Cancer Cell International

RESEARCH Open Access

Identification of a novel immune-related gene signature by single-cell and bulk sequencing for the prediction of the immune landscape and prognosis of breast cancer

Yanlin Gu[1†], Zhengyang Feng[2†], Xiaoyan Xu[3†] and Liyan Jin[1*]

Abstract

Background As a common cause of cancer-related deaths in women, BRCA (breast cancer) shows complexity and requires precise biomarkers and treatment methods. This study delves into the molecular makeup of BRCA, focusing on immune profiles, molecular subtypes, gene expression and single-cell analysis.

Methods XCell was used to assess immune infiltration based on TCGA (the Cancer Genome Atlas) data and the clustering analysis was made. Differentially expressed genes were examined in distinct clusters, and the WGCNA (weighted correlation network analysis) was made to establish co-expression networks. The prognostic models were developed by Cox and LASSO-Cox regression. The clustering analysis, GSEA (Gene set enrichment analysis), GSVA (gene set variation analysis) and communication analysis of the single-cell dataset GSE161529 were performed to investigate the functional relevance. Real-time polymerase chain reaction (RT-PCR) was employed for evaluating gene expression.

Results The results revealed significant differences in immune cell infiltration between two clusters (C1 and C2). C2 had poorer survival outcomes, which was associated with higher expression of immune checkpoints *PD1* and *PD-L1*. The gene modules identified via WGCNA were correlated with the immune-based subtypes. Then, a prognostic model comprising seven genes (*ACSL1*, *ABCB5*, *XG*, *ADH4*, *OPN4*, *NPR3*, *NLGN1*) was used to divide patients into high- and low-risk subgroups. The high-risk group had worse prognosis and higher scores of TIDE (Tumor Immune Dysfunction and Exclusion). The single-cell analysis depicted the immune landscape. Macrophages and endothelial cells exhibited higher AUCell scores. In cellular communication analysis, notably significant ligand-receptor interactions of HLA-DRA-> CD4 and TNFSF13B-> HLA-DPB1 were observed. The proportion of endothelial cells was correlated with risk scores. Finally, RT-PCR results illustrated the expression of seven genes in BRCA specimens.

†Yanlin Gu, Zhengyang Feng and Xiaoyan Xu contributed equally to this work.

*Correspondence:
Liyan Jin
jinyuliangyuan1985@sina.com

Full list of author information is available at the end of the article

Conclusion The integrative analysis provides new insights into molecular complexities of BRCA. Immune profiles and gene signatures hold potential for improving stratification of BRCA patients and guiding the development of personalized immunotherapy strategies.

Keywords Breast cancer, Bioinformatics, Immune infiltration, Molecular subtyping, Prognostic model, Single-cell analysis

Introduction

BRCA (breast cancer) has surpassed lung cancer as the commonest malignant tumor in the world, being the main cause of cancer-related deaths among women. According to a report in 2020, 2.3 million patients are diagnosed with BRCA worldwide annually, accounting for 11.7% of all cancer patients. Of these new BRCA patients, 685,000 patients die, accounting for 6.9% of all cancer deaths [1]. Although remarkable progress has been made in its treatment, staging and molecular biomarkers, the heterogeneity of tumors is still a challenge to treatment results. It is extremely important to solve the heterogeneity problem of breast tumors and it is fundamental to develop personalized treatment strategies for patients with breast tumors of biological heterogeneity. To develop more effective personalized drugs that can accurately target specific molecular changes in individual patients with BRCA, it is necessary to deeply understand the accurate molecular characteristics [2–4]. The development of personalized treatment strategies for patients with tumors constituted by unique biological components is the key to improving the treatment effect and reducing deaths in BRCA patients [5–7]. In addition, in the coming era of precision medicine, cancer treatment will rely on the use of advanced technologies such as genomics, proteomics and artificial intelligence to identify new therapeutic targets and develop customized treatment programs [8–10]. The approaches developed by the most innovative methods will not only accelerate the complete revolution of BRCA treatment, but also pave way for a change of the paradigm towards personalized and targeted treatment in the field of oncology.

Tumor immunotherapy is to fight tumors with the autoimmune system, marking a major breakthrough in medical treatment. This method involves the activation of specific cells in the immune system to identify and destroy tumor cells. Although it shows promising results in some types of tumors, it still faces challenges. Only a few patients benefit from immunotherapy in clinical practice, and this therapy usually leads to immune-related side effects [11]. Further research and clinical trials are required to optimize its effect and determine the appropriate patient population.

With the progress of bioinformatics and single-cell sequencing technology in recent years, there is a growing possibility of revealing the complex molecular landscape of BRCA. The innovative methods provide unprecedented opportunities for researchers to explore the heterogeneity of breast tumors with single-cell resolution. The research results offer valuable insights into the mechanism governing the dynamic process of tumor evolution [12, 13]. In addition, the elucidation of complex interactions among tumor cells, immune effectors and matrix elements has become the focus of research. This kind of comprehensive study not only reveals the complex immune escape mechanism of BRCA cells, but also identifies new immunotherapy targets with clinical significance [14–18]. Similarly, the application of consistent clustering methods promotes successful identification of molecular subtypes and various cancer types with different clinical results. This remarkable achievement highlights the potential of molecular stratification of breast tumors based on the comprehensive molecular analysis [19, 20]. By systematically classifying breast tumors into biologically and clinically relevant subtypes, researchers can tailor the treatment model according to the unique characteristics of each patient. In this way, the treatment effect will be maximized and the treatment results of patients will be improved tremendously [21, 22].

In this study, we employed a comprehensive bioinformatics analysis to dissect the molecular heterogeneity of BRCA. Utilizing publicly available datasets from TCGA (The Cancer Genome Atlas), we performed immune infiltration analysis, consistent clustering, differential gene expression analysis, WGCNA (weighted gene co-expression network analysis), enrichment analysis, and prognostic model construction. Our multi-faceted approach aimed to identify molecular signatures and immune profiles that could inform personalized therapeutic strategies and improve patient prognostication.

Materials and methods

Data download

We obtained gene expression data (TPM) and associated prognosis information from TCGA database (https://portal.gdc.cancer.gov/) using the R package TCGAbiolinks, and utilized the ICGC (International Cancer Genome Consortium) dataset (BRCA-FR) as a validation set for further verification. We also downloaded the single-cell BRCA dataset GSE161529 from the GEO (Gene Expression Omnibus) database (https://www.ncbi.nlm.nih.gov/geo/) dataset using scRNA-seq on the 10X Genomics Chromium platform.

Immune infiltration analysis

The immune microenvironment is a complex integrated system composed of immune cells, inflammatory cells, fibroblasts, stromal tissue, various cytokines, and chemokines. Tissues consist of diverse cell types with unique transcriptional expressions. Deconvolution of gene expression profiles enables the reconstruction of tissue cellular composition. The xCell, developed by the dviraran team, utilizes the ssGSEA algorithm to rank gene expression levels and deconvolute transcriptome expression matrices, thereby estimating the composition and abundance of immune cells within heterogeneous populations for immune infiltration analysis [23]. We utilized xCell to assess the infiltration levels of immune cells, employed the R package IOBR (v0.99.9) for xCell deconvolution [24]. This involved filtering out immune cell data and evaluating immune cell infiltration abundance. Visualization was achieved through box plots and stacked bar charts, created using the R package ggplot2 (v3.4.2) [25]. Additionally, we computed the correlation between immune cells using the Pearson algorithm and represented it as a correlation heatmap using the R package corrplot (v0.92).

Consistent clustering and typing

We utilized the R package ConsensusClusterPlus (v1.62.0) for consensus clustering based on immune infiltration results to differentiate different BRCA subtypes. We repeated sampling of 80% of the total samples 1000 times, with clusterAlg = "km" and distance="euclidean". To validate the relationship between clustered subgroups and OS (overall survival), KM (Kaplan Meier) survival curves were plotted using the survival package (v3.5.3). Subsequently, we examined the expression differences of the immune checkpoint *PD-L1* (*CD274*) and *PD1* (*PDCD1*) within consensus clustering subgroups, visualized as violin plots using the ggplot2 package (v3.4.2).

Differential genes between clustering subgroups were analyzed using the DESeq2 package, with criteria of $|\log_2 FC| \geq 1$ and $P<0.05$. Volcano plots and heatmaps were then generated to visualize significantly differentially expressed immune-related genes. Additionally, we specifically investigated the expression differences of genes closely related to BRCA (*ATM*, *BARD1*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *RAD51D*) among consensus clustering subgroups. We drew the violin plots with the ggplot2 package.

Weighted gene association network analysis

WGCNA is a systems biology method to describe patterns of gene co-expression between different samples [26]. We employed the R package WGCNA (v1.72.1) to analyze differentially expressed genes based on consensus clustering subgroups. We computed the correlation coefficients between genes and constructed a hierarchical clustering tree based on these coefficients. Different branches of the clustering tree represent distinct gene modules, and module significance was subsequently calculated. The minimum module gene count was set to 30, softpower was set to the optimal threshold of 3 and the module merge cut height was set to 0.25. Interested modules were selected based on correlation values, and expression genes highly correlated with the consensus clustering subgroups were identified.

Enrichment analysis

GO (Gene Ontology) enrichment analysis and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis are widely applied for analyzing the functional enrichment of genes across various dimensions and levels. In our study, we focused on expression genes highly correlated with consensus clustering subgroups, employed clusterProfiler (v4.7.1.3) for both GO and KEGG enrichment analyses to identify significantly enriched biological processes and pathways. The enrichment results were visualized using the R packages GOplot (v1.0.2) and enrichplot (v1.18.4), with a significance threshold set at $P<0.05$ for the enrichment analysis.

Prognostic genes were screened and prognostic models were constructed

To identify prognostic genes, we initially conducted univariate Cox regression analysis on genes derived from the consistency clustering results ($P<0.05$). Subsequently, we performed multivariate Cox regression analysis to identify independent prognostic genes. Lasso-Cox regression analysis was carried out via the glmnet package (v4.1.7) in R, where the optimal lambda value was selected. Genes with non-zero coefficients were retained for constructing the prognostic model. A risk score was computed based on gene expression levels and coefficients from the multivariate Cox regression model, according to the following formula:

$$\text{riskscore} = \sum_i \text{Coefficient}^{(\text{hubgene}}{}_i\Big) * mRNA\text{Expression}\left(\text{hubgene}_i\right)$$

Samples were stratified into high- and low-risk groups according to the median risk score. To assess the predictive capacity of the risk score, KM curve analysis was conducted using the survival package (v3.5.3).

To explore the association between risk score and immunotherapy response, we initially scrutinized the expression profiles of immune checkpoint genes (*CD274*, *CD47*, *HAVCR2*, *LAG3*, *IDO1*, *SIRPA*, *TNFRSF4*,

PDCD1, *CTLA4*, *TIGIT*) between high- and low-risk cohorts. Subsequently, we evaluated the TIDE (Tumor Immune Dysfunction and Exclusion) scores across these risk groups [27, 28]. TIDE scores reflect the sensitivity to immune checkpoint inhibitors, serving as a surrogate biomarker predictive of response to immune modulation. Using the TIDE website (https://tide.dfci.harvard.edu/login/), we computed TIDE scores for BRCA samples via the "Predict Response" module, thereby examining disparities between high- and low-risk cohorts. Considering the potential correlation between risk and TIDE scores among patients, we employed the ggExtra package (v0.10.0) in R to construct scatter plots illustrating the relationship between risk scores and TIDE scores, while also fitting correlation curves [29].

Single cell quality control

Prior to delving into the analysis of single-cell gene expression data, it is imperative to ensure the fidelity of UMIs (Unique Molecular Identifiers) by confirming their correspondence to viable cells. QC (Quality control) measures were enacted based on three pivotal cell covariates: count depth per UMI, gene count per UMI, and the mitochondrial gene count score per UMI.

Outlier Management: A robust approach employing the MAD (median absolute deviation) was undertaken to cull cells exhibiting MAD values exceeding 5 across the aforementioned QC metrics. Additionally, cells displaying mitochondrial percentages surpassing 5% were excluded from further analysis.

Bimodal Filtering: Leveraging the scrublet function from the Python package scanpy, we effectively identified and eliminated doublets, defined as instances where two or more cells were coincidentally captured within a single droplet [30, 31]. Subsequent to doublet prediction for each sample, the predicted_doublet attribute facilitated the filtration of the expression matrix.

Single cell deconvolution normalization

The scran package utilizes a pooled size factor estimation deconvolution normalization method, designed to effectively mitigate technical variability between cells while retaining biological diversity [32]. Initially, in Python, we conducted log2 transformation and Leiden clustering using scanpy. Subsequently, in R, the computeSumFactors function was employed to compute size factors for individual cells, followed by their application for normalization. The resultant normalized expression matrix was utilized for subsequent analyses.

Single cell go batch

The autoencoder, an unsupervised neural network, is adept at learning low-dimensional representations of data and reconstructing input data. In this study, we employed the autoencoder functionality within scvi-tools to mitigate batch effects present in single-cell data with Python.

Single-cell clustering

We conducted clustering analysis of single-cell sequencing data with the Python package scanpy. Initially, the sc.pp.neighbors function was applied to construct a cellular neighborhood graph, employing the UMAP algorithm for distance and similarity calculations between cells. Subsequently, the leiden function facilitated Leiden clustering, optimizing modularity via the Leiden algorithm to yield refined clustering outcomes. Visualization of these results depicted the spatial distribution of cells from distinct clusters within the UMAP coordinates.

Single cell annotation

We utilized Enrichment with ORA (Over Representation Analysis) for annotation, leveraging markers provided by the dataset authors. Through one-sided Fisher exact tests based on contingency tables, we assessed the significance of overlap between S and each gene set. Subsequently, we transformed the test *P*-values into logarithmic functional enrichment scores, where higher scores signify greater enrichment. By comparing two annotation methods with previous clustering results, we ultimately chose the more effective ORA annotation as the foundation for subsequent analysis of cell types.

Single cell communication analysis

CellPhoneDB (v2) is a methodology grounded in ligand-receptor inference aimed at elucidating signal transduction dynamics across heterogeneous cell types within single-cell datasets [33]. Leveraging a meticulously curated repository of ligand-receptor interactions and gene expression profiles extracted from single-cell transcriptomic data, this approach enables the prediction of potential communication networks between distinct cellular populations. Implementation of the CellPhoneDB method is facilitated through the cellphonedb module within the Python package Liana, with a particular emphasis on visualizing cell types exhibiting notable perturbations. We utilized Connectome to infer functional connections between cells, NATMI to predict ligand-receptor interactions between cells based on machine learning; SingleCellSignalR to reconstruct signaling networks between cells, CellChat to deduce that communication patterns between cells exhibit high significance in ligand-receptor interactions across different methods, as well as CellPhoneDB [34–36].

Single-cell GSVA

To evaluate functional enrichment among diverse cell types and subgroups within single-cell samples, we

performed GSVA (Gene Set Variation Analysis) on single-cell transcriptomic data using the Python package decoupler [37].

Single cell AUCell analysis

Single-cell AUCell analysis evaluates the activity level of specific gene sets or pathways within each cell, utilizing the AUC (area under the curve) to discern whether key subsets of input gene sets are enriched in the expressed genes of individual cells. This methodology facilitates the identification of cells harboring active gene sets and enables the exploration of relative gene set expression across diverse cell types or states.

Single cell GSEA

The R package clusterProfiler (v4.7.1.3) was employed to conduct GSEA (Gene Set Enrichment Analysis) on differentially expressed genes between subgroups associated with model risk genes and other subgroups. GSEA assesses the distribution trend of predefined gene sets within the gene expression profile ranked by phenotype, thereby determining their contribution to the phenotype. Gene sets from the h.all.v7.4.symbols.gmt collection obtained from the MSigDB (Molecular Signatures Database) were used for this analysis. The parameters were set as follows: 1000 permutations, a minimum of 10 genes per gene set, a maximum of 500 genes per gene set, and BH (Benjamini-Hochberg) correction for p-values. The significance threshold for enrichment was established as a false FDR (discovery rate) value (Q-value) < 0.05.

Cellular composition of TCGA-BRCA was inferred by deconvolution

Cellular interactions constitute an intricate network interfacing the immune system with tumor cells, with discerning the specific immune cell composition within solid tumors being pivotal for prognosticating responsiveness to immunotherapeutic interventions. However, due to the inadequacy of numerous tissue samples for disaggregation into individual cells, direct leveraging of single-cell RNA sequencing techniques is unfeasible. To surmount this obstacle, we applied a deconvolution methodology dubbed MuSiC (v1.0.0), employed the cell-type specificity gleaned from single-cell RNA sequencing data of BRCA as a benchmark gene expression profile [38]. Employing the MuSiC package in R, we inferred the cellular composition within TCGA-BRCA samples and illustrated the outcomes via a stacked bar plot using the ggplot2 package (v3.4.2). Subsequently, we stratified the samples into high-risk and low-risk subcohorts according to median risk score. KM survival analysis was conducred to elucidate survival disparities. Given the association between the proportion of high-risk subgroups and the risk score, we utilized the R-package ggExtra (v0.10.0) to generate a scatter plot depicting this relationship, along with fitting a correlation curve [29].

The expression of prognostic genes in breast cancer tissue

We collected and analyzed the expression of 7 genes in 7 pairs of BRCA and adjacent tissue samples. RNA was extracted using trizol reagent (Invitrogen) and reverse transcribed to cDNA using Hiscript III RT SuperMix for qPCR (Vazyme). Real-time PCR was conducted on an Applied Biosystems platform by Thermo Fisher Scientific using SYBR Green (Vazyme) as the detection method. The primer sequences are provided in the supplementary Table 1. The 2^-ΔΔCt method was employed to calculate gene expression levels, with the adjacent tissue samples serving as the calibrator (assigned a value of 1), and comparative graphs were generated. The Ethics Committee of The Second Affiliated Hospital of Soochow University granted approval for this research (Number: JD-BS-2022-0033).

Statistical analysis

All data computations and statistical analyses were conducted with R (v4.3.2) and Python (v3.12.2). To mitigate the potential for false positives, BH correction was applied for multiple testing, employing FDR correction across multiple tests. For comparisons of continuous variables between the two groups, the Mann-Whitney U test (Wilcoxon rank-sum test) was employed to assess differences in variables that deviated from normal distribution. Survival analysis was executed using the survival package in R (v3.5.3). KM survival curves were utilized to visualize survival disparities, with the Log-rank test applied to assess the significance of survival time differences between the two groups. Univariate and multivariate Cox proportional hazards regression analyses were conducted to delineate independent prognostic factors. All statistical tests were two-sided, and a significance threshold of $P < 0.05$ was adopted.

Results

To elucidate the organizational framework of this manuscript, we have provided a delineation of the Workflow (Fig. 1).

Consensus clustering results based on immune infiltration

To explore the extent of immune cell infiltration in TCGA-BRCA patients, the xCell method was employed to calculate the immune cell infiltration levels for all samples. Based on the results of immune infiltration, samples were clustered into two subtypes (Fig. 2A-B). Specifically, C1 (Cluster 1) comprised 1003 samples, while C2 (Cluster 2) comprised 221 samples (Table 2.1.1). The differences in immune cell infiltration levels between C1 and C2 groups were calculated (Fig. 2C). Notably, our findings revealed

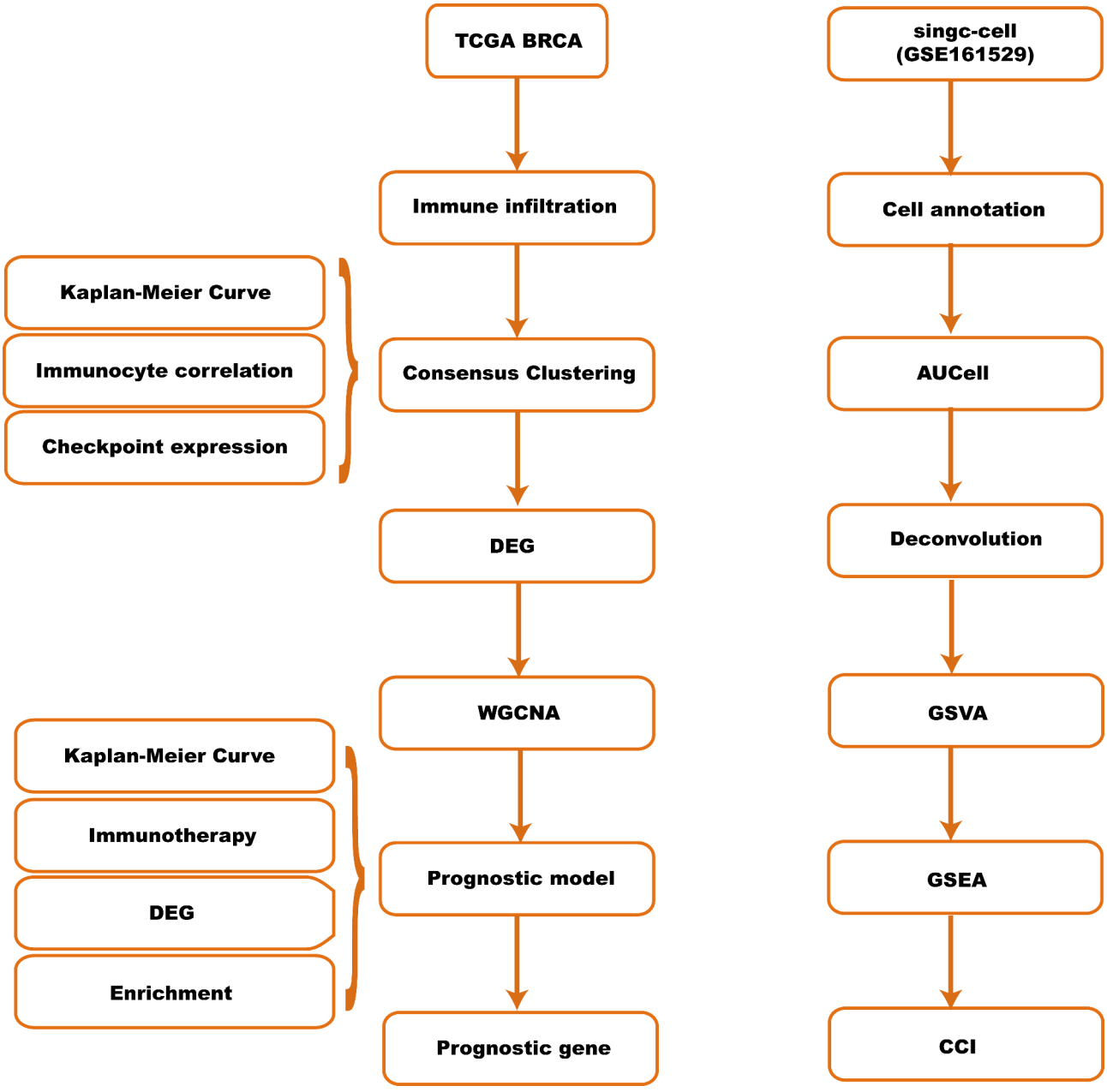

Fig. 1 Workflow. TCGA: The Cancer Genome Atlas, BRCA: Breast Cancer, DEG: Differential Expressed Gene, WGCNA: Weighted correlation network analysis, GSVA: Gene Set Variation Analysis, GSEA: Gene Set Enrichment Analysis, CCI: Cell-Cell Interaction

significant discrepancies across 29 out of 37 immune cell types between the high and low-risk groups. With the exception of mast cells, eosinophils, B cells, class-switched memory B cells, CD8+naïve T cell, Tregs, Natural killer cells, Natural killer T cell, all other cell types exhibited noteworthy differences in immune infiltration levels between the C1 and C2 groups. Specifically, common lymphoid progenitor, granulocyte-monocyte progenitor, basophils, Th1 cells, and Th2 cells demonstrated higher infiltration levels within the C1 group, while the remaining cell types exhibited heightened infiltration within the C2 group. Furthermore, we visualized the proportions of immune cell infiltration between the C1 and C2 groups using a stacked bar plot (Fig. 2D), indicating a reduced proportion of T cell subsets within the C1 group compared to the C2 group.

To delve deeper into the prognostic disparities between the C1 and C2 cohorts, we conducted survival curve analysis for both groups (Fig. 3A). The findings revealed a significant discrepancy in prognosis, with 1003 samples in C1 and 221 samples in C2 possessing survival data, showcasing a notably poorer prognosis in the

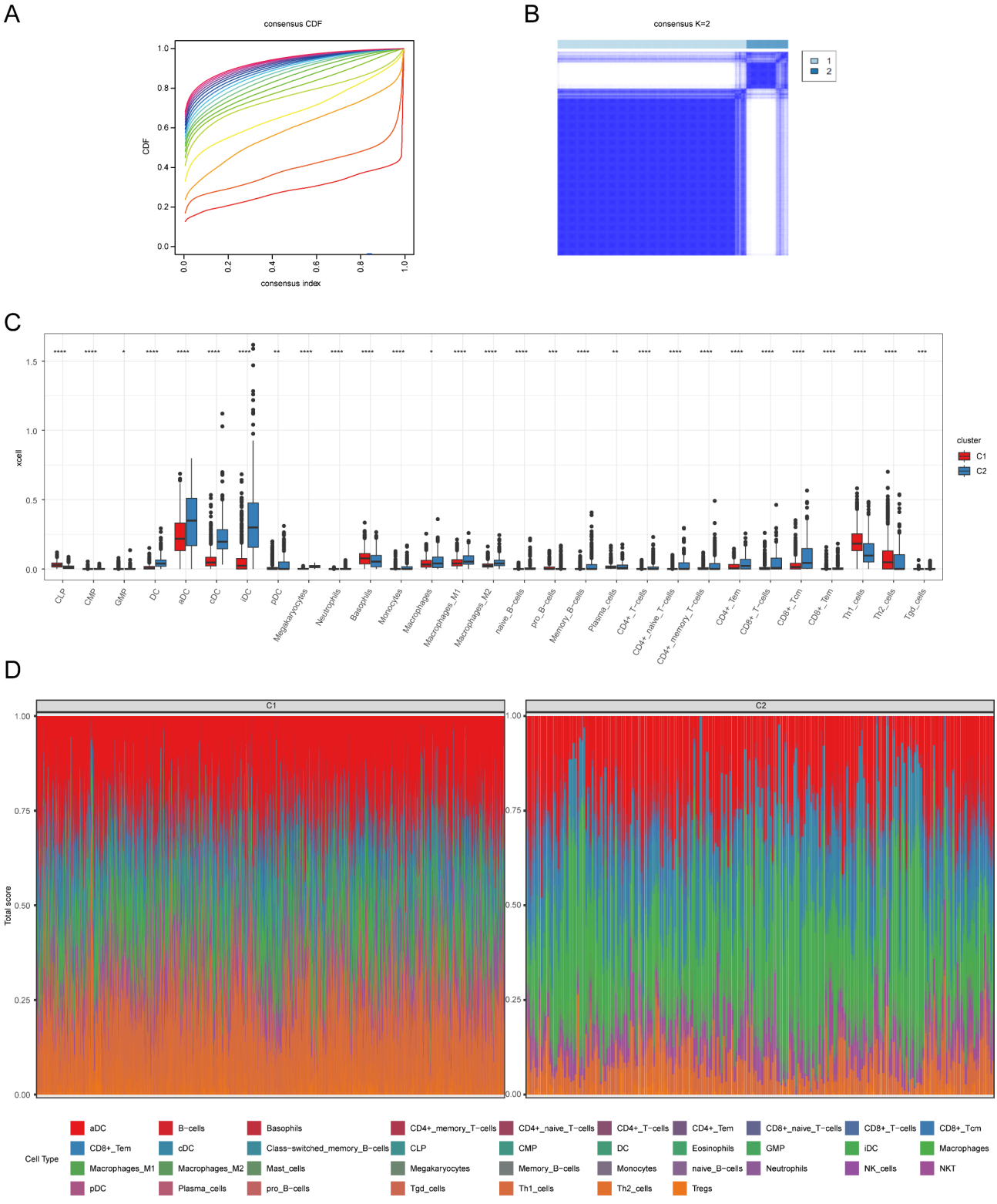

Fig. 2 Consensus Clustering Based on Immune Infiltration. **A**. Plot of cumulative distribution function in consensus clustering. **B**. Results of consistency clustering. **C**. Comparison of immune cell infiltration levels among consistent cluster subtypes. **D**. The distribution of immune cell infiltration levels in TCGA-BRCA samples from xCell analysis results. (* : $P < 0.05$, * * : $P < 0.01$, * * * : $P < 0.001$, * * * * : $P < 0.0001$)

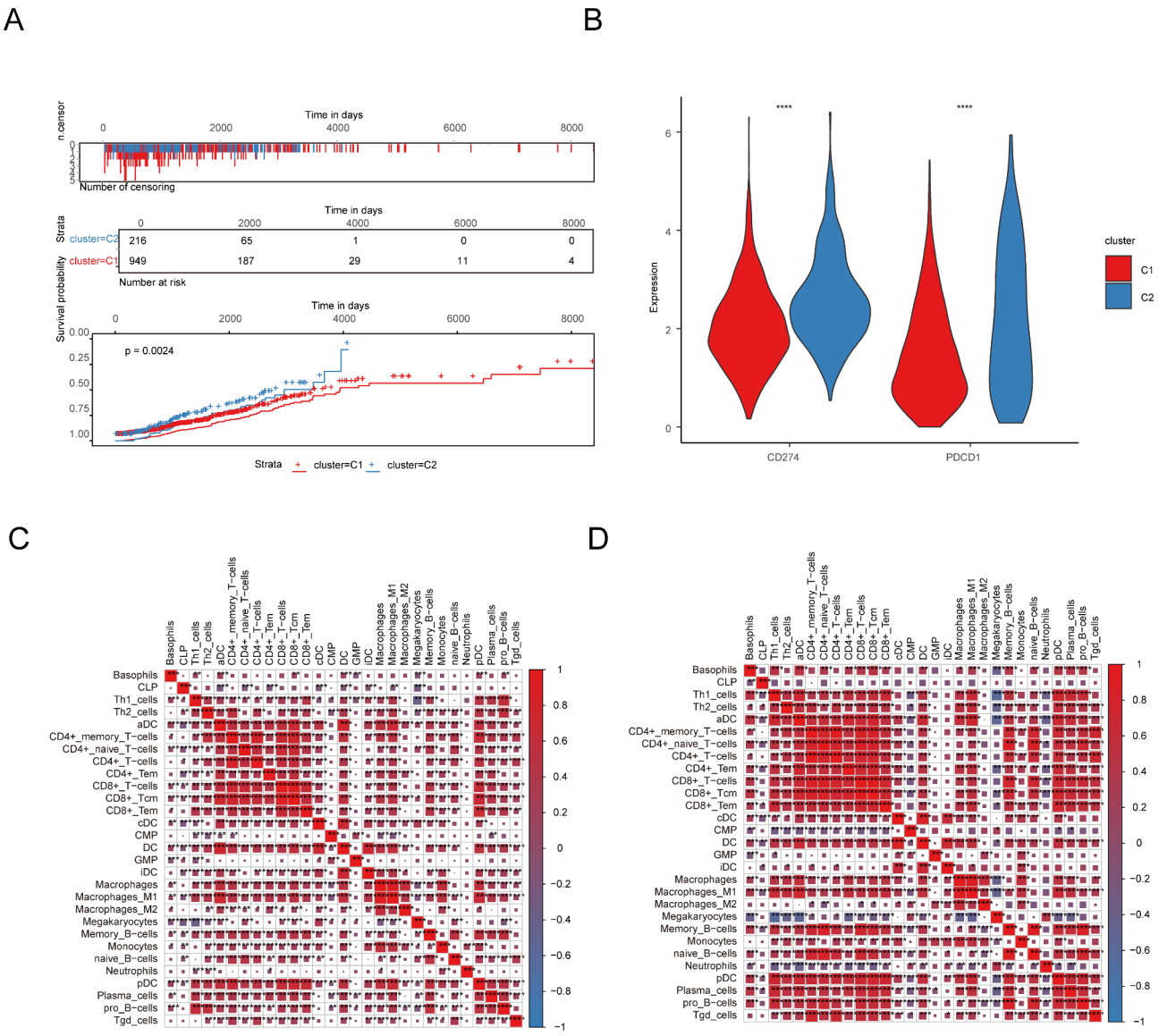

Fig. 3 Immune-cell Correlations of Consensus Clustering Subtypes Based on Immune Infiltration. **A**. Kaplan-Meier analysis of consistent cluster subtypes C1 and C2, with significance *P* values calculated by the log-rank method. **B**. Expression of immune checkpoints *PD-L1* (*CD274*) and *PD1* (*PDCD1*) among cluster subtypes. **C**. The correlation of immune cells within subtype C1 in the consensus cluster. Red represents positive correlation and blue represents negative correlation. **D**. The correlation of immune cells within subtype C2 in the consensus cluster. (* : $P < 0.05$, * * : $P < 0.01$, * * * : $P < 0.001$, * * * * : $P < 0.0001$)

latter. Given the influence of immune checkpoint expression on survival, we scrutinized the variance in *PD-L1* (*CD274*) and *PD1* (*PDCD1*) expression between the two groups (Fig. 3B). The analysis unveiled heightened levels of *PD-L1* (*CD274*) and *PD1* (*PDCD1*) expression within the C2 cohort relative to C1. The overexpression or hyperactivity of immune checkpoint molecules could engender immune suppression, culminating in compromised immunity and heightened susceptibility to cancer. Subsequent analysis involved probing the Pearson correlation of differentially expressed immune cells within the C1 and C2 groups. As illustrated in Fig. 3C, positive correlations predominated among immune cells in the C1 cohort. Figure 3D depicted a similar trend within the C2 cohort, with notably robust correlations observed, particularly between T cell and B cell subsets, alongside elevated correlations among macrophage subsets. While most cells exhibited positive correlations, a subset displayed negative correlations.

Results of weighted gene association network analysis

To investigate the differences between C1 and C2 groups, we conducted PCA principal component analysis) to assess the degree of discrimination between patient types

based on gene expression profiles (Fig. 4A). The results demonstrated noticeable dissimilarities between the samples from C1 and C2 groups (Refer to Table-2.1.2). Subsequently, we performed differential gene expression analysis, classifying genes as differentially expressed with cut-off criteria of $P<0.05$ and $|\log_2 FC| >1$. Our analysis identified a total of 1010 upregulated genes and 1648 downregulated genes. The volcano plot was listed in Fig. 4B, while the heatmap was drawn in Fig. 4C. Furthermore, we investigated the differences in the expression of genes closely associated with BRCA (*ATM*, *BARD1*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, *RAD41D*) between the C1 and C2 groups (Fig. 4D). Notably, *ATM* expression was found to be significantly lower in the C1 group compared to the C2 group, while the expression of *BARD1*, *BRCA1*, *BRCA2*, *CDH1*, *CHEK2*, and *RAD41D* was higher in the C1 group.

To identify gene modules associated with immune infiltration-based clustering, we employed WGCNA on differentially expressed genes in the C1 and C2 groups. We constructed a scale-free network with cut-off criteria of $R^2=0.9$(Fig. 5A). Subsequently, we acquired seven co-expression gene modules with a height cutoff of 0.25 (Fig. 5B). By integrating the expression patterns of module genes with immune infiltration-based clustering information, we identified eight modules (turquoise, brown, green, blue, red, black, yellow, and grey) that displayed correlation with immune infiltration-based clustering subtypes (Fig. 5C). Notably, the turquoise module exhibited the highest correlation with immune infiltration-based clustering subtypes. Utilizing a correlation scatterplot, we investigated the relationship between gene modules and clusters (Fig. 5D). Furthermore, we performed GO and KEGG pathway enrichment analysis on the 979 genes within the turquoise module (Table 2.2.1, Table 2.2.2). The GO analysis revealed enrichment in functions such as ameboidal-type cell migration and vascular process in the circulatory system (Fig. 5E). Enriched KEGG pathways included the AMPK signaling pathway and cAMPK signaling pathway (Fig. 5F).

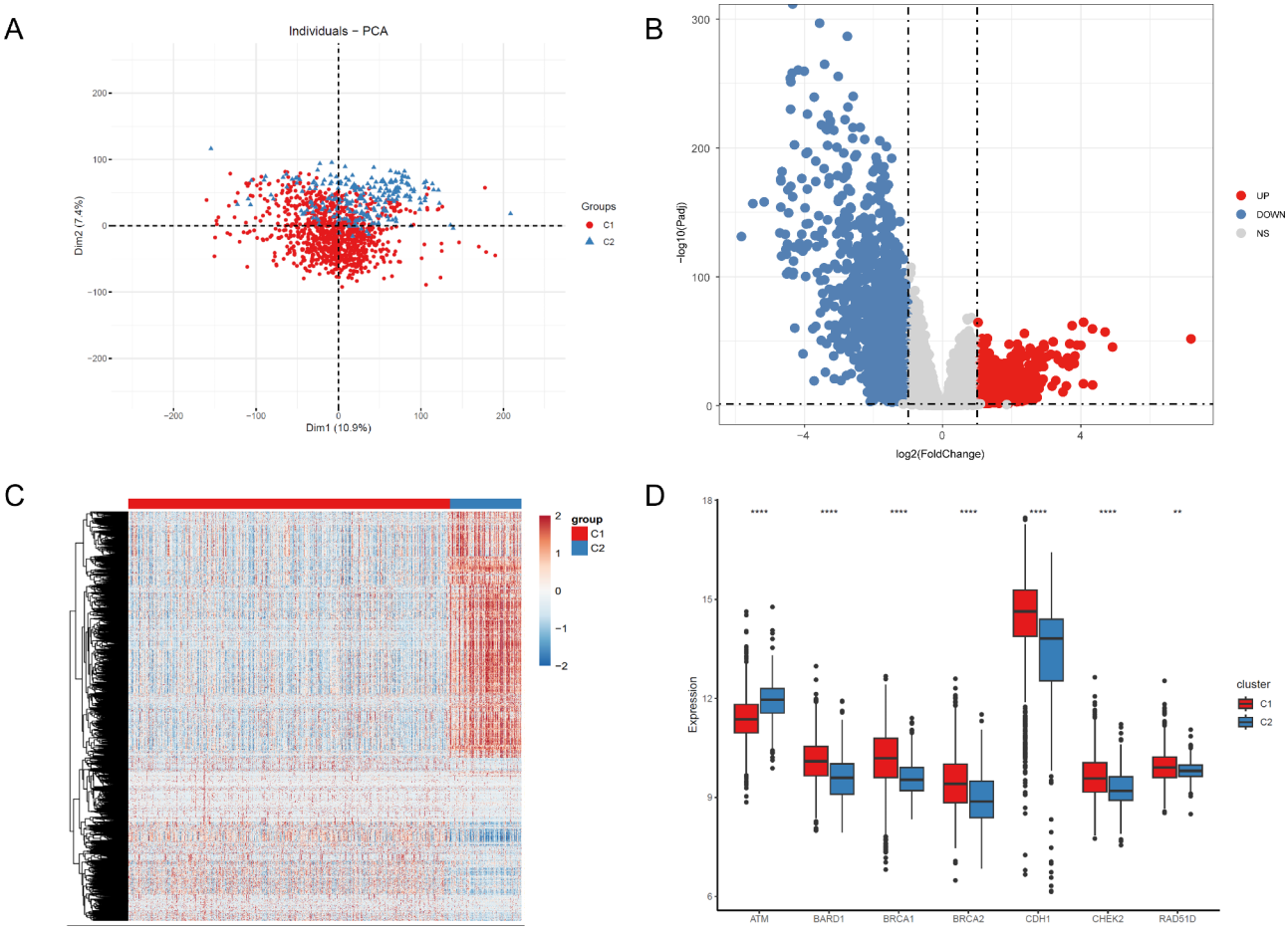

Fig. 4 Differentially Expressed Genes in Consensus Clustering Subtypes Based on immune infiltration. **A**. PCA plot of consensus cluster subtypes in TCGA-BRCA samples. **B**. Volcano plot of differential expression analysis between C1 and C2. **C**. Heatmap between the C1 and C2. **D**. Box plot of BRCA related genes *ATM*, *BARD1*, *BCRA1*, *BCRA2*, *CDH1*, *CHEK2*, *RAD51D* between C1 and C2. (* * : $P<0.01$, * * * * : $P<0.0001$). TCGA: The Cancer Genome Atlas, BRCA: Breast Cancer, NS: Not significant, PCA: Principal Components Analysis

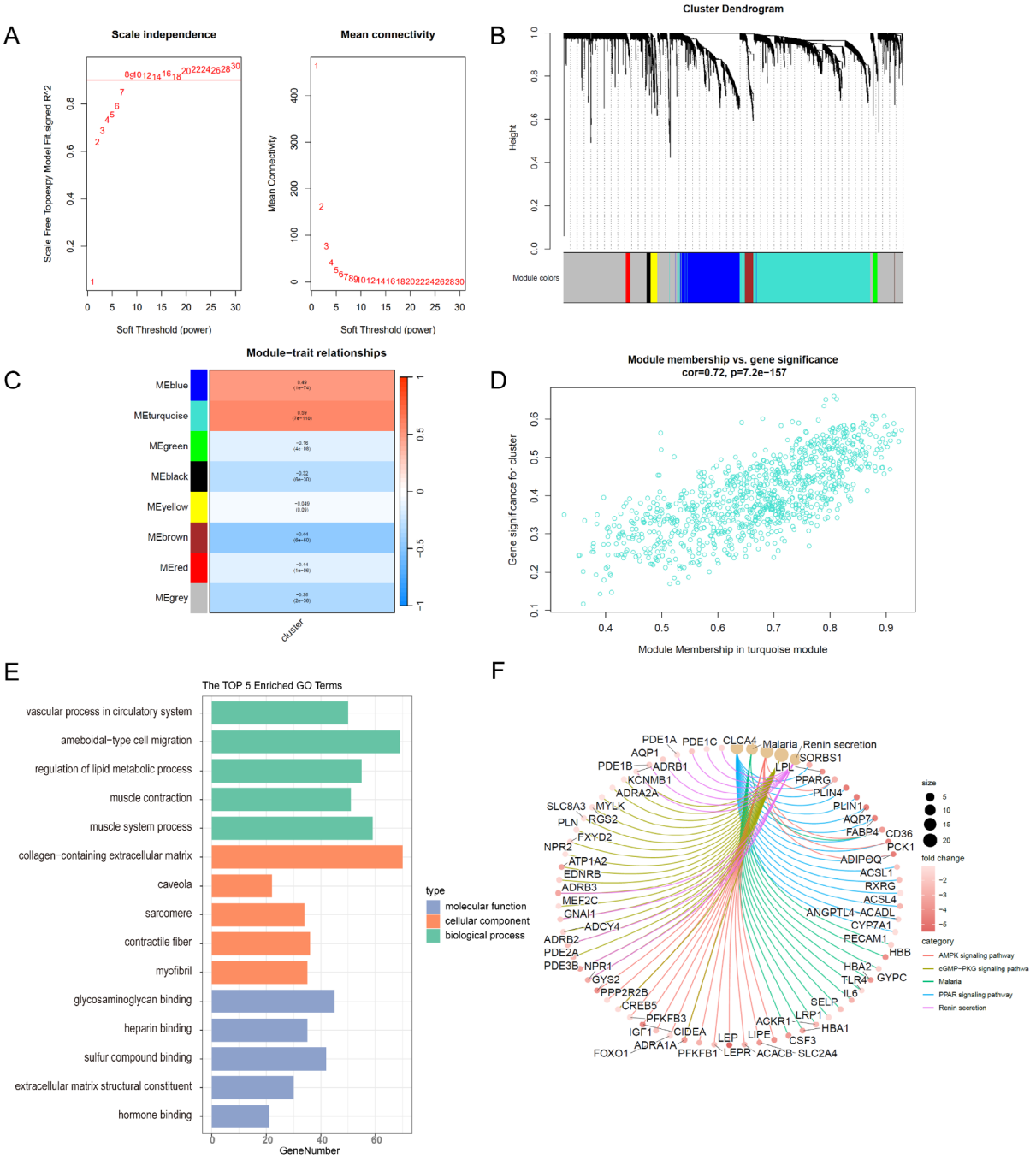

Fig. 5 WGCNA results. **A**. Network topology analysis of different soft threshold powers. **B**. Gene clustering dendrogram obtained by hierarchical clustering based on topological overlap (top) and module colors assigned by different gene clusters (bottom). Each color represents a different module, and the corresponding gene with the same color belongs to the same gene module. **C**. Heatmap of correlation between module and trait. Each row corresponds to a gene module. Each cell contains the corresponding correlation and *P* value; Red indicates a positive correlation and blue indicates a negative correlation. **D**. Scatter plot of GS versus MM in turquoise-colored modules. MM represents the correlation between the expression of eigengenes and genes. The genes in each module are highly correlated with the module to which they are assigned, indicating a high degree of connectivity within the module. GS represents the absolute value of the correlation between a gene and a phenotypic trait. Each point in the graph represents a gene. The abscissal value indicates the correlation between the gene and the module, and the ordinate value indicates the correlation between the gene and the phenotypic trait. There was a highly significant correlation between GS and MM. **E**. Bar chart of enrichment analysis results of BP, CC and MF in GO enrichment results of genes in turquoise-colored modules. **F**. Ring diagram of KEGG enrichment analysis of genes in turquoise-colored modules. GS: Gene significance, MM: Module Membership, GO: Gene Ontology, BP: Biological Process, CC: Cell Component, MF: Molecular Function, KEGG: Kyoto Encyclopedia of Genes and Genomes

Construction of prognostic model

Based on gene modules, we constructed a prognostic model. Initially, we identified 523 prognostic genes with univariate Cox regression analysis (Table-2.3.2). Subsequently, we performed multivariate regression analysis to select independent prognostic genes, 476 genes were identified (Table-2.3.3). To further refine the model, we utilized lasso-cox regression to identify the most relevant independent prognostic genes (Fig. 6A-B, Table-2.3.1), retaining seven genes with non-zero coefficients (*ACSL1*, *ABCB5*, *XG*, *ADH4*, *OPN4*, *NPR3*, *NLGN1*). These genes were used to construct a multifactor prognostic model, which was expressed as follows:

$$\begin{aligned}\text{RiskScore} = {} & 0.00482530570337132 \\ & * \exp(\text{ACSL1}) + 0.201741659673708 \\ & * \exp(\text{ABCB5}) + 0.169530781135497 \\ & * \exp(\text{XG}) + 0.0499602479296823 \\ & * \exp(\text{ADH4}) + 0.515536088167816 \\ & * \exp(\text{OPN4}) + 0.0969600762530821 \\ & * \exp(\text{NPR3}) + 0.190044768175265 \\ & * \exp(\text{NLGN1})\end{aligned}$$

Based on this risk score, the samples were categorized into high-risk and low-risk groups with median value. KM analysis was performed (Fig. 6C), revealing a notable disparity in prognosis between the high-risk and low-risk groups. We utilized 98 samples from the ICGC database for validation, and the results also confirmed that patients in the high-risk group had poorer survival rates (Supplementary Figure). To explore the relationship between the risk groups and immunotherapy, we examined the differential expression of immune checkpoints (*CD274*, *CD47*, *HAVCR2*, *LAG3*, *IDO1*, *SIRPA*, *TNFRSF4*, *PDCD1*, *CTLA4*, *TIGIT*) between the two groups (Fig. 6D). Additionally, we calculated the TIDE scores for each sample and conducted a comparative analysis between the high-risk and low-risk groups (Fig. 6E). The high-risk group exhibited significantly higher TIDE scores than the low-risk group. A higher TIDE score indicates a greater likelihood of immune evasion, suggesting a more unfavorable disease progression and prognosis. Furthermore, we evaluated the correlation between the TIDE and risk scores (Fig. 6F), revealing a positive relationship where an increase in the risk score corresponded to an upward trend in the TIDE score.

Difference analysis and enrichment analysis of high score risk group

To investigate the differential expression profiles between the high-risk and low-risk groups, differentially expressed genes were calculated and visualized with volcano plot and heatmap (Fig. 7A-B, Table-2.4.1). The plots clearly demonstrate a substantial increase in the number of upregulated genes in the high-risk group compared to downregulated genes. In order to gain insight into the potential biological functions of these differentially expressed genes, we conducted GO analysis encompassing BP (biological processes), MF (molecular functions), and CC (cellular components), as well as KEGG enrichment analysis (Table-2.4.2, Table-2.4.3). The results revealed that the upregulated differentially expressed genes are predominantly enriched in GO terms such as GO:0045229, GO:0030198, and GO:0043062 (Fig. 7C-D). Moreover, KEGG analysis highlighted the involvement of pathways such as Viral protein interaction with cytokine and cytokine receptor and Tyrosine metabolism in the upregulated differential gene expression observed in the high-risk group (Fig. 7E). To provide a clearer representation of the top pathways, we visualized them using a network plot (Fig. 7F).

Single cell annotation of breast cancer

The Python package scanpy was utilized to import the Counts matrix of the scRNA-seq dataset (GSE161529) for quality control analysis. The dataset comprised a total of 193,167 cells, but after applying strict filtering based on three QC covariates, we retained 187,333 cells for further analysis.

Normalization and batch correction were performed on the raw data to address differences introduced by experimental batches. Subsequently, we conducted clustering analysis on the batch-corrected data. To construct a graph capturing the cellular neighborhoods and compute cell-cell distances and similarities, we employed the neighbors function. Then, we applied graph-based clustering and embedded the neighborhood graph using the UMAP algorithm with Leiden algorithm (Fig. 8A). To provide cell type annotations, the authors of the GSE161529 dataset provided marker genes. Leveraging these genes and their known cell type mappings, we conducted ORA using the Python package decouperR. This analysis highlighted the most probable cell types: dendritic cells, endothelial cells, fibroblast cells, B cells, macrophages, and T cells (Fig. 8B).

To investigate the expression patterns of the seven genes (*ACSL1*, *ABCB5*, *XG*, *ADH4*, *OPN4*, *NPR3*, *NLGN*1) in the scRNA-seq dataset, we visualized gene-level expression trajectories at the cellular level. *ACSL1* exhibited high expression across almost all cells, while the other genes displayed specific expression distributions among different cell populations (Fig. 8C). Furthermore, we generated average expression heatmaps for the top five genes with the greatest differences in expression levels across distinct cellular subgroups (Fig. 8D). The expression values were normalized using scaling techniques.

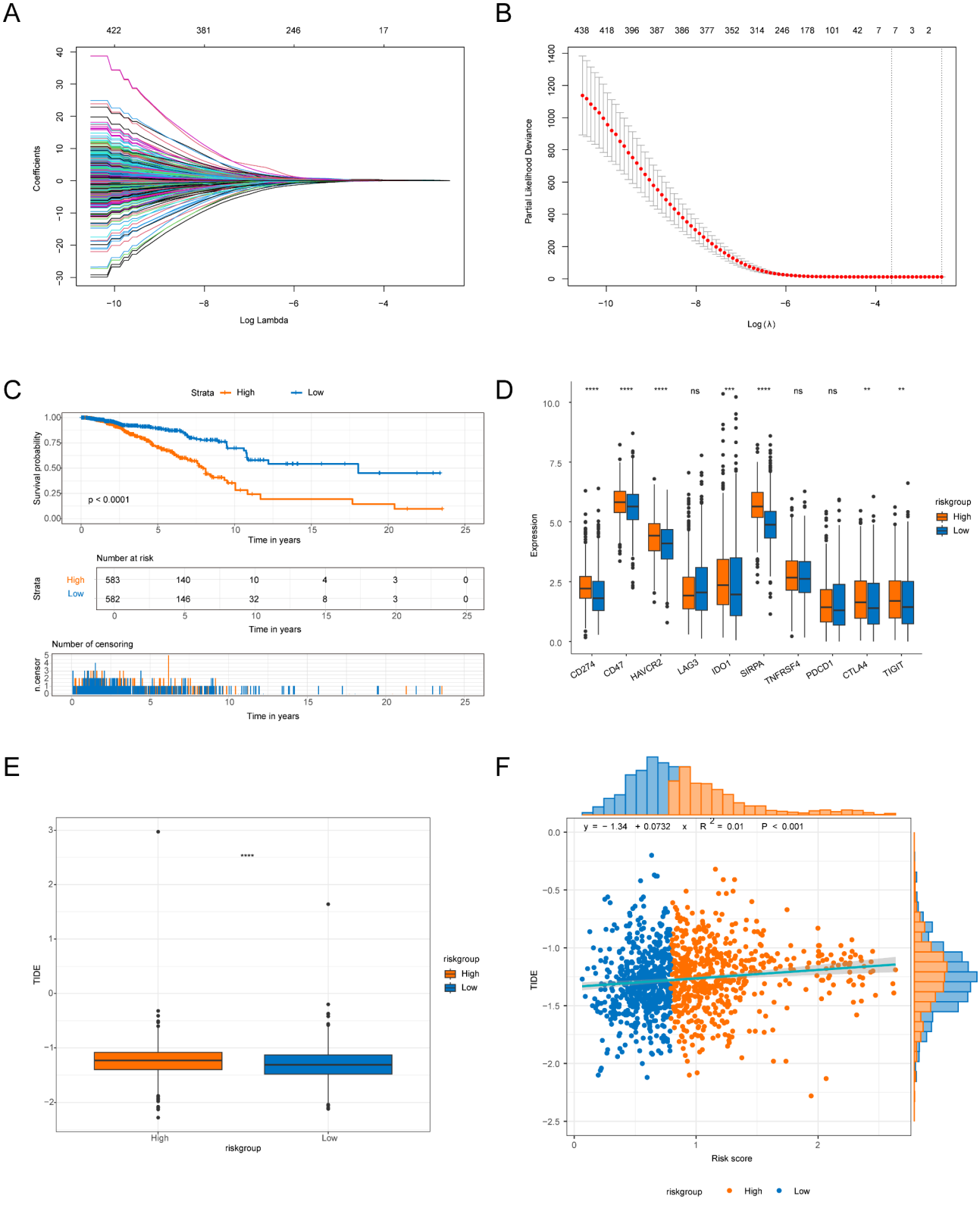

Fig. 6 Construction of the TCGA-BRCA prognostic model. **A**. LASSO-cox regression curve for gene features. **B**. The coefficient plot of LASSO-cox regression. **C**. Kaplan-Meier survival curves between high-risk and low-risk groups. **D**. The expression differences of immune checkpoints *CD274*, *CD47*, *HAVCR2*, *IDO1*, *LAG3*, *SIRPA*, *TNFRSF4*, *PDCD1*, *CTL4A*, *TIGIT* in high and low risk groups, and the horizontal axis is the normalized expression value. **E**. Differences of TIDE score between high and low risk groups. **F**. TIDE and risk score. TCGA: The Cancer Genome Atlas, BRCA: Breast Cancer, LASSO: Least Absolute Shrinkage and Selection Operator, TIDE: Tumor Immune Dysfunction and Exclusion. (ns: $P > 0.05$, * * : $P < 0.01$, * * * : $P < 0.001$, * * * * : $P < 0.0001$)

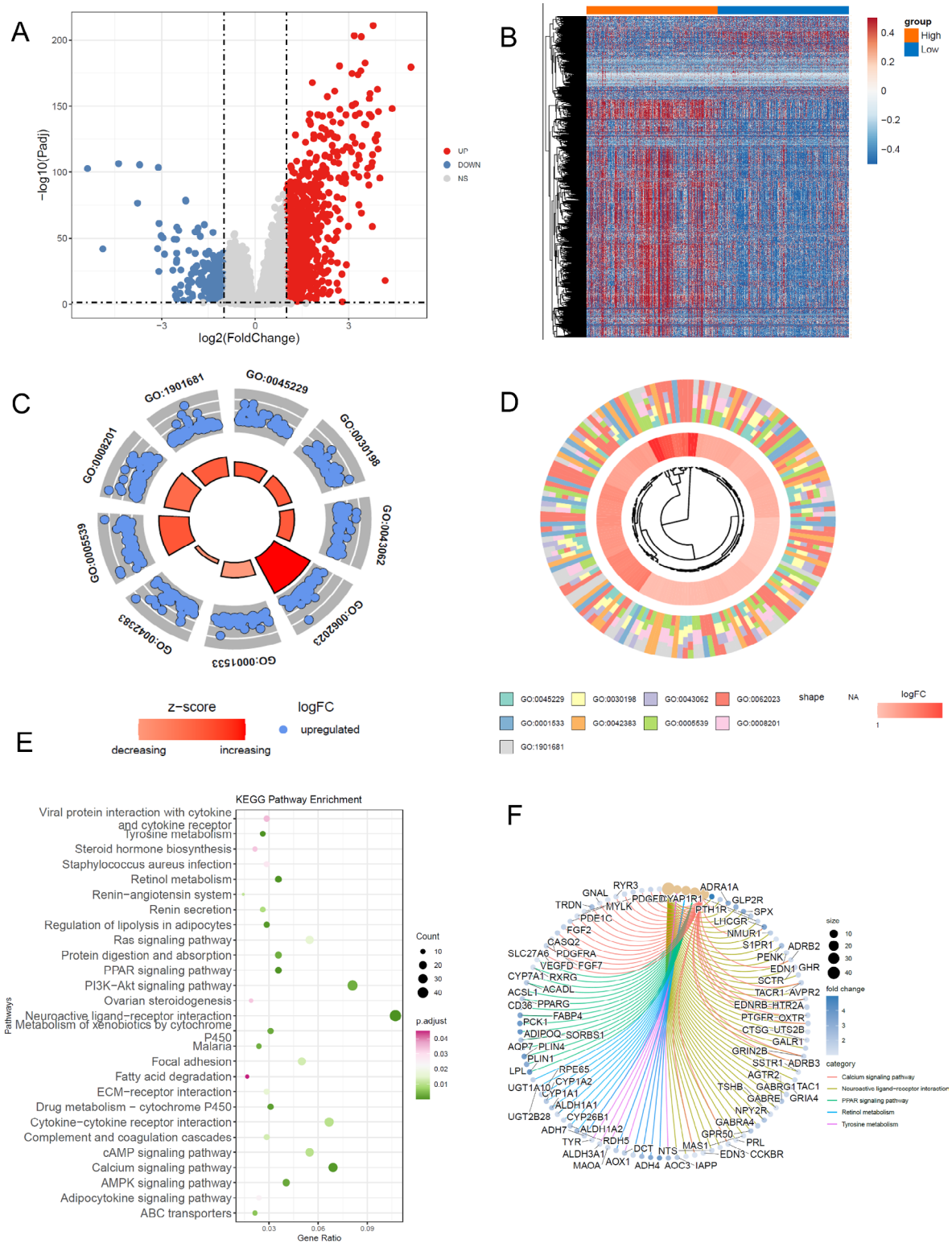

Fig. 7 Differential expression analysis of TCGA-BRCA high and low risk groups. **A**. Volcano plot of differential analysis between high and low risk groups. **B**. Heatmap of differential analysis between high and low risk groups. **C**. Circle diagram of enrichment analysis results of BP, CC and MF in GO enrichment results of up-regulated genes. **D**. Ring clustering dendrogram of enrichment analysis results of BP, CC and MF in GO enrichment results. **E**. KEGG enrichment pathways. F. KEGG enrichment pathway network diagram. TCGA: The Cancer Genome Atlas, BRCA: Breast Cancer, GO: Gene Ontology, BP: Biological Process, CC: Cell Component, MF: Molecular Function, KEGG: Kyoto Encyclopedia of Genes and Genomes

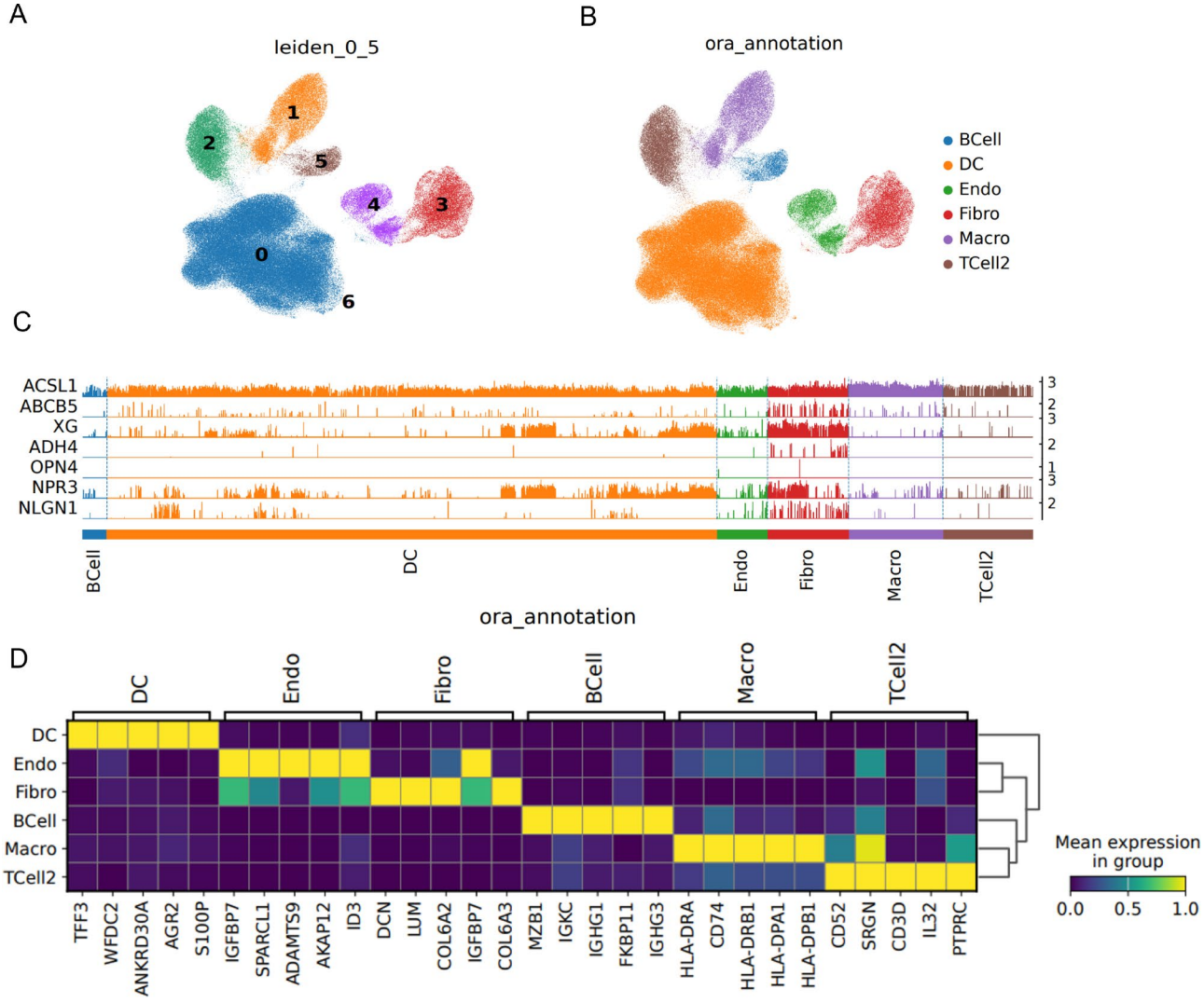

Fig. 8 Annotation of single-cell Cell types. **A**. The UMAP map of Single-cell Leiden cluster. **B**. The UMAP map of over representation analysis annotation. **C**. The trajectory map of prognostic gene. **D**. Heatmap of differentially expressed genes in cell subsets

To investigate the differences in the expression of seven prognostic genes (*ACSL1*, *ABCB5*, *XG*, *ADH4*, *OPN4*, *NPR3*, *NLGN1*) among different cell subgroups, we employed the AUCell algorithm to calculate scores for each subgroup individually. The AUCell scores were visualized with UMAP plots (Fig. 9A) and grouped violin plots (Fig. 9B). Our analysis revealed that macrophages exhibited higher AUCell scores, indicating a significant correlation between macrophages and immune prognosis, thus classifying them as a high-risk subgroup. Additionally, we created grouped violin plots to illustrate the expression levels of the prognostic genes (Fig. 9C-I) and found that *ACSL1* gene demonstrated notably elevated expression specifically within the macrophage subgroup.

Furthermore, we aimed to explore additional biological functions associated with the high-risk macrophage subgroup. To achieve this, we identified genes that were differentially expressed between the macrophage subgroup and other subgroups. Subsequently, we conducted GSEA utilizing fold changes of these genes (Fig. 10A). The GSEA results revealed significant enrichment in various pathways including signal transduction through IL1R, salvador martin pediatric TBD anti TNF therapy nonresponder post treatment up, reactome purinergic signaling in leishmaniasis infection, IL8 CXCR2 pathway, IL12 STAT4 pathway, and dutta apoptosis via NFKB (Fig. 10B-H). These findings provide insights into potential biological functions associated with the high-risk macrophage subgroup.

Single-cell GSVA

To explore the heterogeneity of the HALLMARK gene set across different cell types, we conducted GSVA (Table-2.5.1) on all genes and generated GSVA heatmap

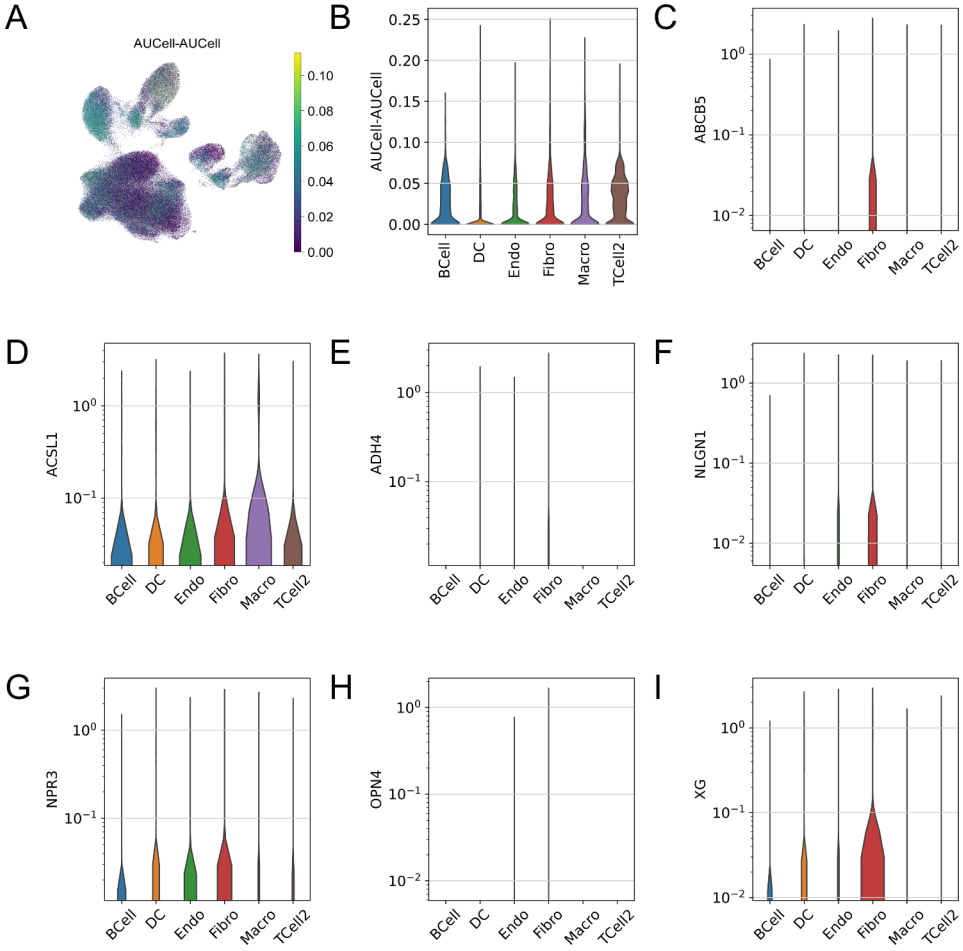

Fig. 9 High-risk Cell subsets in single-cell data. **A**. The UMAP map of single-cell AUCell score, brighter colors represent higher enrichment scores. **B**. The violin plot of single-cell AUCell score. **C**-**I**. Violin plots comparing the expression of prognostic related genes in groups

(Fig. 11A). The findings reveals that the majority of the HALLMARK gene set exhibits higher scores in dendritic cells (DCs), while there is enrichment observed in macrophages (Macros) for pathways such as HALLMARK allograft rejection, HALLMARK complement, and STAT3. Moreover, we utilized the Python package Liana's CellPhoneDB to implement the ligand-receptor method (Fig. 11B), followed by performing consensus cell communication analysis based on different methods using LIANA (Fig. 11C). This analysis highlights the top 20 ligand-receptor interactions, with notable interactions observed predominantly such as HLA-DRA-> CD4 and TNFSF13B-> HLA-DPB1 are particularly significant.

The proportion of single-cell subsets in TCGA-BRCA samples was inferred by deconvolution

The intricate network of cellular interactions governs the interplay between the immune system and tumor cells. Understanding the composition of specific immune cells within solid tumors is paramount for predicting patient responses to immunotherapy. Leveraging subpopulation information from single-cell sequencing data of BRCA, we inferred the cellular composition and proportions of each sample in TCGA-BRCA (Fig. 12A). The results revealed that the proportion of EC (endothelial cells) in the high-risk subgroup was highest across all samples. Subsequently, we categorized TCGA-BRCA samples into high and low EC content groups based on the median proportion of EC cells in the high-risk subgroup and compared the survival analysis between the two groups (Fig. 12B). The findings indicate no significant differences in survival prognosis between patients in the high and low EC groups ($P=0.82$). Furthermore, we analyzed the relationship between EC content and risk scores (Fig. 12C). The results demonstrate an increasing trend in EC content with higher risk scores ($P<0.001$), suggesting a certain degree of correlation between the two. Subsequently, we further examined the joint impact of EC content and risk scores on survival analysis (Fig. 12D). The results reveal the worst survival prognosis for the group with both high risk scores and high EC content. It is noteworthy that although high risk scores and high

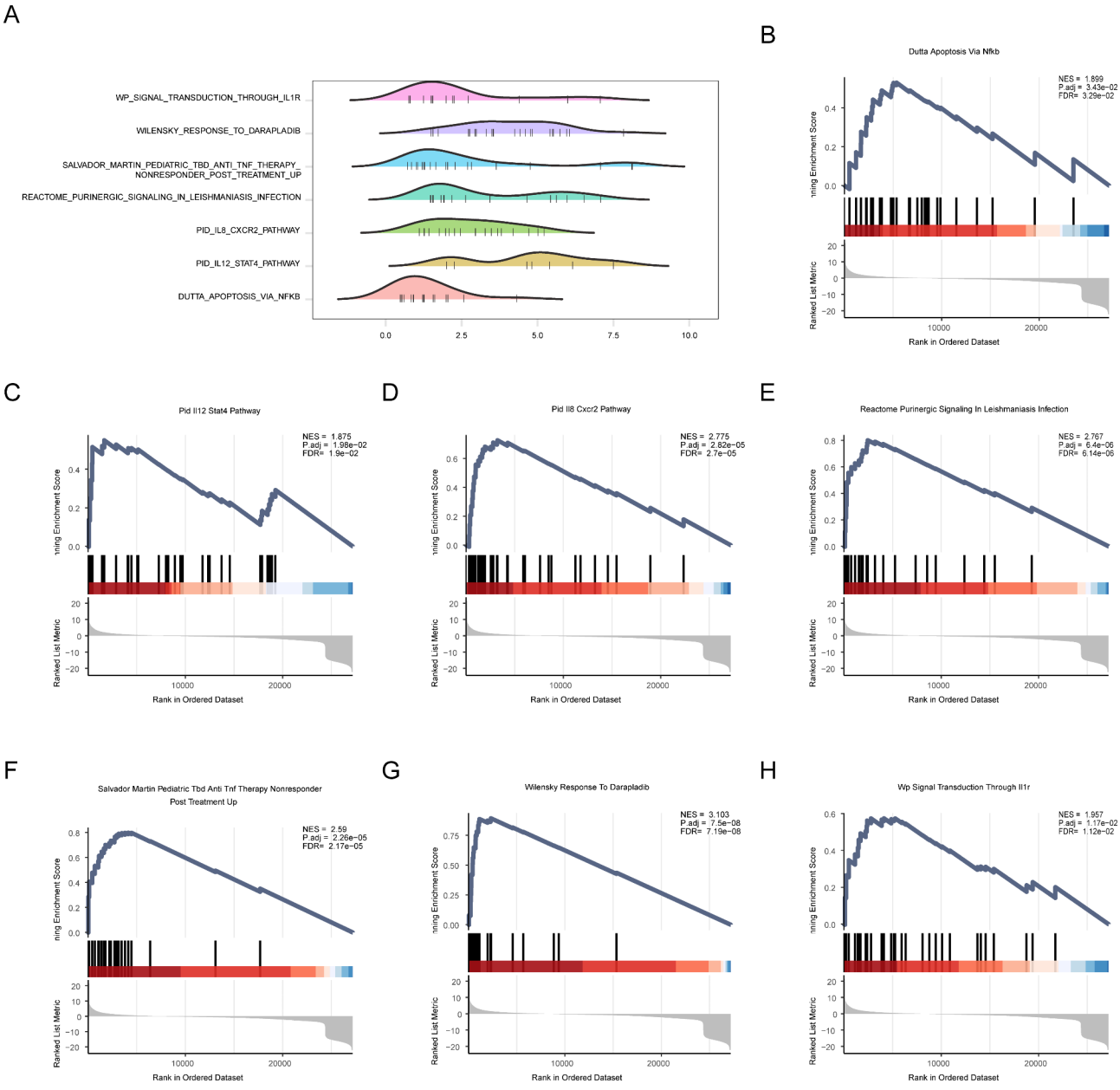

Fig. 10 High-risk cell subpopulation GSEA. **A**. GSEA mountain map of risk cell subpopulation. **B**-**H**. GSEA classic diagram of risk cell subgroup. GSEA: Gene Set Enrichment Analysis

EC content individually have some impact on survival prognosis, there is no significant difference in survival between high and low EC content groups within the low-risk score group, as well as between high and low EC content groups within the high-risk score group.

The expression of prognostic genes in breast cancer tissue

The real-time PCR results revealed substantial differences in gene expression levels between BRCA tissue and adjacent non-cancerous tissue. Specifically, *ABCB5*, *ADH4*, and *NLGN1* exhibited notably reduced expression in BRCA tissue, implying a potential involvement of these genes in tumorigenesis through inactivating mutations (Fig. 13). We have identified significant expression differences of these genes in BRCA tissues, indicating their potential for constructing a robust prognostic model. Further validation with an expanded sample size is necessary to confirm the expression patterns of these genes in BRCA.

Discussion

BRCA is one of the most common malignant tumors detected in women worldwide, having a great adverse impact on their health and quality of life. In recent years,

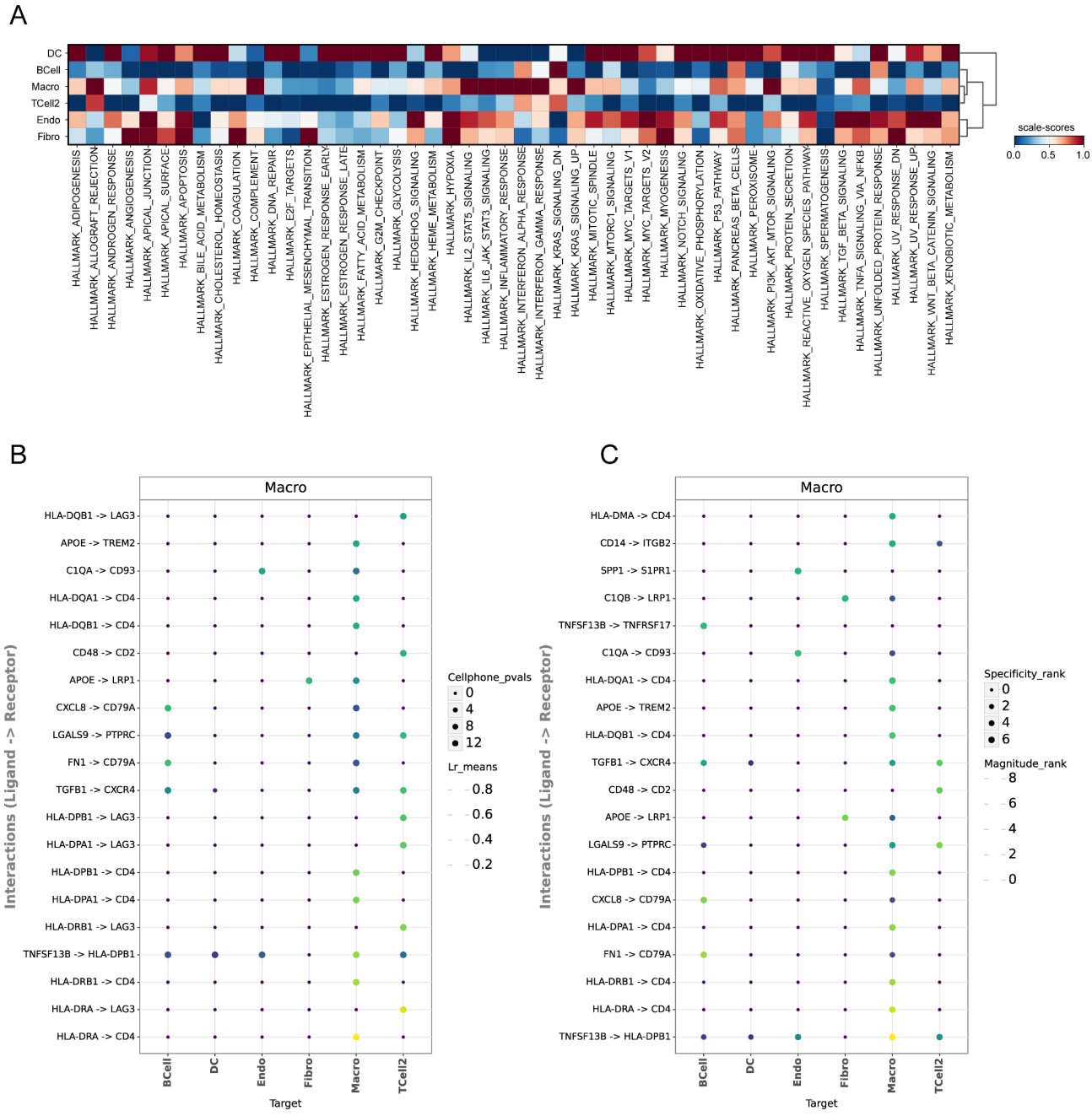

Fig. 11 High risk cell subpopulation GSVA, CCI. **A**. Heatmap of single cell GSVA, red color represents high enrichment. **B**. Bubble map of classical CellPhoneDB cell communication analysis, the size of the dots represents the confidence level and the color from dark to light represents the stronger communication effect. **C**. The consensual-based bubble plot of cell communication analysis implemented by LIANA, the size of the dots represents the confidence level and the color from dark to light represents the stronger communication effect. GSVA: Gene Set Variation Analysis, CCI: Cell-Cell Interaction

remarkable progress has been made in immunotherapy, CDK4/6 inhibitors and ADC (Antibody-drug Conjugate) drugs, which have greatly improved the prognosis of BRCA patients [39, 40]. However, BRCA is still one of the leading causes of cancer-related deaths in women [1]. Therefore, it is urgent to optimize diagnosis methods and treatment strategies of BRCA. The complexity of BRCA is manifested as multiple subtypes with different molecular characteristics, which make it difficult to effectively manage the disease [41]. Through the comprehensive bioinformatics analysis, this study aims to improve the understanding of BRCA and discover new biomarkers and therapeutic targets.

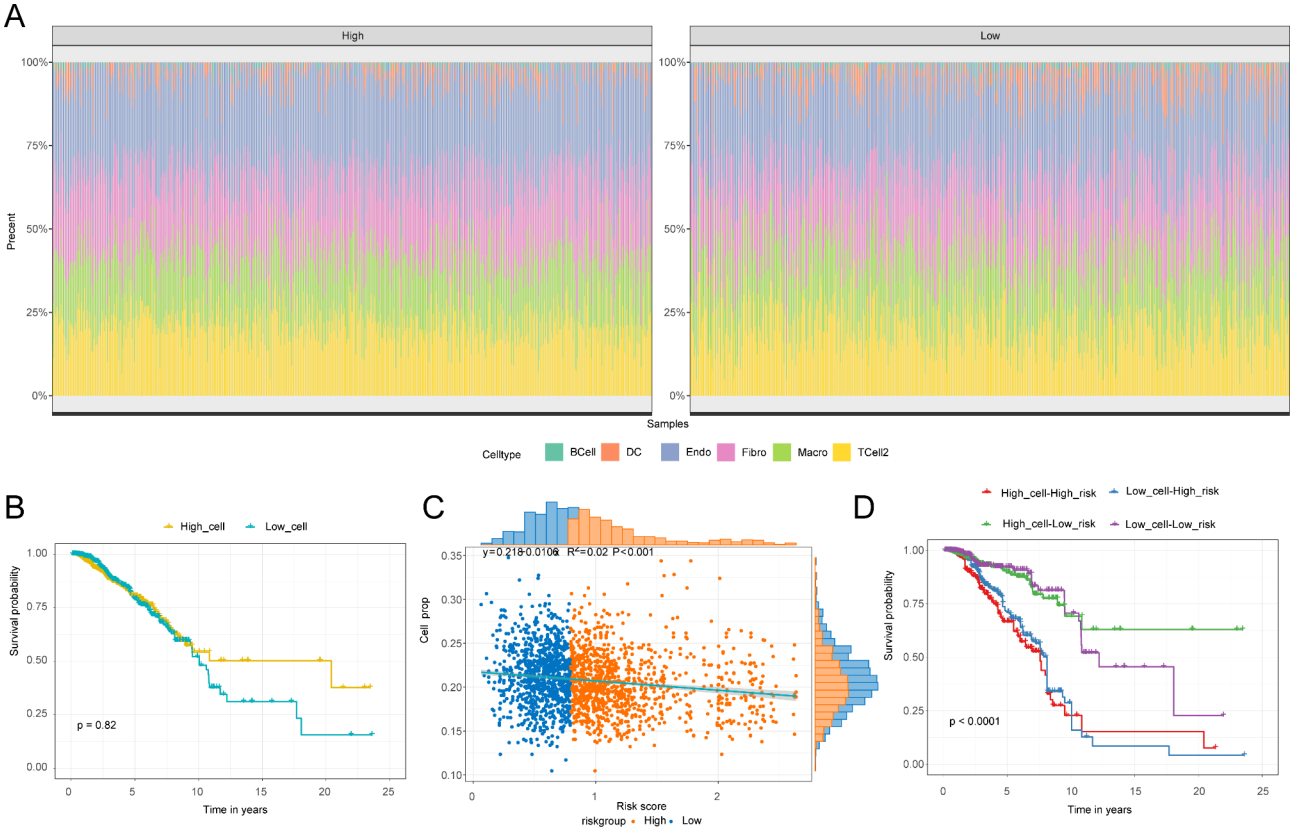

Fig. 12 Deconvolutional Inference of Cell Proportion in TCGA-BRCA Samples. **A**. Bar chart of the proportion of single cells in the sample in the TCGA-BRCA high and low risk groups. The abscissa represents the sample, and the ordinate represents the proportion of the inferred single-cell subpopulation in the TCGA-BRCA sample. **B**. The effect of the proportion of endothelial cells subsets in the sample on the survival prognosis. **C**. Scatter plot of the correlation between the proportion of endothelial cells subsets and the risk score. **D**. The Kaplan-Meier subgroup analysis of risk score group and the endothelial cells proportion. TCGA: The Cancer Genome Atlas, BRCA: Breast Cancer

The application of bioinformatics methods to the study of BRCA phenotypes may change the therapeutic intervention strategy of this disease. The Qualcomm data analysis enables researchers to investigate the molecular heterogeneity of BRCA and the clustering analysis helps supplement the traditional molecular subtypes. The analysis results guide the development of more accurate management methods [42]. In this study, samples were clustered into C1 and C2 groups based on the results of immune infiltration, and the difference in prognosis between the two groups was measured based on the abundance of immune cell infiltration, immune checkpoints and prognosis. At the transcription level, the biological differences between C1 and C2 groups were revealed. The differentially expressed genes in C1 and C2 groups were analyzed by WGCNA and clustered by consistent immune infiltration. Then, seven differentially expressed genes were selected for the single-factor and multi-factor regression analysis of prognosis. In addition, the TIDE values of immune checkpoints were calculated and their functions were analyzed. The association between single-cell sequencing data and immune cells was evaluated. The GSEA, GSVA and prognosis analyses were made to clarify the potential mechanism of BRCA and to identify the factors that might improve the prognosis of patients. The seven differentially expressed genes (*ACSL1*, *ABCB5*, *XG*, *ADH4*, *OPN4*, *NPR3*, *NLGN1*), which showed significant correlations with TIDE scores, were used to build a robust prognosis. The single-cell analysis highlighted the significance of these prognostic genes in cell subtype-specific expression patterns. The cell communication analysis was used to explore the ligand-receptor interaction model, and the signal network between different cell types was revealed, which promoted the understanding of the interaction between cells. The seven differentially expressed genes in clinical tumor specimens had potential value in model construction.

These differentially expressed genes have many functions in various cancer types. First of all, *ACSL1* plays a key enzyme role in fatty acid metabolism, and is closely related to the proliferation and survival of cancer cells. Its expression level is low in normal breast tissue, but it is significantly up-regulated in BRCA tissue [43]. In

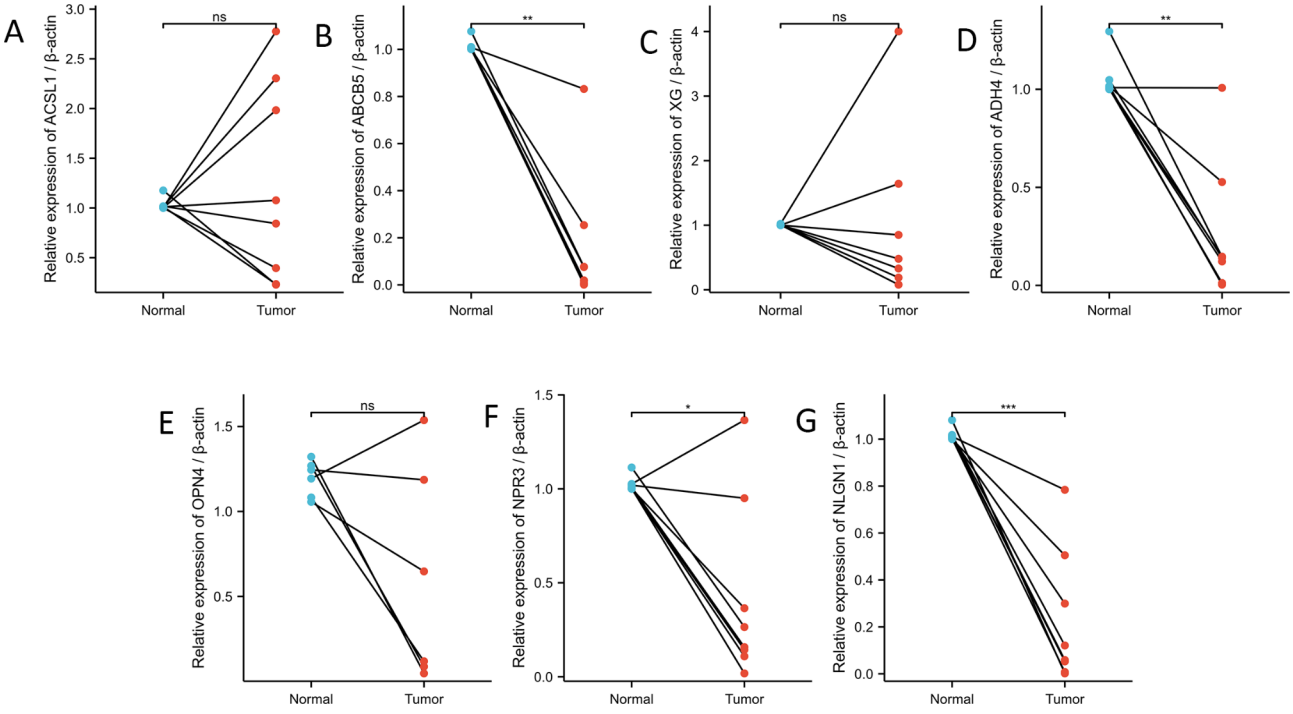

Fig. 13 The expression of prognostic genes in breast cancer tissue. **A**-**G**. RT-PCR revealed the expression of prognostic genes in breast cancer tissue. RT-PCR: Real-time-PCR (ns: $P > 0.05$, * : $P < 0.05$, * * : $P < 0.01$, * * * : $P < 0.001$)

addition, in lung cancer, the combined use of *ACSL1* inhibitors and chemotherapy can reduce drug resistance. It demonstrates the therapeutic potential of *ACSL1* in the treatment and management of BRCA [44]. *ABCB5* is associated with poor prognosis in many cancer types [45–47]. Its expression level is related to the retention of doxorubicin in BRCA cells, which may be involved in the mechanism of chemotherapy resistance [48]. *OPN4* serves as an oncogene in melanoma and affects the cell cycle [49]. In lung adenocarcinoma, the activation of *OPN4* triggers the PKC/BRAF/MEK/ERK signal cascade, and small molecular inhibitors of *OPN4* are effective in inhibiting the proliferation of lung cancer cells [50]. Although there is little research on *OPN4* in BRCA, it has a profound impact on tumor pathobiology and prognosis evaluation of patients. *NPR3* plays a dual role in tumorigenesis. It induces hepatocellular carcinoma cell apoptosis and inhibits tumor growth by inhibiting the PI3K/AKT pathway simultaneously [51, 52]. However, it may also stimulate the proliferation of colon cancer cells [53]. *NLGN1* promotes the invasion and migration of cancer cells in neural networks and enhances the invasion activity of colon cancer cells [54, 55]. It is worth noting that research on the roles of *XG* and *ADH4* in tumorigenesis is still in the primary stage, and there is no comprehensive mechanism clarifying their effects. Their functions are mainly recorded in the framework of prognosis modeling.

The RT-qPCR analysis was performed to assess the expression levels of the seven differentially expressed genes in clinical BRCA specimens. The paired-sample comparative analysis revealed substantial disparities in gene expression in different cancerous tissues. This discrepancy indicates potential gene mutations, possibly manifested as either gene inactivation or overexpression compared with the expression level in normal counterparts, in BRCA pathology.

The analysis of the differentially expressed genes revealed the significant correlation of the key pathways related to BRCA subtypes with immune infiltration characteristics. In particular, the pathways related to immune responses, such as cytokine-receptor interactions, cAMP and AMPK pathways, have been proved to play a key role in the process of the immune system attacking cancer cells [56, 57]. AMPK acts as a central regulator in the cell metabolic pathway, and has a significant impact on tumor cells by regulating energy metabolism and inflammatory responses. The metabolic activity of tumor cells leads to the lack of nutrients in the surrounding immune cells, thus reducing the activity of immune cells. Immune cells need various metabolic pathways to produce effect. AMPK-mediated inflammatory reaction is helpful to the gathering of immune cells in tumor microenvironments and can hinder the occurrence, development and metastasis of tumors. Therefore, AMPK plays a crucial role in connecting cell energy homeostasis, tumor biology and anti-tumor immunity, capable of improving

the treatment and management of cancer patients [58]. CAMP may also inhibit the proliferation of cancer cells, and its effect depends on the environment and tumor type. Tumor-related stromal cells, such as cancer-associated fibroblasts (CAFs) and immune cells, release cytokines and growth factors to stimulate or inhibit the production of cyclic adenosine monophosphate (cAMP) in tumor microenvironments [59]. The imbalance in these pathways may lead to the imbalance of immune responses, thus affecting the progress of tumors.

The integration of single-cell sequencing data in BRCA research provides us with profound insights into tumor microenvironments and tumor cellular heterogeneity. The clustering analysis of single-cell data showed the expression trajectory of all prognostic genes, high expression of *ACSL1* in almost all cells and the cell-specific expression patterns of other genes. This specificity indicates that some cell populations may have a unique impact on tumor behavior and patient prognosis. Macrophages had higher AUCell scores than other cell types, so they were closely related to immune prognosis and considered high-risk cells. This finding is consistent with that of a recent study, which concludes that macrophage polarization plays an important role in cancer and macrophages are a potential therapeutic target [60].

The analysis of the role of genes in intercellular communication and their impact on the immune landscape of BRCA may promote the development of novel therapeutic strategies. The analysis of macrophage cell communication pinpointed that HLA-DRA-> CD4 and TNFSF13B-> HLA-DPB1 were particularly noteworthy. The HLA family of genes plays a crucial role in immune recognition and the presentation of antigens. Research has highlighted a concerning link between the abnormal expression of genes like HLA-DRA and HLA-DPB1 and poor outcomes in various cancers [61, 62]. HLA-DRA plays a vital role in the presentation of antigens to CD4+T cells, its diminished expression may precipitate a cascade of impaired T cell activation, thereby undermining the anti-tumor immune responses [63, 64]. Additionally, the increased interaction between TNFSF13B (also known as BAFF) and HLA-DPB1 may enhance the growth and survival of cancer cells, as higher levels of BAFF have been closely associated with several types of cancer, including breast cancer [65].

GSVA based on single-cell data further elucidated the enrichment of differentially expressed genes in high-risk cells in immune-related pathways. The pronounced enrichment in macrophage subpopulations relative to other cells underscores the significant contributions of macrophages to the immune landscape of BRCA. It improves the understanding of immune therapy responses.

The deconvolution analysis unveiled a positive correlation between the proportion of high-risk cells and risk scores in BRCA samples. The high-risk and high EC content groups had the poorest survival prognosis. We identified numerous ligand-receptor interactions that are intricately linked to the signaling pathways involved in tumor immune suppression via GSEA and GSVA analyses. A key finding is the abnormal activation of the IL1R and IL8-CXCR2 signaling pathways, which can create a chronic inflammatory environment that promotes cancer cell growth and contributes to resistance against treatments. Similarly, the IL12-STAT4 signaling pathway shows increased activity in high-risk patient groups, which may lead to unusual inflammatory responses and help tumors evade the immune system. Our analysis reveals a complex network of cellular communication, highlighted by ligand-receptor pairs like the HLA family genes interacting with T cells, showcasing how various immune evasion mechanisms work together. This intricate web of interactions indicates that tumor cells can manipulate host immune responses through multiple pathways, giving them a significant advantage in their survival.

In recent years, a multitude of articles has graced the immune landscape. Through meticulous dataset analysis and clustering methodologies, BRCA has been delineated into six distinct subtypes. This classification is based on the unique traits of each subtype and serves as a crucial foundation for personalized breast cancer treatment [66]. Charles et al. elucidated common mutation profiles through the lens of PAM50 classification, thereby unveiling potential therapeutic targets that may enhance clinical interventions for breast cancer [67]. Meanwhile, Hu et al. forged a prognostic model rooted in the intricacies of breast cancer stem cell-related genes, with a keen focus on informing clinical decision-making and anticipating immune responses [68]. In contrast, our article explores the subtle differences in immune infiltration by utilizing a model based on clustering analysis that more effectively captures the effects of immune dynamics. Additionally, we conducted extensive multi-dimensional analyses using single-cell datasets to better understand the classification of immune landscapes and cellular interactions, with the goal of introducing new stratification factors to improve the clinical management of breast cancer.

There are certain constraints that may impact the interpretation of the findings of this study. Firstly, there was a lack of wet-lab experiments to validate bioinformatics predictions. Experimental validation is crucial for confirming the biological relevance of computational findings. Secondly, the sample size, although adequate for initial discovery, may be too small to generalize the results across a broader patient population. Thirdly, the clinical validation analysis was not conducted to test

the prognostic models and biomarkers identified, but this analysis is essential for their translation into clinical practice. Additionally, the use of multiple datasets might introduce batch effects, potentially confounding the results. Despite that rigorous bioinformatics approaches were used to minimize such effects, they could not be completely eliminated. Ultimately, while the efficacy of the model has been substantiated through consensus clustering and Kaplan-Meier survival analysis, the sample size of the validation cohort is relatively small, and further confirmation of the generalizability and robustness of the findings is still needed. Given that the consensus clustering analysis was executed on a singular dataset, incorporating validation datasets will require re-clustering and additional analyses, which poses challenges given current research advancements and resource constraints. In future studies, we aim to include more independent validation cohorts to further evaluate the effectiveness and generalizability of the prognostic models, thereby enhancing the credibility and clinical relevance of the results. Concurrently, we plan to integrate new algorithms and a wealth of biological experimental data to explore the potential mechanisms and functions related to our research subjects more thoroughly.

In conclusion, the comprehensive bioinformatics analysis of BRCA yielded significant insights into the molecular landscape, immune infiltration, and potential prognostic biomarkers of the disease. Distinct immune cell correlations and BRCA subtypes with different survival outcomes were analyzed. Differentially expressed genes associated with these subtypes were identified, and a prognostic model that stratified patients into different risk categories with implications for overall survival was constructed. Furthermore, the single-cell analysis provided a detailed cellular annotation and insights into intercellular communication. Despite the limitations of the study, the results validated the targets and their therapeutic potential, paving the way for future research. The prognostic models and biomarkers identified hold promise for improving personalized treatment strategies and outcomes in BRCA patients, particularly in the context of immunotherapy.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12935-024-03589-7.

Supplementary Material 1 (supplementary Table 1)

Supplementary Material 2 (Table 2.1.1)

Supplementary Material 3 (Table 2.1.2)

Supplementary Material 4 (Table 2.2.1)

Supplementary Material 5 (Table 2.2.2)

Supplementary Material 6 (Table 2.3.1)

Supplementary Material 7 (Table 2.3.2)

Supplementary Material 8 (Table 2.3.3)

Supplementary Material 9 (Table 2.4.1)

Supplementary Material 10 (Table 2.4.2)

Supplementary Material 11 (Table 2.4.3)

Supplementary Material 12 (Table 2.5.1)

Supplementary Figure

Acknowledgements
We appreciate the assistance of chatgpt in refining the language.

Author contributions
Conceptualization: Xiaoyan Xu; Methodology: Zhengyang Feng; Software: Yanlin Gu; Validation: Yanlin Gu and Liyan Jin; Formal Analysis: Liyan Jin; Resources: Xiaoyan Xu; Data Curation: Xiaoyan Xu; Writing – Original Draft: Yanlin Gu and Liyan Jin. Writing – Review & Editing: Liyan Jin; Supervision: Zhengyang Feng and Liyan Jin; Funding Acquisition: Liyan Jin.

Funding
This work was supported by grants from the Suzhou Science and Technology Bureau (KJXW2022011), the Beijing City Technology Innovation Fund (KC2021-JF-0167-8), Suzhou Health Young Key Talents"National Tutorial system" training project (Qngg 2023006 and Qngg2024008), State Key Laboratory of Radiation Medicine and Protection (GZK1202408 and GZK12024047).

Data availability
The data analyzed in this study could be obtained from TCGA and GEO database.

Declarations

Ethical approval
This study was conducted in accordance with the principles outlined in the Declaration of Helsinki. The Ethics Committee of The Second Affiliated Hospital of Soochow University granted approval for this research (Number: JD-BS-2022-0033).

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Thyroid and Breast Surgery, The Second Affiliated Hospital of Soochow University, Jiangsu, China
[2]Department of Oncology, The Second Affiliated Hospital of Soochow University, Jiangsu, China
[3]Department of Operating Room, Traditional Chinese Medicine Hospital of Kunshan, Jiangsu, China

Received: 24 July 2024 / Accepted: 26 November 2024
Published online: 03 December 2024

References

1. 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. 2021;71(3):209–49.
2. Caudell JJ, et al. The future of personalised radiotherapy for head and neck cancer. Lancet Oncol. 2017;18(5):e266–73.
3. Ye J, et al. A retrospective prognostic evaluation analysis using the 8th edition of American Joint Committee on Cancer (AJCC) cancer staging system for luminal A breast cancer. Chin J Cancer Res. 2017;29(4):351–60.
4. Koşaloğlu Z, et al. Identification of immunotherapeutic targets by genomic profiling of rectal NET metastases. Oncoimmunology. 2016;5(11):e1213931.

5. Wagner AH, et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. Nat Genet. 2020;52(4):448–57.
6. Mateo L, et al. Personalized cancer therapy prioritization based on driver alteration co-occurrence patterns. Genome Med. 2020;12(1):78.
7. Sa JK, et al. Pharmacogenomic analysis of patient-derived tumor cells in gynecologic cancers. Genome Biol. 2019;20(1):253.
8. Goecks J, et al. How machine learning will transform Biomedicine. Cell. 2020;181(1):92–101.
9. Adrion JR, Galloway JG, Kern AD. Predicting the Landscape of recombination using deep learning. Mol Biol Evol. 2020;37(6):1790–808.
10. Lin PC, et al. Intratumor Heterogeneity of MYO18A and FBXW7 variants Impact the Clinical Outcome of Stage III Colorectal Cancer. Front Oncol. 2020;10:588557.
11. Mandal R, Chan TA. Personalized Oncology meets Immunology: the path toward Precision Immunotherapy. Cancer Discov. 2016;6(7):703–13.
12. Argelaguet R, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):111.
13. Trzupek D, et al. Discovery of CD80 and CD86 as recent activation markers on regulatory T cells by protein-RNA single-cell analysis. Genome Med. 2020;12(1):55.
14. Nirschl CJ, et al. IFNγ-Dependent tissue-Immune Homeostasis is co-opted in the Tumor Microenvironment. Cell. 2017;170(1):127–e14115.
15. Wisdom AJ, et al. Single cell analysis reveals distinct immune landscapes in transplant and primary sarcomas that determine response or resistance to immunotherapy. Nat Commun. 2020;11(1):6410.
16. Liu W, et al. Characterizing the tumor microenvironment at the single-cell level reveals a novel immune evasion mechanism in osteosarcoma. Bone Res. 2023;11(1):4.
17. Zhang Y, Zhang Z. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. Cell Mol Immunol. 2020;17(8):807–21.
18. Chung W, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017;8:15081.
19. Jiang G, et al. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics. 2016;17(Suppl 7):525.
20. Zhang H, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat Genet. 2020;52(6):572–81.
21. Sun Y, et al. Computational approach for deriving cancer progression roadmaps from static sample data. Nucleic Acids Res. 2017;45(9):e69.
22. Yoo SK, et al. Integrative analysis of genomic and transcriptomic characteristics associated with progression of aggressive thyroid cancer. Nat Commun. 2019;10(1):2764.
23. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220.
24. Zeng D, et al. IOBR: Multi-omics Immuno-Oncology Biological Research to Decode Tumor Microenvironment and signatures. Front Immunol. 2021;12:687975.
25. Wickham H. *ggplot2.* 2011. 3(2): pp. 180–185.
26. Wu Z, et al. Integrated analysis identifies oxidative stress genes associated with progression and prognosis in gastric cancer. Sci Rep. 2021;11(1):3292.
27. Xiang L, et al. A potential biomarker of combination of Tumor Mutation Burden and Copy Number Alteration for Efficacy of Immunotherapy in KRAS-Mutant Advanced Lung Adenocarcinoma. Front Oncol. 2020;10:559896.
28. Jiang P, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat Med. 2018;24(10):1550–8.
29. Attali D. and C.J.R.p.v. Baker, *ggExtra: Add marginal histograms to 'ggplot2', and more 'ggplot2'enhancements.* 2019.
30. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
31. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell Transcriptomic Data. Cell Syst. 2019;8(4):281–e2919.
32. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. 2016;5:p2122.
33. Efremova M, et al. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc. 2020;15(4):1484–506.
34. Hou R, et al. Predicting cell-to-cell communication networks using NATMI. Nat Commun. 2020;11(1):5011.
35. Cabello-Aguilar S, et al. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Res. 2020;48(10):e55.
36. Jin S, et al. Inference and analysis of cell-cell communication using CellChat. Nat Commun. 2021;12(1):1088.
37. Badia IMP, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinform Adv. 2022;2(1):vbac016.
38. Wang X, et al. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10(1):380.
39. Modi S, et al. Antitumor Activity and Safety of Trastuzumab Deruxtecan in patients with HER2-Low-expressing advanced breast Cancer: results from a phase ib study. J Clin Oncol. 2020;38(17):1887–96.
40. Johnston SRD, et al. Abemaciclib Combined with Endocrine Therapy for the adjuvant treatment of HR+, HER2-, Node-Positive, High-Risk, early breast Cancer (monarchE). J Clin Oncol. 2020;38(34):3987–98.
41. Dai X, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res. 2015;5(10):2929–43.
42. Curtis C, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.
43. Ma Y et al. The diagnostic value of ACSL1, ACSL4, and ACSL5 and the clinical potential of an ACSL inhibitor in Non-small-cell Lung Cancer. Cancers (Basel), 2024. 16(6).
44. Wang Y, et al. HBXIP up-regulates ACSL1 through activating transcriptional factor Sp1 in breast cancer. Biochem Biophys Res Commun. 2017;484(3):565–71.
45. Tu J, et al. Expression and clinical significance of TYRP1, ABCB5, and MMP17 in sinonasal mucosal melanoma. Cancer Biomark. 2022;35(3):331–42.
46. Yang M, et al. Expression of ABCB5 gene in hematological malignances and its significance. Leuk Lymphoma. 2012;53(6):1211–5.
47. Karas Zella MA et al. Prognostic significance of CD133 and ABCB5 expression in papillary thyroid carcinoma. Eur J Histochem, 2020. 64(4).
48. Sakil HAM, et al. ΔNp73 regulates the expression of the multidrug-resistance genes ABCB1 and ABCB5 in breast cancer and melanoma cells - a short report. Cell Oncol (Dordr). 2017;40(6):631–8.
49. de Assis LVM, et al. Melanopsin (Opn4) is an oncogene in cutaneous melanoma. Commun Biol. 2022;5(1):461.
50. Wang Q, et al. Targeting Opsin4/Melanopsin with a Novel Small Molecule suppresses PKC/RAF/MEK/ERK Signaling and inhibits lung adenocarcinoma progression. Mol Cancer Res. 2020;18(7):1028–38.
51. Qian G, Jin X, Zhang L. LncRNA FENDRR Upregulation promotes hepatic carcinoma cells apoptosis by targeting mir-362-5p Via NPR3 and p38-MAPK pathway. Cancer Biother Radiopharm. 2020;35(9):629–39.
52. Li S, et al. NPR3, transcriptionally regulated by POU2F1, inhibits osteosarcoma cell growth through blocking the PI3K/AKT pathway. Cell Signal. 2021;86:110074.
53. Gu L, et al. Long noncoding RNA BCYRN1 promotes the proliferation of Colorectal Cancer cells via Up-Regulating NPR3 expression. Cell Physiol Biochem. 2018;48(6):2337–49.
54. Pergolizzi M, et al. The neuronal protein neuroligin 1 promotes colorectal cancer progression by modulating the APC/β-catenin pathway. J Exp Clin Cancer Res. 2022;41(1):266.
55. Bizzozero L et al. Tumoral Neuroligin 1 promotes Cancer-nerve interactions and synergizes with the glial cell line-derived neurotrophic factor. Cells, 2022. 11(2).
56. Jia J, et al. AMPK, a Regulator of Metabolism and Autophagy, is activated by Lysosomal Damage via a novel galectin-Directed Ubiquitin Signal Transduction System. Mol Cell. 2020;77(5):951–e9699.
57. Dou AX, et al. Cyclic adenosine monophosphate involvement in low-dose cyclophosphamide-reversed immune evasion in a mouse lymphoma model. Cell Mol Immunol. 2012;9(6):482–8.
58. Wang N, et al. AMPK-a key factor in crosstalk between tumor cell energy metabolism and immune microenvironment? Cell Death Discov. 2024;10(1):237.
59. Zhang H, et al. cAMP-PKA/EPAC signaling and cancer: the interplay in tumor microenvironment. J Hematol Oncol. 2024;17(1):5.
60. Franklin RA, et al. The cellular and molecular origin of tumor-associated macrophages. Science. 2014;344(6186):921–5.
61. Liu D, Hofman P. Expression of NOTCH1, NOTCH4, HLA-DMA and HLA-DRA is synergistically associated with T cell exclusion, immune checkpoint blockade efficacy and recurrence risk in ER-negative breast cancer. Cell Oncol (Dordr). 2022;45(3):463–77.
62. Lyu L, et al. Overexpressed pseudogene HLA-DPB2 promotes Tumor Immune infiltrates by regulating HLA-DPB1 and indicates a better prognosis in breast Cancer. Front Oncol. 2020;10:1245.

63. Wang B et al. Improved Immunotherapy outcomes via Cuproptosis Upregulation of HLA-DRA expression: promoting the aggregation of CD4(+) and CD8(+)T lymphocytes in Clear Cell Renal Cell Carcinoma. Pharmaceuticals (Basel), 2024. 17(6).
64. Rizvi NA, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. 2015;348(6230):124–8.
65. Novak AJ, et al. Expression of BLyS and its receptors in B-cell non-hodgkin lymphoma: correlation with disease activity and patient outcome. Blood. 2004;104(8):2247–53.
66. Wang T, et al. Immunogenomic Landscape in breast Cancer reveals immunotherapeutically relevant Gene signatures. Front Immunol. 2022;13:805184.
67. Comprehensive molecular portraits of human breast tumours. Nature, 2012. 490(7418): pp. 61–70.
68. Hu H, et al. A breast cancer classification and immune landscape analysis based on cancer stem-cell-related risk panel. NPJ Precis Oncol. 2023;7(1):130.

Publisher's note