

Development of T-Cell-Associated Signatures Based on Machine Learning Integration for Predicting Breast Cancer Prognosis

Yan Liu

Shaoguan First People's Hospital, Affiliated Hospital of Shaoguan University

Yuxin Cheng

The Second Affiliated Hospital of Shantou University Medical College

Ting Huang

Shaoguan First People's Hospital, Affiliated Hospital of Shaoguan University

Liang Zhang

Shaoguan First People's Hospital, Affiliated Hospital of Shaoguan University

Zhenwei Yang

Shaoguan First People's Hospital, Affiliated Hospital of Shaoguan University

Mengting Zeng

Shaoguan First People's Hospital, Affiliated Hospital of Shaoguan University

Zhiyang Li

The Second Affiliated Hospital of Shantou University Medical College

Hai Lu

hysa1985@163.com

Shaoguan First People's Hospital, Affiliated Hospital of Shaoguan University

Research Article

Keywords: Breast cancer, immune cell infiltration, single cell sequencing, immunotherapy

Posted Date: September 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10662868/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Breast cancer is the most prevalent cancer worldwide, with prognosis worsening upon progression. However, the role of T cells in breast cancer remains unclear. In this study, we integrated single-cell RNA sequencing and bulk RNA-seq data to identify T-cell-associated signatures (TRS) using multiple machine-learning approaches, and validated them in immunotherapy cohorts. TRS demonstrated strong robustness across seven independent datasets and outperformed conventional clinical traits in predicting prognosis. Patients with low TRS exhibited abundant immune cell infiltration and higher tumor mutational burden. Notably, across three immunotherapy cohorts, low-TRS patients showed better prognosis and greater sensitivity to immunosuppressive therapy. In conclusion, this study developed a robust T-cell-related signature that may facilitate personalized treatment and prognostic prediction in breast cancer.

INTRODUCTION

According to Global Cancer Statistics 2022 breast cancer has become the most common cancer (both in terms of incidence and mortality) in women, with an estimated 2.3 million new cases, accounting for 11.6% of all cancer cases[1]. In China, the incidence of breast cancer is increasing year by year, and has shown a youthfulness[2], and it is the main cause of death for women under 45 years old[3]. As the early symptoms of breast cancer are not obvious, most breast cancer patients have already developed metastasis by the time they visit the doctor, with bone metastasis being the most common site of metastasis[4, 5], which makes the treatment of breast cancer a great challenge.

Single-cell sequencing is a technology that amplifies and sequences DNA or RNA at the single-cell level to address rare cell types, heterogeneous samples, and unculturable microorganisms, and to visualize cellular dynamics at the single-cell level[6]. In breast cancer, single-cell sequencing has been widely used to study the heterogeneity of breast cancer cells, differentiation, EMT and regulatory information on proliferation[7]. Analysis of 1,000 single-cell sequencing of 12 patients with triple-negative breast cancer yielded one to three clonal subpopulations in each tumor, and these subpopulations share a common evolutionary pattern independent of copy number events[8]. Single-cell sequencing of paclitaxel-resistant breast cancer cells showed that resistant cells exhibited higher variability and that single-cell sequencing could track the activation of cancer cell heterogeneity after resistant cells resumed proliferation[9]. Patel et al. performed single-cell sequencing analysis of five primary glioblastomas and found that individual cells differed from genes involved in oncogenic signaling, immune response, proliferation, and other pathways, and most importantly, that the tumor cells exhibited a range of stemness characteristics[10].

Tumor microenvironment refers to the internal environment in which tumor cells arise and live, including tumor cells and their surrounding fibroblasts, immune and inflammatory cells, glial cells, as well as intercellular stroma, microvessels, and biomolecules infiltrated therein, which are characterized by hypoxia, chronic inflammation, and immunosuppression, and which form a very complex network of mechanisms that play an important role in tumor development[11]Time-resolved single-cell analysis of

BRCA1 and P53 knockout mice in a transgenic mouse model of breast cancer identifies the effect of an immunosuppressive environment on tumorigenesis by the specific mechanism of Tregs cell retention and accumulation in tissues[12], Analysis of ductal carcinoma in situ using single-cell technology and multiple spatial analysis showed that compared with normal breast tissues, tumor tissues had a higher total number of leukocytes, more neutrophils, and a significantly lower CD4/CD8 ratio, and, more importantly, a decrease in activated CD8 T cells, an increase in PD-L1,CTLA4 cells expressing immune checkpoint genes, and a higher number of regulatory T cells during the progression to invasive ductal carcinoma[13]. Recruitment of CD206/Tie2 macrophages by CCL2 secreted by cancer and bone marrow cells in precancerous lesions was found to lead to upregulation of Wnt-1, which in turn downregulates E-cadherin and promotes tumor metastasis in a HER2 + mouse model of breast cancer[14]. A mechanism by which macrophages promote cancer cell metastasis was also identified in another mouse model of breast cancer; specifically, Wnt molecules produced by macrophages promote vascular remodeling and facilitate cancer cell metastasis[15].

Successful application of immune checkpoints ignites new hope for solid tumor treatment[16–20]. Ipilimumab, an immunosuppressant of cytotoxic T-lymphocyte-associated antigen (CTLA-4), blocks CTLA-4 to promote anti-tumor immunity in T cells and was approved by the Food and Drug Administration (FDA) in 2011 for the treatment of patients with advanced melanoma[21], Pembrolizumab, a monoclonal antibody against programmed death 1 (PD-1), was found to be superior to platinum-based combination chemotherapy in advanced non-small-cell lung cancer in a randomized clinical study, significantly prolonging progression-free survival and overall survival[22], In a phase 3 clinical trial of patients with triple-negative breast cancer, the combination of atezolizumab and albumin-conjugated paclitaxel significantly prolonged median overall survival relative to placebo plus albumin-conjugated paclitaxel[23].

Here, we first depicted two single-cell datasets of BRCA and established two subtypes based on the T-cell subpopulation differential genes that differed most significantly in tumors versus metastasis; in addition, we analyzed two T-cell subpopulation differential genes and constructed a signature of T-cell-associated genes based on 101 combinations of machine-learning methods, which were validated for their prognostic value in the TCGA and GEO datasets, and their clinical value in the immunotherapy dataset.

MATERIALS AND METHODS

Public data collection and organization

Single-cell data on lymph node metastasis in breast cancer from the Gene Expression Omnibus (GEO) database. where GSE180286[24] includes 5 primary breast cancer tissues and paired 2 lymph node metastatic tissues, GSE158399[25] contains a primary tumor and matching lymph node metastatic sites in a breast cancer patient. We analyzed single-cell data with the R package Seurat[26], The SCTransform method was then used to eliminate the batch effect of samples from both datasets and to

simultaneously find 2000 highly variable genes for downstream analysis. Based on these 2000 highly variable genes, we proceeded to analyze the data by dimensionality reduction using principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE), FindNeighbors and FindClusters functions were used for cluster analysis. We used the R package TCGAbiolinks[27] to download breast cancer transcriptome data and clinical data from The Cancer Genome Atlas (TCGA) database, excluding incomplete clinical information, 1016 samples were included in the later study. Additional transcriptome data were downloaded from the GEO database, The GSE1456[28] dataset contains 159 breast cancer tissues, the GSE20685[29] dataset contains 327 breast cancer tissues, the GSE24450[30] dataset contains 183 breast cancer tissues, the GSE42568[31] dataset contains 104 breast cancer tissues and 17 normal breast tissues, and the GSE58812[32] dataset contains 107 breast cancer tissues, the GSE7390[33] dataset contains 196 breast cancer tissues. Immunotherapy Cohort GSE135222[34] Dataset Contains 27 Small Cell Lung Cancer Patients Receiving PD-1/PD-L1 Therapy, GSE78220[35] dataset contains 28 melanoma patients treated with PD-1, IMvigor210 cohort data were downloaded from <http://research-pub.gene.com/IMvigor210> CoreBiologies.

Consensus Clustering

Based on the results of cell annotation, the cell types that differed the most in tumor vs. tumor metastasis were selected for differential analysis. Cluster analysis in the TCGA-BRCA cohort based on differential genes using the R package ConsensusClusterPlus[36].

Immune infiltration analysis

Enrichment scores for each immune cell per sample in the TCGA-BRCA cohort were assessed by the method of single sample gene set enrichment analysis (ssGSEA) in the R-package GSVA[37] according to 28 immune cell types, The ESTIMATE[38] method was used to assess the immune cell score, stromal cell score, total score, and tumor purity for each sample.

Signature generation based on multiple machine learning methods

Based on the results of cluster analysis grouping differential genes, genes with prognosis were first screened with Univariate COX analysis, followed by the generation of signatures using 10 different machine learning algorithms, and 101 combinations of these 10 machine learning algorithms, These 10 machine learning algorithms include random survival forest (RSF), elastic network (Enet), Lasso, Ridge, stepwise Cox, CoxBoost, partial least squares regression for Cox (plsRcox), supervised principal components (SuperPC), generalised boosted regression modelling (GBM), and survival support vector machine (survival-SVM). All models are validated in six other datasets (GSE1456,GSE20685,GSE24450,GSE42568,GSE58812,GSE7390), Finally, the Harrell's concordance index (C-index) was calculated for each dataset, and the machines with the highest average C-index were counted for inclusion in the follow-up study according to their average C-index. Finally, the accuracy of

the signature was assessed using the area of the ROC curve and the relationship to the clinical subtype in each dataset.

Relationship between signature and tumor microenvironment

We collected immune-related features, including chemokines, immunomodulators, MHC-related genes, and receptors, and observed their expression in high and low signatures. The cancer immune cycle influences the body's immune response to tumor cells and consists of seven steps: Release of cancer cell antigens (Step 1), Cancer antigen presentation (Step 2), Priming and activation (Step 3), trafficking of immune cells to tumors (Step 4),

Infiltration of immune cells into tumors (Step 5), Recognition of cancer cells by T cells (Step 6), Killing of cancer cells (Step 7), Step 4 in turn includes CD4 T cell recruiting, CD8 T cell recruiting, Dendritic cell recruiting, Eosinophil recruiting, Macrophage recruiting, MDSC recruiting, Monocyte recruiting, Neutrophil recruiting, NK cell recruiting, T cell recruiting, TH1 cell recruiting, TH17 cell recruiting, TH2 cell recruiting, TH22 cell recruiting [39]. We used the ssGSEA method to assess the activity of each step for each sample according to the expression of the gene. These activities were then calculated separately between high and low risk groups. We proceeded to download the immune-related gene module from the website <https://www.immport.org/home>, again using the ssGSEA method to calculate the immune-related score for each sample. CIBERSORT [40] is based on the principle of linear support vector regression for back-convolution of transcriptome expression matrices to estimate immune cell composition and abundance in mixed cells. We uploaded the gene expression matrix data (TPM) to CIBERSORT and combined it with the LM22 feature gene matrix to derive the immune cell infiltration matrix. The correlation of the above scores was calculated separately from the risk score. Finally we looked at some common immune cells like CD8T cells, dendritic cells, macrophages, NK cells, and helper T cells separately between the high and low risk groups. We also went on to look at the correlation of some immune checkpoint genes with risk scores.

Multi-omics analysis of signatures

Analysis of somatic mutations in high and low risk groups using the maftools [41] package to calculate tumor mutation compliance (TMB), and observe the genes that are most significantly mutated between risk groups.

The TIDE (Tumor Immune Dysfunction and Exclusion) [42] score reflects how the tumor suppresses cytotoxic T lymphocyte function and infiltration to achieve immune escape. We entered breast cancer expression data into a website <http://tide.dfci.harvard.edu/> to obtain the TIDE score, MDSC, and analyzed differences in high and low risk groups.

Significance of signatures in immunotherapy cohorts

We download expression profiles and clinical information from the GEO database for immunotherapy cohort-related datasets. Mainly study the relationship between risk signatures and various clinical information in non-small cell lung cancer, melanoma, and uroepithelial cancer.

Statistical analysis

All data processing, statistical analysis and graphing are done with R4.2.0. The Wilcoxon test was used to compare two groups, while the Kruskal-Wallis test was used to compare more than two groups. Survival analysis was performed by Kaplan-Meier logit test. The chi-square test or Fisher exact test was used to compare and analyze the statistical significance between two groups of categorical variables, and Spearman's analysis was used to calculate the correlation between coefficients. A two-sided $P < 0.05$ was considered statistically significant, Other differences are significant as follows: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; not significant, $p > 0.05$.

RESULTS

Single-cell landscape of lymph node metastasis versus non-metastasis in breast cancer

Two single-cell datasets on breast cancer metastasis, GSE158399 and GSE180286, were included in the study, and after initial screening, 16,065 cells in the GSE158399 dataset and 20,984 cells in the GSE180286 dataset were advanced to the next step of the study. We apply the method of SCTransform to eliminate the batch-to-batch differences between the two datasets, and then use the tSNE algorithm to analyze the cells by dimensionality reduction, and the results of the dimensionality reduction are shown in Fig. 1A, 1B, from which it can be seen that the batch-to-batch differences between the two datasets have been eliminated. The highly variable genes were then clustered using the FindClusters algorithm, and the cells were grouped into 13 clusters (Fig. 1C), and the cells were annotated using CellMarkers, which showed that clusters of 0, 6, and 12 cells were annotated as tissue stem cells, clusters of 1 and 4 cells were annotated as B cells, clusters of 2, 5, and 11 cells were annotated as T cells, cluster of 3 cells was annotated as epithelial cells, cluster of 8 cells was annotated as macrophages, clusters of 9 and 10 cells were annotated as chondrocytes, and cluster of 7 cells was not annotated to any cell type(Fig. 1D), MFAP5 gene was mainly expressed in tissue stem cells and highly expressed in breast cancer metastatic tissues, CD79A gene was mainly expressed in B cells and highly expressed in breast cancer tissues, CCL5 gene was mainly expressed in T cells and highly expressed in breast cancer metastatic tissues, PPDPF gene was mainly present in epithelial cells and highly expressed in breast cancer tissues, CD14 gene was mainly present in macrophages and highly expressed in breast cancer metastatic tissues, and COL1A1 gene was mainly present in chondrocytes and highly expressed in breast cancer tissues (Fig. 1E). Finally, we observed the differences between the above six cell types in tumor tissues and metastatic tissues, and found that all cells differed in tumor tissues and metastatic tissues, with a significantly higher proportion of T cells in metastatic tissues than

in tumor tissues ($p = 8.72E-181$), and a significantly higher proportion of chondrocytes in tumor tissues than in metastatic tissues ($p = 1.79E-316$) (Fig. 1F). Since we are more concerned with the metastatic mechanisms of tumors, the differential genes of T cells between tumors and metastasis were included in the next study.

Identification of T cell-associated subtypes

The above 534 differential genes (screening criteria were absolute value of logFC greater than 0.5 and corrected p-value less than 0.05) were first analyzed by one-way cox regression to obtain 64 prognostically significant genes, which were subjected to consensus cluster analysis[36]. All the breast cancer samples were finally categorized into k ($k = 2-9$) classes, and according to proportion of ambiguous clustering (PAC), the best classification was obtained when $k = 2$ (Fig. 2A,B). The two consensus clusters (C1 and C2) also differed in terms of immune cell infiltration, broadly speaking, C2 had a higher immune infiltration score than C1 (Fig. 2C), specifically C2 had a higher score of activated B-cells than C1 ($p = 2.0E-09$), C2 had a higher score of activated CD8 T-cells than C1 ($p = 1.3E-15$), and C2 had a higher CD56bright natural killer cell score was higher than C1 ($p = 9.5E-10$), C2's CD56dim natural killer cell score was higher than C1 ($p < 2E-16$), C2's central memory CD4T cell score was higher than C1 ($p = 7.7E-05$), C2's effector memory CD4T cell score was higher than C1 ($p = 0.00041$), C2's eosinophil score was higher than C1 ($p = 3.7E-07$), C2 had a higher Gamma delta T cell score than C1 ($p = 7.6E-05$), Macrophage scores were higher in C2 than C1 ($p = 2.7E-11$), MDSC scores were higher in C2 than C1 ($p = 4.5E-09$), monocyte scores were higher in C2 than C1 ($p < 2E-16$), natural killer cell scores were higher in C2 than C1 ($p = 0.00014$), neutrophil scores were higher in C2 than C1 ($p = 0.02208$), and the C2 follicular helper T cell score was higher than C1 ($p = 0.00058$), C2 had a higher type I helper T cell score than C1 ($p = 3.1E-06$), C2 had a higher type 17 helper T cell score than C1 ($p = 9.7E-05$), While C1 central memory CD8 T cell score was higher than C2 ($p = 6.3E-09$), C1 effector memory CD4 T cell score was higher than C1 ($p = 8.9E-09$), C1 immature dendritic cell score was higher than C2 ($p = 7.4E-10$), C1 memory B cell score was higher than C2 ($p = 3.0E-12$) (Fig. 2D), Finally, we compared the difference in survival between the C1 and C2 clusters, and it can be seen that the overall mortality of C2 was significantly lower than that of C1 (Fig. 2E).

Construction of T cell related signatures(TRS) based on machine learning

We firstly analyzed the C1 and C2 clusters in the TCGA-BRCA dataset for differential analysis (the screening criterion is the absolute value of logFC > 1 , $p < 0.05$) to get 419 differential genes, and the differential genes were firstly subjected to the one-way cox regression analysis to get 52 genes with prognostic significance, and finally, we imported these 52 genes into the 101 combinatorial machine learning methods to get the optimal model, and validated in all datasets according to the C-index, The heatmap shows that the optimal model is StepCox(forward)+Ridge, which has the highest average C-index (0.682), and the C-index for all datasets is the best performing in all datasets (Fig. 3A). Then we

calculated the risk score of each patient based on the optimal model, and then classified the patients into high and low risk groups based on the median risk score. In the GSE1456 dataset, the overall survival rate of the high risk group was lower than that of the low risk group ($p = 0.00089$, Fig. 3B), in the GSE20685 dataset, the overall survival rate of the high risk group was lower than that of the low risk group ($p < 0.0001$, Fig. 3C), in dataset GSE24450, the total survival rate in the high-risk group was lower than that in the low-risk group ($p = 0.0025$, Fig. 3D), in dataset GSE42568, the total survival rate in the high-risk group was lower than that in the low-risk group ($p = 0.049$, Fig. 3E), in dataset GSE58812, the total survival rate in the high-risk group was lower than that in the low-risk group ($p = 0.0014$, Fig. 3F), in the dataset GSE7390 dataset, the overall survival rate was lower in the high-risk group than in the low-risk group ($p = 0.0011$, Fig. 3G), and in the TCGA-BRCA dataset, the overall survival rate was lower in the high-risk group than in the low-risk group ($p < 0.0001$, Fig. 3H).

Validation of the TRS model

ROC curves are used to analyze the effectiveness of the model, In GSE1456, the AUC values for 1-, 3-, and 5-years were 0.704, 0.688, and 0.651 (Fig. 4A), in GSE20685, the AUC values for 1-, 3-, and 5-years were 0.839, 0.722, and 0.711 (Fig. 4B), and in GSE24450, the AUC values for 1-, 3-, and 5-years were 0.799, 0.796, 0.709 (Fig. 4C), in GSE42568, the AUC values for 1-, 3-, and 5-years were 0.535, 0.604, 0.674 (Fig. 4D), in GSE58812, the AUC values for 1-, 3-, and 5-years were 0.906, 0.727, and 0.743 (Fig. 4E), and in GSE7390, the 1-, 3-, 5-year AUC values were 0.651, 0.674, 0.694 (Fig. 4F), and in TCGA, 1-, 3-, 5-year AUC values were 0.669, 0.769, 0.758 (Fig. 4G).

Then we compared the C-index of TRS with other clinical traits, as shown in the figure, in GSE1456, the accuracy of TRS was significantly better than subtype (Fig. 5A), in GSE20685, the accuracy of TRS was also better than age and M staging (while there was no statistically significant difference in N staging, and T staging, but the C-index of TRS was still higher than that of N,T staging) (Fig. 5B), in GSE7390, the accuracy of TRS was significantly better than age and staging (Fig. 5C), in TCGA, the accuracy of TRS was significantly better than age, sex, M,N,T staging, race, and staging (Fig. 5D), and there was no statistically significant difference between TRS and clinical trait comparisons in GSE42568, GSE58812 (Supplementary Fig. 1A,B). (Clinical trait information was not available for the GSE24450 dataset), We next analyzed the relationship between TRS values and clinical traits in each dataset specifically. In GSE1456, there was a statistically significant difference in TRS for all subtypes ($p = 1.1E-05$, Fig. 5E), In GSE20685, the TRS in M1 was significantly higher than that in M0 ($p = 0.0026$, Fig. 5F), In GSE20685, as T staging went up, TRS significantly followed ($p = 0.00044$, Fig. 5G) (In GSE20685, although there was no difference in TRS in N staging, we also observed that there was an elevation in TRS with increasing N staging, Supplementary Fig. 1C), Similarly, in GSE42568, the higher the grading, the higher the TRS was ($p = 0.0061$, Fig. 5H), In GSE7390, also the TRS increased with increasing grading ($p = 7.4E-05$, Fig. 5I), in TCGA, the difference in TRS between M stages was significant ($p = 0.0095$, Fig. 5J), in T stage, the TRS was highest in T4 and lowest in T1 ($p = 7E-07$, Fig. 5K), in N stage, with increasing number of lymph node metastases, the TRS subsequently increased ($p = 0.02$, Fig. 5L), and in staging, as the staging increased, the TRS was correspondingly higher ($p = 0.0014$, Fig. 5M).

We continued to investigate whether TRS was an independent prognostic factor, in GSE1456, one-way COX regression analysis showed that subtype and TRS were independent prognostic factors (Supplementary Fig. 2A), and multifactorial cox regression analysis showed that subtype, TRS remained an independent prognostic factor (Fig. 6A), and in GSE20685, one-way cox regression analysis showed that M,N,T, TRS were independent prognostic factors (Supplementary Fig. 2B), multifactorial cox regression analysis showed that N,TRS were independent prognostic factors (Fig. 6B), in GSE42568, unifactorial cox regression analysis showed that grading and TRS were independent prognostic factors (Supplementary Fig. 2C), and multifactorial cox regression analysis showed that grading and TRS remained independent prognostic factors (Fig. 6C), the In GSE58812, one-way cox regression analysis showed that age and TRS were independent prognostic factors (Supplementary Fig. 2D), and multifactor cox regression analysis showed that age and TRS remained independent prognostic factors (Fig. 6D), and in GSE7390, one-way cox regression analysis showed that only TRS was an independent prognostic factor (Supplementary Fig. 2E), and multifactor cox regression analysis showed that TRS was an independent prognostic factor (Fig. 6E), in TCGA, one-way cox regression analysis showed that age, TRS,N,T,staging were independent prognostic factors (Supplementary Fig. 2F), and multifactorial cox regression analysis showed that TRS and age, staging were independent prognostic factors (Fig. 6F)

Immunoassay of TRS in BRCA

In order to analyze the relationship between TRS and immunomodulation, we analyzed the relationship between chemokines, immunomodulators, MHC molecules, and receptors in the high and low TRS groups, and the results showed that the expression of most of the MHC molecules was significantly higher in the low than in the high TRS group, suggesting that the antigen-presenting ability of the low TRS group was higher than that of the high TRS group. Monocyte macrophage chemokines CCL5, CCL7, CCL13, CCL17 have the ability of monocyte macrophage to act at the site of inflammation, and the expression of these molecules in the low TRS group was also higher than that in the high TRS group. CD40 can activate B cells after binding to CD40 ligand, and the interaction of CD86 with the inducer CD28 and inhibitor CTLA4 is a key factor for inducing the proliferation and the production of ICDs by T-lymphocytes. lymphocyte proliferation and IL-2 production, IL6 binding to IL6 receptor can stimulate B, T cell proliferation. Tumor necrosis factor family TNFSF can cause tumor cell necrosis and apoptosis by binding to TNFRSF, TNFSF18, TNFSF13, TNFSF14, TNFRSF17, TNFRSF13B.

TNFRSF4, TNFRSF14, TNFRSF25, TNFRSF9, TNFRSF8 These molecules were all elevated in the low TRS group (Fig. 7A). The cancer immune cycle (CIC) consists of a series of events required for immune-mediated control of tumor growth, and this cycle can be divided into seven major steps, beginning with the release of antigens from cancer cells and ending with the killing of cancer cells[39]. In the low TRS group, cancer immune cycle activity was elevated, including cancer cell antigen release (step 1, $p = 0.00068$), tumor antigen presentation (step 2, $p = 2.7E-07$), initiation and activation (step 3, $p = 1.5E-10$), immune cell delivery to the tumor (step 4) (B cell recruitment, CD4T cell recruitment, CD8T cell recruitment, dendritic cell recruitment, macrophage recruitment, monocyte recruitment, neutrophil recruitment, NK cell recruitment, T cell recruitment, TH1 cell recruitment, TH17 cell recruitment, TH2 cell

recruitment, regulatory cell recruitment), T cell recognition of cancer cells (step 6, $p = 0.043$), killing of cancer cells (step 7, $p = 2.5E-09$) (Fig. 7B). We then analyzed the relationship between TRS and immune cells and immune-related pathways. TRS was negatively correlated with all immune-related pathways, with the strongest correlation being chemokine receptors (correlation of -0.34), and TRS was also negatively correlated with the majority of immune cells, with the exception of type 2 T helper cells and Gamma delta T cells, with the strongest correlation being eosinophils (correlation of -0.38) (Fig. 7C). We went on to investigate the expression of signature genes of CD8T cells, dendritic cells, macrophages, NK cells, and TH1 cells, which are the main effector cells of tumor immune-infiltrating cells, in the high and low TRS groups, and showed that the majority of immune-infiltrating cells were highly expressed in the low TRS group (Fig. 7D). Immune checkpoints play an important role in organismal escape immunosurveillance[20], We investigated the correlation between TRS and common immune checkpoints and showed that TRS was negatively correlated with most of the immune checkpoint genes (Fig. 7E), and the expression of immune checkpoint genes was higher in the low TRS group than in the high TRS group, with the best-known genes, PD-1 ($p = 3.39E-16$), PD-L1 ($p = 3.56E-06$), CTLA-4 ($p = 0.0003$), LAG-3 ($p = 0.002$), TIM-3 ($p = 0.03$), TIGIT ($p = 1.25E-07$) (Supplementary Fig. 3D)

Multi-omics analysis of high and low TRS groups

In order to further handover the function of high and low TRS groups, we performed a multi-omics analysis of the high and low TRS groups, and the mutation analysis showed that the most significantly mutated genes in the low TRS group were PIK3CA, CDH1, TP53, TTN, GATA3, and the mutation frequencies were 43%, 22%, 20%, 14%, and 11%, respectively (Fig. 8A). The most significantly mutated genes in the high-TRS group were TP53, PIK3CA, TTN, GATA3, and MUC16, with mutation frequencies of 49%, 25%, 20%, 14%, and 10%, respectively (Fig. 8B), and there was a significant covariance between these mutated genes (Fig. 8C). TIDE analysis showed that TIDE was significantly higher in the low TRS group than in the high TRS group ($p = 1.4E-36$, Fig. 8D), whereas MDSC was significantly higher in the high TRS group than in the low TRS group ($p = 4.2E-29$, Fig. 8E). Tumor mutation load showed that TMB was significantly higher in the high TRS group than in the low TRS group ($p = 7.3E-10$, Fig. 8F).

Predictive value of TRS in immunotherapy cohorts

To investigate the value of TRS in immunotherapy, we found that in GSE78220, GSE135000, IMVigor210, the overall survival rate of the high TRS group was lower than that of the low TRS group ($p < 0.0001$, $p = 0.0034$, $p < 0.0001$, Fig. 9A-C), and ROC analysis demonstrated that TRS accurately predicted overall survival, where the AUC values for 0.5, 1, and 1.5 years were 0.855, 0.904, and 0.973, respectively, in GSE78220 (Fig. 9D), and 0.886, 0.977, respectively, for 0.5, and 1 year in GSE135222 (Fig. 9E), and in IMVigor210, the AUC values for 0.5, 1, and 1.5 years were 0.684, 0.720, and 0.754, respectively (Fig. 9F). Finally, we analyzed the relationship between clinical subtype information and TRS, and the results showed that in GSE78220, the TRS was significantly lower in complete remission (CR) than in progressive disease (PD) ($p = 0.0017$), and significantly lower in partial remission (PR) than in progressive disease (PD) ($p = 0.00047$, Fig. 9G), and in GSE135222, the TRS of the responder group had a

significantly lower TRS than the non-responding group ($p = 0.0002$, Fig. 9H), in IMVigor210, the TRS of the immunoinflammatory type (Inflamed) was significantly lower than that of the immune desert type (Desert) ($p = 2.1 \times 10^{-5}$), the TRS of the immunoinflammatory type (Inflamed) was significantly lower than that of the immune barrier type (Excluded) ($p = 0.0011$, Fig. 9I), CR/PR was significantly higher in the low TRS group than in the high TRS group, and SD/PD was significantly lower in the low TRS group than in the high TRS group ($p < 0.001$, Fig. 9J).

DISCUSSION

Tumor staging plays an important role in the treatment of breast cancer, which not only determines the surgical approach to breast cancer, but also has great significance in guiding postoperative treatment and judging prognosis[43]. However, breast cancer is again one of the most heterogeneous tumors[44], which dictates that relying solely on tumor staging may delay patient treatment, and to address this issue, we analyzed T-cell related signature profiles, prognosis, and applications in immunotherapy.

In this study, we first identified the T-cell populations that changed most significantly in tumor and metastasis from breast cancer single-cell data, and constructed TRS based on the expression profiles of subtype-differentiated genes using a combination of 101 machine learning methods. validated in 6 independent datasets, yielding the optimal model as stepwise COX (forward) and ridge regression, and ROC and C-index analyses demonstrated that TRS in the 7 datasets maintained high accuracy and could be used as a breast cancer prognostic indicator

Two clusters were identified based on the differential gene of T cells between metastases and tumors, and the C2 cluster had a better prognosis, probably due to the high immune infiltration score of the C2 cluster, which is classified as a "hot" tumor from an immunological point of view, and tumors infiltrated by a large number of T cells are defined as "hot". Tumors infiltrated by a large number of T cells are defined as "hot" tumors[45], It has been shown that radiotherapy can transform "cold" tumors to "hot" tumors by recruiting T cells to infiltrate "cold" tumors, resulting in the transformation of "cold" tumors to "hot" tumors. Refractory tumors are sensitive to immune checkpoint inhibitors, improving patient prognosis[46, 47]. Immune infiltration analysis between the two clusters showed that the majority of immune cells were highly enriched in C2, confirming that C2 tends to be a "hot" tumor with a higher prognosis than C1.

It is well known that aging is one of the major risk factors for the development of breast cancer, with studies showing that the average age of diagnosis of breast cancer in the United States is 62, and that the majority of deaths from breast cancer occur at ages 65 and above[48], Molecular typing of breast cancer is also fundamental to the diagnosis and treatment of breast cancer and plays an important role in personalized therapy[49]. Tumor grading is also an important prognostic factor in breast cancer, and the latest AJCC staging manual, which incorporates grading as a prognostic criterion, remains a key pathologic feature in the diagnosis and treatment of breast cancer.[50]Our TRS was superior to age,

staging and grading in predicting prognosis, and the TRS scores were statistically different in staging, grading M staging, T staging, which side by side illustrates the robustness of the TRS we constructed.

In addition to assessing the robustness of TRS, we also made an in-depth investigation in the clinical application of TRS. First, both unifactorial and multifactorial cox regression analyses showed that TRS can be used as an independent prognostic factor. In studies that collected tumor endpoints, mutant tumors were larger and heavier than controls, possibly because mutant tumors had fewer stromal fibroblasts and myofibroblasts. Our study showed that the mutation frequency in the high TRS group was higher than that in the low TRS group, suggesting that the genome of the high TRS group is unstable, and the prognosis was worse. And the frequency of TP53 mutation was significantly higher in the high TRS group than in the low TRS group. It is well known that TP53-mutated tumors are more aggressive than unmutated tumors[51], and the mechanism may be that TP53 mutations are associated with the formation of TPYM signaling complexes, which are implicated in a variety of tumorigenesis[52]. Tumor mutation load is the total number of, somatic gene coding errors, base substitutions, gene insertion or deletion errors detected per million bases, and some studies have shown that high mutation load is associated with a poor prognosis in ER+ breast cancer[53], and some investigators have used tumor mutation load to monitor antiangiogenic therapy in animal models of breast cancer lung metastasis, and the results showed that if the antiangiogenic therapy is ineffective, the tumor load of the lung will be significantly increased[54], and our results showed that patients with high TRS had higher tumor mutation load, indicating a worse prognosis. We then went on to analyze the immune infiltration in patients with high and low TRS. Patients with low TRS showed higher immune cell infiltration and expression of antigen-presenting cell genes, presenting a "hot tumor" phenotype and a better prognosis, which is consistent with the finding that high TMB and immune cell infiltration in patients with breast cancer ("hot tumor") is associated with prolonged survival in breast cancer patients[55], In contrast, patients with high TRS have higher tumor purity and lower immune cell infiltration, poorer prognosis.

Tumor immunotherapy is currently an active area of oncology research[56], and it is no exception in breast cancer, with studies confirming that CIK immunotherapy improves the prognosis of breast cancer patients, and for the first time, PD-L1 expression has been found to serve as an indicator of postoperative CIK adjuvant therapy in patients[57], So we continued to study the response of TRS in immunotherapy, and patients with low TRS had higher immune checkpoint expression, suggesting a greater benefit in immunotherapy. Similarly, our constructed TRS signature performed well in immunotherapy cohorts of other tumors, with a significantly higher prognosis for patients with low TRS than for patients with high TRS, and a distant second in predicting overall survival, and in terms of clinical application, TRS scores for responders to immunosuppressants were significantly lower than those for non-responders, which is consistent with the better prognosis for patients with low TRS.

Although the TRS signature is a very good indicator for assessing the prognosis of breast cancer, we should also recognize its limitations, firstly, the samples applied in this study are retrospective, and the TRS should be validated later with more multicenter samples, and secondly, the mechanism of the genes

included in the model is not clear in breast cancer, and further in vivo and in vitro experiments need to be used to validate their function.

Limitations

Several limitations of this study should be acknowledged. First, all data analyzed were obtained from publicly available retrospective cohorts, which may introduce inherent selection bias and limit the generalizability of our findings. Second, although we validated the TRS signature in seven independent datasets, the sample sizes of some cohorts were relatively small, particularly in the immunotherapy validation sets (GSE78220, $n = 28$; GSE135222, $n = 27$). Third, the mechanistic roles of the genes included in the signature have not been experimentally validated in vitro or in vivo, and further functional studies are required to elucidate their biological significance in breast cancer progression and immunotherapy response. Fourth, the lack of prospective clinical trial data limits the direct clinical application of the TRS model. Despite these limitations, our comprehensive bioinformatics analysis provides robust evidence supporting the potential utility of TRS as a prognostic and predictive biomarker in breast cancer.

Conclusion

This study combines single-cell and multi-group transcriptomic data using sophisticated machine learning methods to construct robust and powerful features for assessing the prognosis and clinical performance of breast cancer patients. This TRS model is an effective tool for decision-making and clinical guidance programs for breast cancer patients.

Declarations

Funding: This work was supported by the Basic and Applied Basic Research in Guangdong Province (grant No. 2023A1515220244), the Construction of Social Development Technology Collaborative Innovation System in Shaoguan City (grant No. 230317218038313), and the Shaoguan Science and Technology Program (grant No. 251011178038256). All funding was awarded to Hai Lu.

Ethics approval and consent to participate: Not applicable. The present study is a retrospective analysis of publicly available datasets. All data were de-identified and obtained from open-access repositories. Therefore, no ethical approval or informed consent was required.

Availability of data and materials: The datasets analyzed in the present study are publicly available from the following repositories:

Gene Expression Omnibus (GEO):

GSE180286 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE180286>)

GSE158399 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158399>)

GSE1456 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1456>)

GSE20685 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20685>)

GSE24450 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24450>)

GSE42568 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42568>)

GSE58812 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58812>)

GSE7390 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7390>)

GSE135222 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE135222>)

GSE78220 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78220>).

The Cancer Genome Atlas (TCGA): <https://portal.gdc.cancer.gov> (breast cancer transcriptome and clinical data).

IMvigor210 cohort: <http://research-pub.gene.com/IMvigor210> (Core Biologies).

All relevant data generated or analyzed during this study are included in this article and its supplementary information files.

Authors' contributions: Yan Liu and Yuxin Cheng performed data curation, formal analysis, and drafted the manuscript. Ting Huang, Liang Zhang, Zhenwei Yang, and Mengting Zeng contributed to data collection and interpretation. Hai Lu and Zhiyang Li conceived and designed the study, acquired funding, supervised the project, and revised the manuscript. All authors read and approved the final manuscript.

Patient consent for publication: Not applicable.

Competing interests: The authors declare that they have no competing interests.

Use of artificial intelligence tools: Not applicable.

References

1. Bray F, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J Clin*. 2024. 10.3322/caac.21834.
2. Liang AL, et al. MiRNA-10b sponge: An anti-breast cancer study in vitro. *Oncol Rep*. 2016;35:1950–8. 10.3892/or.2016.4596.
3. Chen W et al. Cancer statistics in China, 2015. *CA: a cancer journal for clinicians* 66, 115–132. 10.3322/caac.21338 (2016).

4. Minn AJ, et al. Genes that mediate breast cancer metastasis to lung. *Nature*. 2005;436:518–24. 10.1038/nature03799.
5. Weigelt B, Peterse JL, t Veer, L. J. Breast cancer metastasis: markers and models. *Nat Rev Cancer*. 2005;5:591–602. 10.1038/nrc1670. van.
6. Nawy T. Single-cell sequencing. *Nat Methods*. 2014;11:18–18.
7. Akrap N, et al. Identification of Distinct Breast Cancer Stem Cell Populations Based on Single-Cell Analyses of Functionally Enriched Stem and Progenitor Pools. *Stem cell Rep*. 2016;6:121–36. 10.1016/j.stemcr.2015.12.006.
8. Gao R, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet*. 2016;48:1119–30. 10.1038/ng.3641.
9. Lee MC, et al. Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing. *Proc Natl Acad Sci USA*. 2014;111:E4726–4735. 10.1073/pnas.1404656111.
10. Patel AP, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Sci (New York N Y)*. 2014;344:1396–401. 10.1126/science.1254257.
11. de Visser KE, Joyce JA. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer Cell*. 2023;41:374–403. 10.1016/j.ccell.2023.02.016.
12. Bach K, et al. Time-resolved single-cell analysis of Brca1 associated mammary tumourigenesis reveals aberrant differentiation of luminal progenitors. *Nat Commun*. 2021;12:1502. 10.1038/s41467-021-21783-3.
13. Gil D, Alcazar CR, et al. Immune Escape in Breast Cancer During In Situ to Invasive Carcinoma Transition. *Cancer Discov*. 2017;7:1098–115. 10.1158/2159-8290.Cd-17-0222.
14. Linde N, et al. Macrophages orchestrate breast cancer early dissemination and metastasis. *Nat Commun*. 2018;9. 10.1038/s41467-017-02481-5.
15. Ojalvo LS, Whittaker CA, Condeelis JS, Pollard JW. Gene expression analysis of macrophages that facilitate tumor invasion supports a role for Wnt-signaling in mediating their activity in primary mammary tumors. *Journal of immunology (Baltimore, Md.: 1950)* 184, 702–712. 10.4049/jimmunol.0902360 (2010).
16. Topalian SL, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med*. 2012;366:2443–54. 10.1056/NEJMoa1200690.
17. Robert C, et al. Pembrolizumab versus Ipilimumab in Advanced Melanoma. *N Engl J Med*. 2015;372:2521–32. 10.1056/NEJMoa1503093.
18. Joyce JA, Fearon DT. T cell exclusion, immune privilege, and the tumor microenvironment. *Sci (New York N Y)*. 2015;348:74–80. 10.1126/science.aaa6204.
19. Tumeh PC, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*. 2014;515:568–71. 10.1038/nature13954.

20. Sharma P, Allison JP. The future of immune checkpoint therapy. *Sci (New York N Y)*. 2015;348:56–61.
21. Hodi FS, et al. Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med*. 2010;363:711–23. 10.1056/NEJMoa1003466.
22. Reck M, et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *N Engl J Med*. 2016;375:1823–33. 10.1056/NEJMoa1606774.
23. Schmid P, et al. Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer. *N Engl J Med*. 2018;379:2108–21. 10.1056/NEJMoa1809615.
24. Xu K, et al. Single-cell RNA sequencing reveals cell heterogeneity and transcriptome profile of breast cancer lymph node metastasis. *Oncogenesis*. 2021;10:66. 10.1038/s41389-021-00355-6.
25. Xu K, et al. Integrative analyses of scRNA-seq and scATAC-seq reveal CXCL14 as a key regulator of lymph node metastasis in breast cancer. *Hum Mol Genet*. 2021;30:370–80. 10.1093/hmg/ddab042.
26. Hao Y, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184:3573–e35873529. 10.1016/j.cell.2021.04.048.
27. Colaprico A, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44:e71–71.
28. Pawitan Y, et al. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast cancer research: BCR*. 2005;7:R953–964. 10.1186/bcr1325.
29. Kao KJ, Chang KM, Hsu HC, Huang AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer*. 2011;11:143. 10.1186/1471-2407-11-143.
30. Heikkinen T et al. Variants on the promoter region of PTEN affect breast cancer progression and patient survival. *Breast cancer research: BCR* 13, R130. 10.1186/bcr3076 (2011).
31. Clarke C, et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis*. 2013;34:2300–8. 10.1093/carcin/bgt208.
32. Jézéquel P, et al. Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. *Breast cancer research: BCR*. 2015;17. 10.1186/s13058-015-0550-y.
33. Desmedt C, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin cancer research: official J Am Association Cancer Res*. 2007;13:3207–14. 10.1158/1078-0432.Ccr-06-2765.
34. Jung H, et al. DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load. *Nat Commun*. 2019;10:4278. 10.1038/s41467-019-12159-9.
35. Hugo W, et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell*. 2016;165:35–44. 10.1016/j.cell.2016.02.065.

36. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26:1572–3.
37. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:1–15.
38. Yoshihara K, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. 2013;4:2612. 10.1038/ncomms3612.
39. Chen DS, Mellman I. Oncology meets immunology: the cancer-immunity cycle. *Immunity*. 2013;39:1–10. 10.1016/j.immuni.2013.07.012.
40. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. *Cancer Syst Biology: Methods Protocols*, 243–59 (2018).
41. Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res*. 2018;28:1747–56.
42. Jiang P, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med*. 2018;24:1550–8.
43. Ginter PS, et al. Histologic grading of breast carcinoma: a multi-institution study of interobserver variation using virtual microscopy. *Mod pathology: official J United States Can Acad Pathol Inc*. 2021;34:701–9. 10.1038/s41379-020-00698-2.
44. Rennhack JP, et al. Integrated analyses of murine breast cancer models reveal critical parallels with human disease. *Nat Commun*. 2019;10:3261. 10.1038/s41467-019-11236-3.
45. Qin G, et al. NPM1 upregulates the transcription of PD-L1 and suppresses T cell activity in triple-negative breast cancer. *Nat Commun*. 2020;11:1669. 10.1038/s41467-020-15364-z.
46. Ott PA, Hodi FS, Kaufman HL, Wigginton JM, Wolchok JD. Combination immunotherapy: a road map. *J Immunother Cancer*. 2017;5:16. 10.1186/s40425-017-0218-5.
47. Teng F, Kong L, Meng X, Yang J, Yu J. Radiotherapy combined with immune checkpoint blockade immunotherapy: Achievements and challenges. *Cancer Lett*. 2015;365:23–9. 10.1016/j.canlet.2015.05.012.
48. Muss HB, et al. Randomized Trial of Standard Adjuvant Chemotherapy Regimens Versus Capecitabine in Older Women With Early Breast Cancer: 10-Year Update of the CALGB 49907 Trial. *J Clin oncology: official J Am Soc Clin Oncol*. 2019;37:2338–48. 10.1200/jco.19.00647.
49. Liu M, et al. An Aptamer-Based Probe for Molecular Subtyping of Breast Cancer. *Theranostics*. 2018;8:5772–83. 10.7150/thno.28949.
50. Rakha EA, et al. Prognostic significance of Nottingham histologic grade in invasive breast carcinoma. *J Clin oncology: official J Am Soc Clin Oncol*. 2008;26:3153–8. 10.1200/jco.2007.15.5986.
51. Muller PA, Vousden KH. Mutant p53 in cancer: new functions and therapeutic opportunities. *Cancer Cell*. 2014;25:304–17. 10.1016/j.ccr.2014.01.021.

52. Lu S, et al. Signal-Oriented Pathway Analyses Reveal a Signaling Complex as a Synthetic Lethal Target for p53 Mutations. *Cancer Res.* 2016;76:6785–94. 10.1158/0008-5472.Can-16-1740.
53. Haricharan S, Bainbridge MN, Scheet P, Brown PH. Somatic mutation load of estrogen receptor-positive breast tumors predicts overall survival: an analysis of genome sequence data. *Breast Cancer Res Treat.* 2014;146:211–20. 10.1007/s10549-014-2991-x.
54. Zhou Y, Shao G, Liu S. Monitoring Breast Tumor Lung Metastasis by U-SPECT-II/CT with an Integrin $\alpha(v)\beta(3)$ -Targeted Radiotracer(^{99m}Tc-3P-RGD(2)). *Theranostics.* 2012;2577–88. 10.7150/thno.4443.
55. Fancello L, Gandini S, Pelicci PG, Mazzarella L. Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. *J Immunother Cancer.* 2019;7:183. 10.1186/s40425-019-0647-4.
56. Lee SJ, et al. Enhanced anti-tumor immunotherapy by silica-coated magnetic nanoparticles conjugated with ovalbumin. *Int J Nanomed.* 2019;14:8235–49. 10.2147/ijn.S194352.
57. Zhou ZQ, et al. PD-L1 expression is a predictive biomarker for CIK cell-based immunotherapy in postoperative patients with breast cancer. *J Immunother Cancer.* 2019;7:228. 10.1186/s40425-019-0696-8.

Figures

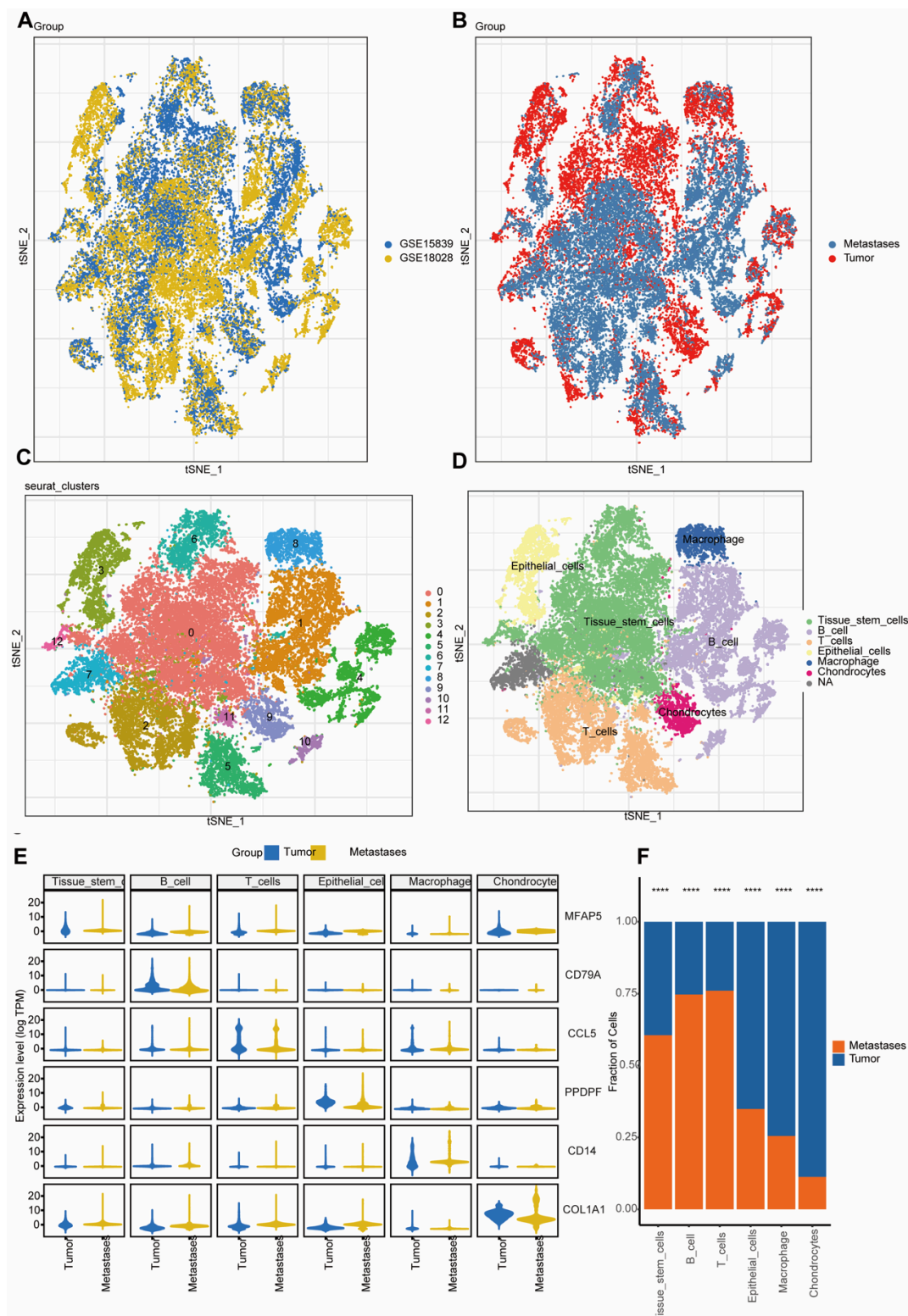


Figure 1

The single-cell landscape of breast cancer. A: tSNE plots of single cells from both datasets after correction for batch effects. B: tSNE plots of single cells of tumors and metastases after correction for batch effects. C: tSNE plots of single cells from 12 breast cancer samples. D: tSNE mapping of six cell clusters. E: Violin maps of representative genes of cell clusters. F: Scale diagram of cell clusters in tumors and metastasis

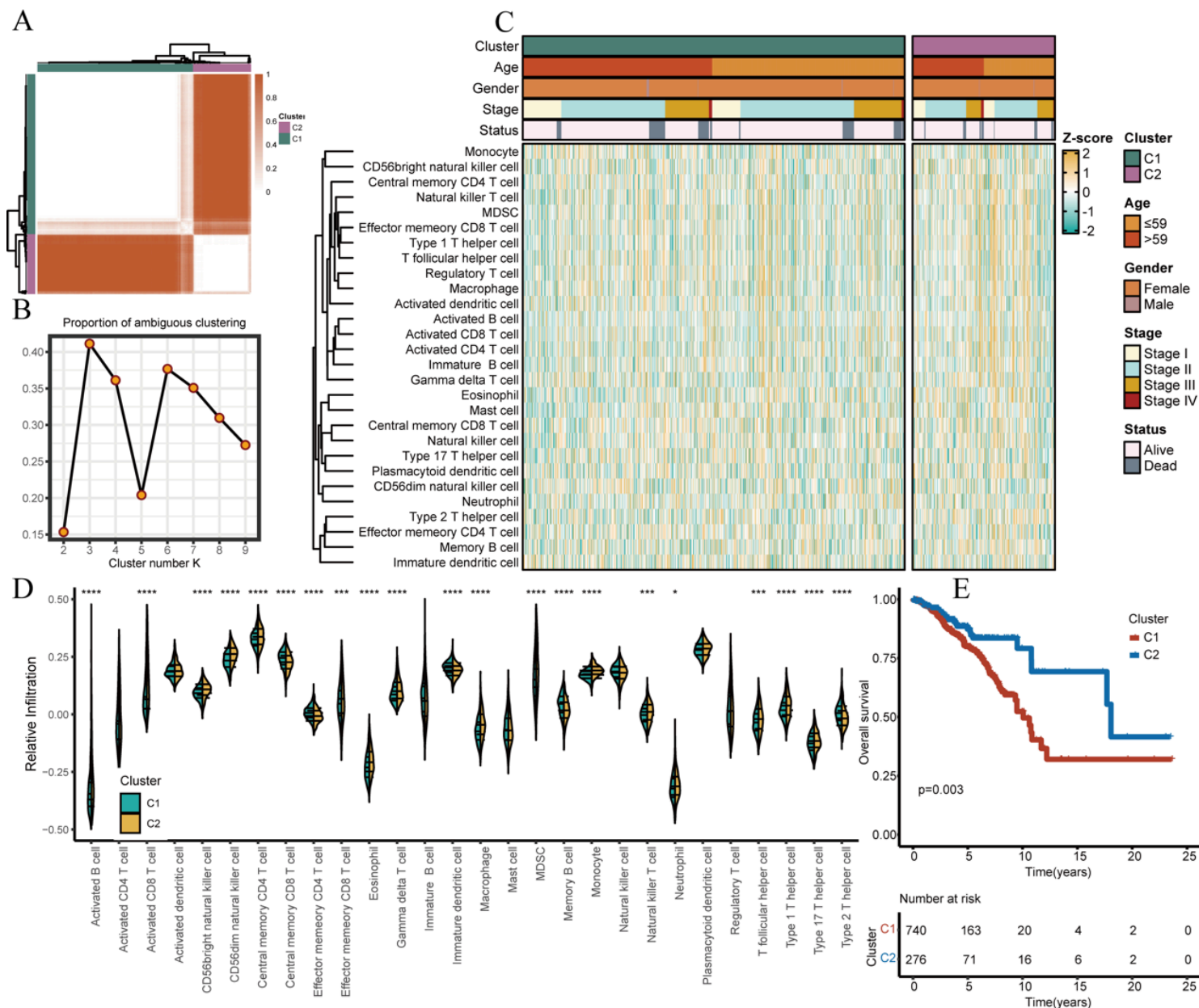


Figure 2

Identification of T cell-associated subtypes. A: Consensus score matrix for all samples when $k = 2$. B: Homophenotypic correlation coefficients apply to the optimal number of subtypes. C: Heatmap demonstrates differences in infiltration of 28 immune cells in two clusters. D: Special violin plot demonstrating 28 types of immune cell infiltration between two clusters. E: Kaplan-Meier curves of overall survival for the two clusters

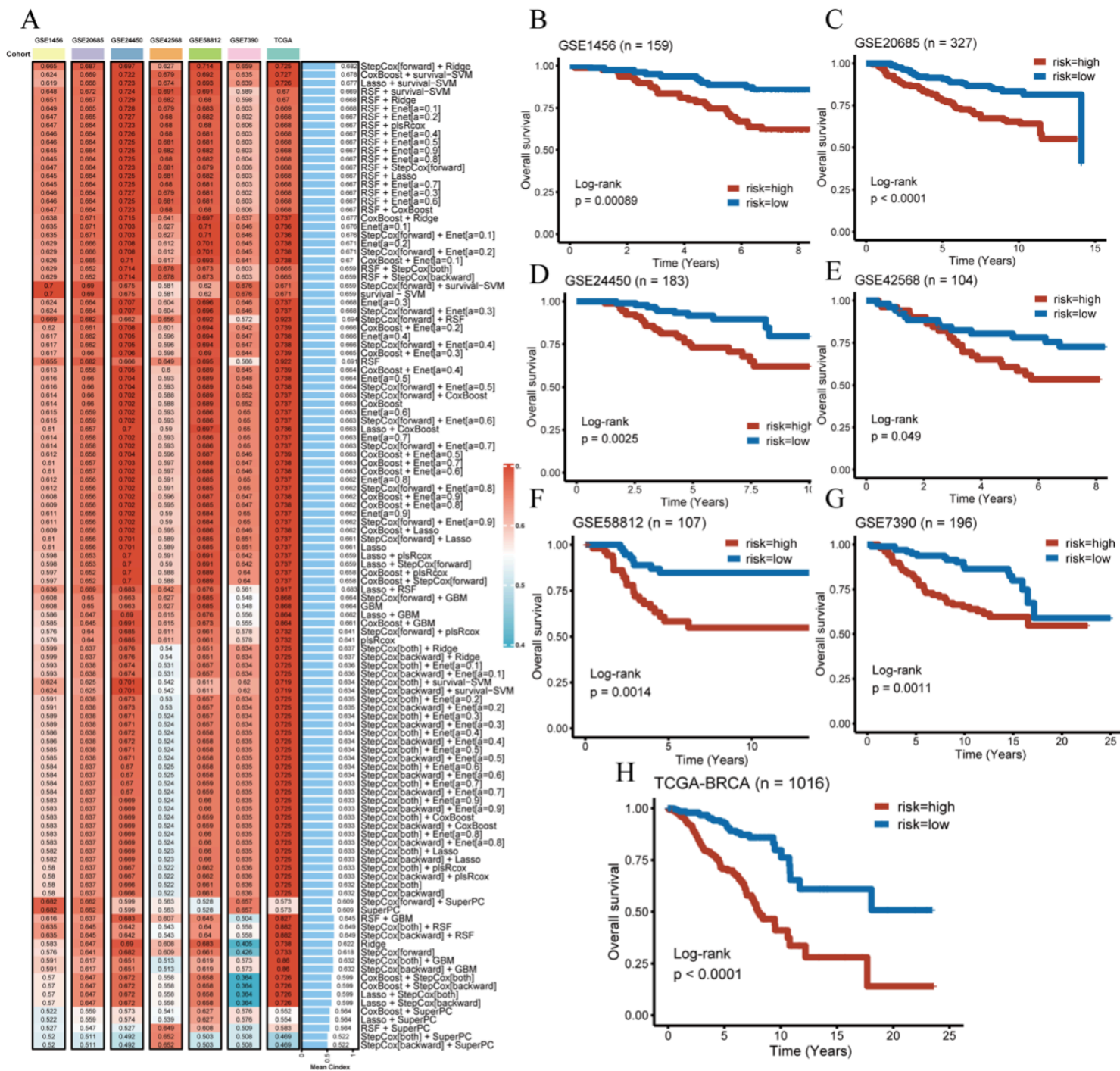


Figure 3

Construction and validation of the T-cell-related signature (TRS) based on machine learning. A: Heatmap of C-index based on 117 machine learning predictive models in 7 datasets. B: Kaplan-Meier curves for TRS in the GSE1456 dataset. C: Kaplan-Meier curves for TRS in the GSE20685 dataset. D: Kaplan-Meier curves for TRS in the GSE24450 dataset. E: Kaplan-Meier curves for TRS in the GSE42568 dataset. F: Kaplan-Meier curves for TRS in the GSE58812 dataset. G: Kaplan-Meier curves for TRS in the GSE7390 dataset. H: Kaplan-Meier curves for TRS in the TCGA-BRCA dataset

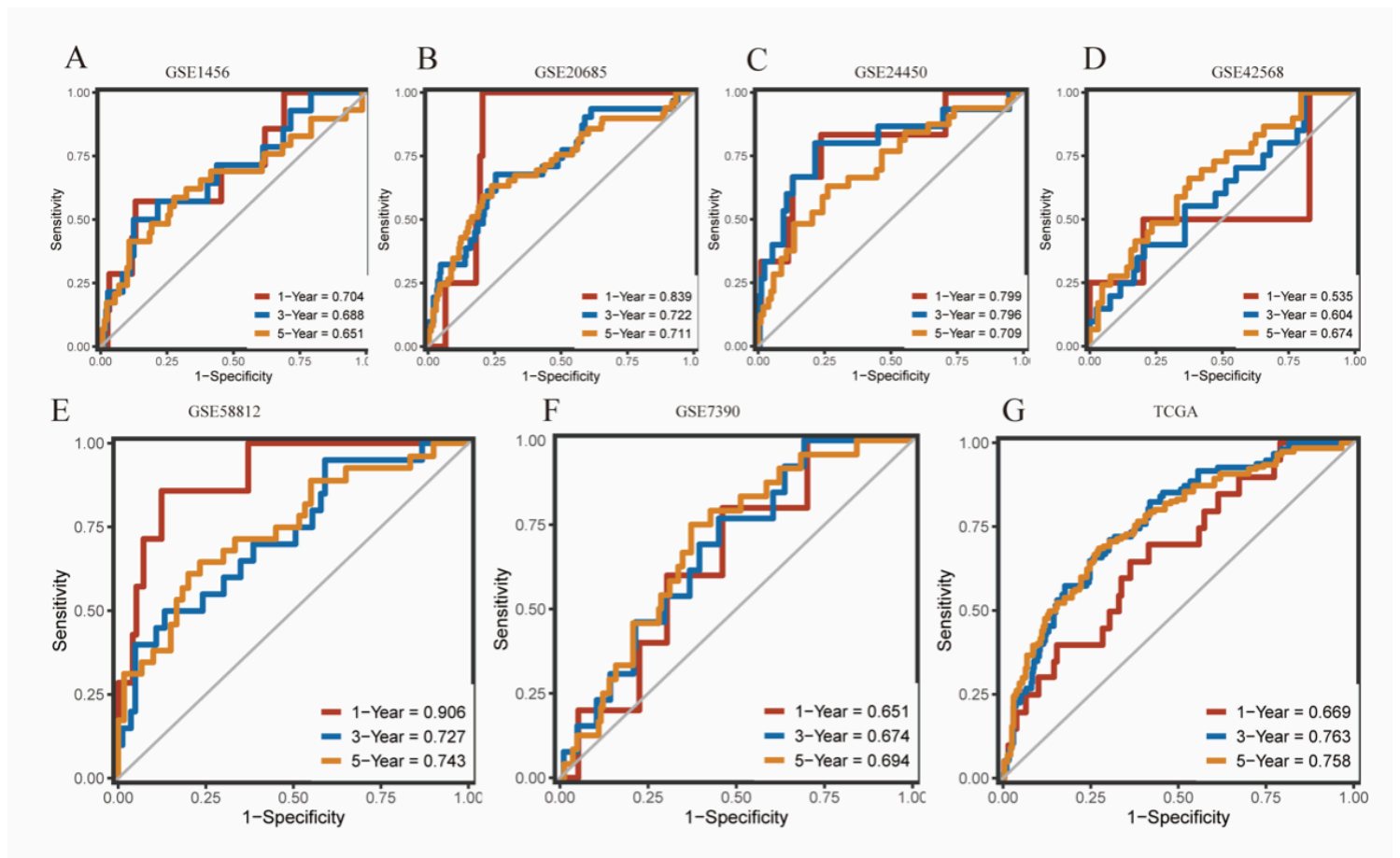


Figure 4

Validation of the TRS model. A: Time-dependent ROC analysis of TRS in the GSE1456 dataset. B: Time-dependent ROC analysis of TRS in the GSE20685 dataset C: Time-dependent ROC analysis of TRS in the GSE24450 dataset D: Time-dependent ROC analysis of TRS in the GSE42568 dataset E: Time-dependent ROC analysis of TRS in the GSE58812 dataset F: Time-dependent ROC analysis of TRS in the GSE7390 dataset G: Time-dependent ROC analysis of TRS in the TVGA-BRCA dataset

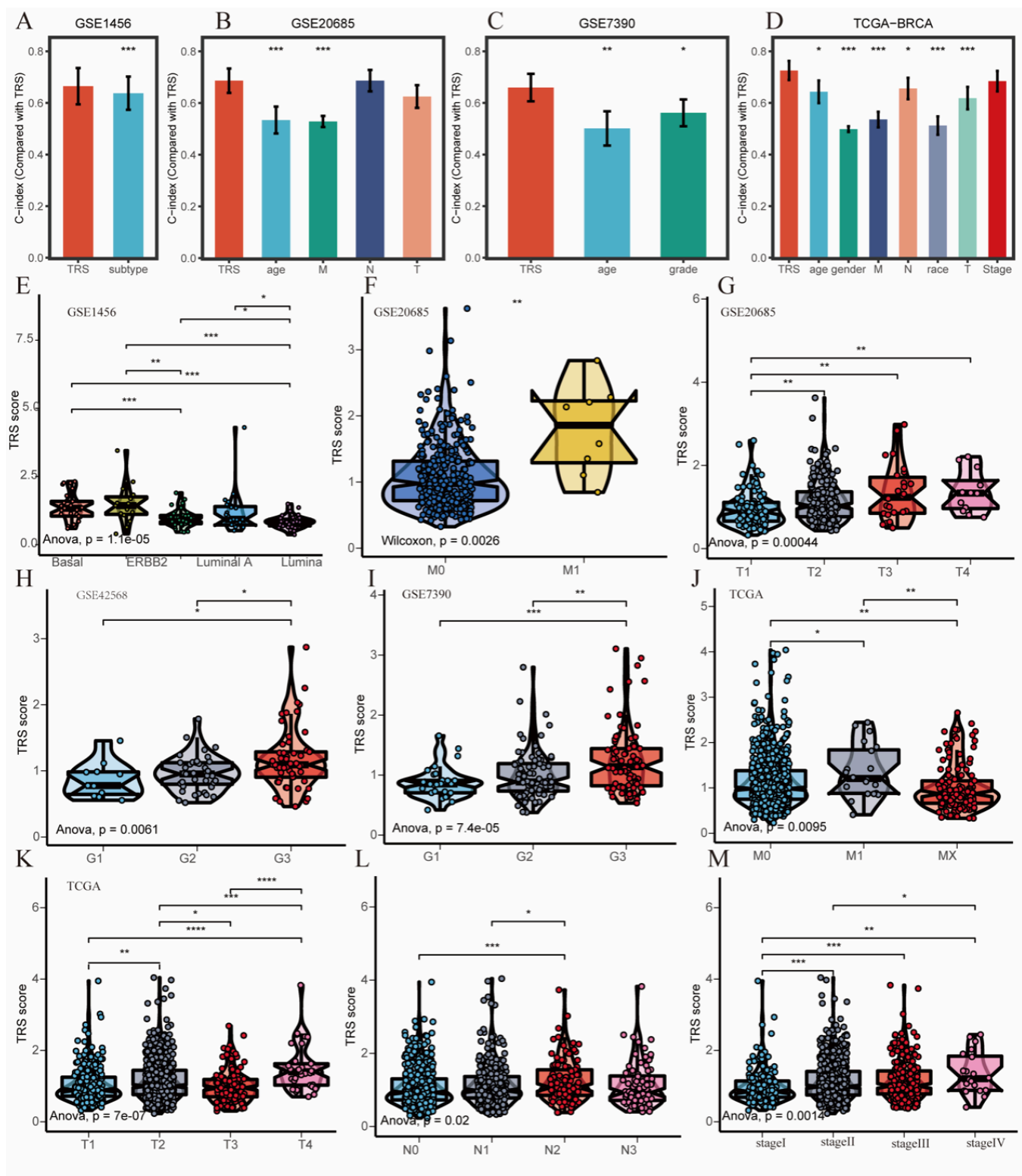


Figure 5

Predictive Performance Evaluation of TRS. A: Histogram of the C-index of TRS compared to other clinical traits in the GSE1456 dataset. B: A: Histogram of the C-index of TRS compared to other clinical traits in the GSE20685 dataset. C: Histogram of the C-index of TRS compared to other clinical traits in the GSE7390 dataset. D: Histogram of the C-index of TRS compared to other clinical traits in the TCGA-BRCA dataset. E: Distribution of TRS in different Grade in the GSE1456 dataset. F: Distribution of TRS in

different M Stage in the GSE20685 dataset. H: Distribution of TRS in different T Stage in the GSE20685 dataset. I: Distribution of TRS in different Grade in the GSE42568 dataset. J: Distribution of TRS in different M Stage in the TCGA-BTCA dataset. K: Distribution of TRS in different T Stage in the TCGA-BRCA dataset. L: Distribution of TRS in different N Stage in the GSE42568 dataset. M: Distribution of TRS in different Stage in the TCGA-BRCA dataset. *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001

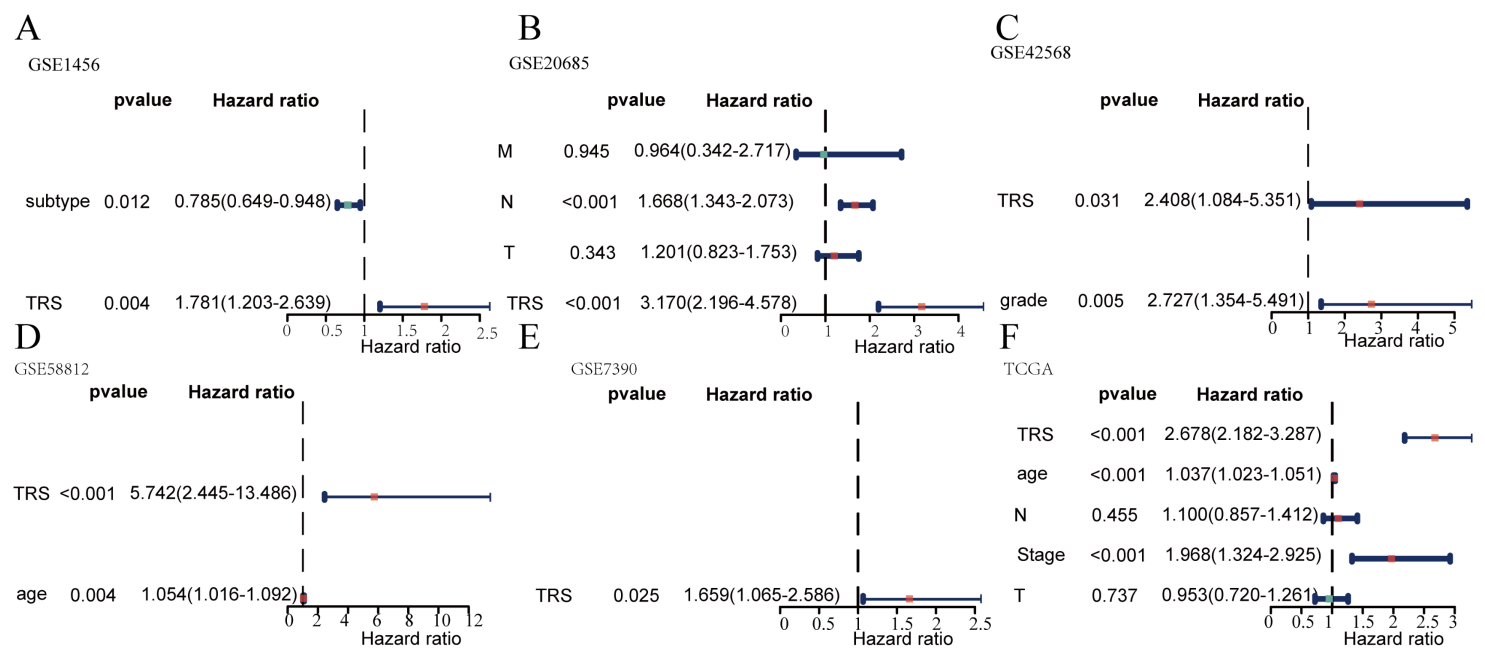


Figure 6

Independent prognostic analysis of TRS. A: Forest plot showing multifactorial cox analysis of TRS and clinical traits in the GSE1456 dataset. B: Forest plot showing multifactorial cox analysis of TRS and clinical traits in the GSE20685 dataset. C: Forest plot showing multifactorial cox analysis of TRS and clinical traits in the GSE42568 dataset. D: Forest plot showing multifactorial cox analysis of TRS and clinical traits in the GSE58812 dataset. E: Forest plot showing multifactorial cox analysis of TRS and clinical traits in the GSE7390 dataset. F: Forest plot showing multifactorial cox analysis of TRS and clinical traits in the TCGA dataset



Figure 7

TME analysis of the TRS signature. A: Differences in expression of immunomodulators (chemokines, immunostimulants, MHC, receptors) between high and low TRS groups in BRCA. B: Differences between individual steps of the cancer immunization cycle in the high and low TRS groups. C: Correlation between TRS signature and 28 immune cells and immune-related pathways. D: Differences in immune

effector cells between high and low TRS groups. E: Correlation between TRS signature and immune checkpoints. $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$

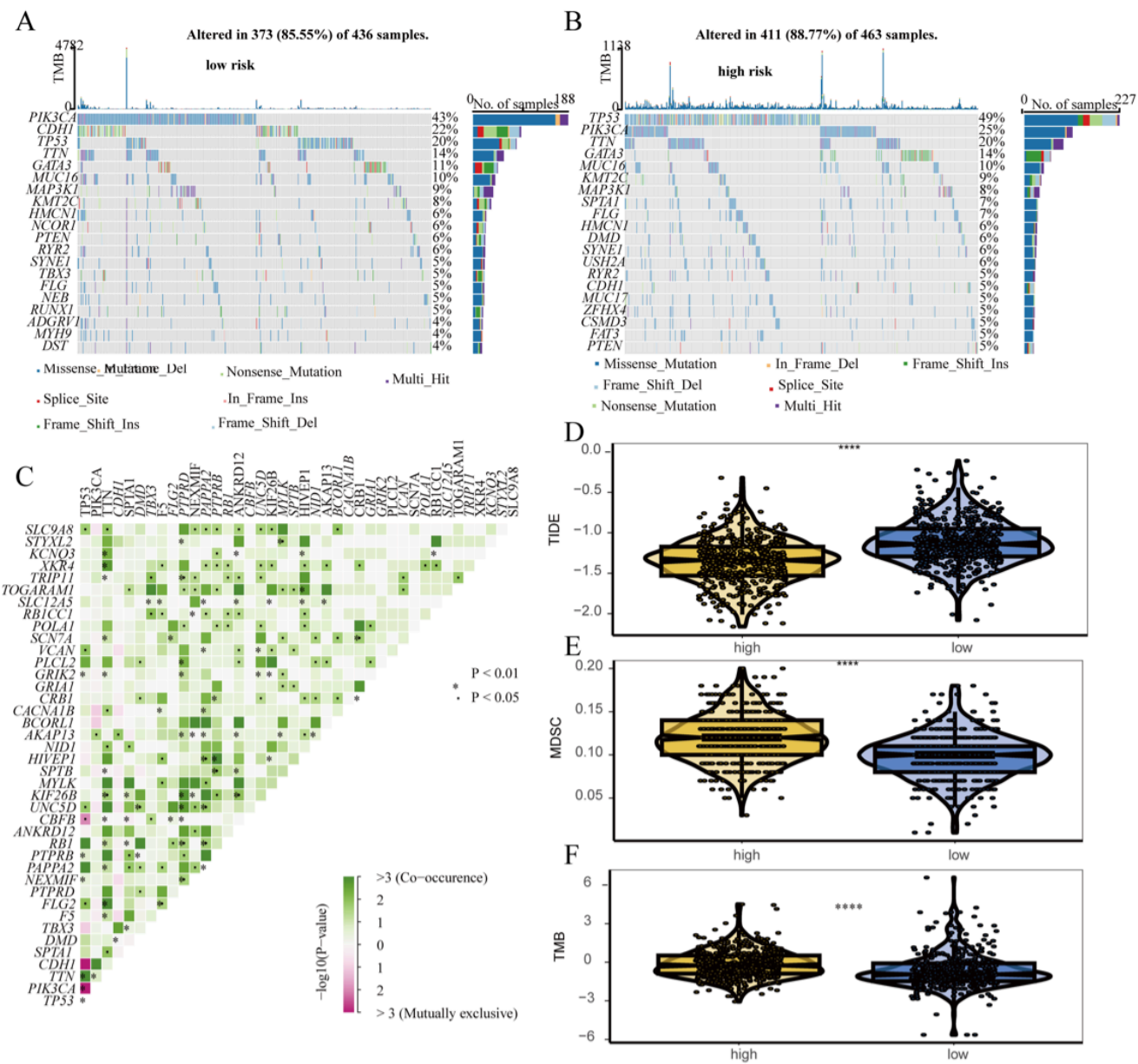


Figure 8

Multi-omics analysis between high and low TRS groups. AB: Mutational mapping of the high and low TRS groups. C: Interaction of differential mutations in high and low TRS groups. D: Differences in TIDE between high and low TRS groups. E: Differences in MDSC between high and low TRS groups. F: Differences in TMB between high and low TRS groups. $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$; **** $P < 0.0001$

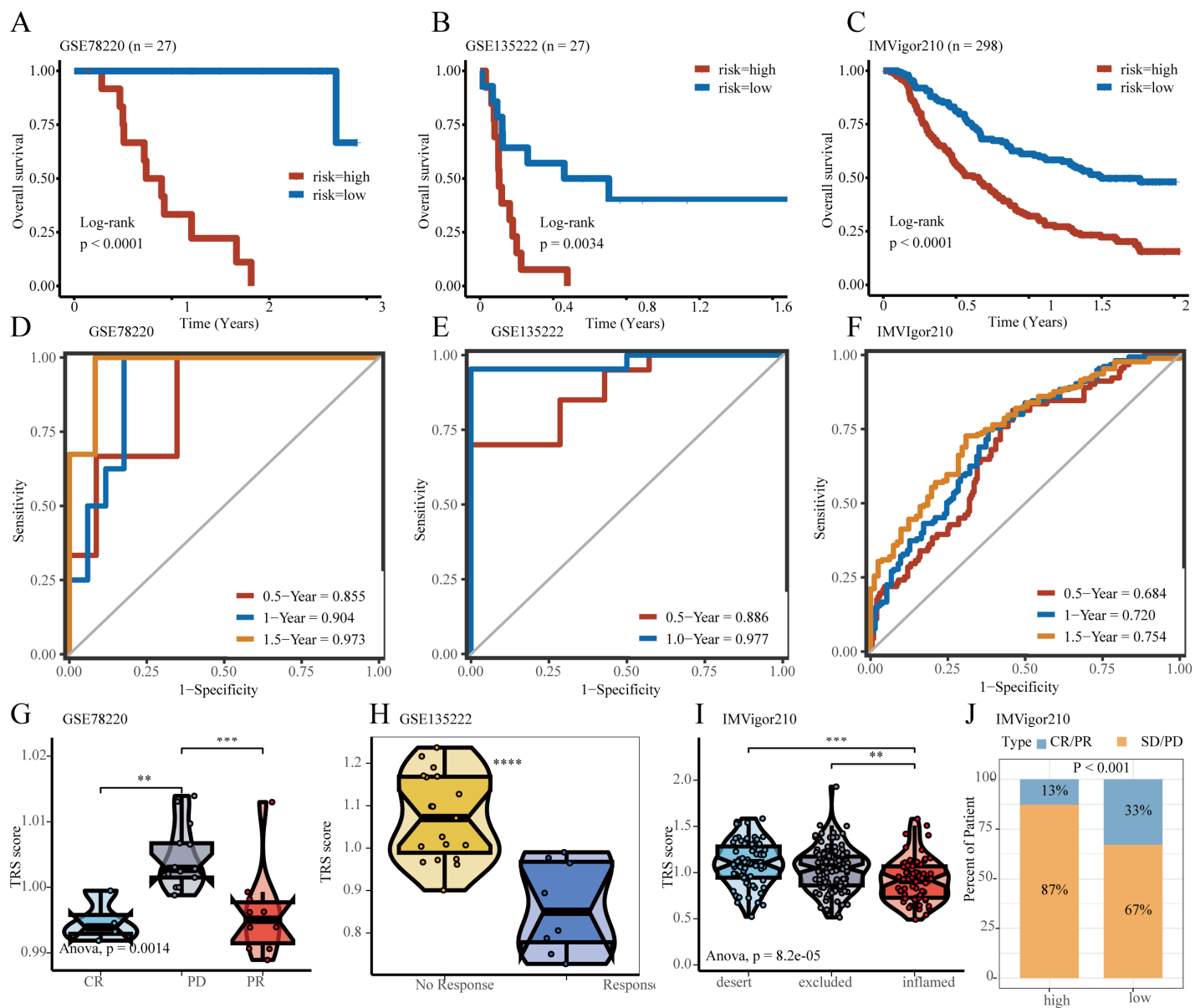


Figure 9

Predictive value of TRS in immunotherapy cohorts. A: Kaplan-Meier curves for TRS in the GSE78220 dataset. B: Kaplan-Meier curves for TRS in the GSE135222 dataset. C: Kaplan-Meier curves for TRS in the IMVigor210 dataset. D: Time-dependent ROC analysis of TRS in the GSE78220 dataset. E: Time-dependent ROC analysis of TRS in the GSE135222 dataset. F: Time-dependent ROC analysis of TRS in the IMVigor210 dataset. G: TRS distribution between CR,PD,PR in GSE78220 dataset. H: TRS distribution with responders, non-responders in GSE135222. I: TRS distribution between desert, excluded, inflamed in IMVigor210. J: Distribution of high and low TRS groups between CR/PR,SD/PD. P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SUPPEMENTALDATA.docx](#)