.01$; and $***p < 0.001$. ImmCI, immune cell infiltration; OS, overall survival.

package. The “corrplot” package was used to draw correlation plot. In all our analyses, statistical significance was defined as a two-tailed p-value <0.05.

3. Results

3.1. Tumor microenvironment landscape and immune cell infiltration in sarcomas

The datasets used in our initial study consisted of three independent cohorts: GSE72545, GSE17118, and TCGA-SARC (training cohort). The proportion of 22 immune cells in the tumor microenvironment of sarcomas was quantified using the CIBERSORT algorithm. After removing samples with CIBERSORT returned p-value greater than 0.05, 330 patients were included in subsequent analyses. The findings indicated that the predominant components within the sarcoma TME were M0, M1, and M2 macrophages, along with activated mast cells, regulatory T cells (Tregs), and $CD8^{+}$ T cells. Thus, these immune cell types became the focus of our research in sarcomas. The relationship among immune infiltration cells in the landscape is displayed in Fig. 2A. The proportion of $CD8^{+}$ T cells was found to be negatively related to M0 and M2 macrophages but positively associated with M1 macrophages. The correlation between M0 macrophages and M1/M2 macrophages and resting mast cells was negative. Moreover, the “StromalScore” was negatively correlated with M0 macrophages and activated mast cells while positively correlated with $CD8^{+}$ T cells and M1 macrophages. At the same time, the “ImmuneScore” was negatively correlated with M0/M2 macrophages and resting mast cells but positively associated with $CD8^{+}$ T cells and M1 macrophages. These results reveal the broad interactions and functional associations of different immune cells. In order to detect patterns of ImmCI within the TME, we employed the "ConsensusClusterPlus" R package to ascertain the most suitable number of clusters. Despite $k = 3$ clusters returns the most consistent results according to Figure S1 H-I, Figure S1 B shows a higher inter-subtype correlation and a relatively low within-subtype correlation for $k = 3$. Combining the above results and results of the K-M curves, we chose the Consensus Clustering result for $k = 2$. Subsequently, the combined cohort was divided into two distinct subgroups (initial clustering): ImmCI clusters A and B (Fig. 2B, Fig. S1). ImmCI cluster A With 255 samples and ImmCI cluster B with 75 samples. After excluding samples with missing survival information or a survival time of 0, Kaplan-Meier survival analysis of the two ImmCI subgroups revealed that ImmCI cluster B had a poorer prognosis compared to the ImmCI A group, which showed better survival outcomes ($p < 0.001$) (Fig. 2C). A heatmap was generated to illustrate the differences among ImmCI subtypes (initial clustering) and visualize the intrinsic variations among immune cells in the TME (Fig. 2D). Cluster A was correlated with high infiltration levels of $CD8^{+}$ T cells, active NK cells, and M1 macrophages, whereas cluster B was correlated with high infiltration of M0 macrophages. Moreover, cluster A was associated with higher “StromalScore” and “ImmuneScore” (Fig. 2E).

To delve deeper into the potential for immunotherapy in sarcomas, we assessed the differences in ICB gene expression between the two ImmCI subgroups, including *PD-1, CTLA4, PD-L2, ICOS, GITRL, LIGHT, CD27, CD28, CD40, CD80, IDO1, LAG3.* (A detailed explanation of gene name abbreviations is provided in the list of abbreviations, which also applies to abbreviations used elsewhere in the text.) The results showed that ICB-related genes exhibited significantly lower expression levels in ImmCI cluster B and higher expression levels in ImmCI cluster A ($P < 0.001$), substantiating the distinct effect of ICB treatment (Fig. 3A–L).

3.2. Identification of ImmCI-related gene-driven clusters

In order to examine potential variations in gene expression related to immune cell infiltration across ImmCI subtypes (initial clustering), we utilized the R package "limma," which identified 157 genes as being differentially expressed (DEGs) (Table S1). In order to enhance objectivity and increase the sample size, additional samples that were initially excluded in the first cluster analysis due to a p-value greater than 0.05 from the CIBERSORT algorithm were incorporated. Using the same unsupervised clustering method as for ImmCI clusters, we categorized these cohorts into two distinct gene-driven clusters (secondary clustering): gene-driven cluster A, comprising 178 samples, and gene-driven cluster B, comprising 185 samples. (Fig. 4A, Fig. S2). The DEGs were categorized based on their levels of expression. A total of 58 genes, referred to as type A genes, showed up-regulated expression in gene-driven cluster A, while 99 genes, referred to as type B genes, showed up-regulated expression in gene-driven cluster B. The clinical phenotypes, ImmCI clusters, and gene-driven cluster signatures among independent gene types A/B are displayed in Fig. 4B. A noticeable distinction emerged between gene-driven clusters A and B and gene types A and B. Within gene type A, gene-driven cluster A demonstrated lower expression levels, whereas gene-driven cluster B displayed higher expression levels. Conversely, in gene type B, gene expression was higher in gene-driven cluster A compared to gene-driven cluster B. Prognostic analysis demonstrated better survival rates associated with gene-driven cluster A, while gene-driven cluster B was linked to poorer survival outcomes ($p < 0.001$) (Fig. 4C). Fig. 4D illustrates significant differences in immune infiltrating characteristics, “StromalScore”, and “ImmuneScore” between the two gene-driven clusters. Gene-driven cluster A was characterized by elevated infiltration of $CD8^{+}$ T cells, M1 macrophages, and monocytes, as well as higher levels of “StromalScore” and “ImmuneScore”.

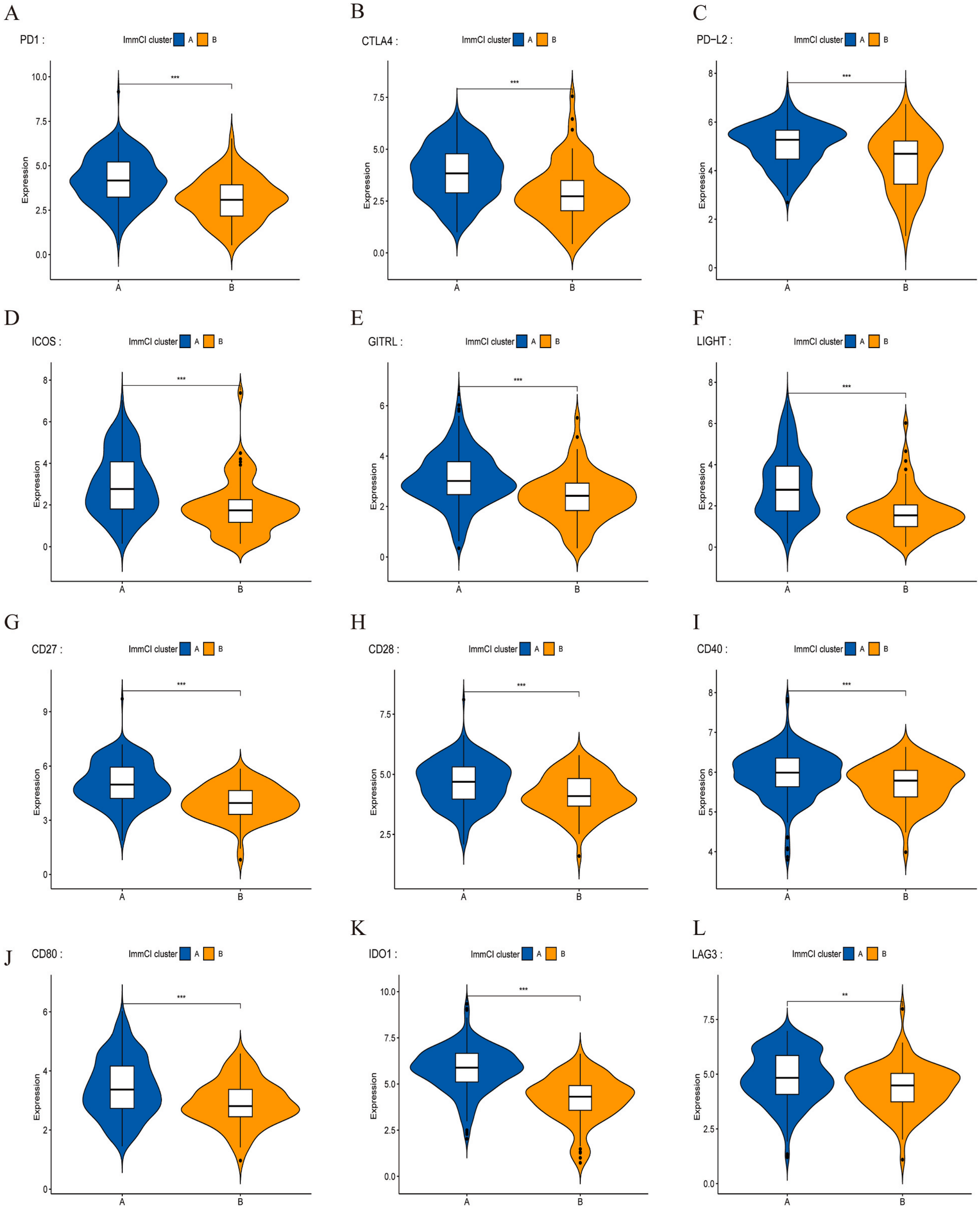
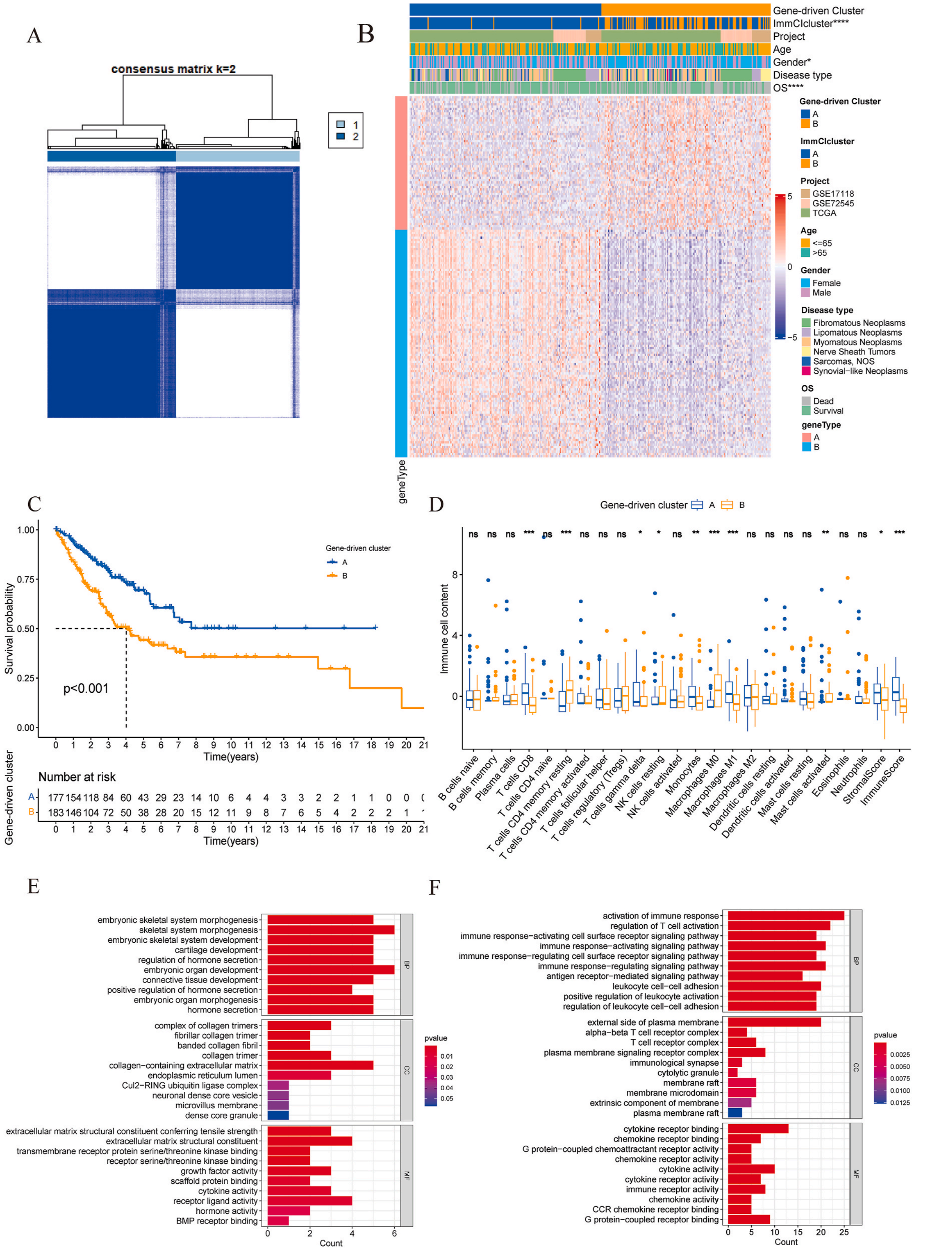


Fig. 3. Immune checkpoint blockade-related gene expression status in different ImmCI clusters of sarcomas (initial clustering): *PD-1*(A), *CTLA4*(B), *PDL2*(C), *ICOS*(D), *GITRL*(E), *LIGHT*(F), *CD27*(G), *CD28*(H), *CD40*(I), *CD80*(J), *IDO1*(K), *LAG3*(L). There is a significant difference between the two independent ImmCI subtypes. ImmCI cluster A has significantly higher gene expression levels in all twelve independent ICB genes. *$p < 0.05$; **$p < 0.01$; and ***$p < 0.001$. ICB, immune checkpoint blockade; ImmCI, immune cell infiltration.



(caption on next page)

Fig. 4. Establishment of ImmCI-related gene subtypes (secondary clustering) and enrichment analysis. (A) According to the unsupervised clustering results obtained for ImmCI-related differentially expressed genes (DEGs), the samples were divided into two independent gene-driven clusters. (B) Unsupervised clustering of project, gender, age, ImmCI-related gene-driven cluster, and survival status in sarcomas. (C) OS curve in different ImmCI-related gene subgroups of sarcomas. (D) The degree of immune cell infiltration and "StromalScore" and "Immune Score" in different ImmCI gene-driven clusters. (E, F) Enrichment analysis of type A genes (E) and type B genes (F). The ordinate represents the Gene Ontology (GO) term, the abscissa represents the number of enriched genes, and the color represents the significance of the correlation, where red indicates positive correlation and blue indicates negative correlation. ImmCI, immune cell infiltration; OS, overall survival.

On the purpose of locating the possible molecular functions of type A/B genes, R package "clusterProfiler" was applied for GO analysis. Type A genes mostly involved in the development of skeletal system, such as "embryonic skeletal system morphogenesis" and "cartilage development". Meanwhile, "collagen-containing extracellular matrix" and "complex of collagen trimers" were enriched in cellular component. In molecular function, "extracellular matrix structural constituent" and "cytokine activity" were significantly enriched (Fig. 4E). The tumor cells within gene-driven cluster A exhibited a closer resemblance to the original immature cells and demonstrate a heightened level of malignancy. Furthermore, the secretion of extracellular matrix by these cells might play a crucial role in tumor angiogenesis, growth, invasion, and the facilitation of an immunosuppressive microenvironment. In type B genes, results showed they have a close relationship with "activation of immune response" and "regulation of T cell activation", and realized the cellular component of "T cell receptor complex", at the same time molecular function of "immune receptor activity" was significantly enriched (Fig. 4F).

Furthermore, we analyzed the expression status of ICB genes within the gene-driven clusters to confirm the concordance between ImmCI classification (initial clustering) and gene profiling (secondary clustering). Similar to ImmCI typing results, ICB genes were upregulated in gene-driven cluster A and downregulated in gene-driven cluster B (Fig. 5A–L). These findings suggest that sarcoma patients with gene-driven cluster A might respond better to ICB, while those with gene-driven cluster B exhibit a low response rate. The congruence of ICB gene expression across various unsupervised cluster classifications substantiated the validity of ImmCI categorization in sarcomas.

3.3. Computation of ImmCI scores and prognostic signature

As described in the Methods section, the "Boruta" algorithm and PCA algorithm were used to obtain the ImmCI score for each individual. The Boruta algorithm was employed to conduct a more refined screening of genes closely associated with the ImmCI cluster. Ultimately, 111 genes were incorporated into the calculation of ImmCI Scores. The importance score of each gene was presented in Table S2. The datasets were subsequently divided into subcategories based on the ImmCI score, leading to the creation of a high-ImmCI score subgroup (comprising 290 samples) and a low-ImmCI score subgroup (consisting of 70 samples). A Sankey diagram was employed to illustrate the connections between gene-driven clusters, ImmCI score subcategories, and survival status.

The majority of samples in gene-driven cluster A and surviving patients were classified under the high-ImmCI score subgroup (H-ImmCI). Compared with the L-ImmCI, the H-ImmCI exhibited better survival outcomes ($P < 0.001$). Specifically, 65 percent of patients in the H-ImmCI survived, compared to 47 percent in the L-ImmCI (Fig. 6A). Furthermore, it was demonstrated that patients in the L-ImmCI had worse prognostic outcomes (Fig. 6B). The correlation was estimated using immunostimulatory, chemokine gene, and ICB gene signatures (combined gene set: *IDO1,CD274,HAVCR2,PDCD1,CTLA4,LAG3,CD8A,CXCL10,CXCL9,GZMA, GZMB,PRF1,IFNG, TBX2,TNF,CD27,CD28,CD40,CD80,CTLA4,TNFSF18,ICOS,IDO1,LAG3,TNFSF14,PDCD1,PDCD1LG2,HPD,CD276*), where the H-ImmCI exhibited higher expression levels for nearly all immune checkpoint signatures and chemokines, except for *TBX2* (Fig. 6C). Multi-GSEA analyses were carried out based on ImmCI score gene signature data to achieve the different functional pathway analysis. The results showed that immune related pathways just like antigen processing and presentation, T cell receptor signaling pathway, leukocyte transendothelial migration, natural killer cell mediated cytotoxicity, B cell receptor signaling pathway and T cell receptor signaling pathway were enriched in the High-ImmCIs. The results were highly focused on the immune functions of different immune cells, suggesting that the anti-tumor immune response in the High-ImmCIscore group was more active, leading to better prognostic outcomes. Meanwhile, nucleotide excision repair and basal transcription factors pathways were enriched in Low-ImmCIs (Fig. 6D). It was suggested that tumor cells in the Low-ImmCIscore group may enhance their resistance to the body's immune response and treatment by activating different DNA repair pathways and altering metabolic processes. The TIDE score for the H-ImmCI was notably lower compared to the L-ImmCI ($P < 0.001$), indicating a better immune response efficacy to ICB treatment in the H-ImmCI and a stronger immune exclusion and escape ability in the L-ImmCI (Fig. 6E).

3.4. Relationship between ImmCI score and TMB

We observed that the high-TMB subgroup had a lower survival rate than the low-TMB subgroup (Fig. 7A). The stratified analysis revealed that patients with high TMB and low ImmCI scores had the worst survival outcome among the four subgroups (Fig. 7B). Similarly, when examining patients with the same TMB status but different ImmCI scores, the H-ImmCIs demonstrated higher survival rates independent of TMB's influence (Fig. 7B). Further correlation analysis between the ImmCI scores and TMB levels was displayed in Supplementary Fig. S3. Overall, there was no significant interaction between the ImmCI score and TMB results. In addition, the R package "maftools" was applied to explore the distribution of mutational gene in these two ImmCI score subgroups. The top 20 highest mutational genes including *TP53, ATRX, MUC16, TTN, RB1* were illustrated in Fig. 7C and D. The top 20 highest mutational genes in both the high TMB group and the low TMB group were identical.

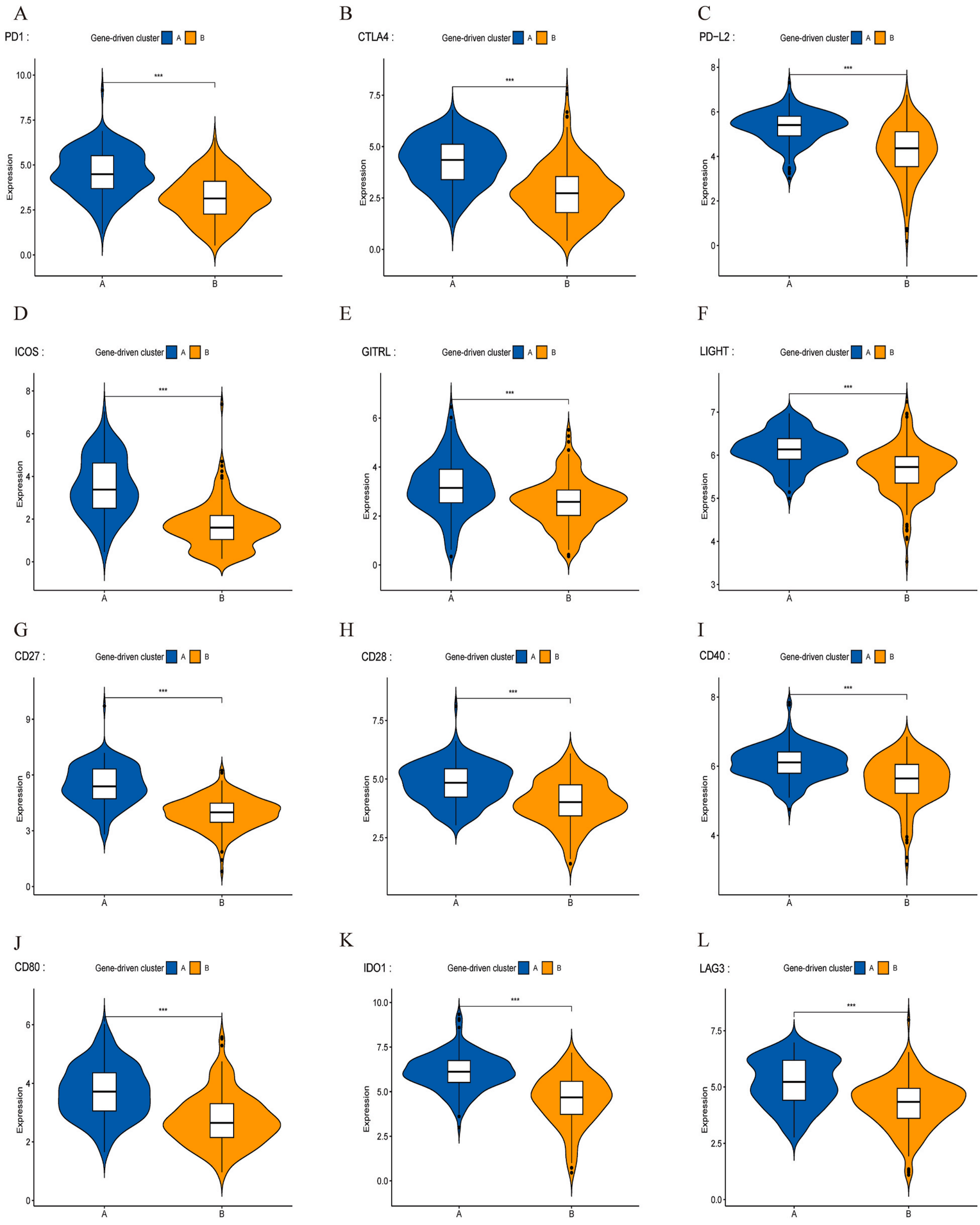


Fig. 5. Immune checkpoint blockade-related gene expression differences between ImmCI-related gene-driven clusters (secondary clustering) were consistent with those in ImmCI subtypes (initial clustering): *PD-1*(A), *CTLA4*(B), *PDL2*(C), *ICOS*(D), *GITRL*(E), *LIGHT*(F), *CD27*(G), *CD28*(H), *CD40* (I), *CD80*(J), *IDO1*(K), *LAG3*(L). *p < 0.05; **p < 0.01; and ***p < 0.001. ImmCI, immune cell infiltration.

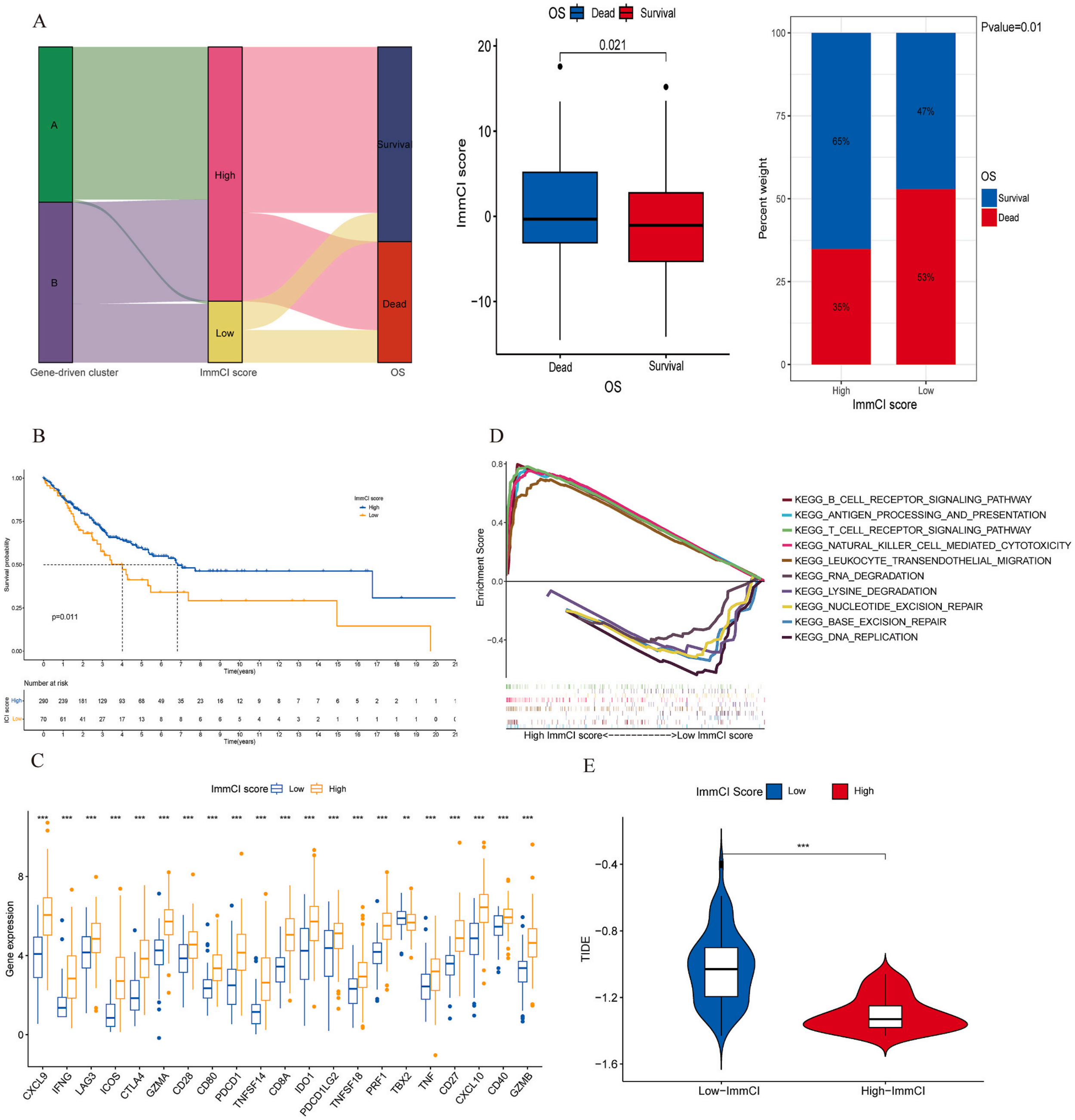


Fig. 6. Establish and verify ImmCI scores. (A) Sankey diagram showing the connection among gene-driven clusters, ImmCI high or low score groups, and prognosis endpoint. ImmCI scores distribution for different survival statuses. ($p = 0.021$). (B) OS curve in different ImmCI score subgroups. (Kaplan-Meier, $p = 0.011$). (C) The consistent expression tendency of immunostimulatory, chemokine, and ICB genes between high and low ImmCI score subgroups. (D) Multi-GSEA results between high and low ImmCI score subgroups. (E) Tumor immune dysfunction evaluation (TIDE) score between ImmCI score groups. *$p < 0.05$; **$p < 0.01$; and ***$p < 0.001$. ICB, immune checkpoint blockade; GSEA, gene set enrichment analysis; TIDE, tumor immune dysfunction evaluation.

3.5. Validation

To validate the ImmCI score in representing survival outcomes and ICB therapy response in sarcomas, which are known for their heterogeneity, we performed a re-analysis focusing on the six major subtypes of sarcoma in internal cohort. Additionally, to examine the reliability of the analysis results that may arise from the internal cohort, we selected two independent external sarcoma cohorts: the osteosarcoma cohort (TARGET-osteosarcoma and GSE21257) and the Ewing sarcoma cohort (ICGC-ES) for simultaneous verification.

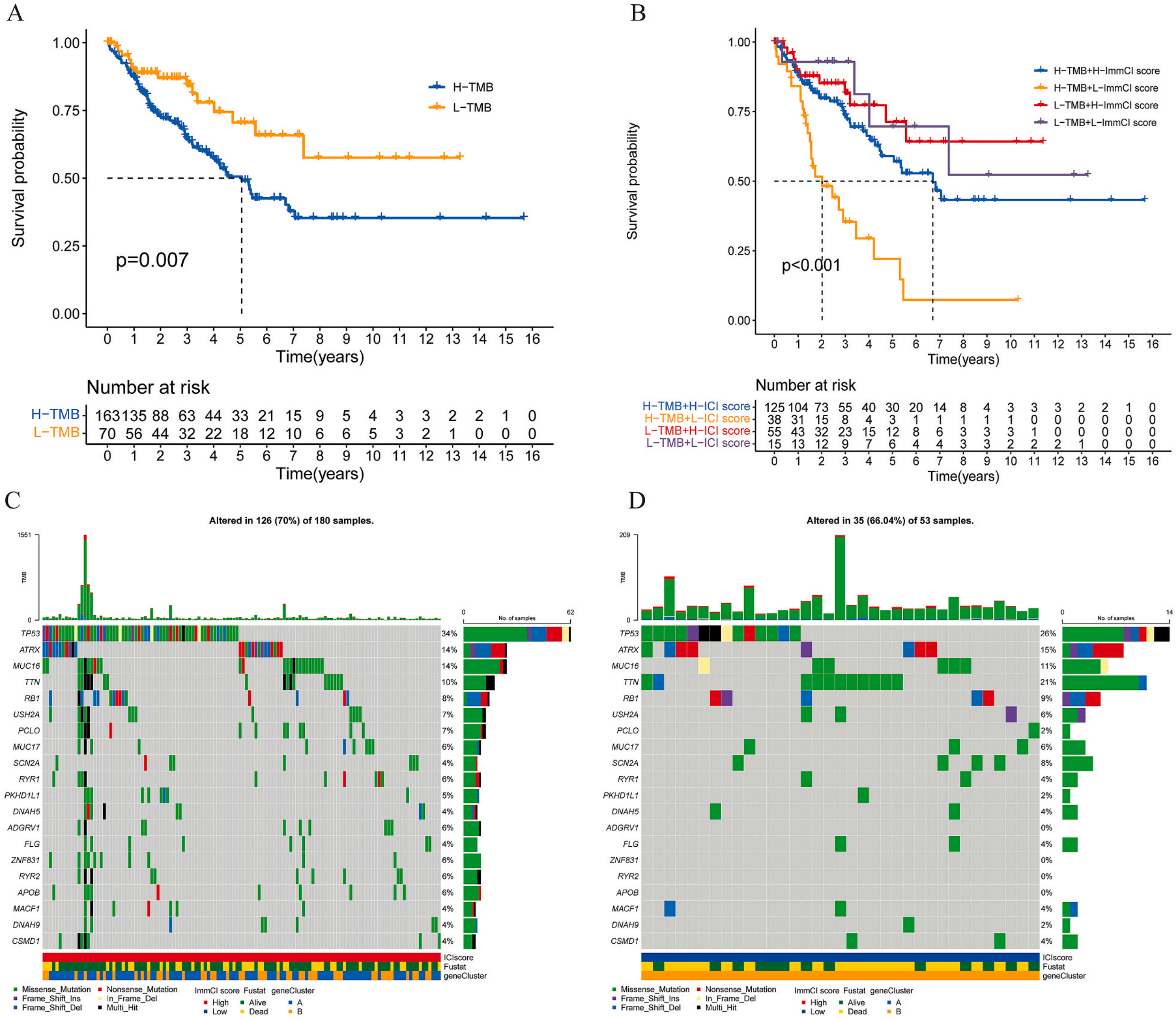


Fig. 7. Stratification according to the TMB score was conducted to verify the feasibility of the ImmCI score and mutation analysis. (A) OS curve for the high- and low-TMB score groups. (Kaplan–Meier, $p = 0.007$). (B) OS curve for samples stratified by TMB score and ImmCI score. (Kaplan–Meier, $p < 0.001$). (C, D) Top 20 mutation genes in high-ImmCI score and low-ImmCI score. Rows represent gene names, and columns represent patients. TMB, tumor mutation burden; ImmCI, immune cell infiltration; OS, overall survival.

These eight validation cohorts yielded consistent results with the training cohorts. Fig. 8 shows that the prognosis of the H-ImmCI was elevated in all eight cohorts ($p < 0.05$). TIDE analysis also yielded consistent results. We confirmed the predictive utility of the ImmCI score for immunotherapy in all validation sarcoma patients except for nerve sheath tumors: the H-ImmCI was associated with a lower ability of immune escape and a better response to ICB treatment, similar to the training cohort (Fig. 9). The expression differences of ICB genes between high/low-ImmCIs also supported this finding. Almost all ICB genes and chemokines displayed high expression in the H-ImmCI, including *GZMB, CD40, IFNG, CD274, CD80, HAVCR2, CXCL9, GZMA, CD8A, CTLA4, ICOS, CXCL10, CD28* (Fig. 10). These findings provided compelling evidence that the ImmCI score exhibited a good correlation with the immunotherapy response. In sarcoma samples originating from the same tissue, various pathological diagnosis types can be present. To ensure consistency, we focused on the most prevalent pathological diagnosis type in each tissue for stratified analysis. This included leiomyosarcoma, dedifferentiated liposarcoma, fibromyxosarcoma, and undifferentiated sarcoma. Other pathological types were not included in the stratified analysis due to limited sample size. As shown in Fig. 11, the analysis of intra-subtype heterogeneity yielded the same conclusions that have been previously drawn: ImmCI scores reflect prognosis and immunotherapy (Fig. 11).

3.6. Experimental verification

To identify the pivotal biomarkers, Friends analysis and Protein-Protein-Interaction (PPI) analysis were employed [38,39]. Friends

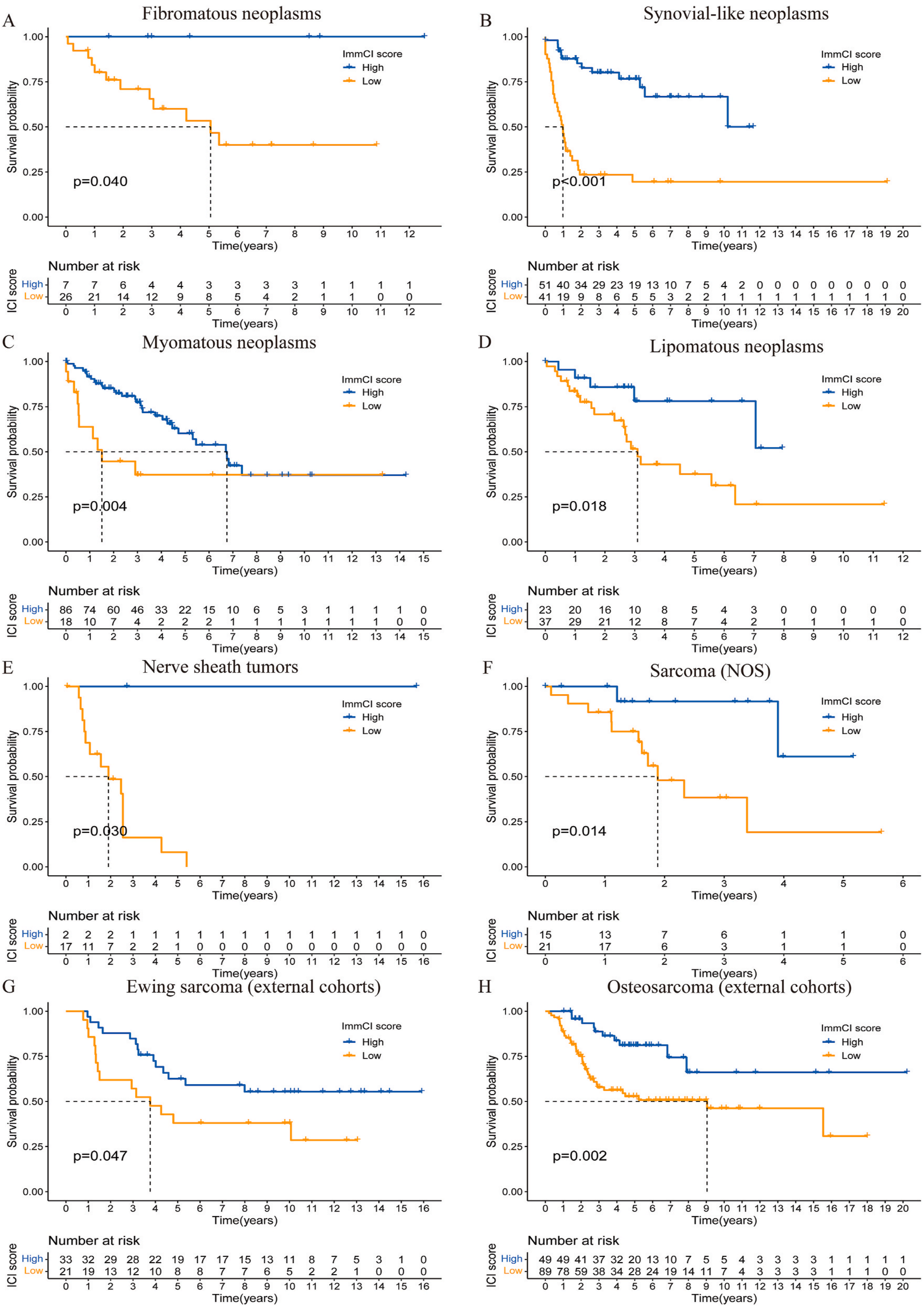
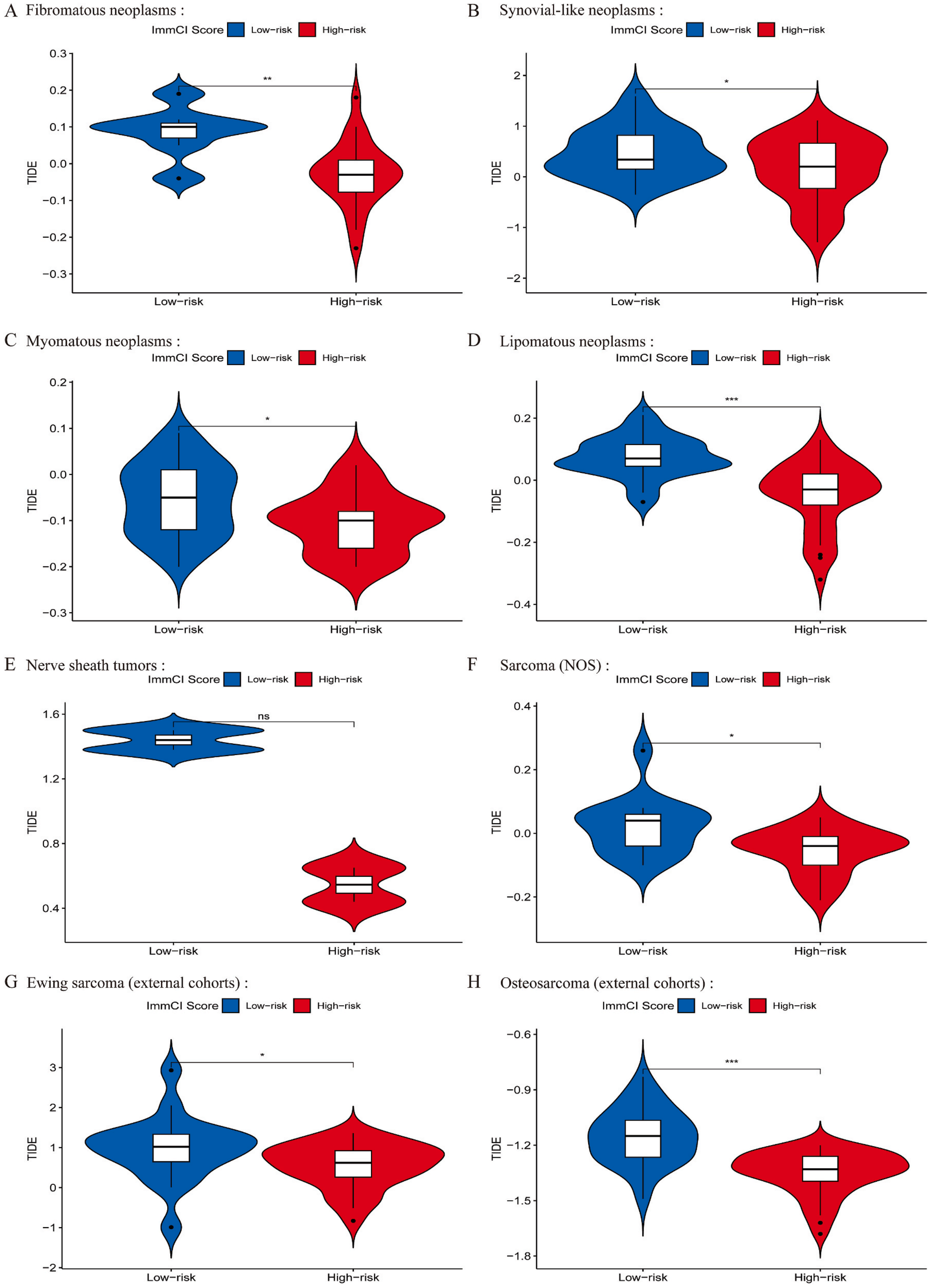


Fig. 8. OS curve in different ImmCI score subgroups in eight sarcoma subtypes (including internal and external cohorts). (A). Fibromatous neoplasms. (B). Synovial-like neoplasms. (C). Myomatous neoplasms. (D). Lipomatous neoplasms. (E). Nerve sheath tumors. (F). Sarcoma (NOS). (G). Ewing sarcoma (external cohorts). (H). Osteosarcoma (external cohorts). OS, overall survival.



(caption on next page)

Fig. 9. TIDE score between ImmCI high score subgroup and ImmCI low score subgroup in eight sarcoma subtypes (including internal and external cohorts). (A). Fibromatous neoplasms. (B). Synovial-like neoplasms. (C). Myomatous neoplasms. (D). Lipomatous neoplasms. (E). Nerve sheath tumors. (F). Sarcoma (NOS). (G). Ewing sarcoma (external cohorts). (H). Osteosarcoma (external cohorts). TIDE, tumor immune dysfunction evaluation. $*p < 0.05$; $**p < 0.01$; and $***p < 0.001$.

analysis demonstrated that *CTLA4* was the most important biomarker among the twelve biomarkers, and the importance was ranked in descending order, as shown in Fig. 12A and B. Based on the biological functions of the biomarkers themselves, we grouped these twelve markers into four categories: *a. PDCD1, PDCD1LG2; b. CTLA4, CD28, CD80, ICOS; c. TNFSF18, CD27, CD40, TNFSF14; d. IDO1, LAG3*. Combining the Friends analysis results and their biological functions, we selected representative biomarkers from these four categories for experimental validation, including *CTLA4, PD-1 (PDCD1)*, and *CD80*. The PPI analysis revealed that 12 biomarkers exhibited strong interaction relationships within the interaction network (Fig. 12 C). To test our conclusions generated from big data bioinformatics analysis in another dimension, we conducted experiments to verify the results. The stacked column chart indicated that ImmCI cluster A and gene-driven cluster A were associated with better survival outcomes (Fig. 12D and E). We collected clinical information and pathology slides from 12 sarcoma patients and performed immunohistochemistry to validate the biomarkers from ImmCI cluster A and gene-driven cluster A. Based on this indicator, we divided the patients into high or low survival based on mean survival time (Table 1). We conducted immunohistochemical staining experiments for further validation. Fig. 13 displays typical IHC images of CTLA4, PD-1, and CD80 in two patients from the high-survival and low-survival groups, respectively. IHC statistical results demonstrated that the selected biomarkers were expressed relatively higher in the survival-high groups than in the survival-low and immunotherapy-low groups (Fig. 12 F).

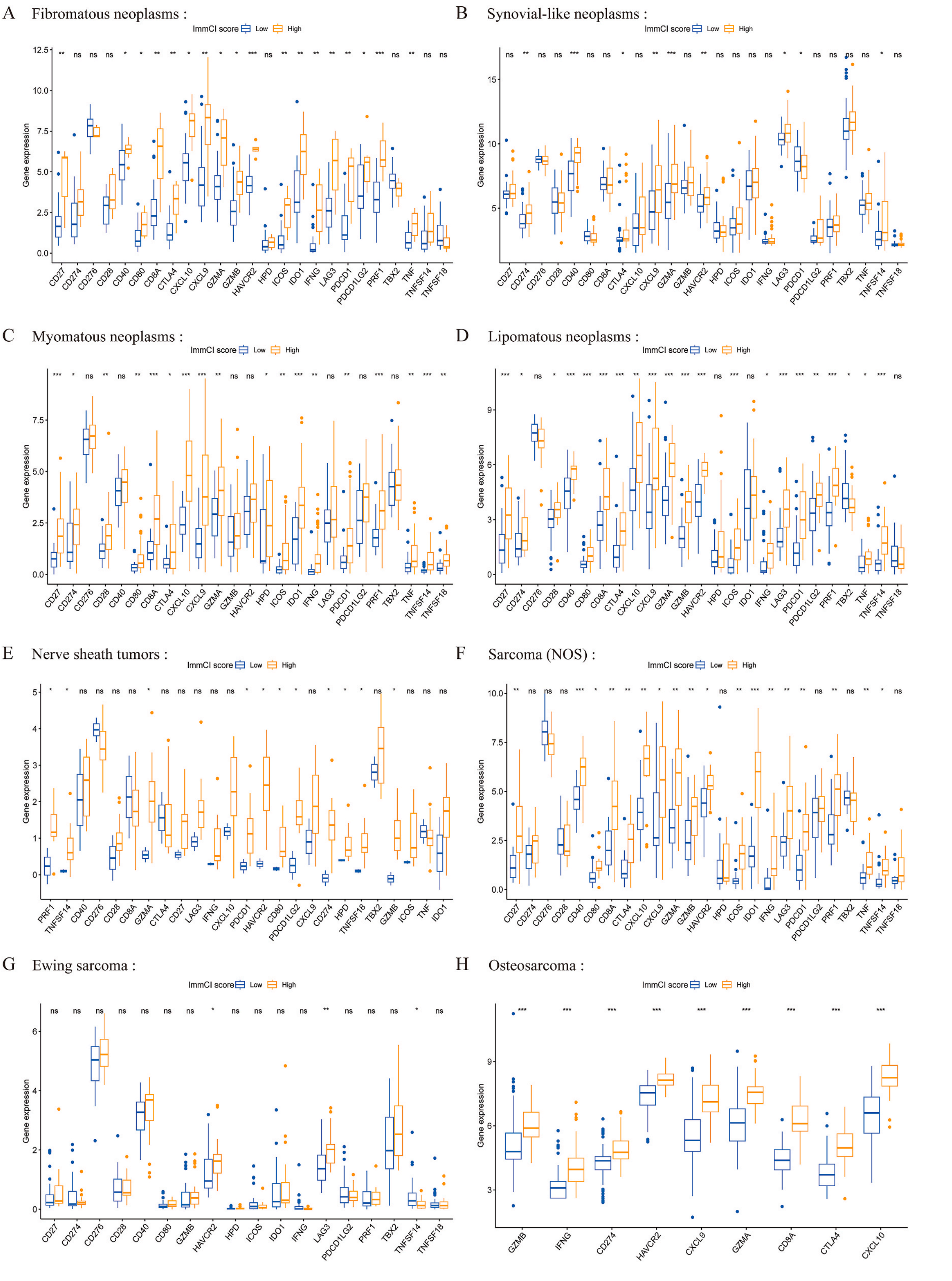
4. Discussion

In this fast-developing biomedical era, we have witnessed epoch-making approvals from the Food and Drug Administration (FDA) and European Medicines Agency (EMA) for ipilimumab against metastatic melanoma [40]. Other clinical trials have also yielded encouraging results, including nivolumab (anti-PD1) plus ipilimumab against advanced melanoma, lung cancer, metastatic esophagogastric cancer, and so on [41–43]. Nevertheless, due to the strong heterogeneity of sarcomas, identifying which patients are sensitive to immunotherapy in sarcomas is challenging [8,44,45]. Therefore, it is essential to establish a signature that could predict the efficacy of immunotherapy. Recent research has indicated that the polygenic immune score signature could be harnessed as a reliable prognostic marker in sarcomas [46,47]. Nonetheless, few studies have focused on the ImmCI characteristics of sarcoma patients. Our research addresses this gap and sheds light on the immunotherapy response of sarcomas, especially from the perspective of autoimmunity.

Indeed, a deeper understanding of ImmCI function within the TME is vital to identify patients sensitive to immunotherapy. It has been established that patients with a high degree of immune infiltration experience better therapeutic efficacy [48]. The intrinsic mechanism includes activating T cells, then impelling T cells to execute tumor cells [6,49]. We developed an approach to characterize sarcoma samples with ImmCI patterns and categorized them into four consistent clusters: ImmCI clusters A/B and gene-driven clusters A/B. These two classification methods exhibit consistent patterns in prognostic outcomes (same advantage in ImmCI/gene-driven cluster A), immune characteristics (same characterization of TME in ImmCI/gene-driven cluster A), and ICB treatment response (same advantage in ImmCI/gene-driven cluster A), which further validated the rationality of our binary classification.

The researcher Galon proposed for the first time an immune-based classification of tumors: ’hot’ (highly infiltrated) and ’cold’ (non-infiltrated) tumors [50,51]. Interestingly, prognostic outcomes and immune characteristics of the clusters A/B (initial and secondary clustering) based on bioinformatics data seem to have certain connections with the biological classification "hot/cold tumor". Hot tumors are distinguished by a substantial presence of immune cells within the stroma, along with elevated immune and stromal scores and a high mutation burden, which makes checkpoint inhibitors effective, leading to a potent antitumor effect [51,52]. In our study, both ImmCI cluster A and gene-driven cluster A consistently exhibited higher levels of $CD8^+$ T cell infiltration, as well as elevated “ImmuneScore” and “StromalScore”. Furthermore, the response to ICB inhibitors demonstrated a consistent tendency, similar to that of hot tumors known for their favorable response to immunotherapy. Moreover, the majority of gene-driven cluster A samples corresponded to the H-ImmCI, which showed a higher mutation load and a more favorable immunotherapy response. These findings suggest that sarcoma patients classified in cluster A may potentially experience a more effective immune treatment response, similar to the behavior observed in hot tumors that exhibit a strong response to immunotherapy. Cold tumors are extremely difficult to mobilize an immune response due to the few T cells in their microenvironment. In our research, ImmCI cluster B and gene-driven cluster B exhibited decreased infiltration of $CD8^+$ T cells and low Immune/Stromal Score. The differences in immunotherapy response and survival outcome were consistent with the biological behaviors of cold tumors. The alignment between our ImmCI clusters and the "hot or cold" immune phenotypes further reinforces the rationale behind our typing methodology, where ImmCI cluster A represents a hot tumor subtype and ImmCI cluster B corresponds to a cold tumor subtype [Fig. 2D–F; Fig. 3].

In the present study, we could effectively identify patients sensitive to immunotherapy based on the ImmCI scores consistent with the "cold" and "hot" tumor theory. Specifically, patients with high ImmCI scores ("hot") experienced a better immunotherapy response and a superior survival outcome. Besides, we found that the ImmCI score was a significant independent prognostic signature in sarcoma. To validate the reliability and independence of our ImmCI score prediction model, we used TMB, a well-established independent biomarker for tumor characterization, as a standard in stratified analysis. The results confirmed the feasibility and independence of our ImmCI score in predicting the prognosis and immunotherapy response of sarcoma patients. Furthermore, we validated the accuracy



A Fibromatous neoplasms :

B Synovial-like neoplasms :

C Myomatous neoplasms :

D Lipomatous neoplasms :

E Nerve sheath tumors :

F Sarcoma (NOS) :

G Ewing sarcoma :

H Osteosarcoma :

(caption on next page)

Fig. 10. The consistent expression of immunostimulatory, chemokine, and ICB genes between high and low ImmCI score subgroups in eight sarcoma subtypes (including internal and external cohorts). (A). Fibromatous neoplasms. (B). Synovial-like neoplasms. (C). Myomatous neoplasms. (D). Lipomatous neoplasms. (E). Nerve sheath tumors. (F). Sarcoma (NOS). (G). Ewing sarcoma. (H). Osteosarcoma. *$p < 0.05$; **$p < 0.01$; and ***$p < 0.001$. ICB, immune checkpoint blockade; ImmCI, immune cell infiltration. Due to the differences in samples and detection platforms in different data sets, the expression level of some genes may not be well detected or retained in all the data sets.

and efficacy of the ImmCI score in external cohorts. The results consistently showed that a high ImmCI score is associated with a better prognosis and a more favorable immunotherapy response. Our multivariate GSEA analysis revealed that the top 5 enriched pathways in H-ImmCIs are immune-related pathways. The activation of these pathways provides a biological basis for the superior immunotherapy response observed in patients with high ImmCI scores.

One of the primary challenges in current sarcoma research is disease heterogeneity. Our analysis was conducted on a series of heterogeneous sarcoma cases, encompassing six subtypes of sarcoma: fibromatous neoplasms, lipomatous neoplasms, myomatous neoplasms, nerve sheath tumors, soft tissue tumors, and sarcomas, NOS, as well as synovial-like neoplasms. This heterogeneity could potentially impact the reliability of our analysis. To address this concern, we performed individual analyses for each subtype and found that the conclusions remained consistent with those drawn from the overall sarcoma analysis. In addition, we sought to validate our findings by using osteosarcoma and Ewing's sarcoma, two sarcomas derived from external databases (TARGET and ICGC), in addition to TCGA database. The results from these external cohorts aligned with our initial conclusions, further supporting the reliability of our analysis. In addition to heterogeneity between subgroups, there is also heterogeneity within subgroups, i.e., individual variability of patients between sarcomas of the same subtype. To address this, we conducted further analysis within specific subtypes. For instance, our conclusions regarding ImmCI scores and their association with prognosis remained consistent in leiomyosarcoma, dedifferentiated liposarcoma, fibromyxosarcoma and undifferentiated sarcoma. We have reason to believe that these findings are not coincidental, as our ImmCI score employs a secondary clustering method with a dual assessment of ImmCI and Gene classification (initial and secondary clustering), which passed the consistency check. Experimental validation further confirmed the validity of this classification. The ImmCI score underwent a scientifically rigorous dimensionality reduction process, wherein the secondary clustering and dimensionality reduction helped mitigate biases and reduce the likelihood of chance-driven data analysis. This approach is particularly well-suited for handling heterogeneous clusters, showcasing the power of machine learning in uncovering underlying patterns and relationships that might be overlooked or underestimated by human analysis. However, it is essential to acknowledge the limitations of our analysis concerning heterogeneity. Since each patient has unique traits, a classification based solely on the histology of the disease might not encompass all variations [53].

To enhance the robustness of our findings, a cohort was obtained comprising patients that attended our center. Owing to the heterogeneity of sarcomas, we collected samples from eight different sarcoma subtypes for experimental validation, including fibrosarcoma, synovial sarcoma, liposarcoma, rhabdomyosarcoma, nerve sheath tumor, osteosarcoma, Ewing sarcoma, and chondrosarcoma. Patients were included for each subtype, totaling 12 patients in the clinical cohort. To validate our *in-silico* findings, we performed IHC to validate the biomarkers distinguishing ImmCI cluster A/B and gene-driven cluster A/B. Based on the corresponding clinical information, samples were divided into high and low-survival groups. However, a limitation of our study is the lack of whole gene sequencing on each sample. Given that both ImmCI/A and gene/A were strongly associated with high survival, biomarkers for ImmCI/A and gene/A should exhibit comparable significance in the high survival group. Among the twelve biomarkers, *PD1* and *PDL2* are part of the *PD1-PDL1-PDL2* system; *CTLA4, CD28, CD80,* and *ICOS* are part of the *CTLA4-CD80* system; *GITRL, CD27, CD40*, and *LIGHT* belong to the *TNF* superfamily; and *IDO1* and *LAG3* are categorized together. Based on the results of the Friends analysis, we selected *PD1, CTLA4, CD80* as the representative biomarkers. 12 patients with 3 pending-validation indications comprise our experimental verification matrix. The results revealed that these representative biomarkers exhibited higher expression levels in the high survival group. This finding aligns with the results obtained from our bioinformatics analysis. Importantly, this represents a biologically significant outcome. On the one hand, the high expression of immune checkpoint genes indicates that these patients likely have hot tumors, leading to a better prognosis and improved response to immunotherapy. On the other hand, all these immune checkpoints have established inhibitors or activators, some of which have received FDA drug approval. Consequently, it highlights the importance of considering immunotherapy for sarcoma patients. Due to the heterogeneity of sarcomas, some patients may be in an immune activation state, making immunotherapy clinically worthwhile. Even though, a larger sample cohort would be necessary for the validation of our findings. Expanding the cohort size would enable us to further corroborate our conclusions and better account for the individual variability within the sarcoma subtypes.

5. Conclusion

To sum up, our research portrayed the ImmCI landscape and emphasized the clinical significance of ImmCI scores reflecting prognosis in sarcomas. Importantly, we established an independent prediction model based on ImmCI, which has potential clinical significance in screening sarcoma patients sensitive to immunotherapy.

Ethics statement

The study was supervised by the Declaration of Helsinki and was approved by the Ethics Evaluation Committee of the Changsha Central Hospital (2021-S0100, 2021-08-20). All participants provided written informed consent.

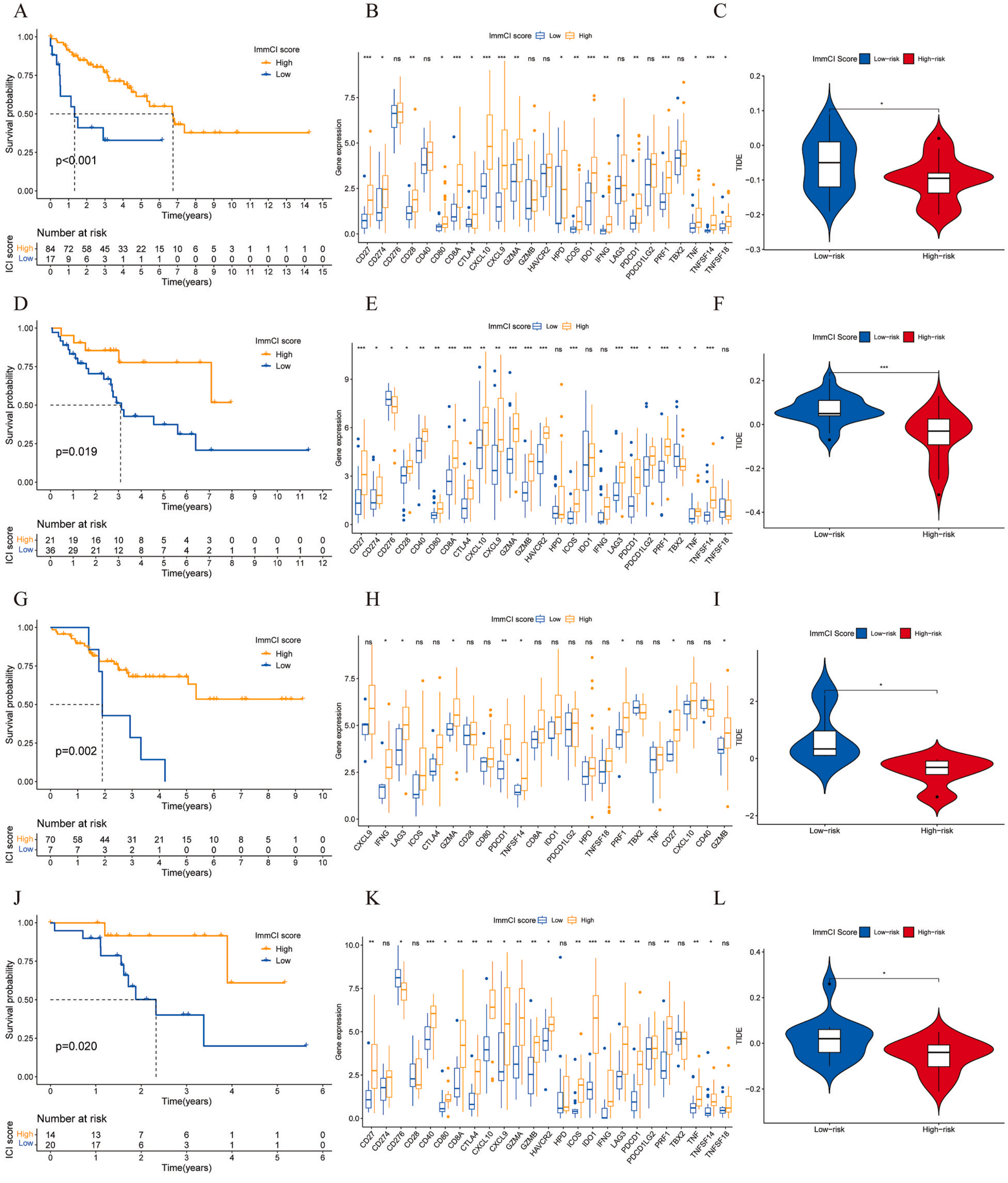
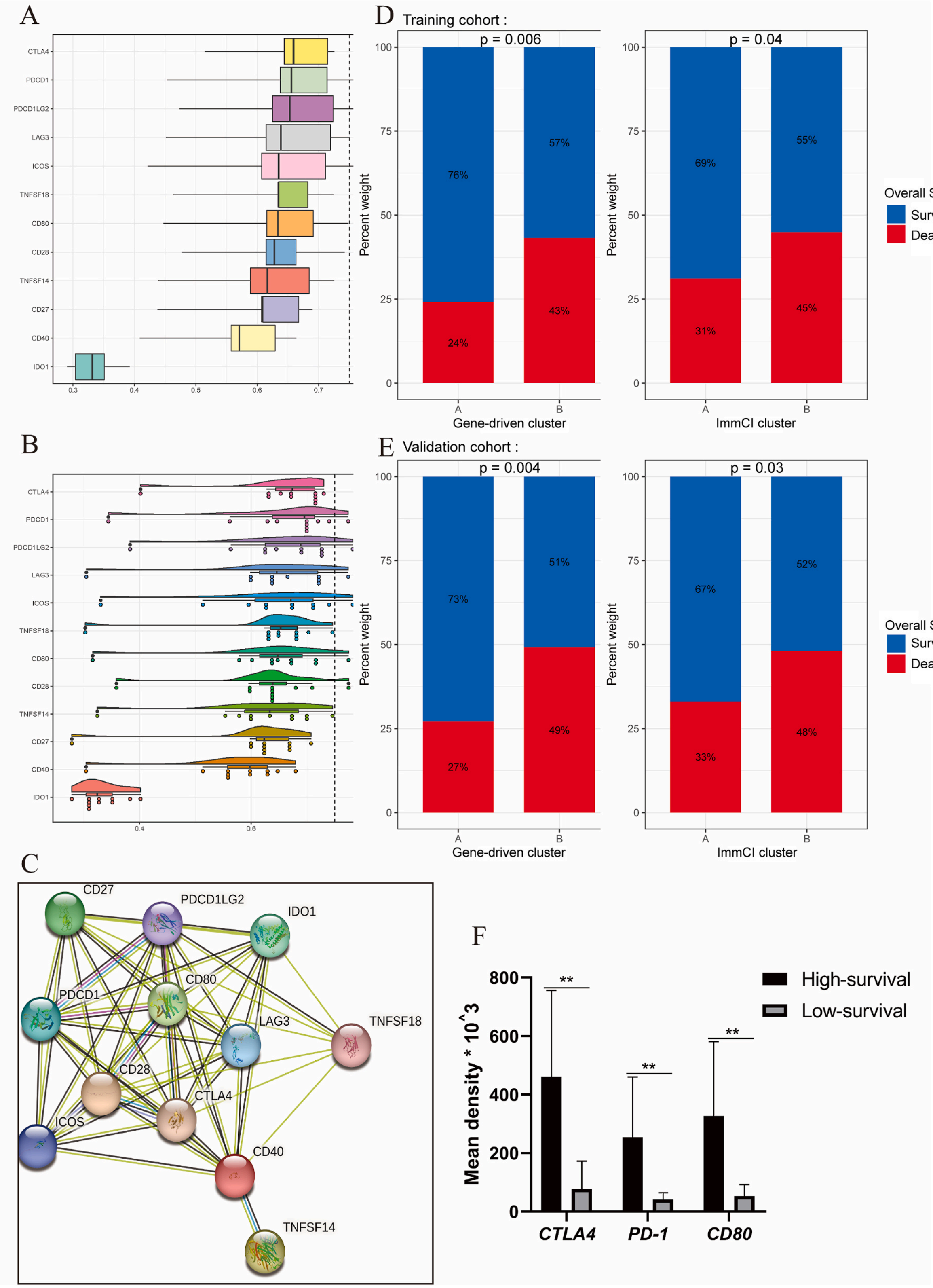


Fig. 11. ImmCI scores reflect prognosis and immunotherapy in leiomyosarcoma (A,B,C), dedifferentiated liposarcoma (D,E,F), fibromyxosarcoma (G,H,I), and undifferentiated sarcoma (J,K,L). OS curve in different ImmCI score subgroups in eiomyosarcoma (A), dedifferentiated liposarcoma (D), fibromyxosarcoma (G) and undifferentiated sarcoma (J). The consistent expression of immunostimulatory, chemokine, and ICB genes between high and low ImmCI score subgroups in leiomyosarcoma (B), dedifferentiated liposarcoma (E), fibromyxosarcoma (H), and undifferentiated sarcoma (K). TIDE score between ImmCI high score subgroup and ImmCI low score subgroup in leiomyosarcoma (C), dedifferentiated liposarcoma (F), fibromyxosarcoma (I), and undifferentiated sarcoma (L). ImmCI, immune cell infiltration.



(caption on next page)

Fig. 12. (A–C). Mining key characters based on gene function through Friends analysis and Protein-Protein-Interaction analysis. (A). Box plot of the distribution of similarity of each gene to other genes, ordered from top to bottom according to the magnitude of similarity, with the gene at the top indicating the gene with the greatest similarity to other genes. (B). The raincloud plot represents the distribution of similarity between each gene and other genes ordered from top to bottom according to the magnitude of similarity, with the gene at the top indicating the gene with the greatest similarity to other genes. (C). Protein-Protein-Interaction analysis. (D–E). The stacked column chart illustrated the relationship between gene-driven clusters, ImmCI clusters, and survival outcomes in both the training (D) and validation (E) cohort. (F). Differences in expression of representative biomarkers between two survival groups. *$p < 0.05$; **$p < 0.01$; and ***$p < 0.001$.

Table 1
Characteristics of the collected clinical cohort.

Characteristics	Number(percentage)
Gender	
Male	6(50 %)
Female	6(50 %)
Age (year)	31 ± 17.3
Subtype	
osteosarcoma	1
rhabdomyosarcoma	1
fibrosarcoma	2
synovial sarcoma	2
Ewing sarcoma	2
liposarcoma	2
chondrosarcoma	1
MPNST	1
Survival	
Mean survival time (month)	25.3
Staining	
CTLA4	
positive	6(50 %)
low positive	2(16.7 %)
negative	4(33.3 %)
PD1	
positive	5(41.7 %)
low positive	2(16.7 %)
negative	5(41.7 %)
CD80	
positive	4(33.3 %)
low positive	4(33.3 %)
negative	4(33.3 %)

Data availability statement

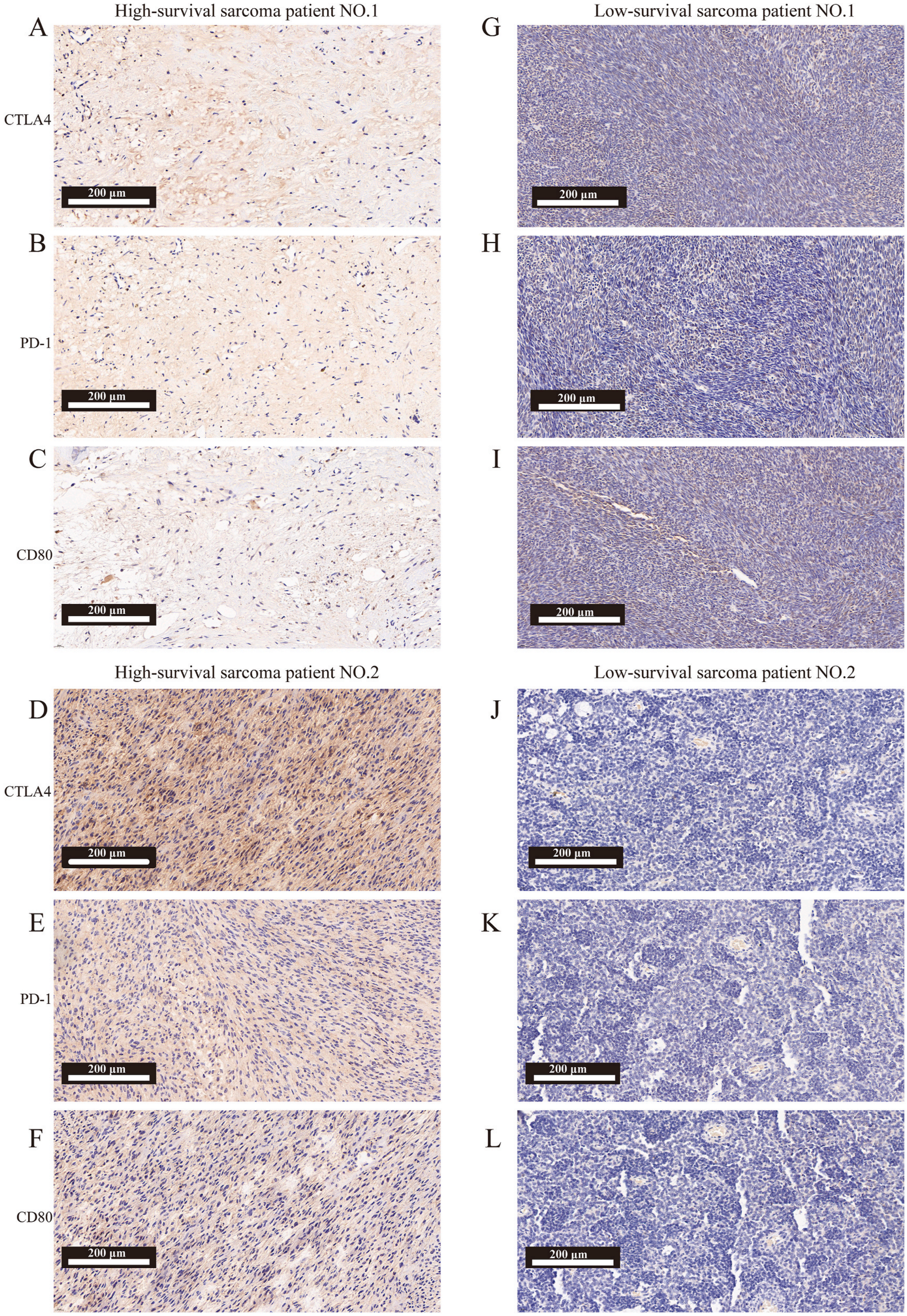
The datasets supporting the conclusions of this article are available in the: TCGA: https://portal.gdc.cancer.gov/repository; GEO: https://www.ncbi.nlm.nih.gov/geo/; TARGET: https://ocg.cancer.gov/programs/target/data-matrix; ICGC: https://dcc.icgc.org/projects; KEGG: https://www.genome.jp/kegg/; GO: http://geneontology.org/docs/go-enrichment-analysis/. Clinical case information and pathology slides can be obtained by contacting the corresponding author on reasonable request.

Funding

This article was funded by the key project of Changsha Central Hospital, University of South China (YNKY202206). The funding doesn't influence study design, collection, analysis and interpretation of data; the writing of the report; and the decision to submit the article for publication.

CRediT authorship contribution statement

Ao-Yu Li: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Jie Bu:** Writing – original draft, Visualization, Validation, Methodology. **Hui-Ni Xiao:** Writing – original draft, Validation. **Zi-Yue Zhao:** Visualization, Validation, Methodology. **Jia-Lin Zhang:** Validation, Methodology. **Bin Yu:** Writing – original draft, Visualization, Methodology. **Hui Li:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Jin-Ping Li:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Tao Xiao:** Writing – review & editing, Supervision, Formal analysis, Conceptualization.


High-survival sarcoma patient NO.1
Low-survival sarcoma patient NO.1
A
CTLA4
200 µm
B
PD-1
200 µm
C
CD80
200 µm
G
200 µm
H
200 µm
I
200 µm
High-survival sarcoma patient NO.2
Low-survival sarcoma patient NO.2
D
CTLA4
200 µm
E
PD-1
200 µm
F
CD80
200 µm
J
200 µm
K
200 µm
L
200 µm

(caption on next page)

Fig. 13. Immunohistochemistry validation of *CTLA4*, *PD-1* and *CD80* expression level in sarcomas of different survival statuses by clinical samples. Scale bar of 20 μm is located in the lower left corner of the image. (A). *CTLA4* expression status in high-survival sarcoma patient NO.1. (B). *PD-1* expression status in high-survival sarcoma patient NO.1. (C). *CD80* expression status in high-survival sarcoma patient NO.1. (D). *CTLA4* expression status in high-survival sarcoma patient NO.2. (E). *PD-1* expression status in high-survival sarcoma patient NO.2. (F). *CD80* expression status in high-survival sarcoma patient NO.2. (G). *CTLA4* expression status in low-survival sarcoma patient NO.1. (H). *PD-1* expression status in low-survival sarcoma patient NO.1. (I). *CD80* expression status in low-survival sarcoma patient NO.1. (J). *CTLA4* expression status in low-survival sarcoma patient NO.2. (K). *PD-1* expression status in low-survival sarcoma patient NO.2. (L). *CD80* expression status in low-survival sarcoma patient NO.2. The area and color depth of the brown part in the image represent the protein expression.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank our sample patients for their efforts to advance medicine. We appreciate the databases of TCGA, TARGET, GEO, ICGC, KEGG and GO for the availability of the data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e38253.

References

[1] A.C. Gamboa, A. Gronchi, K. Cardona, Soft-tissue sarcoma in adults: an update on the current state of histiotype-specific management in an era of personalized medicine, CA Cancer J Clin 70 (3) (2020) 200–229.

[2] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA Cancer J Clin 69 (1) (2019) 7–34.

[3] M.E. Kallen, J.L. Hornick, The 2020 WHO classification: what's new in soft tissue tumor pathology? Am. J. Surg. Pathol. 45 (1) (2021) e1–e23.

[4] S.M. Norberg, S. Movva, Role of genetic and molecular profiling in sarcomas, Curr. Treat. Options Oncol. 16 (5) (2015) 24.

[5] R.L. Siegel, et al., Cancer statistics, 2022, CA Cancer J Clin 72 (1) (2022) 7–33.

[6] P.M. Anderson, Immune therapy for sarcomas, Adv. Exp. Med. Biol. 995 (2017) 127–140.

[7] V. Siozopoulou, et al., Immune checkpoint inhibitory therapy in sarcomas: is there light at the end of the tunnel? Cancers 13 (2) (2021).

[8] A.J. Wisdom, et al., Single cell analysis reveals distinct immune landscapes in transplant and primary sarcomas that determine response or resistance to immunotherapy, Nat. Commun. 11 (1) (2020) 6410.

[9] P. Tsagozis, et al., Sarcoma tumor microenvironment, Adv. Exp. Med. Biol. 1296 (2020) 319–348.

[10] M. Ehnman, et al., The tumor microenvironment of pediatric sarcoma: mesenchymal mechanisms regulating cell migration and metastasis, Curr. Oncol. Rep. 21 (10) (2019) 90.

[11] C.H. Lee, et al., Prognostic significance of macrophage infiltration in leiomyosarcomas, Clin. Cancer Res. 14 (5) (2008) 1423–1430.

[12] N. Oike, et al., Prognostic impact of the tumor immune microenvironment in synovial sarcoma, Cancer Sci. 109 (10) (2018) 3043–3054.

[13] J.N. Kather, et al., CD163+ immune cell infiltrates and presence of CD54+ microvessels are prognostic markers for patients with embryonal rhabdomyosarcoma, Sci. Rep. 9 (1) (2019) 9211.

[14] Y. Inagaki, et al., Dendritic and mast cell involvement in the inflammatory response to primary malignant bone tumours, Clin. Sarcoma Res. 6 (2016) 13.

[15] D. Berghuis, et al., Pro-inflammatory chemokine-chemokine receptor interactions within the Ewing sarcoma microenvironment determine CD8(+) T-lymphocyte infiltration and affect tumour progression, J. Pathol. 223 (3) (2011) 347–357.

[16] H. Fujii, et al., CD8$^+$ tumor-infiltrating lymphocytes at primary sites as a possible prognostic factor of cutaneous angiosarcoma, Int. J. Cancer 134 (10) (2014) 2393–2402.

[17] L.B. Alexandrov, et al., Signatures of mutational processes in human cancer, Nature 500 (7463) (2013) 415–421.

[18] S.K. Kim, et al., PD-L1 tumour expression is predictive of pazopanib response in soft tissue sarcoma, BMC Cancer 21 (1) (2021) 336.

[19] F. Petitprez, et al., B cells are associated with survival and immunotherapy response in sarcoma, Nature 577 (7791) (2020) 556–560.

[20] L. Qi, et al., Deciphering the role of NETosis-related signatures in the prognosis and immunotherapy of soft-tissue sarcoma using machine learning, Front. Pharmacol. 14 (2023) 1217488.

[21] L. Qi, et al., Identification of anoikis-related molecular patterns to define tumor microenvironment and predict immunotherapy response and prognosis in soft-tissue sarcoma, Front. Pharmacol. 14 (2023) 1136184.

[22] Q. Li, X. Xu, X. Jiao, Prognostic implication of cuproptosis related genes associates with immunity in Ewing's sarcoma, Transl Oncol 31 (2023) 101646.

[23] W. Weng, et al., The immune subtypes and landscape of sarcomas, BMC Immunol. 23 (1) (2022) 46.

[24] D. Shi, et al., Pan-sarcoma characterization of lncRNAs in the crosstalk of EMT and tumour immunity identifies distinct clinical outcomes and potential implications for immunotherapy, Cell. Mol. Life Sci. 79 (8) (2022) 427.

[25] X. Feng, et al., Comprehensive immune profiling unveils a subset of leiomyosarcoma with "hot" tumor immune microenvironment, Cancers 15 (14) (2023).

[26] A.M. Newman, et al., Robust enumeration of cell subsets from tissue expression profiles, Nat. Methods 12 (5) (2015) 453–457.

[27] K. Yoshihara, et al., Inferring tumour purity and stromal and immune cell admixture from expression data, Nat. Commun. 4 (2013) 2612.

[28] M.D. Wilkerson, D.N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking, Bioinformatics 26 (12) (2010) 1572–1573.

[29] G.K. Smyth, Limma: linear models for microarray data, in: Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, 2005, pp. 397–420.

[30] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Software 36 (2010) 1–13.

[31] G. Yu, et al., clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS 16 (5) (2012) 284–287.

[32] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res 28 (1) (2000) 27–30.

[33] M. Kanehisa, Toward understanding the origin and evolution of cellular organisms, Protein Sci. 28 (11) (2019) 1947–1951.

[34] P. Jiang, et al., Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response, Nat Med 24 (10) (2018) 1550–1558.
[35] T.M. Horton, et al., PAM staining intensity of primary neuroendocrine neoplasms is a potential prognostic biomarker, Sci. Rep. 10 (1) (2020) 10943.
[36] F. Varghese, et al., IHC Profiler: an open source plugin for the quantitative evaluation and automated scoring of immunohistochemistry images of human tissue samples, PLoS One 9 (5) (2014) e96801.
[37] J. Shu, et al., Statistical colour models: an automated digital image analysis method for quantification of histological biomarkers, Biomed. Eng. Online 15 (2016) 46.
[38] G. Yu, et al., GOSemSim: an R package for measuring semantic similarity among GO terms and gene products, Bioinformatics 26 (7) (2010) 976–978.
[39] D. Szklarczyk, et al., STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Res 47 (D1) (2019). D607-d13.
[40] P. Specenier, Ipilimumab in melanoma, Expert Rev. Anticancer Ther. 16 (8) (2016) 811–826.
[41] J. Larkin, et al., Five-year survival with combined nivolumab and ipilimumab in advanced melanoma, N. Engl. J. Med. 381 (16) (2019) 1535–1546.
[42] Y.Y. Janjigian, et al., CheckMate-032 study: efficacy and safety of nivolumab and nivolumab plus ipilimumab in patients with metastatic esophagogastric cancer, J. Clin. Oncol. 36 (28) (2018) 2836–2844.
[43] T.K. Owonikoko, et al., Nivolumab and ipilimumab as maintenance therapy in extensive-disease small-cell lung cancer: CheckMate 451, J. Clin. Oncol. 39 (12) (2021) 1349–1359.
[44] H.A. Tawbi, et al., Pembrolizumab in advanced soft-tissue sarcoma and bone sarcoma (SARC028): a multicentre, two-cohort, single-arm, open-label, phase 2 trial, Lancet Oncol 18 (11) (2017) 1493–1501.
[45] B.A. Wilky, et al., Axitinib plus pembrolizumab in patients with advanced sarcomas including alveolar soft-part sarcoma: a single-centre, single-arm, phase 2 trial, Lancet Oncol 20 (6) (2019) 837–848.
[46] J. Wang, et al., Immune-related prognostic genes signatures in the tumor microenvironment of sarcoma, Math. Biosci. Eng. 18 (3) (2021) 2243–2257.
[47] C. Zhang, et al., Profiles of immune cell infiltration and immune-related genes in the tumor microenvironment of osteosarcoma, Aging (Albany NY) 12 (4) (2020) 3486–3501.
[48] C.C. Wu, et al., Immuno-genomic landscape of osteosarcoma, Nat. Commun. 11 (1) (2020) 1008.
[49] S.P. D'Angelo, et al., Antitumor activity associated with prolonged persistence of adoptively transferred NY-ESO-1 (c259)T cells in synovial sarcoma, Cancer Discov. 8 (8) (2018) 944–957.
[50] J. Galon, et al., Type, density, and location of immune cells within human colorectal tumors predict clinical outcome, Science 313 (5795) (2006) 1960–1964.
[51] J. Galon, D. Bruni, Approaches to treat immune hot, altered and cold tumours with combination immunotherapies, Nat. Rev. Drug Discov. 18 (3) (2019) 197–218.
[52] S. Maleki Vareki, High and low mutational burden tumors versus immunologically hot and cold tumors and response to immune checkpoint inhibitors, J Immunother Cancer 6 (1) (2018) 157.
[53] J. McClellan, M.C. King, Genetic heterogeneity in human disease, Cell 141 (2) (2010) 210–217.