scientific reports

OPEN

Use of machine learning-based integration to develop an immune-related signature for improving prognosis in patients with gastric cancer

Jingyuan Ning[1,5], Keran Sun[1,5], Xiaoqing Fan[1,5], Keqi Jia[2], Lingtong Meng[1], Xiuli Wang[3], Hui Li[4], Ruixiao Ma[4], Subin Liu[4], Feng Li[4] & Xiaofeng Wang[1,4]✉

Gastric cancer is one of the most common malignancies. Although some patients benefit from immunotherapy, the majority of patients have unsatisfactory immunotherapy outcomes, and the clinical significance of immune-related genes in gastric cancer remains unknown. We used the single-sample gene set enrichment analysis (ssGSEA) method to evaluate the immune cell content of gastric cancer patients from TCGA and clustered patients based on immune cell scores. The Weighted Correlation Network Analysis (WGCNA) algorithm was used to identify immune subtype-related genes. The patients in TCGA were randomly divided into test 1 and test 2 in a 1:1 ratio, and a machine learning integration process was used to determine the best prognostic signatures in the total cohort. The signatures were then validated in the test 1 and the test 2 cohort. Based on a literature search, we selected 93 previously published prognostic signatures for gastric cancer and compared them with our prognostic signatures. At the single-cell level, the algorithms "Seurat," "SCEVAN", "scissor", and "Cellchat" were used to demonstrate the cell communication disturbance of high-risk cells. WGCNA and univariate Cox regression analysis identified 52 prognosis-related genes, which were subjected to 98 machine-learning integration processes. A prognostic signature consisting of 24 genes was identified using the StepCox[backward] and Enet[alpha = 0.7] machine learning algorithms. This signature demonstrated the best prognostic performance in the overall, test1 and test2 cohort, and outperformed 93 previously published prognostic signatures. Interaction perturbations in cellular communication of high-risk T cells were identified at the single-cell level, which may promote disease progression in patients with gastric cancer. We developed an immune-related prognostic signature with reliable validity and high accuracy for clinical use for predicting the prognosis of patients with gastric cancer.

Abbreviations

WGCNA	Weighted correlation network analysis
GC	Gastric cancer
PD-1	Programmed cell death 1
CTLA4	Cytotoxic T lymphocyte antigen 4
ICIs	Immune checkpoint inhibitors
TME	Tumor immune microenvironment
TCGA	The Cancer Genome Atlas
GEO	Gene expression omnibus

[1]Department of Immunology, Immunology Department of Hebei Medical University, Shijiazhuang, People's Republic of China. [2]Department of Pathology, Shijiazhuang People's Hospital, Shijiazhuang, People's Republic of China. [3]Department of Laboratory, The Second Hospital of Hebei Medical University, Shijiazhuang, People's Republic of China. [4]Department of Oncology, Shijiazhuang Fourth Hospital, Shijiazhuang, People's Republic of China. [5]These authors contributed equally: Jingyuan Ning, Keran Sun and Xiaoqing Fan. ✉email: wxf9992021@163.com

C-index	Concordance index
ssGSEA	Single sample gene set enrichment analysis
GS	Gene significance
MM	Module membership
IRS	Immune-related signature
GO	Gene ontology
ROC	Receiver operating characteristic
AUC	Area under the curve
RSF	Random survival forest
Enet	Elastic network
plsRcox	Partial least squares regression for Cox
SuperPC	Supervised principal components
GBM	Generalised boosted regression
survival-SVM	Survival support vector machine
C-index	Concordance index
OS	Overall survival
PFS	Progression-free survival

Gastric cancer (GC) is among the most common malignant tumors. Its incidence in China ranks second among that of all malignant tumors[1]. Worldwide, GC is the fifth leading cause of all cancers and the fourth leading cause of cancer-related mortality[2]. Currently, there are various treatment methods for GC[3], such as surgery[4,5], chemotherapy[6], radiotherapy[7], targeted therapy[8], and immunotherapy[9,10]. However, these approaches are not effective in prolonging the life of most patients[11]. Additionally, traditional treatment methods are not effective for patients with advanced stages of GC[12]. At present, anti-programmed cell death 1 (PD-1) monoclonal antibodies, anti-cytotoxic T lymphocyte antigen 4 (CTLA4) monoclonal antibodies, and other immune checkpoint inhibitors (ICIs) are considered innovative treatment strategies for advanced GC[13]. Although studies Keynote-059[14], Keynote-061[15], Keynote-062[16], and attraction-02[17] have shown a good efficacy of immunotherapy against GC, it seems to be more effective for subgroups with high mutation load, positive Epstein Barr virus, or high microsatellite instability[18]. Many factors, including the tumor immune microenvironment (TME), affect the effectiveness of immunotherapy. There are only a few accurate biomarkers that can predict the response to immunotherapy[19]. Identification of potential prognostic markers and the development of immunotherapy guidelines can aid in designing personalized immunotherapy for patients with GC. Some researchers have suggested that a more in-depth analysis of the complexity of the TME can help reveal efficacious biomarkers that can identify patient populations responsive to immunotherapy[20]. Unfortunately, we still know little about the TME in GC, and we urgently need to identify effective prognostic signatures.

Based on machine learning, prognostic models have been shown to have predictive value in various diseases, including renal cancer[21,22] colon adenocarcinoma[23], and endometriosis[24]. Although previous studies have screened immune-related genes of GC to predict the prognosis characteristics, their prediction accuracy is not high[25,26]. In our study, we used a combination of 98 machine learning algorithms to determine the best immune-related prognostic signature for GC and performed external prognostic prediction validation using multiple datasets. Finally, we collected 93 prognostic signatures for comparison. The results showed that our signature was the most effective prognostic biomarker compared to other signatures.

Methods.

Acquisition and pre-processing of transcriptome data.

We downloaded the transcriptome data of 32 normal gastric tissues and 375 GC tissues from The Cancer Genome Atlas (TCGA) website (https://portal.gdc.cancer.gov/). The fragments per kilobase million values were transformed into transcripts per million. Concomitantly, the clinical information corresponding to all patients was downloaded for subsequent analysis.

Weighted correlation network analysis (WGCNA).

The WGCNA approach was employed to build coexpression networks of genes. To establish a scale-free network, we calculated an optimal soft threshold β. The weighted adjacency matrix was then converted into a topological overlap matrix (TOM), and its corresponding dissimilarity (1-TOM) was computed. The dynamic tree cutting method was utilized to identify modules of coexpressed genes.

Enrichment analysis.

Enrichment analysis of differential genes was performed using the "GSEABase" package, "ClusterProfiler" package and "org.Hs.eg.db" package. The database used for the enrichment analysis was derived from the Gene Ontology (http://geneontology.org/). Use the EnrichGO function for enrichment. If $P < 0.05$, the pathway was considered to be significantly enriched. "ggplot2" package, "ggpubr" package for visualization.

Machine learning to build prognostic signatures.

In the R(4.2.1) environment, a total of 10 machine learning algorithms, including random survival forest (RSF), elastic network (Enet), Lasso, Ridge, stepwise Cox, CoxBoost, partial least squares regression for Cox (plsRcox), supervised principal components (SuperPC), generalized boosted regression (GBM), and survival support vector machine (survival-SVM) were used. In the process, we used one algorithm to filter the variables and another algorithm to build the prognostic signature. When the final prognostic signature contained less than 5 genes, the signature was considered an invalid signature. A total of 98 combinations of machine learning algorithms were eventually integrated. Finally, Harrell's concordance index (C-index) was calculated for each signature, and the signature with the highest average

C-index value was considered to be the best signature. After calculating the risk score for each patient using the predict function, the optimal cutoff value for the risk score is determined using the surv_cutpoint function in the "srvminer" package. Based on the optimal cutoff value of the risk score, patients are divided into high-risk and low-risk groups.

Acquisition and pre-processing of single-cell transcriptome data.

Single-cell transcriptome data were obtained from the GEO database (GEO registration number: GSE163558; https://www.ncbi.nlm.nih.gov/geo/). Quality control was performed in R(4.1.2) environment using standard single cell processing procedures. The count matrix were read using the Read10X function from the Seurat package (Version 4.0.4), and the latter was further converted to dgCMatrix format. The merge function was used to integrate all individual objects into an aggregate object, and the RenameCells function was used to ensure that all cell labels were unique. We filtered low quality cells with the following filtering criteria: when a gene was expressed in less than 3 cells, the gene was deleted. When the number of genes expressed in a cell was less than 200, the cell was deleted. A global-scaling normalization method ("LogNormalize") was employed to ensure that the total gene expression in each cell was equal, and the scale factor was set to 10,000. The top 2000 variably expressed genes were returned for downstream analysis using the FindVariableFeatures function. The ScaleData function, "vars.to.regress" option UMI, and percent mitochondrial content were used to regress out unwanted sources of variation. Principal component analysis (PCA) incorporating highly variable features reduced the dimensionality of this dataset, and the first 30 PCs were identified for analysis. Harmony method[27] was used to remove batch effects between samples. Cells were down-dimensioned using the UMAP method. Clustering analysis was performed based on the edge weights between any two cells, and a shared nearest-neighbor graph was produced using the Louvain algorithm, which was implanted in the FindNeighbors and FindClusters functions. The parameter of resolution in the FindClusters function was tried repeatedly between 0.1 and 1. Cell clustering trees at different resolutions were observed using the clustree function, and the results showed that the clearest clustering results were obtained when the resolution was 0.5. To annotate the cell clusters, differentially expressed markers of the resulting clusters were identified with the FindAllMarkers function using the default nonparametric Wilcoxon rank sum test with Bonferroni correction. All cells were annotated according to cell surface markers and annotated genes used in the relevant literature and CellMarker database[28] (http://xteam.xbio.top/CellMarker/).

Identification of high-risk-related phenotypic cells.

Scissor algorithm from the "Scissor" package[29] (2.0.0). By leveraging bulk data and phenotype information, this algorithm automatically selects cell subpopulations from single-cell data that are most responsible for the differences of phenotypes. The novelty of Scissor is that it utilizes phenotype information from bulk data to identify the most highly disease-relevant cell subsets. In our study, high-risk patients and low-risk patients identified in TCGA were treated as two different phenotypes. Based on the transcriptomic data of high- and low-risk phenotypes for all patients, the "Scissor" function was used to associate both phenotypes with each cell in the single-cell data.

Cellular communication network.

Cell–cell interaction analysis was performed based on the "CellChat" (v1.0.0) R package[30]. CellChat has a public repository of ligands, receptors, cofactors and their interactions (http://www.cellchat.org/). The CellChat R package is a versatile and easy-to-use toolkit for inferring, analyzing, and visualizing cell–cell communication from any given scRNA-seq data. The ligand and receptor genes expressed by each cell were projected into a manually selected reference communication network and the probability of communication in each pathway was inferred by gene expression. Finally use the netVisual_bubble function for visualization, with all parameters as default.

Statistical analysis.

All statistical analyses were carried out using R (4.1.2). The statistical methods were all set up according to the corresponding R software.$P<0.05$ was considered statistically significant. *$P<0.05$, **$P<0.01$, ***$P<0.001$.

Ethics approval and consent to participate.

All methods were carried out in accordance with relevant guidelines and regulations.

Results

Cluster analysis of immune subtypes.

We calculated multiple immune cell scores for each patient with GC using the single sample Gene Set Enrichment Analysis (ssGSEA) method, and based on the scores, we clustered all the patients (Fig. 1A). Furthermore, analysis of all the patients using the T-SNE clustering algorithm showed that the clustering was highly stable in all patients (Fig. 1B). As shown in a heat map, all patients could be divided into two subtypes: one subtype of patients had a higher immune cell score (Immunity_H) and the other subtype had a lower immune cell score (Immunity_L) (Fig. 1C). To demonstrate that our immune cell clustering results are not subject to algorithmic bias, we first used ESTIMATE to verify the plausibility of the ssGSEA results. The results showed that patients in the Immunity_H group had a higher immune score and lower tumor purity compared to patients in the Immunity_L group (Fig. 1C). We then used six algorithms based on TIMER, CIBERSORT, CIBERSORT-ABS, QUANTISEQ, MCPCOUNTER, XCELL, and EPIC for immune cell content assessment. The results all showed that patients in the Immunity_H group had significantly higher immune cell content than patients in the Immunity_L group, which was highly consistent with our clustering results based on the ssGSEA algorithm (Fig. 1D). These findings demonstrate the high stability of our clustering results.

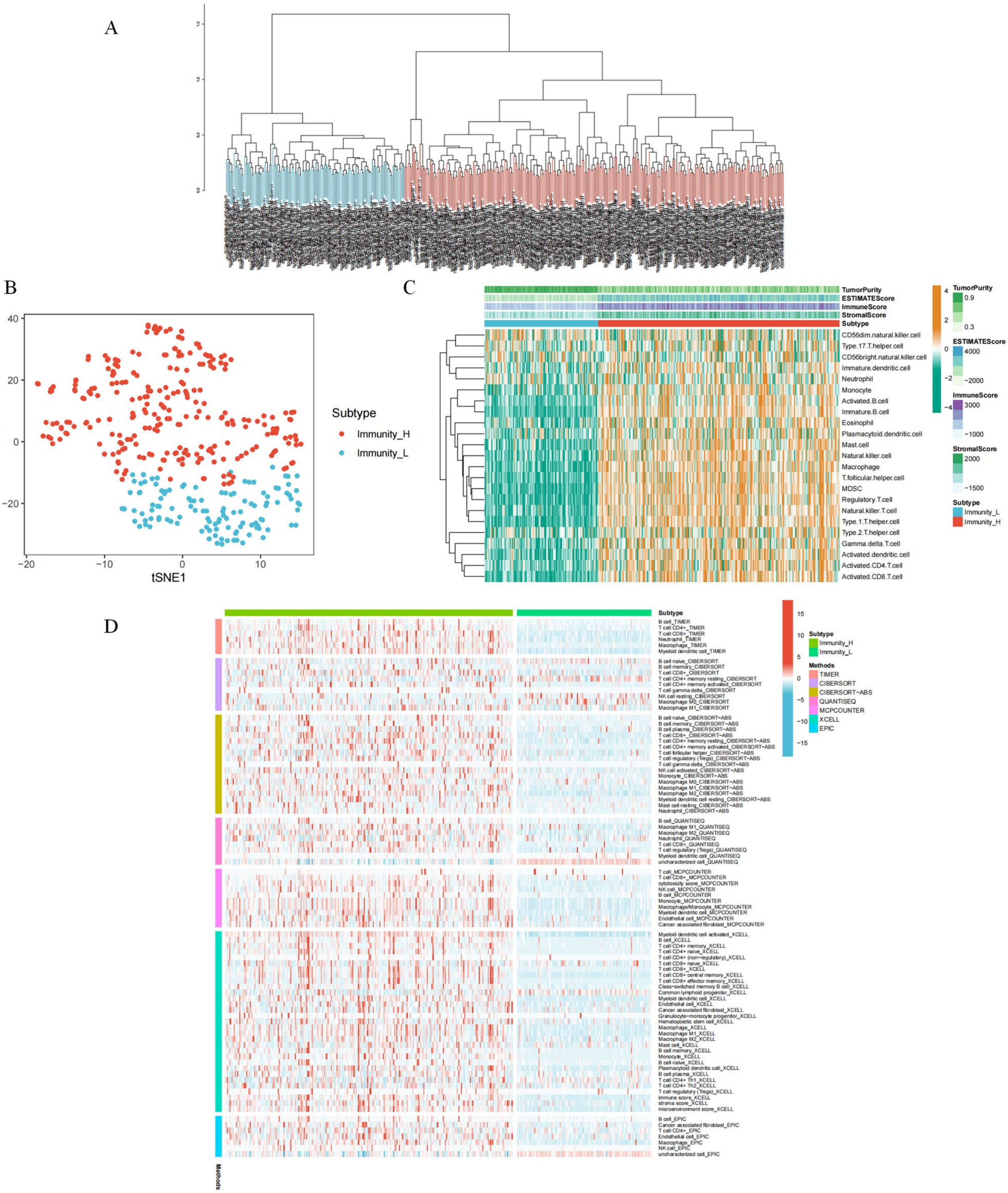

Figure 1. Subtype analysis of patients with TCGA gastric cancer. (**A**) Clustering analysis based on immune cell content. This analysis identified two different patient immunotypes. (**B**) Reduced dimensional analysis of T-SNE based on immune cell content. This figure further validates the immunophenotyping of two different patients. (**C**) Immune cell content calculated by the ssGSEA and ESTIMATE algorithms. (**D**) six algorithms based on TIMER, CIBERSORT, CIBERSORT-ABS, QUANTISEQ, MCPCOUNTER, XCELL, and EPIC for immune cell content assessment. The results show two types of immune patients typed without the bias of the algorithm.

Identification of immune-related modules. In the weighted correlation network analysis (WGCNA), the soft threshold β was set to 8 (Fig. 2A), which provided a suitable power value for the construction of a coexpression network. We identified a total of 14 gene modules, each of which was represented using a different color (Fig. 2B). The correlation between each module and the patient's clinical traits, including sex, grade, stage, and subtype that we clustered based on the ssGSEA method, was evaluated. Among all modules, the correlation between the magenta module and subtype was the highest (Fig. 2C). The correlation coefficient between gene significance (GS) and module membership (MM) reached 0.73 (Fig. 2D), which suggested that the quality of the magenta module construction was superior. Based on these results, we defined the 1104 genes in the magenta module as immune subtype-related genes. To further determine the correlation between these genes and immunity, we performed an enrichment analysis of immune subtype-related genes. The Gene ontology (GO) enrichment analysis results showed that these genes were enriched in T cell activation, leukocyte cell–cell

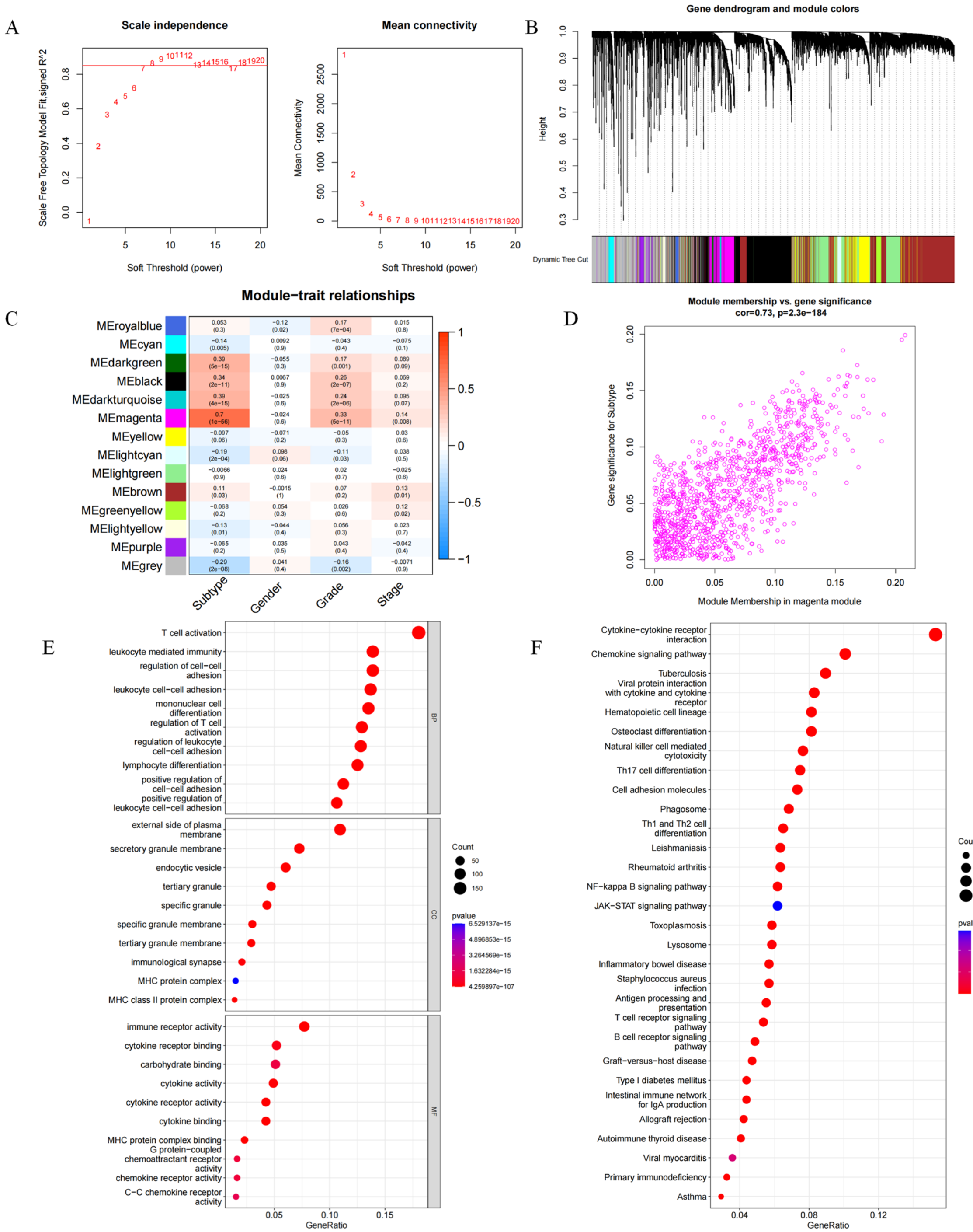

Figure 2. The weighted correlation network analysis. (**A**) Determination of soft thresholds. (**B**) Identification of gene clustering modules. (**C**) Correlation meter analysis between gene modules and phenotypes. Memagenta modules are highly correlated with subtypes (**D**) The correlation coefficient between gene significance (GS) and module membership (MM). (**E**) GO enrichment analysis. (**F**) KEGG enrichment analysis.

adhesion, regulation of T cell activation, and positive regulation of leukocyte cell–cell adhesion (Fig. 2E). Kyoto Encyclopedia of Genes and Genomes[31] (KEGG) enrichment analysis results showed that these genes were significantly enriched in cytokine-cytokine receptor interaction, Th1 and Th2 cell differentiation, antigen processing and presentation, B cell receptor signaling pathway, and T cell receptor signaling pathway(Fig. 2F). Together, these results demonstrate a high correlation between immune subtype-related genes and the immune system.

Generation of signature based on machine learning integration. We performed a univariate Cox regression analysis on the 1104 immune subtype-related genes and identified 52 prognosis-related genes (Fig. 3A), including 6 protective genes (HR < 1) and 46 risk genes (HR > 1). These 52 genes were subjected to our machine learning integration process for establishing immune-related signatures. Specifically, we removed patients with a survival time of fewer than 30 days, after which the remaining 335 TCGA patients with GC were used as the total cohort for subsequent analysis. Meanwhile, to determine the accuracy of our signature, we randomly divided the 335 patients into two cohorts named test 1 and test 2 in a 1:1 ratio. In the total cohort, we fitted 98 prognostic prediction signatures, and the C-index values were calculated in the total, test 1, and test 2 cohorts for each signature. Interestingly, StepCox[backward] + Enet[alpha = 0.7] had the highest mean C-index value (0.726) among all signatures (Fig. 3B). The signature consisted of 24 genes and we named it immune-related signature (IRS). Based on the expression of these 24 genes, we calculated the risk score for all patients. Risk scores = (− 0.2076025 × *TNFAIP2* expression) + (0.1472951 × *SLC37A2* expression) + (0.1714101 × *RGS1* expression) + (− 0.3396804 × *ZNF101* expression) + (− 0.6150795 × *TM6SF1* expression) + (− 0.3724076 × *CRHBP* expression) + (− 0.9322991 × *AKAP5* expression) + (− 0.3562780 × *CRYBB1* expression) + (0.5096191 × *S100Z* expression) + (0.4818817 × *ACSM5* expression) + (0.2350907 × *NTAN1* expression) + (0.5508151 × *IL5RA* expression) + (0.1979289 × *ABCG1* expression) + (0.8638607 × *CAMK4* expression) + (0.2211539 × *MCEMP1* expression) + (0.2176318 × *SLC2A3* expression) + (0.1990519 × *RENBP* expression) + (0.1643554 × *BASP1* expression) + (0.2138879 × *KYNU* expression) + (− 0.2643476 × *CTLA4* expression) + (− 0.2688278 × *FCGR2B* expression) + (− 0.1302975 × *ENTPD8* expression) + (0.2952993 × *DDO* expression) + (− 0.1408996 × *FCN1* expression).

Accuracy and validity assessment of IRS. Notably, a large number of machine learning-based prognostic prediction signatures have been developed in recent years. There is a diversity of these signatures in terms of research perspectives, such as pyroptosis, cuproptosis, ferroptosis, EMT conversion, hypoxia, metabolism, aging, and immune response. To determine the superiority of the IRS, we collected 93 published prognostic signatures and calculated the C-index values for these signatures (Fig. 4A). Importantly, IRS showed the highest C-index values for the total, test 1, and test 2 cohorts compared to these signatures. This suggests that the IRS has robust accuracy. Next, we determined the optimal cutoff value based on the risk score of each patient using the "survminer" package. Kaplan–Meier analysis showed that the overall survival of high-risk patients in the total, test 1, and test 2 cohorts was significantly worse than that of low-risk patients (Fig. 4B). Additionally, we found that progression-free survival (PFS) was significantly worse in high-risk patients in the total, test 1, and test 2 cohorts compared to that in low-risk patients, demonstrating the value of IRS for predicting PFS as well (Fig. 4C).The receiver operating characteristic curve (ROC) analysis showed that the area under the curve (AUC) values for the total cohort were 0.728, 0.798, and 0.791 at 1, 3, and 5 years, respectively. Meanwhile, the AUC values for the test 1 cohort were 0.726, 0.726, and 0.806, and for the test 2 cohort were 0.733, 0.838, and 0.766 at 1, 3, and 5 years, respectively (Fig. 4D). The results of univariate Cox regression (Fig. 4E) and multivariate Cox regression analysis (Fig. 4F) showed that risk score and stage were the two independent factors affecting prognosis, with risk score having the largest HR value. For the purpose of external validation, we introduced the GSE84437 and GSE84433 datasets for external validation of the prognostic effect. The results showed that in both GSE84437 and GSE84433 datasets, high-risk patients had worse overall survival (OS) than low-risk patients (Fig. 4G,H). Finally, we assessed the value of IRS in immunotherapy. AUC values were higher than 0.75 in two anti-PD-1 treatments and one anti-PD-1/anti-CTLA-4 (Fig. 4I), suggesting that IRS is relevant for predicting sensitivity to immunotherapy in patients with gastric cancer. In conclusion, our results suggest that IRS has excellent stability and validity. To improve the clinical value of this study, we created a nomogram based on risk scores and clinical characteristics to facilitate clinical translation (Fig. 4J). The calibration curve showed that the nomogram had good accuracy at 1, 2, 3, and 5 years.

IRS combined with single-cell analysis identifies communication perturbations in high-risk cells. Cell-to-cell communication plays a crucial role in understanding the complexity of the tumor immune microenvironment. For instance, in cancer, tumor cells interact with various immune cells and stromal cells in the tumor microenvironment. These interactions can shape disease progression and response to therapy. However, this is not captured by bulk transcriptomic data. Luckily, single-cell sequencing technology enables the possibility of uncovering intricate and complex interactions between cells. Single-cell analysis of tumors and immune cells can provide an in-depth understanding of the molecular mechanisms of tumor-immune cell interactions, which can offer information for the development of new immunotherapies. Hence, we acquired and processed single-cell sequencing data from the tumor sites of three patients with GC.Here, a total of 10,234 cells passed quality control. All cells were annotated according to cell surface markers (Fig. 5A). Five cell types were present, namely B cells, epithelial cells, myeloid cells, stroma cells, and T cells. For epithelial cells, we identified 2634 malignant cells using the "SCEVAN" package[32], which has been shown to be significantly more accurate than “inferCNV” and “copyKAT”. Figure 5B shows the landscape of copy number variation in normal and tumour cells. Cell down-dimensioning and visualization were performed using T-SNE (Fig. 5C). Among the immune cells, T cells were the most numerous. To identify the T cells that contribute to the high-risk disease

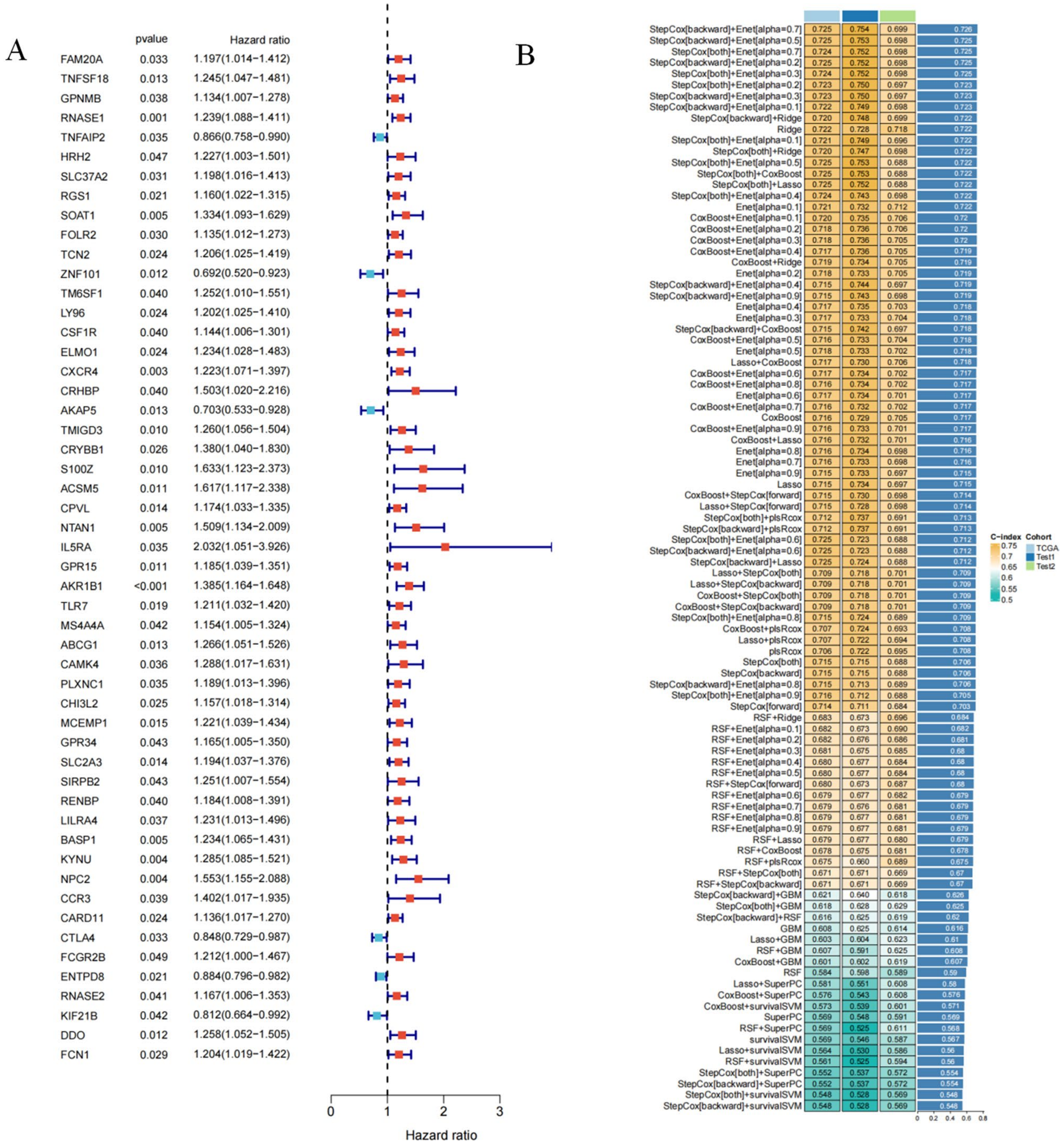

Figure 3. Machine learning integration to build prognostic models. (**A**) Univariate Cox regression identifying 52 prognosis-related genes. (**B**) 98 machine learning integrated prognostic models and their C-index values. Determine StepCox[backward] + Enet[alpha = 0.7] as the best signature.

phenotype, the “scissor” package was used to correlate bulk sequencing data with single-cell sequencing data. This method uses single-cell data and phenotypic information to identify subpopulations of cells. Using bulk sequencing data and its annotated information with various phenotypes, the algorithm automatically selects cells that are highly correlated with the phenotype. We considered high-risk and low-risk in patients as two phenotypes, associating both phenotypes with T cells that contribute to the high-risk disease phenotype. We successfully identified a total of 507 high-risk cells and 365 low-risk cells (Fig. 5D). We then used the "cellchat" package to analyze the differences in cellular communication networks between the high-risk and low-risk cells (Fig. 5E). Multiple signaling perturbations were found between high-risk T cells, low-risk T cells, tumor cells, and normal epithelial cells (Fig. 5F). JAG1-NOTCH1 and TNFSF15-TNFRSF25 signals were present between tumor cells and low-risk T cells, and absent between tumor cells and high-risk T cells. The autocrine SELPLG-SELL signaling present in low-risk T cells was lost in high-risk T cells. Furthermore, we found some alterations

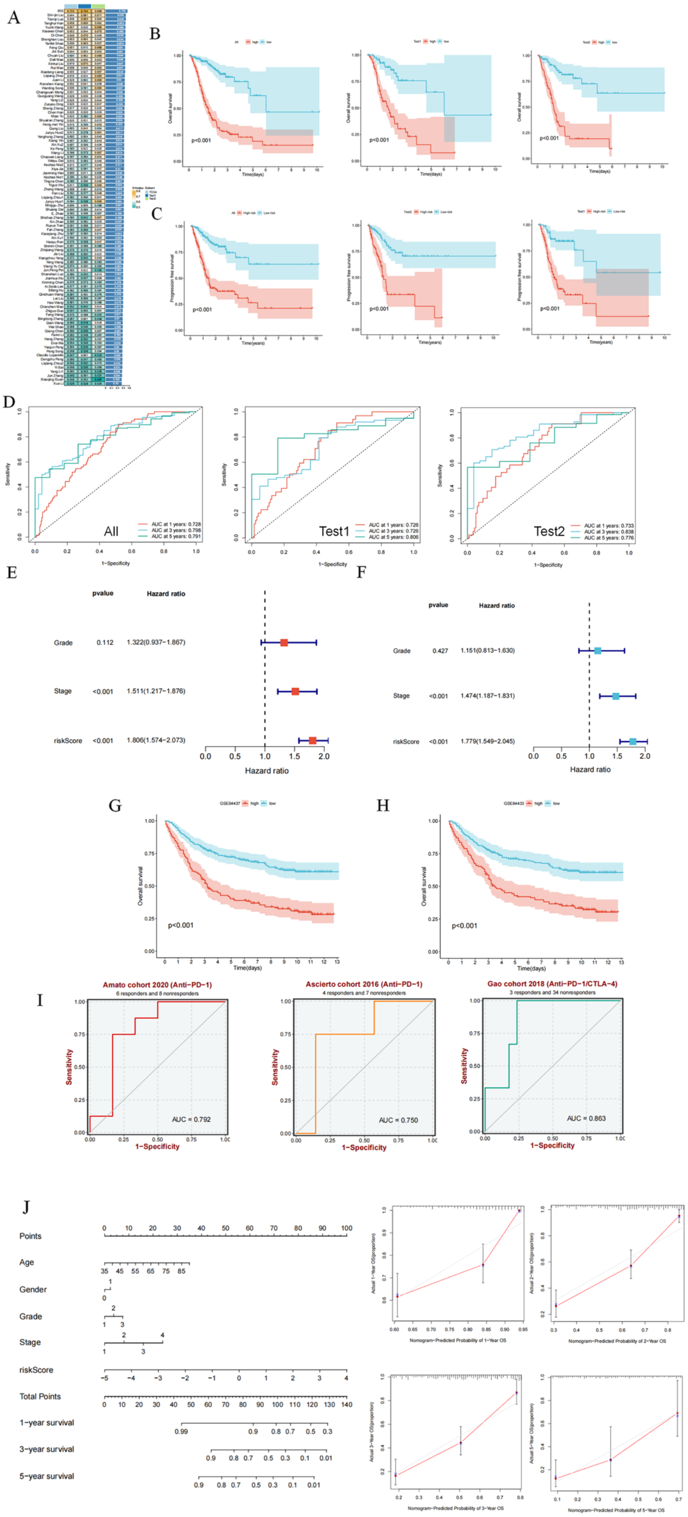

Figure 4. Analysis of the accuracy and validity of IRS. (**A**) C-index values of IRS compared with 93 published prognostic models for gastric cancer. The results demonstrate the outperformance of IRS over published signatures (**B**) OS analysis of IRS in the total cohort, test1 cohort, and test2 cohort. Poorer prognosis for high-risk patients compared to low-risk patients. (**C**) PFS analysis of IRS in the total cohort, test1 cohort, and test2 cohort. Poorer prognosis for high-risk patients compared to low-risk patients. (**D**) 1, 3, and 5 year ROC analysis of IRS in total cohort, test1 cohort, and test2 cohort. (**E**) Univariate Cox regression analysis of IRS and clinical characteristics. (**F**) Multivariate Cox regression analysis of IRS and clinical characteristics. (**G**) OS analysis of IRS in GSE84437. (**H**) OS analysis of IRS in GSE84433. (**I**) ROC analysis of IRS differentiating between responding and non-responding patients in an immunotherapy cohort. IRS has predictive value for patient response to immunotherapy. (**J**) Risk scores combined with clinical characteristics of the nomogram. The calibration curve proves the accuracy of this nomogram.

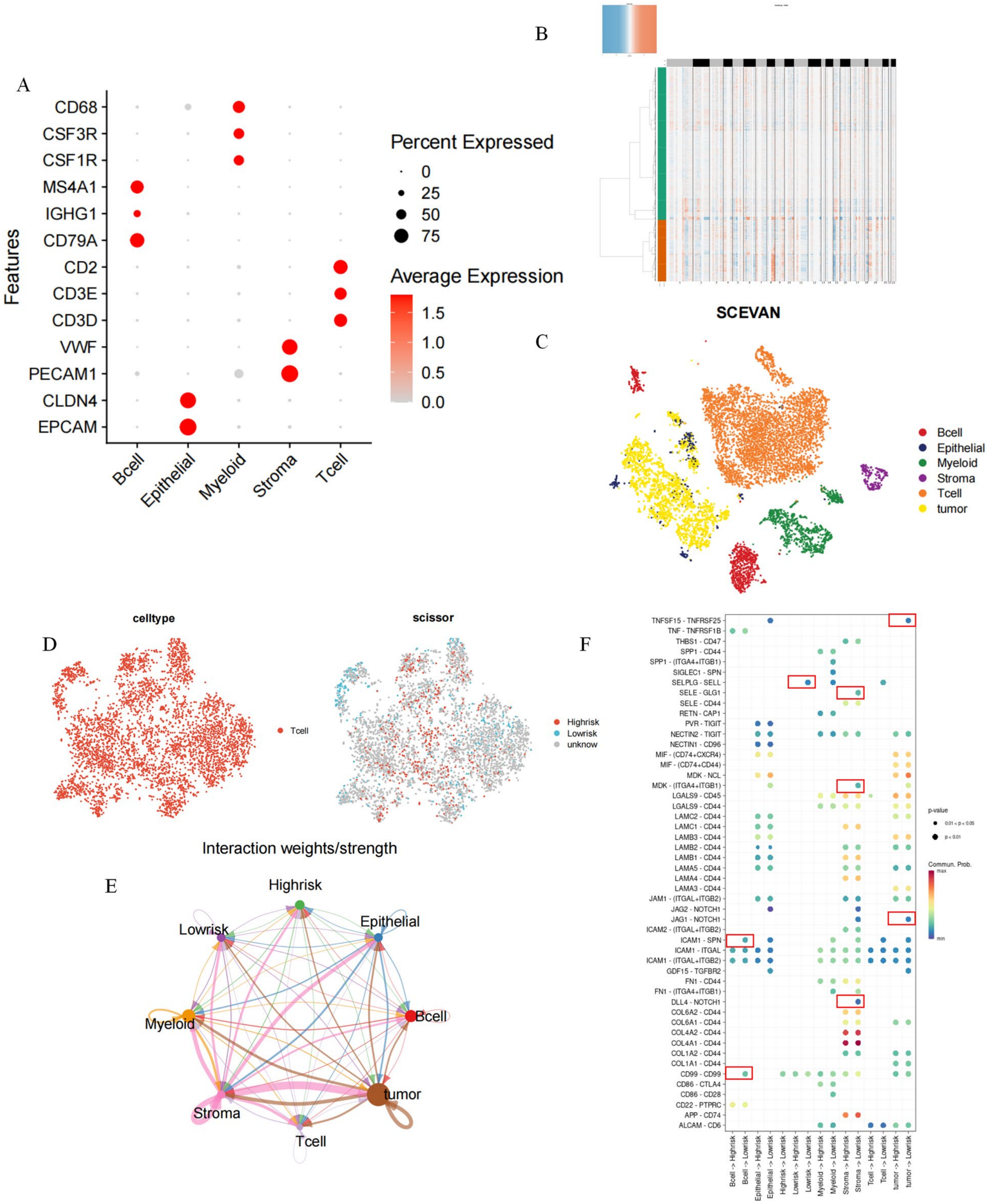

Figure 5. Single-cell analysis of IRS. (**A**) Expression of cellular annotated genes. (**B**) Copy number changes in tumor cells and normal cells. (**C**) Cell types after annotation of all cells. (**D**) Identification of high-risk and low-risk T cells. (**E**) Cellular communication analysis of the landscape. (**F**) Analysis of ligand-receptor interactions between different cell types.

between immune cells. For example, disappearance of DLL4-NOTCH1, MDK-(ITGA4+ITGB1), and SELE-GLG1 signaling between stromal cells and high-risk T cells.Disappearance of ICAM1-SPN signaling with CD99-CD99 signaling in B cells and high-risk T cells.

Discussion

With the application of immunotherapy to GC, the treatment of GC has entered a new era. However, not all patients with GC can benefit from immunotherapy. Several studies have attempted to identify better immune-related characteristic genes that affect the prognosis of patients but have not been very successful[25,26]. As early as last century, it was proposed that the immune microenvironment of gastric cancer is the key factor affecting the prognosis of gastric cancer patients[33]. The level of infiltration of T cells, macrophages and various immune cells affects the prognosis of patients with gastric cancer[34,35]. Since 2019, many studies have been devoted to the establishment of immune related signature for gastric cancer. These signature not only affect the prognosis of gastric cancer patients, but also affect the efficacy of chemotherapy, immunotherapy and other treatments for patients with gastric cancer[36–38]. Therefore, the current study aimed to find the optimal immune-related prognostic signature for patients with GC.

WCGNA analysis was used to identify 14 gene modules. The evaluation of the correlation between each module and the patient's clinical characteristics showed that among all modules, the magenta module had the highest correlation with the subtype. Therefore, 1104 genes in the magenta module were defined as immune subtype-related genes. Subsequently, these 1104 immune subtype-related genes were used for univariate Cox regression analysis, and 52 prognosis-related genes were identified to establish immune-related signatures. In the total cohort, we fitted 98 prognosis prediction signatures. For each signature, we calculated the C-index value in the total, test 1, and test 2 cohorts, and the signature with the highest average C-index value was considered the best signature. Among all the signatures, StepCox [backward] + Enet [alpha = 0.7] showed the highest average C-index value (0.726). This signature, named in this study as IRS, is composed of 24 genes: *TNFAIP2*[39,40], *SLC37A2*, *RGS1*[41], *ZNF101*, *TM6SF1*, *CRHBP*, *AKAP5*[42], *CRYBB1*, *S100Z*[43], *ACSM5*, *NTAN1*[44], *IL5RA*, *ABCG1*, *CAMK4*, *MCEMP1*[45–47], *SLC2A3*[48–50], *RENBP*, *BASP1*[51,52], *KYNU*[53,54], *CTLA4*[55], *FCGR2B*[56], *ENTPD8*, *DDO*, *FCN1*[57]. All of these genes have been mentioned in previous studies to affect the prognosis of patients with GC. Especially, *CTLA4* has been used in the clinic as a target for mature tumor-targeted therapy[58], which indicates the accuracy of our signature.

A large number of prognosis prediction signatures based on machine learning have been reported in recent literature. In terms of research, these signals have diversity, such as pyroptosis, cuproptosis, ferroptosis, EMT conversion, hypoxia, metabolism, aging, and immune response[36,37,59–64]. We collected 93 published prognostic signatures and calculated C-index values for these signatures. Compared with these signatures, the IRS showed a higher C-index value in the total, test 1, and test 2 cohorts, indicating its high accuracy. Higher TMB has been demonstrated to be associated with better prognosis in patients with GC, which is consistent with our findings[65]. Immune interaction is the key feature of tumorigenesis and the therapeutic target of GC. Stromal cells and immune cells are the main components of TME, and immune and matrix scores are related to the clinical features and prognosis of GC[66,67]. Our results also confirmed that tumor immunity is the most important factor affecting the prognosis of GC patients.

Subsequently, we processed the tumor site single-cell sequencing data of patients with GC and identified 2634 malignant tumor cells. Upon cell dimension reduction and visualization, we noticed that the number of T cells is the largest among immune cells. Recent studies emphasize that several types of tumor infiltrating lymphocytes (TIL) are associated with better disease outcomes for various human cancers[68,69], It indicates that more CD3+, CD8+ or CD45RO + T cells in tumor tissue are significantly associated with lower frequency of lymph node metastasis, disease recurrence or longer survival of patients. However, tumors have developed many different strategies to escape immune surveillance, such as loss of tumor antigen expression, expression of Fas ligand (Fas-L) or *CD200* that can induce apoptosis of activated T cells, and immunosuppressive cytokine secretion, such as IL-10 or TGF-β, Or production of regulatory T cells, and downregulation or loss of MHC[70]. The change of HLA class I expression occurs in gastric cancer[71], and may play a role in the clinical process of disease by making tumor cells escape T cell mediated immune response[72]. This intercellular communication may be the main reason for the different prognoses of different patients with GC.

In our study, we used a combination of various machine learning algorithms to construct immune-related prognostic signatures for gastric cancer, and validated the stability and effectiveness of the immune-related signature (IRS) using multiple datasets. We compared IRS with 93 previously published prognostic signatures, and demonstrated that IRS was the most effective prognostic signature. Through the evaluation of IRS, doctors can better understand the patient's prognosis and consider it in their treatment plan, helping to develop more personalized treatment plans and maximize the patient's survival rate. Additionally, we further discovered the predictive value of IRS for the response to immune checkpoint therapy, which is based on the patient's immune gene expression profile and can predict the patient's response to immunotherapy. In clinical practice, the application of IRS can providing more accurate and personalized guidance for patient treatment and management. Importantly, we further revealed cellular communication between high-risk and low-risk T cells at the single-cell level, which provides important reference value for the study of tumor immune microenvironment in gastric cancer patients. However, our study still has some shortcomings. IRS needs to be re-validated in real-world cohorts, and the combination of IRS with clinical features that affect patient prognosis may further improve accuracy. In summary, IRS is a promising tool for clinical prognosis prediction and immune therapy decision-making for GC patients.

Conclusions

We developed an immune-related prognostic signature with reliable validity and high accuracy for clinical use for predicting the prognosis of patients with gastric cancer.

Data availability

Single-cell transcriptome data were obtained from the GEO database (GEO registration number: GSE163558; https://www.ncbi.nlm.nih.gov/geo/). Bulk transcriptome sequencing data were obtained from the TCGA database (https://portal.gdc.cancer.gov/) and GEO database (GSE84437, GSE84433).

Received: 5 December 2022; Accepted: 27 April 2023
Published online: 29 April 2023

References

1. Cao, M., Li, H., Sun, D. & Chen, W. Cancer burden of major cancers in China: A need for sustainable actions. *Cancer Commun. Lond. Engl.* **40**, 205–210 (2020).
2. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
3. Smyth, E. C., Nilsson, M., Grabsch, H. I., van Grieken, N. C. & Lordick, F. Gastric cancer. *Lancet* **396**, 635–648 (2020).
4. Tan, Z. Recent advances in the surgical treatment of advanced gastric cancer: A review. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* **25**, 3537–3541 (2019).
5. Caruso, S. & Scatizzi, M. Laparoscopic gastrectomy for gastric cancer: Has the time come for considered it a standard procedure?. *Surg. Oncol.* **40**, 101699 (2022).
6. Lu, S. *et al.* A phase III trial of neoadjuvant intraperitoneal and systemic chemotherapy for gastric cancer with peritoneal metastasis. *Future Oncol.* **18**, 1175–1183 (2022).
7. Ng, S. P. & Leong, T. Role of radiation therapy in gastric cancer. *Ann. Surg. Oncol.* **28**, 4151–4157 (2021).
8. Nakamura, Y., Kawazoe, A., Lordick, F., Janjigian, Y. Y. & Shitara, K. Biomarker-targeted therapies for advanced-stage gastric and gastro-oesophageal junction cancers: An emerging paradigm. *Nat. Rev. Clin. Oncol.* **18**, 473–487 (2021).
9. Kole, C. *et al.* Immunotherapy for gastric cancer: A 2021 update. *Immunotherapy* **14**, 41–64 (2022).
10. Takei, S., Kawazoe, A. & Shitara, K. The new era of immunotherapy in gastric cancer. *Cancers* **14**, 1054 (2022).
11. Joshi, S. S. & Badgwell, B. D. Current treatment and recent progress in gastric cancer. *CA. Cancer J. Clin.* **71**, 264–279 (2021).
12. Wilke, H. *et al.* Ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (RAINBOW): A double-blind, randomised phase 3 trial. *Lancet Oncol.* **15**, 1224–1235 (2014).
13. Kang, Y.-K. *et al.* Nivolumab in patients with advanced gastric or gastro-oesophageal junction cancer refractory to, or intolerant of, at least two previous chemotherapy regimens (ONO-4538-12, ATTRACTION-2): A randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet* **390**, 2461–2471 (2017).
14. Fuchs, C. S. *et al.* Pembrolizumab versus paclitaxel for previously treated PD-L1-positive advanced gastric or gastroesophageal junction cancer: 2-year update of the randomized phase 3 KEYNOTE-061 trial. *Gastric Cancer* **25**, 197–206 (2022).
15. Smith, R. J. & Bryant, R. G. Metal substitutions incarbonic anhydrase: A halide ion probe study. *Biochem. Biophys. Res. Commun.* **66**, 1281–1286 (1975).
16. Chen, L.-T. *et al.* A phase 3 study of nivolumab in previously treated advanced gastric or gastroesophageal junction cancer (ATT RACTION-2): 2-year update data. *Gastric Cancer* **23**, 510–519 (2020).
17. Chaganty, B. K. R. *et al.* Trastuzumab upregulates PD-L1 as a potential mechanism of trastuzumab resistance through engagement of immune effector cells and stimulation of IFNγ secretion. *Cancer Lett.* **430**, 47–56 (2018).
18. Kono, K., Nakajima, S. & Mimura, K. Current status of immune checkpoint inhibitors for gastric cancer. *Gastric Cancer* **23**, 565–578 (2020).
19. Nishino, M., Ramaiya, N. H., Hatabu, H. & Hodi, F. S. Monitoring immune-checkpoint blockade: Response evaluation and biomarker development. *Nat. Rev. Clin. Oncol.* **14**, 655–668 (2017).
20. Binnewies, M. *et al.* Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **24**, 541–550 (2018).
21. Liu, Y. *et al.* AC010973.2 promotes cell proliferation and is one of six stemness-related genes that predict overall survival of renal clear cell carcinoma. *Sci. Rep.* **12**, 4272 (2022).
22. Wei, X. *et al.* Construction of circRNA-based ceRNA network to reveal the role of circRNAs in the progression and prognosis of metastatic clear cell renal cell carcinoma. *Aging* **12**, 24184–24207 (2020).
23. Wu, D. *et al.* Identification of novel autophagy-related lncRNAs associated with a poor prognosis of colon adenocarcinoma through bioinformatics analysis. *Sci. Rep.* **11**, 8069 (2021).
24. Yu, L. *et al.* Multi-omics analysis reveals the interaction between the complement system and the coagulation cascade in the development of endometriosis. *Sci. Rep.* **11**, 11926 (2021).
25. Yu, M. *et al.* A risk model of eight immune-related genes predicting prognostic response to immune therapies for gastric cancer. *Genes* **13**, 720 (2022).
26. Xu, X. *et al.* A signature of seven immune-related genes predicts overall survival in male gastric cancer patients. *Cancer Cell Int.* **21**, 117 (2021).
27. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
28. Zhang, X. *et al.* Cell Marker: A manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47**, D721–D728 (2019).
29. Sun, D. *et al.* Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. *Nat. Biotechnol.* **40**, 527–538 (2022).
30. Jin, S. *et al.* Inference and analysis of cell-cell communication using cell chat. *Nat. Commun.* **12**, 1088 (2021).
31. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
32. De Falco, A., Caruso, F., Su, X.-D., Iavarone, A. & Ceccarelli, M. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. *Nat. Commun.* **14**, 1074 (2023).
33. Compare, D., Rocco, A. & Nardone, G. Risk factors in gastric cancer. *Eur. Rev. Med. Pharmacol. Sci.* **14**, 302–308 (2010).
34. Wei, M. *et al.* The progress of T cell immunity related to prognosis in gastric cancer. *BioMed Res. Int.* **2018**, 3201940 (2018).
35. Lin, C. *et al.* Tumour-associated macrophages-derived CXCL8 determines immune evasion through autonomous PD-L1 expression in gastric cancer. *Gut* **68**, 1764–1773 (2019).
36. Zeng, D. *et al.* Tumor microenvironment characterization in gastric cancer identifies prognostic and immunotherapeutically relevant gene signatures. *Cancer Immunol. Res.* **7**, 737–750 (2019).
37. Liu, Y. *et al.* Development and validation of a hypoxia-immune-based microenvironment gene signature for risk stratification in gastric cancer. *J. Transl. Med.* **18**, 201 (2020).
38. Qing, X. *et al.* Molecular characteristics, clinical significance, and cancer immune interactions of angiogenesis-associated genes in gastric cancer. *Front. Immunol.* **13**, 843077 (2022).

39. Guo, F. *et al.* Correlation between TNFAIP2 gene polymorphism and prediction/prognosis for gastric cancer and its effect on TNFAIP2 protein expression. *Front. Oncol.* **10**, 1127 (2020).
40. Xu, Y. *et al.* The miR-184 binding-site rs8126 T>C polymorphism in TNFAIP2 is associated with risk of gastric cancer. *PLoS ONE* **8**, e64973 (2013).
41. Zhu, T., Lou, Q., Shi, Z. & Chen, G. Identification of key miRNA-gene pairs in gastric cancer through integrated analysis of mRNA and miRNA microarray. *Am. J. Transl. Res.* **13**, 253–269 (2021).
42. Zhong, Z. *et al.* Low expression of A-kinase anchor protein 5 predicts poor prognosis in non-mucin producing stomach adenocarcinoma based on TCGA data. *Ann. Transl. Med.* **8**, 115 (2020).
43. Wang, C. *et al.* Distinct prognostic roles of S100 mRNA expression in gastric cancer. *Pathol. Res. Pract.* **215**, 127–136 (2019).
44. Zhang, J. *et al.* Bioinformatic analysis of cancer-associated fibroblast related gene signature as a predictive model in clinical outcomes and immune characteristics of gastric cancer. *Ann. Transl. Med.* **10**, 698 (2022).
45. Huang, P., Liu, Y. & Jia, B. The expression, prognostic value, and immunological correlation of MCEMP1 and its potential role in gastric cancer. *J. Oncol.* **2022**, 8167496 (2022).
46. Wang, D. *et al.* MCEMP1 is a potential therapeutic biomarker associated with immune infiltration in advanced gastric cancer microenvironment. *Gene* **840**, 146760 (2022).
47. Hu, G., Sun, N., Jiang, J. & Chen, X. Establishment of a 5-gene risk model related to regulatory T cells for predicting gastric cancer prognosis. *Cancer Cell Int.* **20**, 433 (2020).
48. Yao, X. *et al.* SLC2A3 promotes macrophage infiltration by glycolysis reprogramming in gastric cancer. *Cancer Cell Int.* **20**, 503 (2020).
49. Lin, L. *et al.* Prognostic value of the ferroptosis-related gene SLC2A3 in gastric cancer and related immune mechanisms. *Front. Genet.* **13**, 919313 (2022).
50. Chen, D. *et al.* MicroRNA-129-5p regulates glycolysis and cell proliferation by targeting the glucose transporter SLC2A3 in gastric cancer cells. *Front. Pharmacol.* **9**, 502 (2018).
51. Li, L., Meng, Q., Li, G. & Zhao, L. BASP1 suppresses cell growth and metastasis through inhibiting wnt/β-catenin pathway in gastric cancer. *BioMed Res. Int.* **2020**, 8628695 (2020).
52. Xin, D., Man, Y., Yang, Y. & Wang, F. A novel prognostic and therapeutic target biomarker based on necroptosis-related gene signature and immune microenvironment infiltration in gastric cancer. *Front. Genet.* **13**, 953997 (2022).
53. Zheng, X. *et al.* Construction and analysis of the tumor-specific mRNA–miRNA–lncRNA network in gastric cancer. *Front. Pharmacol.* **11**, 1112 (2020).
54. Gong, Y. *et al.* Development of a prognostic metabolic signature in stomach adenocarcinoma. *Clin. Transl. Oncol.* **24**, 1615–1630 (2022).
55. Katoh, M. Multi-layered prevention and treatment of chronic inflammation, organ fibrosis and cancer associated with canonical WNT/β-catenin signaling activation (review). *Int. J. Mol. Med.* **42**, 713–725 (2018).
56. Tang, W. *et al.* Epstein-barr virus infected gastric adenocarcinoma expresses latent and lytic viral transcripts and has a distinct human gene expression profile. *Infect. Agent. Cancer* **7**, 21 (2012).
57. Khan, M. *et al.* A novel necroptosis-related gene index for predicting prognosis and a cold tumor immune microenvironment in stomach adenocarcinoma. *Front. Immunol.* **13**, 968165 (2022).
58. Chen, Y. *et al.* The immune subtypes and landscape of gastric cancer and to predict based on the whole-slide images using deep learning. *Front. Immunol.* **12**, 685992 (2021).
59. Wei, J., Zeng, Y., Gao, X. & Liu, T. A novel ferroptosis-related lncRNA signature for prognosis prediction in gastric cancer. *BMC Cancer* **21**, 1221 (2021).
60. Huo, J., Wu, L. & Zang, Y. Eight-gene prognostic signature associated with hypoxia and ferroptosis for gastric cancer with general applicability. *Epigenomics* **13**, 875–890 (2021).
61. Cheong, J.-H. *et al.* Development and validation of a prognostic and predictive 32-gene signature for gastric cancer. *Nat. Commun.* **13**, 774 (2022).
62. Chen, W. *et al.* Identification of ferroptosis-related long noncoding RNA and construction of a novel prognostic signature for gastric cancer. *Dis. Markers* **2021**, 7724997 (2021).
63. Cai, W.-Y. *et al.* Identification of a tumor microenvironment-relevant gene set-based prognostic signature and related therapy targets in gastric cancer. *Theranostics* **10**, 8633–8647 (2020).
64. Dai, W. *et al.* Identification of an EMT-related gene signature for predicting overall survival in gastric cancer. *Front. Genet.* **12**, 661306 (2021).
65. Cai, H. *et al.* Mutational landscape of gastric cancer and clinical application of genomic profiling based on target next-generation sequencing. *J. Transl. Med.* **17**, 189 (2019).
66. Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nat. Med.* **19**, 1423–1437 (2013).
67. Pitt, J. M. *et al.* Targeting the tumor microenvironment: removing obstruction to anticancer immune responses and immunotherapy. *Ann. Oncol.* **27**, 1482–1492 (2016).
68. Wang, X. C., Zhang, J. Q., Shen, Y. Q., Miao, F. Q. & Xie, W. Loss of heterozygosity at 6p21.3 underlying HLA class I downregulation in gastric cancer. *J. Exp. Clin. Cancer Res.* **25**, 115–119 (2006).
69. Takahashi, A. *et al.* Elevated caspase-3 activity in peripheral blood T cells coexists with increased degree of T-cell apoptosis and down-regulation of TCR zeta molecules in patients with gastric cancer. *Clin. Cancer Res.* **7**, 74–80 (2001).
70. Dunn, G. P., Old, L. J. & Schreiber, R. D. The three Es of cancer immunoediting. *Annu. Rev. Immunol.* **22**, 329–360 (2004).
71. Galon, J. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313**, 1960–1964 (2006).
72. Gao, Q. *et al.* Intratumoral balance of regulatory and cytotoxic T cells is associated with prognosis of hepatocellular carcinoma after resection. *J. Clin. Oncol.* **25**, 2586–2593 (2007).

Author contributions

J.N., K.S., X.F. and X.W. conceived and wrote the paper. J.N., K.S., X.F., X.W., K.J., L.M., X.W., H.L., R.M., S.L. and F.L. analyzed the materials and drafted the manuscript. Xiaofeng Wang revised the whole paper. All authors have reviewed the final version of the manuscript and approved to submit to your journal. All authors have agreed to publish this manuscript.

Funding

The authors did not receive support from any organization for the submitted work.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.W.

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

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