

Integration of Bulk and Single-Cell Transcriptomics Reveals BCL2L14 as a Novel IGKC+ T Cell-Associated Therapeutic Target in Breast Cancer

Jiaming He^{1,2}, Aiman Akhtar¹, Jing Li², Qiang Wei³, Yang Yuan¹, Jianhua Ran⁴, Yongping Ma¹, Dilong Chen^{2,5,6}

¹Department of Biochemistry and Molecular Biology, Basic Medical College, Molecular Medicine & Cancer Research Center, Chongqing Medical University, Chongqing, 400016, People's Republic of China; ²Laboratory of Stem Cells and Tissue Engineering, Department of Histology and Embryology, Chongqing Medical University, Chongqing, 400016, People's Republic of China; ³Department of Laboratory Medicine, The First Affiliated Hospital of Chongqing Medical University, Chongqing, 400016, People's Republic of China; ⁴Department of Anatomy, Laboratory of Neuroscience and Tissue Engineering, Basic Medical College, Chongqing Medical University, Chongqing, 400016, People's Republic of China; ⁵Chongqing Key Laboratory of Development and Utilization of Genuine Medicinal Materials in Three Gorges Reservoir Area, Chongqing Three Gorges Medical College, Chongqing, 404120, People's Republic of China; ⁶NMPA Key Laboratory for Quality Monitoring of Narcotic Drugs and Psychotropic Substances, Chongqing Institute for Food and Drug Control, Chongqing, 401120, People's Republic of China

Correspondence: Yongping Ma; Dilong Chen, Email mayongping@cqmu.edu.cn; chendilong@aliyun.com

Background: The tumor microenvironment and biomarkers play a pivotal role in breast cancer research, yet there remains a pressing need for effective biomarkers. This study focuses on identifying a novel IGKC+ T Cell subpopulation and its related biomarkers to pave the way for innovative targeted therapies and improved clinical outcomes.

Methods: We first performed single-cell RNA sequencing (scRNA-seq) analysis to characterize immune cell heterogeneity within the tumor microenvironment, leading to the identification of series cell subpopulation. Then, by performing univariate analysis to correlate cell proportions with patient prognosis, we identified a novel IGKC+ T cell subpopulation. Next, we applied bulk RNA-seq deconvolution algorithms to estimate the abundance of this subpopulation across breast cancer cohorts. Differential expression analysis and weighted gene co-expression network analysis (WGCNA) were employed to identify genes associated with the IGKC+ T cell population. To pinpoint key regulatory genes, we applied machine learning algorithms. Based on the hub genes identified, we constructed a prognostic risk model and developed a nomogram to aid clinical decision-making. Immune infiltration patterns were further assessed in high- vs low-risk groups defined by the model. Finally, functional validation was performed through overexpression of BCL2L14 in vitro, and downstream signaling pathways were examined.

Results: We identified the novel IGKC+ T cell subpopulation and core genes. Machine learning pinpointed BCL2L14, IGHD, MAPT-AS1, NT5DC4, and TNIP3 as key regulators of breast cancer progression in this subpopulation. The model stratified patients into high- and low-risk groups, with high-risk patients showing worse prognosis and weaker immune infiltration. Overexpression of BCL2L14 was experimentally demonstrated to accelerate breast cancer progression, linked to enhanced phosphorylation of the NF- κ B pathway.

Conclusion: Our results underscore BCL2L14 as a potential driver within the novel T-cell subpopulation and a critical biomarker for breast cancer diagnosis. These findings provide a basis for developing advanced diagnostic tools and targeted therapies, which may ultimately enhance patient prognosis.

Keywords: single cell sequencing, breast cancer, scPagwas, Mendelian randomization, bioinformatics

Introduction

Breast cancer (BRCA) is the most prevalent malignancy among women and ranks as the fifth leading cause of cancer mortality globally, with approximately 685,000 deaths reported annually.^{1,2} Characterized by substantial molecular and histological heterogeneity, breast cancer is traditionally classified into three major subtypes: hormone receptor-positive breast cancer (ER+ or PR+), HER2-positive breast cancer (HER2+), and triple-negative breast cancer (TNBC), which lacks ER, PR, and HER2 expression.³ Treatment strategies and risk profiles differ across these subtypes. In addition to surgical excision, treatment

options include hormone therapy, radiotherapy, chemotherapy, and immunotherapy.⁴ However, the heterogeneity of the breast cancer tumor microenvironment poses significant challenges for targeted therapies. This underscores the urgent need to deepen our understanding of the tumor microenvironment's composition.

In recent years, increasing research efforts have been directed toward exploring the tumor microenvironment and identifying potential biomarkers for breast cancer treatment.⁴ The advent of single-cell technologies has enabled more precise molecular subtyping of breast cancer patients, thereby advancing the feasibility of personalized treatment strategies through a deeper understanding of the tumor microenvironment.^{5–7} For example, *Jiang YZ et al* categorized primary triple-negative breast cancer (TNBC) into four molecular subtypes and identified their respective biomarkers.⁵ Additionally, studies have modeled the evolutionary transition from normal to malignant subtypes, emphasizing the clinical relevance of the luminal progenitor (LP) subtype of breast cancer.⁸ Furthermore, novel immune subpopulations within the breast cancer microenvironment have been identified as potential therapeutic targets in multiple studies.^{7–9} Consequently, a comprehensive understanding of the breast cancer microenvironment is pivotal for the development of precise and effective biological therapies.

Recent large-scale sequencing analyses of solid tumors have revealed substantial spatial and temporal intratumoral heterogeneity, which has been increasingly recognized as a critical driver of therapeutic resistance and treatment failure.¹⁰ Advances in single-cell RNA sequencing (scRNA-seq) and spatially resolved single-cell omics technologies have markedly enhanced our understanding of cellular heterogeneity across diverse pathological tissues.^{11,12} These high-throughput approaches have been employed to characterize the transcriptomic landscapes of both breast cancer and normal breast tissues.¹³ Nevertheless, a comprehensive understanding of the subpopulation cells within the immune microenvironment, their gene expression profiles, and their specific roles in the pathogenesis of breast cancer remains elusive. Moreover, the clinical implications of T-cell subtypes and their prognostic significance in breast cancer have yet to be elucidated. In this study, we hypothesize that an integrative scRNA-seq analysis of breast cancer cells will provide a deeper and more systematic characterization of breast cancer cell subtypes, offering novel insights into their biological functions and the signaling pathways activated within breast tumor tissues, potentially influencing clinical outcomes.

Machine learning, a data analysis method that automates the construction of analytical models, has been extensively utilized in cancer research.^{14,15} Evidence from previous studies highlights its significant potential in drug discovery, pathology identification, and predictive model development.^{16,17} In recent years, machine learning has been increasingly applied to the diagnosis and treatment of a wide range of diseases, including cancer, cardiovascular diseases, and diabetes.^{17–19} Among the commonly used methods, least absolute shrinkage and selection operator (LASSO) logistic regression, a regularized linear regression technique that is particularly suitable for high-dimensional data analysis.²⁰ Furthermore, support vector machine-recursive feature elimination (SVM-RFE) is a powerful approach for selecting the optimal combination of variables, owing to its nonlinear discriminative capability and flexibility in modeling diverse variables.²¹ Thus, the identification of breast cancer biomarkers and the construction of predictive models using machine learning algorithms are crucial for advancing breast cancer research and treatment.

In this study, we conducted a systematic classification of breast cancer cell populations and identified a novel T-cell subpopulation within the tumor microenvironment. We characterized the molecular and biological features of this subpopulation and, through deconvolution analysis and WGCNA, identified co-expression gene modules associated with the IGKC+ T cell subpopulation. Using machine learning approaches, we further identified prognostic signature genes linked to this cell subtype and developed a clinical prediction model, offering valuable tools for disease prognosis and therapeutic intervention. Furthermore, our experimental validation demonstrated that overexpression of BCL2L14 activates the NF- κ B signaling pathway, thereby exacerbating breast cancer cell proliferation. These findings establish the biomarker potential of this T-cell subpopulation and its implications for breast cancer progression.

Materials and Methods

Online Data

RNA-sequencing matrix, related clinical information, and gene mutation data regarding BRCA were derived from The Cancer Genome Atlas (TCGA) database, which contained 1109 BRCA cases and 113 normal cases (downloaded on 11 March 2023, <http://portal.gdc.cancer.gov>). The Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) database was used to

extract the bulk- (ID: GSE58812) and single cell RNA-seq data (ID: GSE252175). Genome-wide association study (GWAS) data for BRCA can be downloaded from the IEU OPEN GWAS project (<https://gwas.mrcieu.ac.uk/>, ID: ieu-a-1126), which included 122,977 cases and 105,974 controls, totaling 228,951 samples, numbers of 10,680,257 SNPs (Single nucleotide polymorphisms).

Single Cell Sequencing Analysis

Quality control (QC) was initially conducted to assess data quality, with key metrics such as gene count, RNA count, and the proportions of mitochondrial and ribosomal genes visualized via violin plots. Data normalization was carried out using the “LogNormalize” method, and 2,000 highly variable genes were identified for downstream analyses. The top 10 highly variable genes were selected for visualization to illustrate variability within the dataset. Dimensionality reduction techniques, including PCA (Principal Component Analysis), t-SNE (t-distributed stochastic neighbor embedding), and UMAP (Uniform Manifold Approximation and Projection), were applied to visualize the initial distribution of cell clusters, with the first six PCs chosen for analysis. Following batch correction, a KNN (k-Nearest neighbors)-based cell-neighbor graph was constructed. And, clustering was performed using various resolutions to identify the optimal clustering strategy. A resolution of 1.2 was selected as the optimal parameter. A KNN graph is a graph structure built based on the KNN algorithm, where each node represents a data point (such as a single cell sample), and the edges between nodes represent the similarity or proximity between these data points. KNN plots are often used to capture cell-to-cell interrelationships in single-cell transcriptome data. Cell populations were annotated based on the “HumanPrimaryCellAtlas” reference using the “SingleR” tool, and a heatmap of the annotation results was generated. UMAP and t-SNE were then employed to visualize the annotated cell types and their distributions. Finally, differential expression analysis of module genes across clusters was performed, with manual annotation and visualization of specific cell groups.

scPagwas Analysis

We performed GWAS integrated with single-cell sequencing data using R packages such as “Seurat”, “scPagwas”, and “ggplot2”. The GWAS data was loaded with “readVcf”, converted to the “Granges” format, and processed into a data frame using “dplyr::as_tibble()”. Key columns like chromosome location, SNP ID, effect size, and standard error were extracted, and minor allele frequency (MAF) was set to 0.6. The data were saved in two files: one for GWAS summary statistics and another for allele frequency. The “scPagwas” package was used to integrate GWAS and single-cell data, generating single-cell genome-wide association results through multiple iterations. We visualized the p-values of the bootstrap results with bar and estimation plots, evaluating result stability and significance. Cell clustering across different types was visualized with “DimPlot”, and the distribution of TRS (trait-relevant score) scores was mapped with “FeaturePlot”. TRS refers to a measure of the strength or significance of the association between each SNP and a specific gene expression pattern in this association analysis. Usually, a higher score indicates a stronger or more significant relationship between the SNP and gene expression. By calculating these scores, it is possible to screen out genetic variants closely associated with changes in the expression of specific genes that may play an important role in biological function, disease susceptibility, or treatment response. By calculating TRS, it is possible to help identify subsets of cells that play a key role in genetic traits. A violin plot generated by “VlnPlot” showed TRS score distributions in various cell types, and a scatter plot displayed the heritability of genes linked to traits. This analysis systematically identified genes and pathways associated with phenotypes and cell types, providing a basis for exploring genetic variations and cell-specific features.

Analysis of Clinical Traits Based on Cell Clusters

To investigate how gene expression affects survival in different cell types, we processed survival data and performed survival curve plotting and Cox regression analysis using “survival”, “survminer”, “dplyr”, “rms”, and “PredictABEL” packages. We visualized differences in Theta values for different cell types across groups using violin and box plots and annotated the results for statistical significance. We grouped the data based on the median expression of IGKC+ T cell and visualized differential gene expression in TCGA breast cancer RNA-seq data using a volcano plot. Estimate scores,

including Stromal, Immune, Estimate, and Purity, were calculated for the tumor samples and visualized using heatmaps after Z-score normalization. Immune-related gene expression differences between high and low expression groups were analyzed and visualized. TIDE (Tumor Immune Dysfunction and Exclusion) is a computational algorithm designed to predict cancer patient responses to immune checkpoint inhibitors, such as PD-1/PD-L1 blockade. By analyzing the association between gene expression profiles and cytotoxic T lymphocyte (CTL) infiltration levels, as well as their impact on patient survival in a modeling cohort, TIDE constructs a predictive framework to identify gene signatures associated with T cell dysfunction in other tumor cohorts. The TIDE tool was used to calculate immune therapy-related scores, including immune suppression and T-cell dysfunction scores, to examine the effects of immune cell populations on immunotherapy outcomes. Drug sensitivity predictions were made using the “OncoPredict” package, with GDSC2 gene expression and drug response data to train the model. Drug response patterns were analyzed and visualized across groups. Lastly, “maftools” was used to analyze SNV data from tumor samples, visualizing differential mutation patterns based on sample grouping.

WGCNA Analysis

This analysis focuses on the modular partitioning of gene expression data and its association with clinical traits or other characteristics. Initially, the code reads the gene expression dataset, filters out lowly expressed genes and samples, and selects 10,000 genes for further analysis, with additional filtering based on differential expression results. Following this, sample clustering is conducted to identify any anomalous samples, and a clustering dendrogram is generated for visualization. The soft thresholding power is then calculated, set to 0.9, and a topological overlap matrix (TOM) is constructed for gene clustering. The dynamic tree cutting method, set to a threshold of 0.25, is employed to partition genes into multiple modules, and the gene characteristics of each module are analyzed. Finally, the modules are associated with clinical traits, with the correlation between the module eigengenes (ME) and clinical data calculated. The results are visualized in a module-trait association plot, and detailed information regarding gene significance (GS) and module membership (MM) for the genes in each module is provided.

Core Gene Acquisition and Survival Analysis

We used the “ggvenn” package to generate a Venn diagram, incorporating differential genes from TCGA breast cancer data, genes from the purple module in WGCNA, marker genes of the identified cell subgroup, and PCC genes. Using Cox regression analysis in conjunction with survival data, we assessed the impact of these genes on patient survival prognosis. The prognostic effects of each gene were visualized using Kaplan-Meier survival curves, with corresponding statistical results (such as Cox regression hazard ratios (HR) and p-values) annotated on the plot. Next, we applied the “parLapply” package to perform Spearman correlation analysis between gene expression data and convolution cell matrix data, calculating correlation coefficients and p-values, with multiple testing corrections applied. For each gene-cell pair, we generated scatter plots and regression curves.

Summary-Data-Based Mendelian Randomization

SMR (Summary-data-based Mendelian Randomization) analysis can identify genetic loci causally associated with certain outcomes without relying on intermediate traits. By using eQTLs (expression quantitative trait loci) as instrumental variables, SMR analysis evaluates the relationship between gene expression levels and outcomes, based on summary-level data from GWAS and eQTL studies.²² In this study, we used SMR software (version 1.03, <https://cnsgenomics.com/software/smr/#Overview>) for allele coordination and analysis. For each gene, the primary analysis selected the most strongly associated variant within its cis-eQTL as a single genetic instrument. To increase the reliability of the results and minimize potential confounding effects, we performed heterogeneity-dependent tool (HEIDI) tests on multiple SNPs. Only results that met both the criteria of SMR $P < 0.05$ and HEIDI P-value > 0.05 were retained for further analysis. We filtered for common SNPs (MAF $> 1\%$, $P < 5.0e-8$) that were significantly associated with the expression of the selected genes in breast cancer. This study only included cis-eQTLs, which are located within 1 Mb of the target gene, to generate genetic instruments. SNPs with MAF $< 5\%$, INFO score < 0.8 , missing rate > 0.02 , or HWE exact test $P < 1.0e-6$ were excluded from the analysis.

Predictive Feature Construction Based on Integrated Machine Learning

In the extensive RNA-seq data from TCGA, we performed differential expression analysis between normal and tumor samples using the limma package in R. The filtering criteria were $|\logFC| > 0.5$ and $P < 0.05$. We then conducted an intersection analysis of the differentially expressed genes (DEGs) with the genes identified in the IGKC+ T cell-related module from WGCNA. These intersecting genes were considered to be closely related to the characteristic genes of the new breast cancer cell subgroups at both large-scale and single-cell transcriptomic levels. Therefore, we referred to these genes as new breast cancer IGKC+ T cell subgroup-related genes (TIRS genes). The construction method is as follows: First, we performed univariate Cox regression analysis to screen for potential prognostic genes in the TCGA-BRCA dataset. Second, the TCGA-BRCA dataset was used as the training set, and GEO cohort (ID: GSE58812) were used as the validation set to ensure balanced distribution of clinical features between the two groups. Third, the LASSO algorithm, a machine learning method, was applied. And, the constructed model was evaluated. TIRS score is the result of multiplying the genes and risk coefficients from the LASSO regression model. The cohort could be divided into low-risk and high-risk groups by calculating the mean value of TIRS score.

Survival Analysis and Prediction Nomogram Construction

Based on the median risk score, we divided the BRCA patients in the TCGA training set and GEO validation set into high-risk and low-risk groups. Next, we performed Kaplan-Meier (KM) survival curve analysis using the survminer package in R to compare overall survival (OS) between the high-risk and low-risk groups. The Log rank test was used to determine if there were significant differences ($P < 0.05$). Additionally, we used the “timeROC” package to plot ROC curves to assess the sensitivity and specificity of the risk score in predicting OS for BRCA patients, and compared the area under the curve (AUC) with other clinical features.

Furthermore, we explored the correlation between the risk score and multiple clinical features, including age, tumor stage, T and N stages, and tumor grade. To validate the independent prognostic value of the risk score, we performed both univariate and multivariate Cox regression analyses on the TCGA-BRCA and GEO datasets to evaluate whether the risk score is an independent prognostic factor for survival in BRCA patients.

To improve the accuracy and clinical applicability of the predictive model, we developed a nomogram incorporating clinical features (T, N, Stage, Age) to quantify the expected survival of BRCA patients. Finally, we assessed the predictive accuracy, discriminative ability, and calibration of the nomogram using ROC curves, and calibration curves.

Drug Sensitivity Analysis

To achieve personalized treatment, we used the R package “pRRophetic” to predict the chemotherapy sensitivity of breast cancer patients with different TIRS risk scores.²³ This method fits the gene expression profiles of patient tissues to those of cancer cells, calculating the half-maximal inhibitory concentration (IC_{50}). We used the Wilcoxon test to assess differences in drug IC_{50} values between the high-risk and low-risk groups, setting $P < 0.05$ as statistically significant. Additionally, we utilized the Connectivity Map (CMap) database gene expression profiles to screen for potential compounds that activate or inhibit specific pathways. Based on the upregulation and downregulation patterns of genes in the high-risk and low-risk groups, we identified drug components that could reverse the phenotype of the high-risk group, thus providing potential candidate drugs for personalized therapy.

Cell Culture

All cells were preserved and provided by the Stem Cell and Tissue Engineering Laboratory of Chongqing Medical University. The cells were cultured in incubators set at 37°C with 5% CO₂ and maintained at a constant humidity. MDA-MB-231 and MDA-MB-468 cells were cultured in 1640 medium (Gibco BRL, USA) supplemented with 10% FBS (SP01002, SPERIKON, Chengdu, China) and 1% penicillin-streptomycin (C100C5, NCM Biotech, Suzhou, China).

Transient Transfection

The vector and pECMV-3×FLAG-BCL2L14 plasmids were purchased from Miaoling Biology Co., Ltd (Wuhan, China). And, plasmids were introduced into the cells by PEI (40816ES, YEASEN, Shanghai, China) following the instructions provided by the manufacturer.

Cell Viability Assay and EdU Assay

To assess cell viability, the Cell counting Kit-8 (MA0218, Meilunbio, Dalian, China) was employed. Following various procedures, the cells were placed in 96-well plates with a concentration of 2×10^4 cells per well for 24, 48, and 72 hours, correspondingly. Next, the medium containing 10% CCK8 concentration was replaced according to the provided guidelines. The measurement of absorbance was observed at a wavelength of 450 nm.

The EdU cell proliferation kit (MA0424, Meilunbio, Dalian, China) was purchased to examine cell proliferation by the manufacturer. Images were captured and analyzed using a fluorescence microscope (Leica, Munich, Germany).

Scratch Assay

Cells were introduced into 12-well culture dishes. After the cells had completely overgrown the well plate, it was scratched vertically with a sterilized 200 μ L head. After 24 h, one random selection from each group was selected for observation, and visual field photographs were taken.

Western Blot

Western blot analysis was performed as described previously.²⁴ Western blot analysis was performed with anti-I κ B α (#A19714), anti-Phospho-I κ B α (#AP0707), anti-NF- κ B p65 (#A19653), anti-Phospho-NF- κ B p65(#AP1460), anti-BCL2L14 (#A19292), antibodies purchased from Abclonal Technology (Wuhan, China). Goat Anti-Rabbit IgG H&L (#511203) antibody was purchased from ZEN-BIOSCIENCE (Chengdu, China).

Statistical Analysis

All bioinformatics statistical analyses were performed using R software (version: R 4.4.1). The Chi-square test was used to compare the differences in clinical features between the training set and the internal validation set. For two non-normally distributed variables, the Wilcoxon test was applied as a non-parametric method to assess their differences. Kaplan-Meier survival analysis combined with the Log rank test was used to compare the overall survival (OS) of patients in different subgroups. Univariate and multivariate Cox regression analyses were performed to explore independent prognostic factors. Spearman correlation analysis was used to evaluate the correlation between the risk score and immune cell infiltration. Statistical significance was determined as values of $P < 0.05$. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.001$, ns: not statistically significant.

Results

Single-Cell Analysis of Breast Cancer and Acquisition of Subsets

We downloaded the single-cell dataset from the GEO gene database (ID: GSE252175) for comprehensive analysis. First, quality control (QC) of the sequencing data was performed, and violin plots were generated using the VlnPlot function to display the QC metrics of the samples, including nFeature_RNA (the number of genes per cell), nCount_RNA (the total RNA count per cell), percent.mt (the proportion of mitochondrial genes), and percent.rb (the proportion of ribosomal genes). These metrics helped identify potential low-quality cells (Figure 1A). A correlation plot between different QC metrics was created using the FeatureScatter function to further filter out low-quality cells (Figure 1B). Subsequently, data normalization was carried out, and the FindVariableFeatures function was used to identify highly variable genes (genes with the most biological variation). PCA and RunTSNE functions were used to perform dimensionality reduction. Cluster analysis was then conducted using the “FindNeighbors” and “FindClusters” packages. A dendrogram was generated using the clustree function (Figure S1) to help evaluate the clustering results at different resolutions, and a resolution of 1.2 was chosen. For each cell cluster, the “FindAllMarkers” function was used to identify genes with

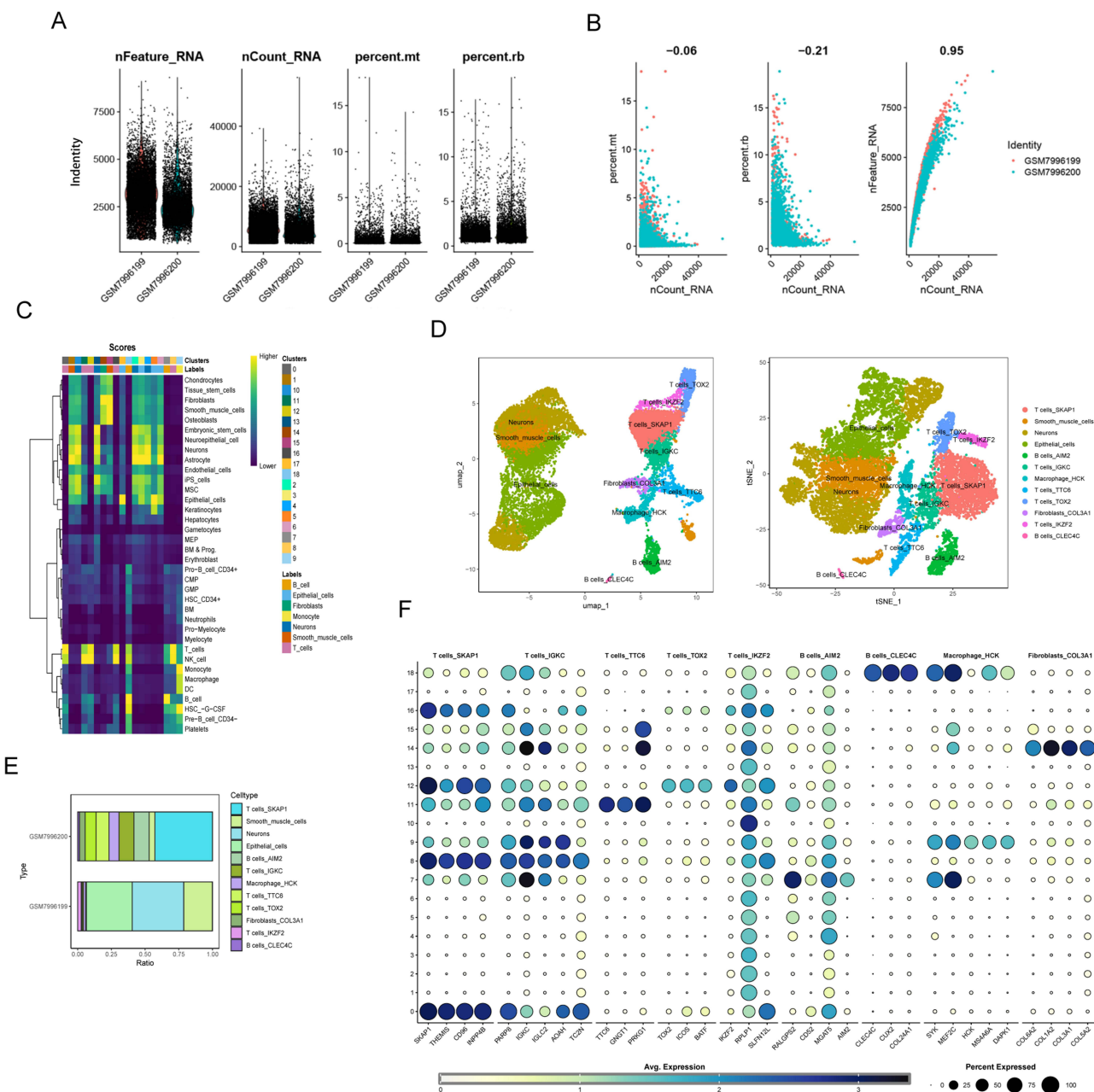


Figure 1 Single-cell analysis of breast cancer tissues. **(A)** The violin plot shows the sample QC metrics, including nFeature_RNA (number of genes per cell), nCount_RNA (total RNA per cell), percent.mt (proportion of mitochondrial genes) and percent.rb (proportion of ribosomal genes). These metrics help identify potentially low-quality cells. **(B)** Correlations between different QC indicators were plotted. **(C)** Heat map was drawn to show the effect of cell clustering under the setting resolution. **(D)** UMAP (left) and t-sne (right) analyses were performed to reduce the dimensionality of the data and visualize the data. **(E)** The bar chart shows the cell types enriched in different samples. **(F)** Bubble Plots show manually annotated genes significantly differentially expressed in different cell subsets.

significantly differential expression within each group, helping to define the biological features of various cell populations. Cell type annotation was performed using the “SingleR” package (Figure 1C and D). We then manually annotated 9 new cell subgroups and displayed the proportion of annotated cell types enriched in different samples using a bar chart (Figure 1E) and a bubble plot of cell type-specific genes (Figure 1F). As we are particularly interested in immune cells and fibroblasts in the tumor microenvironment, here we present the nine newly annotated cell subgroups, which are named SKAP+ T cells, IGKC+ T cells, TTC6+ T cells, TOX2+ T cells, IKZF2+ T cells, AIM2+ B cells, CLEC4C+ B cells, HCK+ Macrophage, and COL3A1+ Fibroblasts.

BayesPrism Associated Cell Proportions with Clinical Data

Deconvolution analysis was used to estimate the relative proportions of different cell types in the tissue. This method confirmed the correlation between cell populations, which was visualized in a heatmap, and tested for significant correlations between cell populations while excluding unnecessary genes (such as ribosomal protein genes and mitochondrial genes) to improve the accuracy of the analysis. **Figure 2A** shows the aberrant genes from both scRNA and bulk RNA sequencing. A gene correlation scatter plot was created using the “heritability_cor_scatterplot” function to display the genetic correlation of phenotype-associated genes, where we focused on *PIK3CD*, *CD247*, *PTPRC*, *FYB1*, *ARHGAP15*, *RIPOR2*, *IKZF1*, *SKAP1*, *PARP8*, and *CD96* (**Figure 2B**, [Table S1](#)). In the scPagwas algorithm score, SKAP1+ T cells and IGKC+ T cell populations were significantly higher than other cell types, which piqued our interest (**Figure 2C**, [Table S2](#)). Subsequently, we calculated the differentially significant genes in TCGA-BRCA (**Figure 2D**). Additionally, we observed that protein-coding genes were the most consistent group in the analysis ($R = 0.544$, **Figure 2E**), meaning that protein-coding genes were the focus of our deconvolution analysis in this study. We then performed univariate analysis, using cell proportions as independent variables and linking them to patient prognosis (**Figure 2F**). We identified four potential prognosis-related cell types: IGKC+ T cells, AIM+ B cells, HCK+ Macrophage, and Neurons. Based on these findings, we anchored our research focus on IGKC+ T cells.

Clinical Phenotypes and Therapeutic Analysis Based on New T Cell Typing

Based on the potential prognostic functions of the IGKC+ T cell subpopulation, we further explored the differentially expressed genes associated with this cell group. As shown in **Figure 3A**, upregulated genes included *CSN2* and *CNMD*, while downregulated genes included *FGA* and *XIRP2* ([Table S3](#)). We then calculated the immune infiltration scores based on high and low expression of this cell population, observing a significant upregulation in Stromal, Immune, and Estimate scores, while the Purity score was downregulated (**Figure 3B**, [Table S4](#)). Additionally, we presented a detailed heatmap showing the expression of certain genes, with *FGA*, *XIRP2*, *KRT1*, and *LACRTRNU1-56P* significantly downregulated, and *AL356276.1*, *PLA2G2D*, *CR2*, *IGKV1OR22-5*, and *FCER2* significantly upregulated (**Figure 3B**). Interestingly, we also observed upregulated expression of AIM+ B cells, HCK+ Macrophage, and Neurons in the IGKC+ T cells group (**Figure 3B**).

Furthermore, we found that patients in the high expression IGKC+ T cells group had higher expression of CD8 and CD274 (PD-L1), as well as higher T cell dysfunction and TIDE scores (**Figure 3C–F**). These results suggest that patients with higher expression of this target cell population may have better responses to immunotherapy. We then showcased 20 potential therapeutic drugs for high-expression patients in this cell group, including Obatoclox Mesylate, PRT062607, JAK1_8709, XAV939, Niraparib, Olaparib, Dabrafenib, LJI308, and Venetoclax (**Figure 3G**, [Table S5](#)). Additionally, we presented the SNV results for high and low expression patients of this cell population. The top-ranking variant classifications were Missense Mutation, Nonsense Mutation, Frame shift deletion, and Splice Site. The most frequent Variant type was SNP, and the most common Class changes were C>T, C>G, and C>A. The most frequently mutated genes were *PIK3CA*, *TP53*, *TTN*, *CDH1*, *GATA3*, *MUC16*, *MAP3K1*, *KMT2C*, *HMCN1*, and *FLG* (**Figure 3H**). In summary, we believe that the expression of this cell population can effectively indicate the survival prognosis of patients and provide potential therapeutic drug candidates.

Exploration of Core Genes in IGKC+ T Cells Subpopulations

Next, we further identified core genes within this cell population. We applied the WGCNA algorithm to establish co-expression modules for further analysis using hierarchical clustering trees. Through the soft-thresholding method, we achieved an ideal network structure with a correlation coefficient greater than 0.9 (**Figure 4A**). Subsequently, using the most suitable soft thresholding method and average linkage hierarchical clustering, we identified 27 modules (**Figure 4B**, [Table S6](#)). By calculating the Pearson correlation coefficient between each module and sample characteristics, we observed a stable correlation between the purple module and IGKC+ T cells (**Figure 4C**). Therefore, we performed further analysis on the genes within this module.

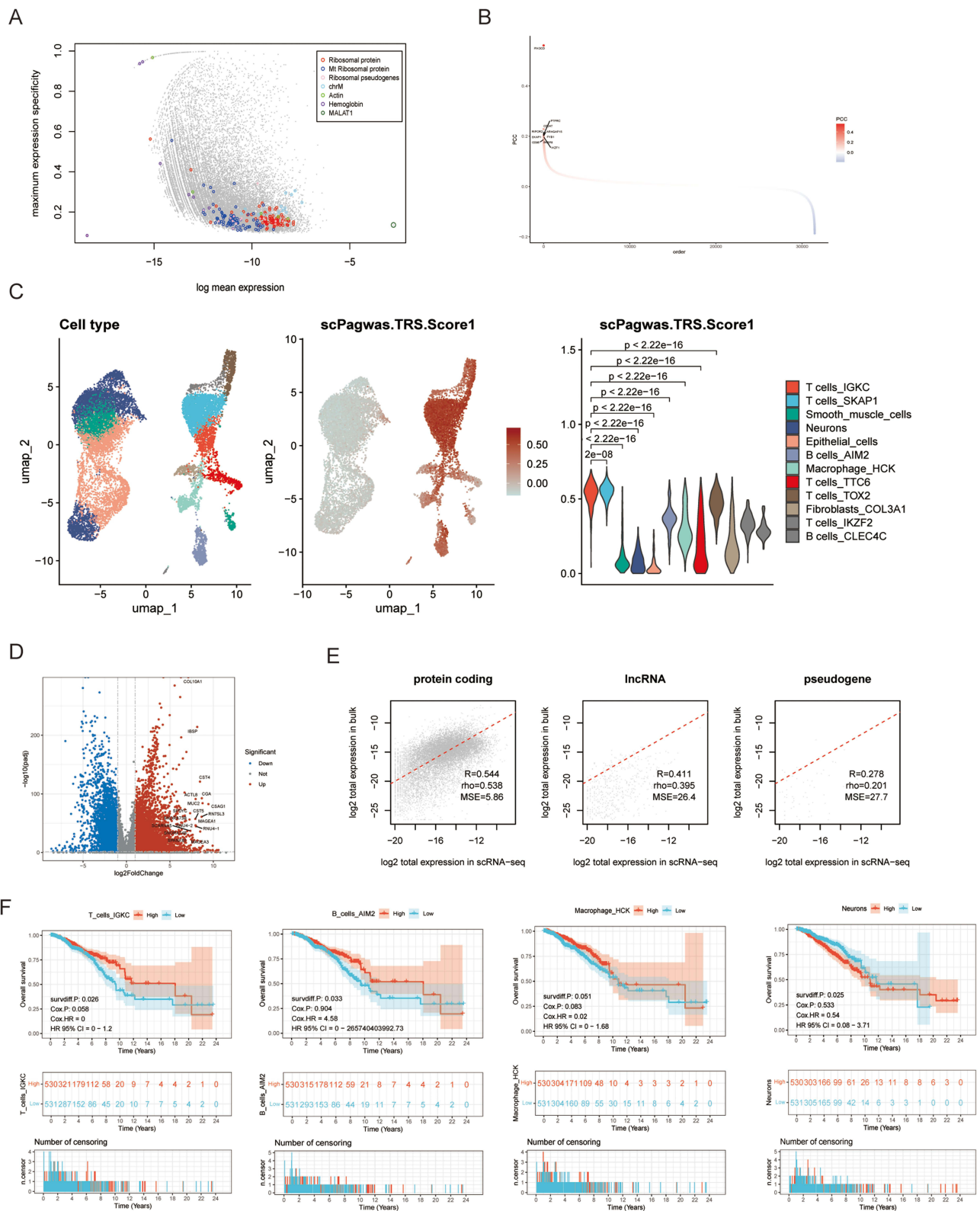


Figure 2 BayesPrism deconvolution correlates cell proportions to clinical data. **(A)** The quality control process of BayesPrism removed noise. **(B)** GWAS data were combined to identify trait-related genes. **(C)** GWAS data were combined to identify trait-related cell subsets. **(D)** Volcano map of differentially expressed genes in TCGA-BRCA cohort. **(E)** The consistency of gene expression across different types of genes. **(F)** The KM survival curves demonstrated significant survival differences between indicated high and low cells.

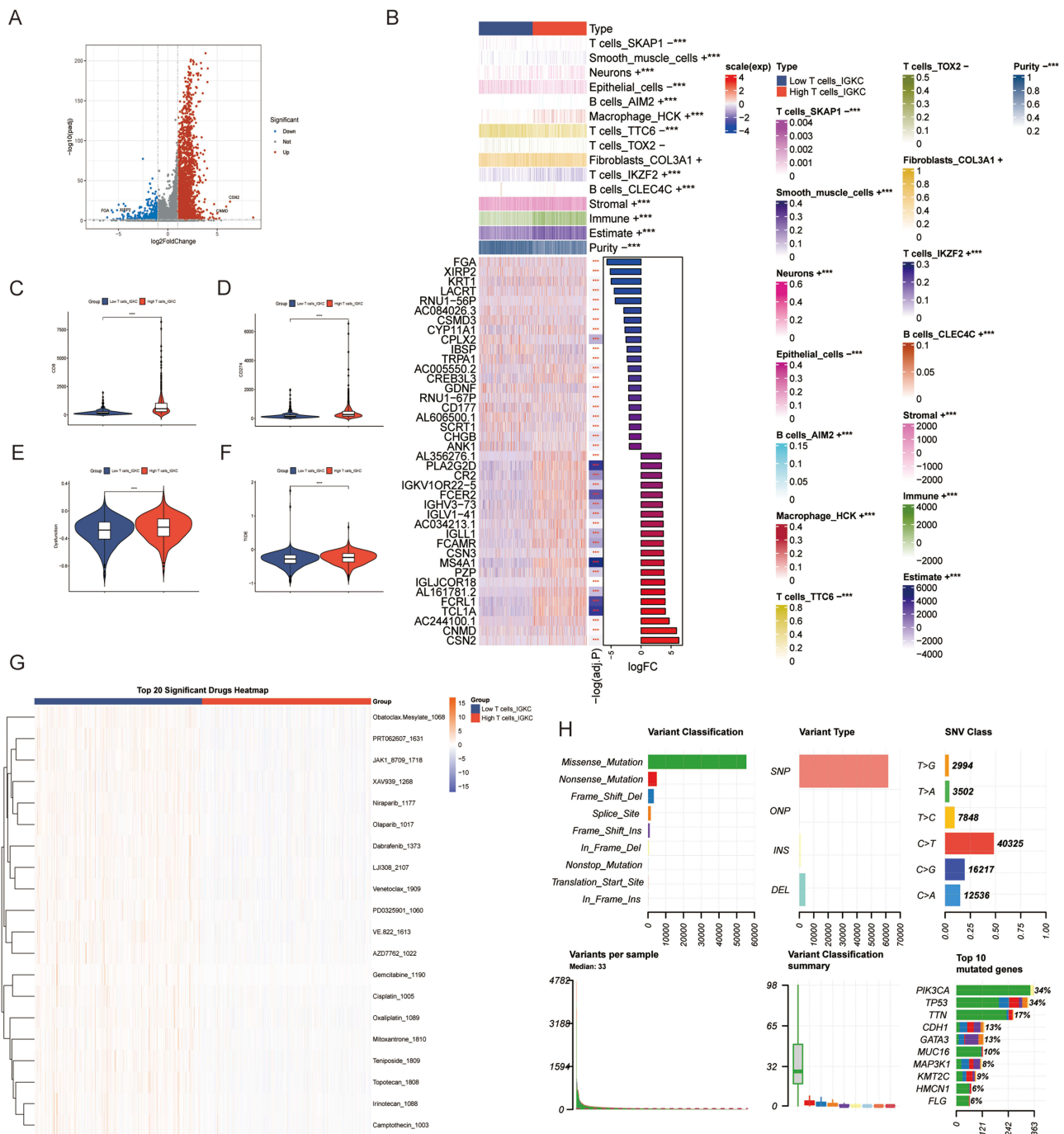


Figure 3 Analysis of clinical traits based on cell subsets. **(A and B)** Transcriptional differences between high and low groups of convolutional cell populations as well as estimate immune scores were specified. **(C–F)** TIDE immunotherapy analysis of specified convolutional cells. **(G)** The differences in drug sensitivity of the convolutional cells are specified, and the top 20 drugs are shown. **(H)** SNV differences for the convolutional cells are specified. The significance levels are indicated as follows: *** $P < 0.001$, **** $P < 0.001$.

To identify the core gene group, we intersected genes from four datasets (including TCGA-BRCA differential genes, single-cell marker genes, WGCNA purple module genes, and deconvolution PCC genes), resulting in 20 genes (Figure 4D). These genes are considered potential survival-related genes within this cell population. Finally, we performed Kaplan-Meier analysis on these genes (Figure 4E–M), progressively narrowing down the list to 9 genes: *CD6*, *LTB*, *SHISAL2A*, *SH3BP1*, *SELL*, *TRBC1*, *IL2RG*, *GZMK*, and *LIMD2*. In conclusion, we identified the core genes of the IGKC+ T cells subgroup from multiple perspectives, providing a foundation for our subsequent research.

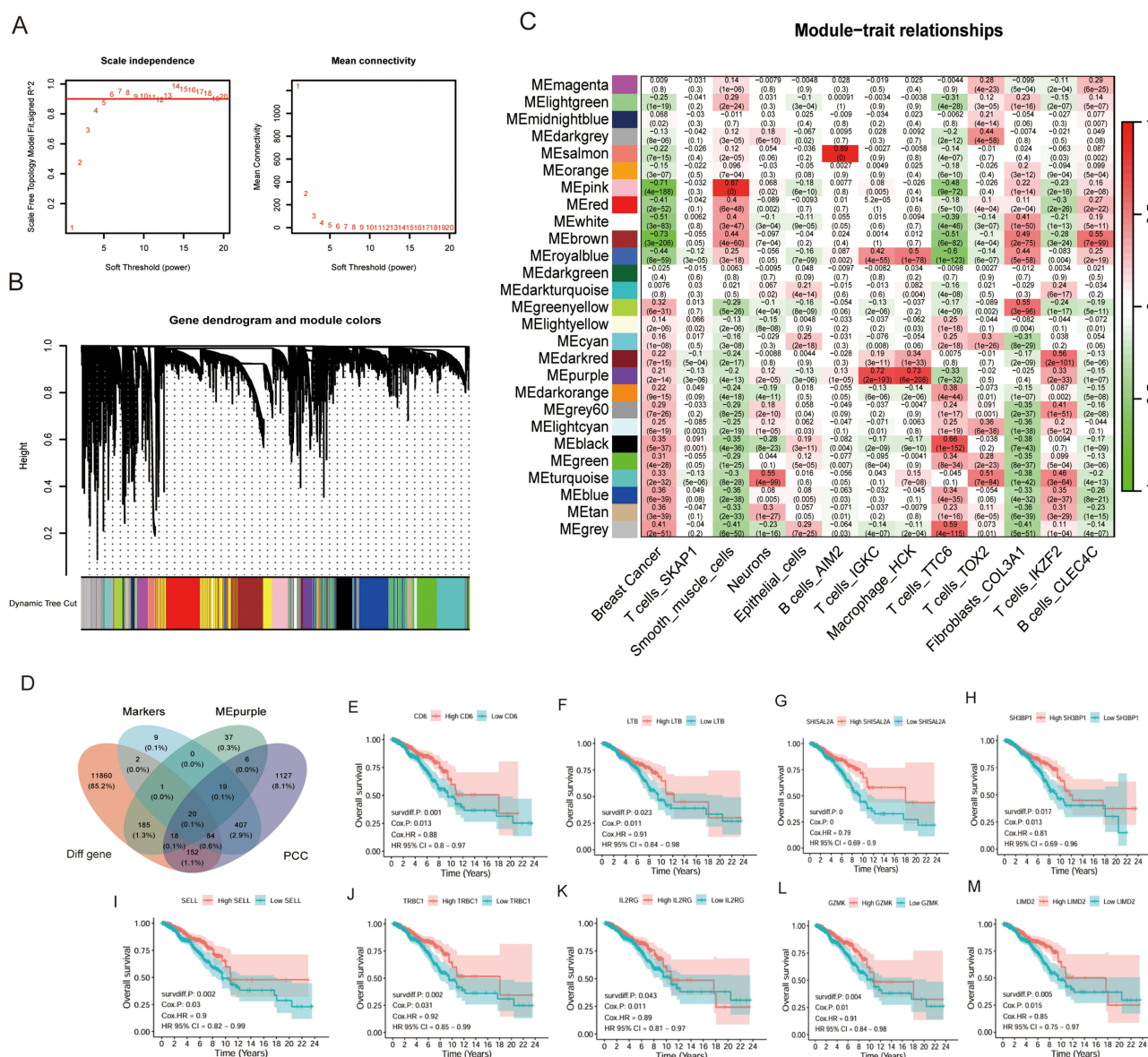


Figure 4 Coregene identification of IGKC+ T cells. **(A)** In the left graph, the horizontal line indicates that the threshold value is 0.9. **(B)** Cluster dendrogram of the WGCNA analysis. **(C)** Module-trait heatmap showing that the MEmurple module was closely related to the IGKC+ T cells. **(D)** Venn plot showing the intersecting types of genes. **(E-M)** The KM survival curve showed the genes with significant survival differences in hubgenes.

Mendelian Randomization Analysis and Identification of Therapeutic Genes for BRCA

Furthermore, this study used SMR analysis to identify pathogenic genes associated with breast cancer. The results of the SMR analysis are presented in [Table S7](#). By sorting according to the HEIDI P-value, we identified the top six genes (*CEP68*, *KIF16B*, *DGCR9*, *RP11*, *LIN52*, *ZDHHC11*) that exhibit causal relationships with breast cancer ([Figure 5A-F](#)). For these genes, the SMR analysis also provided information on related genetic loci. The results suggest that the development of breast cancer may be associated with SNPs such as rs2723063, rs4814508, rs17021270, rs12882888, and rs11745487 ([Table S7](#)).

Prognostic Model Construction Based on Core Genes and Immune Infiltration Scoring

We then began searching for potential core regulatory genes in the IGKC+ T cells cell group ([Table S8](#)). Based on the previously identified genes, we constructed a PRGcluster model, which shows how these genes differentiate between different patient subgroups and their associated risk ([Figure S2](#)). Further, we performed both univariate and multivariate

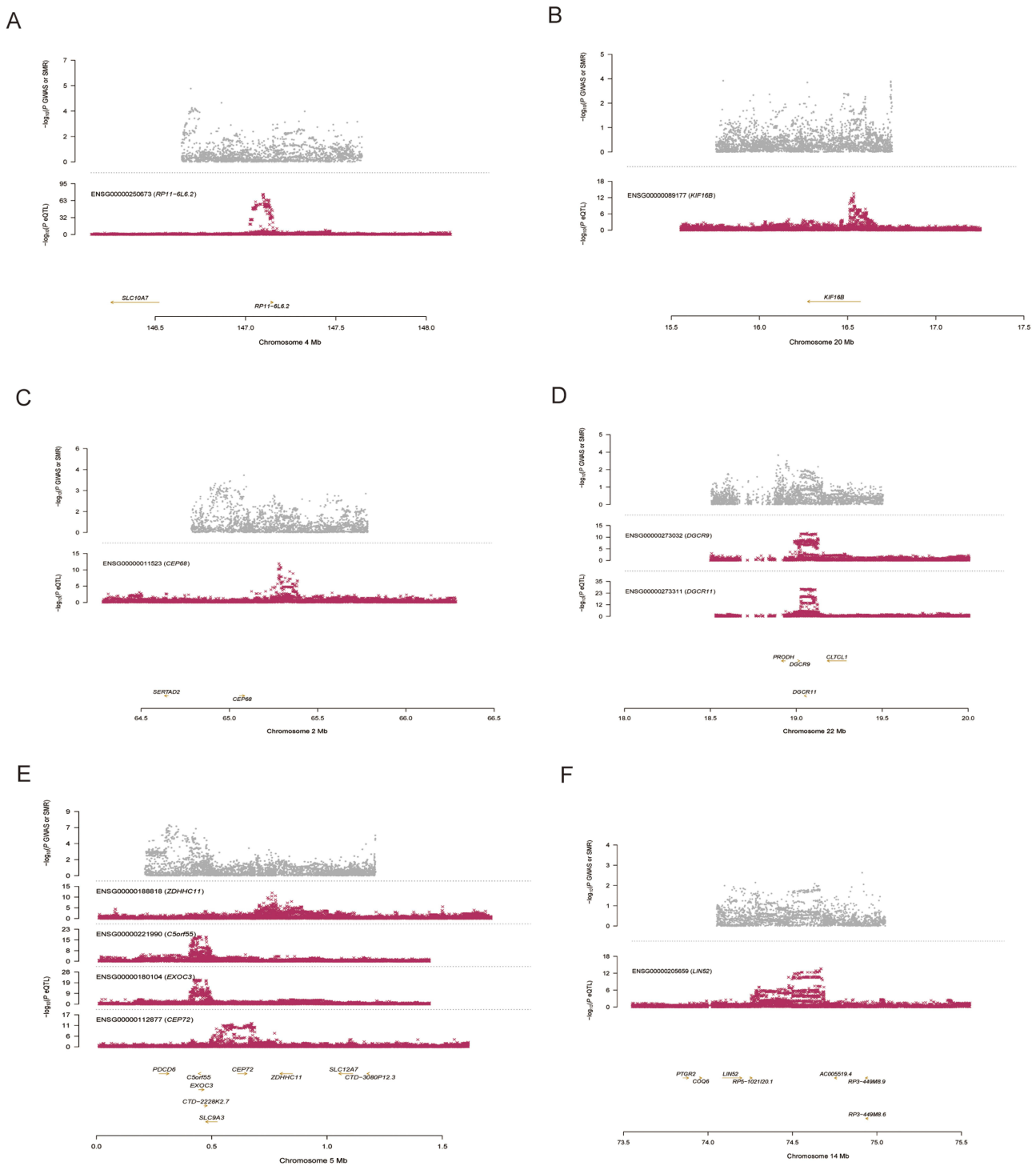


Figure 5 SMR reveal disease-causing genes. (A–F) The risky genes locus for BRCA.

regression analysis and constructed a LASSO model using these genes to predict patient survival prognosis. The model included five genes: *BCL2L14*, *IGHD*, *MAPT-AS1*, *NT5CD4*, and *TNIP3* (Figure 6A and B, Figure S3). The risk score for this model is calculated as: Total risk score = $-0.112 * TNIP3 + -0.177 * NT5DC4 + -0.144 * BCL2L14 + -0.08 * IGHD + -0.210 * MAPT-AS1$.

The higher the score, the greater the risk, and we refer to this classification as geneCluster, which we also call IGKC+ T cells related score. At the same time, we illustrated the relationship between patient subgroups in both models, displaying it

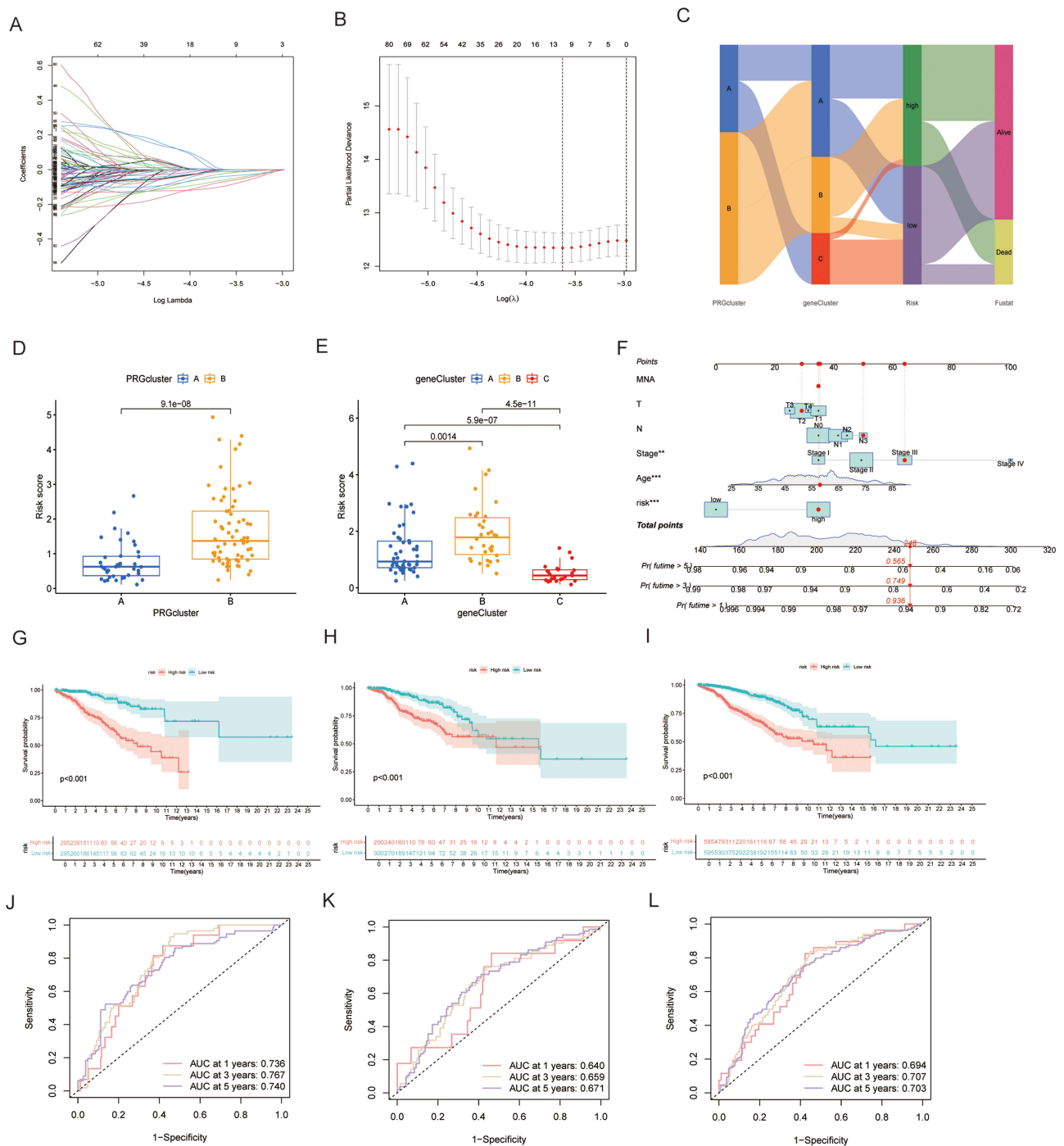


Figure 6 Development and validation of prognostic models. **(A)** Visualization of LASSO regression in the TCGA-BRCA cohort. **(B)** The optimal λ was obtained when the partial likelihood deviance reached the minimum value. Regression coefficients of 6 genes obtained in Cox regression. **(C)** Sankey diagram was used to visualize the correspondence between patient subgroups across the two models. **(D)** Presentation of group risk scores based on PRGcluster. **(E)** Presentation of group risk scores based on geneCluster. **(F)** Construction of the nomogram based on the TIRS and clinical characteristics, including age, stage, risk score, N, and T. **(G-I)** Kaplan-Meier curves of OS according to the TIRS in the TCGA training set, TCGA internal validation set, and GEO cohort, based on the Log rank test. **(J-L)** ROC curves showing the specificity and sensitivity of TIRS in predicting 1, 3, and 5-year OS in the TCGA training set, TCGA internal validation set, and GEO cohort.

as a Sankey diagram (Figure 6C). In Figure 6D and E, we show the risk scores for patients in different subgroups. In the PRGcluster model, subgroup B has the highest risk, and in the geneCluster model, subgroup B also has the highest risk, followed by subgroups A and C. Based on this model and the TCGA breast cancer patient cohort, we constructed a nomogram to calculate patient risk scores (Figure 6F). Subsequently, based on the risk score, we found that compared to individuals with

low-risk scores, those with high-risk scores had significantly worse overall survival (OS) (Figures 6G–I). To evaluate and confirm the predictive performance of this feature, ROC curve analysis and the area under the curve (AUC) were used for validation (Figures 6J–L). The AUC values at 1, 3, and 5 years were all close to 0.7, respectively, indicating that our model had excellent predictive power for clinical information and risk scores (Figure S4).

Based on our model, we mined KEGG enrichment in the GSE58812 and TCGA-BRCA cohorts and displayed the top 20 enriched signaling pathways (Figure 7A). We also calculated immune infiltration in different subgroups. Interestingly, the immune score in subgroup A was significantly higher than in subgroup B (Figure 7B). Furthermore, we displayed the correlation between individual genes in the model and immune cell scores (Figure 7C), and we observed that the genes in the model were correlated with several immune-infiltrating cells. Additionally, we computed the expression differences in immune scores between high- and low-risk groups (Figure 7D) and microsatellite instability scores (Figure 7E). These results suggest that the high-risk group has lower immune scores, lower microsatellite instability, and poorer immune therapy outcomes, potentially indicating a “cold tumor”.

BCL2L14 Upregulates the NF-κB Signaling Pathway to Promote the Proliferation and Migration of Breast Cancer Cells

Based on the above results, we observed that BCL2L14 is positively correlated with T cells follicular helper, T cells CD4 memory activated, and resting dendritic cells, and negatively correlated with resting mast cells (Figure 7C). Compared to

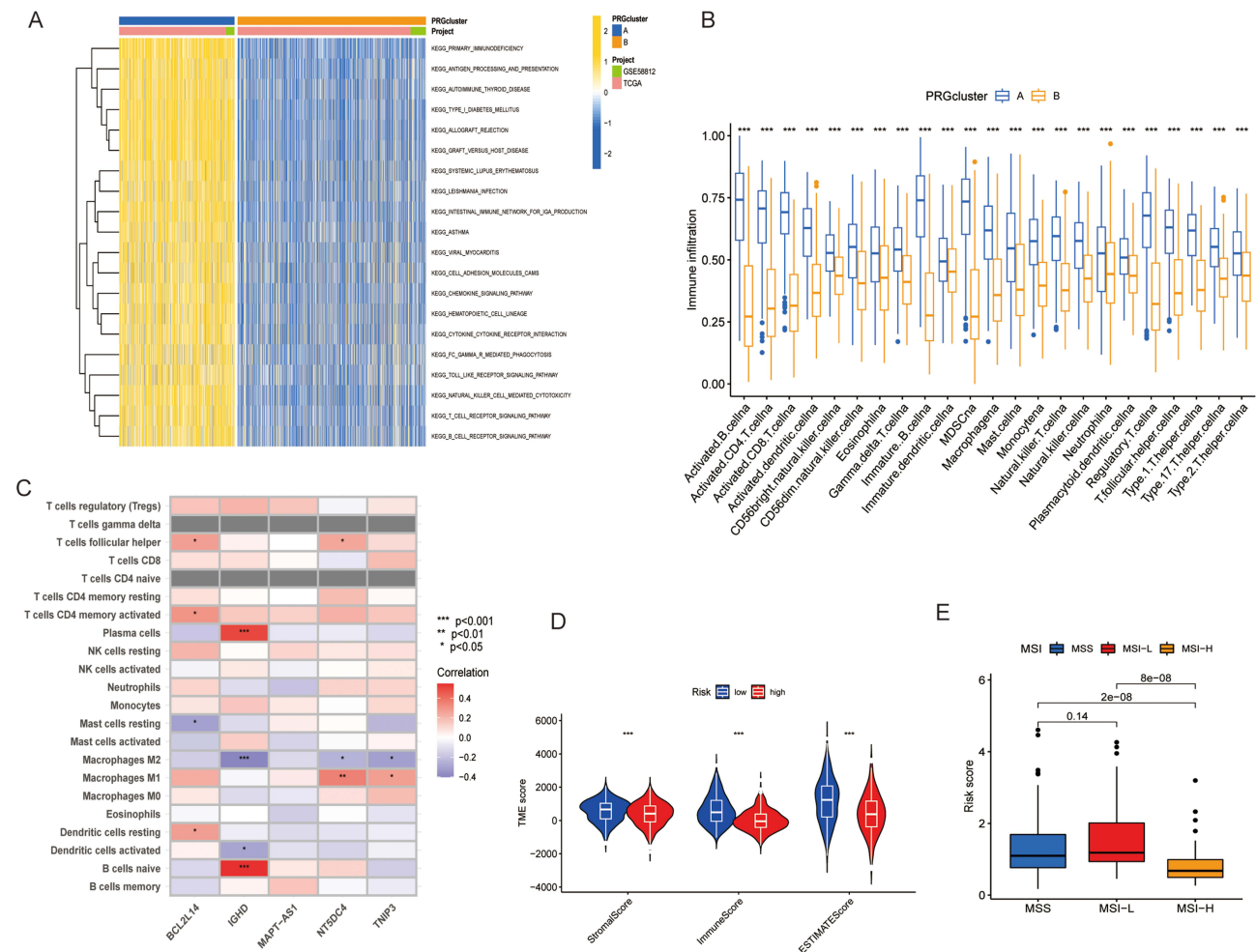


Figure 7 The Immunophenotypes and Therapeutic Analysis Based on Core Gene Models. **(A)** Pathway analysis based on risk score grouping. **(B)** Differential analysis of immune infiltrating cells based on risk score grouping. **(C)** Expression levels of model genes in different immune cells. **(D)** Expression differences in immune scores between high and low-risk groups. **(E)** Differences in risk scores in microsatellite instability. The significance levels are indicated as follows: *P <0.05, **P <0.01, ***P <0.001.

other genes in the model, we believe BCL2L14 plays a more critical role. Therefore, we further investigated the biological function of BCL2L14. Through the CCLE database, we confirmed that the expression of BCL2L14 is lower in breast cancer cells ([Table S9](#), [Figure S5](#)). Thus, we chose to overexpress BCL2L14 in MDA-MB-231 and MDA-MB-468 cells and conducted relevant experimental validation ([Figure 8A](#)). Through CCK-8 and EdU assays, we found that overexpression of BCL2L14 accelerated the proliferation of breast cancer cells ([Figure 8B](#) and [C](#)). In Scratch assays, we found that BCL2L14 overexpression increased cell migration ability ([Figure 8D](#)). In Transwell assays, we observed that

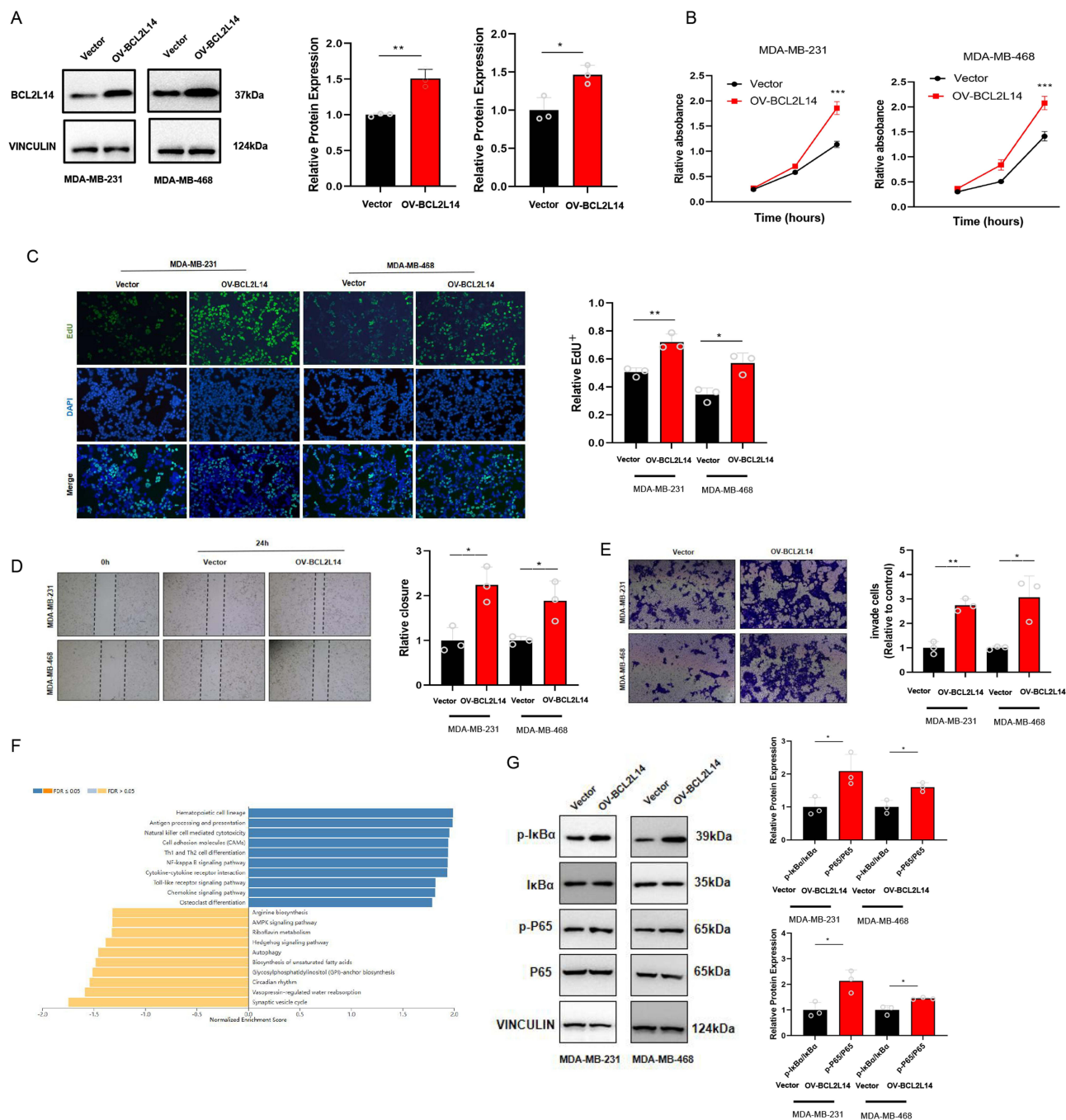


Figure 8 BCL2L14 upregulates the NF- κ B signaling pathway to promote the proliferation and migration of breast cancer cells. **(A)** Validation of BCL2L14 overexpression in MDA-MB-231 and MDA-MB-468 cells verified by Western blot. **(B)** CCK8 was used to detect the cell growth rate at 24 h, 48 h and 72 h, respectively. **(C)** EdU was used to detect the growth rate of cells which overexpressed of BCL2L14 or empty vector, respectively. **(D)** and **(E)** Scratch assay and invasion assay were used to examine the effect of BCL2L14 overexpression on cell invasion ability, respectively. **(F)** GSEA bioinformatics enrichment analysis. **(G)** Western blot was used to detect the expression of NF- κ B signaling pathway-related proteins under different treatment. The significance levels are indicated as follows: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

overexpression of BCL2L14 enhanced the invasive ability of breast cancer cells (Figure 8E). Furthermore, using the GSEA algorithm from the LinkedOmics database (<http://www.linkedomics.org>), we found that BCL2L14 expression was positively correlated with the NF- κ B pathway (Figure 8F). Functional gain-of-function experiments revealed that overexpression of BCL2L14 increased the phosphorylation of NF- κ B signaling pathway proteins (Figure 8G). In conclusion, the experimental results suggest that overexpression of BCL2L14 enhances the proliferation and invasion of breast cancer cells and is positively correlated with the activation of the NF- κ B signaling pathway. These findings also suggest that BCL2L14 is a novel therapeutic target for breast cancer.

Discussion

Most BRCA patients do not have metastases at diagnosis. However, the median overall survival of patients with metastatic BRCA is approximately 5 years or less.³ Although surgery is an effective treatment for localized tumors, there are limited options for treating. Therefore, it is essential to understand the underlying molecular mechanisms involved in BRCA, and identify effective targets for therapeutic strategies.

Previous single-cell RNA sequencing (scRNA-seq) studies have investigated the heterogeneity of the microenvironment-resident cell populations in breast cancer tissues, providing valuable insights.^{11–13} Through our extensive characterization of multiple cell subpopulations in the microenvironment, we successfully identified a unique T cell subtype, named IGKC+ T cells. We characterized its molecular features and identified novel markers for related subpopulations. We identified *BCL2L14*, *IGHD*, *MAPT-ASI*, *NT5DC4*, and *TNIP3* as core enriched genes for this T cell subtype; however, their roles within the tumor microenvironment remain elusive and require further exploration. Additionally, our findings show that tissues with high expression of IGKC+ T cells exhibit increased infiltration of CD8 and CD274, as well as higher T cell dysfunction and TIDE scores. We also observed higher Stromal, Immune, and Estimate scores, and lower Purity scores, suggesting that patients with high expression of IGKC+ T cells may have immune activation effects. These scRNA-seq data reveal the dynamic molecular features of IGKC+ T cells during development and provide valuable resources for further research.

Subsequently, this study investigated whether pathogenic factors were associated with the specific gene expression of malignant cells at the genetic level. Therefore, we chose to use Mendelian randomization analysis to investigate the susceptibility genes associated with breast cancer. Through SMR analysis, this study ultimately identified six potential genes: CEP68, KIF16B, DGCR9, RP11-6L6.2, LIN52, ZDHHC11. We found that research on RP11-6L6.2, LIN52, and ZDHHC11 is quite limited. LIN52, as a subunit of the MuvB complex, forms a complex with four other proteins: LIN9, LIN37, LIN54, and RBAP48. The components of the MuvB complex are evolutionarily conserved across animals and ciliates and are important regulators of cell cycle-dependent gene expression programs.²⁵ The role of LIN52 in tumors remains unclear, although its MuvB complex has been linked to imatinib-induced apoptosis.²⁶ ZDHHC11 has been shown to be highly correlated with the progression of B-cell non-Hodgkin lymphoma.^{27,28} CEP68 is required for centrosome cohesion, with fibers extending from the proximal end of the centrioles and dissociating during mitosis.²⁹ Currently, research on CEP68 in tumors is scarce. However, some data mining studies have found that CEP68 expression is decreased in breast cancer patients, suggesting it as a potential therapeutic target.³⁰ Genetic variations in KIF16B have been found to be associated with the survival rates of non-small cell lung cancer.³¹ Other studies have shown that depletion of KIF16B in macrophages reduces tumor invasiveness, and KIF16B-driven recycling pathways are key regulatory elements in the tumor microenvironment.³² In a multi-cohort study, a model comprising four lncRNAs including DGCR9 was identified by machine learning for non-invasive early detection of gastric cancer.³³ Moreover, a proteomics study found that DGCR9 encodes peptides that are overexpressed in tumors compared to normal samples, suggesting new therapeutic strategies targeting this gene.³⁴ In conclusion, our Mendelian analysis has identified potential new therapeutic targets for breast cancer, and intriguingly, these genes have rarely been studied in breast cancer, if at all.

In the model we constructed, we used univariate Cox analysis and LASSO analysis to identify five core genes: *BCL2L14*, *IGHD*, *MAPT-ASI*, *NT5DC4*, and *TNIP3*, and validated their accuracy and reliability using the TCGA dataset. We found that breast cancer patients in the high-risk group had a worse survival rate. The results from the nomogram suggest that incorporating the values from the multivariate Cox regression model into the risk score calculation can accurately predict the patient's prognostic risk. At the same time, we observed that the immune infiltration and related immune scores in the high-risk group were significantly lower than those in the low-risk group. In the patient stratification based on risk, we also found

that the KEGG enrichment analysis for different subtypes was immune-related, including pathways such as Primary immunodeficiency, Antigen processing and presentation, Natural killer cell-mediated cytotoxicity, etc. Additionally, we observed enrichment in inflammation-related signaling pathways, such as the Chemokine signaling pathway, Toll-like receptor signaling pathway, etc. Furthermore, the malignant potential of breast cancer is closely associated with the infiltration of various immune cells, and more aggressive triple-negative breast cancer (TNBC) tends to have lower immune infiltration, displaying immune checkpoint inhibition characteristics and is referred to as “cold tumors” Tumor-associated macrophages (TAMs) can promote the proliferation, metastasis, and drug resistance of TNBC cells by secreting transforming growth factor- β 1 (TGF- β 1), which activates hepatocellular leukemia factor (HLF) activity.³⁵ Meanwhile, the activated cancer-associated fibroblasts (CAFs) induced by macrophage signaling can enhance TAM activity and create a positive feedback loop to support cancer growth and suppress immune responses within the tumor microenvironment (TME).³⁶ The tumor immune microenvironment is highly complex, with various immune cells, including macrophages, exhibiting different functional effects under different conditions.³⁷ Therefore, molecular stratification of patients or the exploration of new microenvironmental cell subtypes is crucial for improving immune therapy.

The genes in our model include BCL2L14, IGHD, MAPT-AS1, NT5DC4, and TNIP3. BCL-Gonad (BCLG, also known as BCL2-like 14) is an atypical protein of the BCL2 family, as its long isoform (BCL-GL) consists of BH2 and BH3 domains without the BH1 motif.³⁸ BCL2L14 is primarily expressed in normal testes and various organs of the gastrointestinal tract.³⁸ Studies have found that the BCL2L14-ETV6 fusion gene enhances the migration and invasiveness of triple-negative breast cancer (TNBC) cells, and promotes taxol resistance when ectopically expressed in non-metastatic, chemotherapy-sensitive TNBC cell line models.³⁹ Medullary breast carcinoma (MBC), a rare subtype of TNBC, exhibits specific genomic characteristics, and BCL2L14 is the most highly overexpressed gene in MBC.⁴⁰ Other research has identified BCL2L14 through RNA interference screening, revealing that its inhibition increases the survival rate of renal epithelial cells in vitro models under oxygen and glucose deprivation.⁴¹ However, some studies have also shown that the deletion of BCL2L14 accelerates the progression of colitis-associated cancers,⁴² which may be linked to its high expression in intestinal tissues. Therefore, BCL2L14's role in breast cancer may promote tumorigenesis, while its function in other diseases warrants further investigation. The genes encoding the variable region of the human immunoglobulin heavy chain are formed by the recombination of variable genes (IGHV), diversity genes (IGHD), and joining genes (IGHJ). Since IGHD provides an important part of the antigen-binding loop in immunoglobulins, it plays a crucial role in determining antigen-binding specificity.⁴³ Studies have shown that specific amino acids in the nuclear IGHD gene segment significantly influence the absolute number of developing and mature B cell subgroups, antibody production, epitope recognition, protection against pathogen attack, and sensitivity to autoreactive antibody production.⁴⁴ In a risk prediction model for breast cancer constructed by Wang Z et al, IGHD was also included,⁴⁵ which further highlights the importance of this gene. MAPT-AS1 is a long non-coding RNA (LncRNA) that has been confirmed as a reliable diagnostic biomarker for breast cancer due to its overexpression in both tissues and serum. MAPT-AS1 regulates BC cell proliferation and metastasis through the activation of the Wnt/ β -catenin signaling pathway.⁴⁶ However, some studies suggest that MAPT-AS1 is overexpressed in breast cancer but not in TNBC, and that higher MAPT-AS1 expression correlates with better patient survival.⁴⁷ Research by Pan Y et al found that MAPT-AS1, through antisense pairing with MAPT, is related to cell growth, invasiveness, and taxol resistance in ER-negative breast cancer cells. MAPT-AS1 may serve as a potential therapeutic target for ER-negative breast cancer.⁴⁸ Therefore, we believe that MAPT-AS1 plays different roles in different subtypes of breast cancer, which warrants further investigation. The NT5DC family is an evolutionarily conserved 5'-nucleotidase family that catalyzes the hydrolysis of nucleotides in cells.⁴⁹ Dysfunction of NT5DC family members is associated with immune system diseases, sensitivity to cancer treatments, and metabolic disorders.⁵⁰ Mutations in NT5DC4 are rare, but its high mRNA expression suggests poor prognosis in pancreatic cancer patients.⁴⁰ Another predictive model has also suggested that the inclusion of NT5DC4 could predict survival prognosis in esophageal squamous cell carcinoma patients.⁵¹ TNFAIP3 interacting protein 3 (TNIP3), initially discovered during *Listeria monocytogenes* infection in human monocyte-like macrophages,⁵² has recently been found to assist the ubiquitination and degradation of LATS2, leading to YAP activation, which alleviates cell death and inflammation. This makes TNIP3 a promising therapeutic target for liver I/R injury.⁵³ Additionally, TNIP3 plays a critical role in isoproterenol-induced cardiac remodeling and ventricular arrhythmias, which is mediated by the inhibition of the PI3K/Akt/NF- κ B signaling pathway.⁵⁴ However, research on this gene in tumors is still limited, and further studies are needed.

Conclusion

In summary, our study suggests that the prognostic model based on the new cell subset IGKC⁺ T cells can effectively predict the prognosis of breast cancer patients. Furthermore, *BCL2L14*, the core gene of this cell subset, modulates breast cancer characteristics by activating the NF- κ B signaling pathway and may serve as a potential therapeutic target. However, due to the lack of in vivo experimental validation, further experimental research is required to confirm and refine the results of this study. Additionally, more data is needed to enhance the accuracy and performance of the model. In the future, further analysis of *BCL2L14*'s function will be necessary to uncover its potential broader impact in breast cancer.

Data Sharing Statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Ethics Approval and Informed Consent

The use of the TCGA and GEO databases in this study was approved by the ethics committee of First Hospital of Chongqing Medical University (ID: 2023-792).

Acknowledgments

This study was supported by Chongqing Education Commission “Natural Medicine Anti- tumor” innovative research (No. CXQT20030, to Dilong Chen). Chongqing Talent Program, (No. cstc2022ycjh- bgzxm0226, to Dilong Chen), Selection and Development of indigenous medicinal materials in the Three Gorges Reservoir Area. And, the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJZD- M202202701, to Dilong Chen).

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Disclosure

The authors declare that there are no conflicts of interest.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* 2020;70:7–30. doi:10.3322/caac.21590
2. Sung H, Ferlay J, Siegel RL, 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):209–249. doi:10.3322/caac.21660
3. Waks AG, Winer EP. Breast cancer treatment: a review. *JAMA.* 2019;321(3):288–300. doi:10.1001/jama.2018.19323
4. Pal B, Chen Y, Vaillant F, et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J.* 2021;40(11):e107333. doi:10.15252/embj.2020107333
5. Jiang Y-Z, Ma D, Suo C. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer Cell.* 2019;35(3):428–440.e5. doi:10.1016/j.ccell.2019.02.001
6. Gambardella G, Viscido G, Tumaini B, Isacchi A, Bosotti R, Di Bernardo D. A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response. *Nat Commun.* 2022;13(1):1714. doi:10.1038/s41467-022-29358-6
7. Xu L, Saunders K, Huang SP, et al. A comprehensive single-cell breast tumor atlas defines epithelial and immune heterogeneity and interactions predicting anti-PD-1 therapy response. *Cell Rep Med.* 2024;5(5):101511. doi:10.1016/j.xcrm.2024.101511
8. Gao ZJ, Fang H, Sun S, et al. Single-cell analyses reveal evolution mimicry during the specification of breast cancer subtype. *Theranostics.* 2024;14(8):3104–3126. doi:10.7150/thno.96163
9. Nalio Ramos R, Missolo-Koussou Y, Gerber-Ferder Y, et al. Tissue-resident FOLR2⁺ macrophages associate with CD8⁺ T cell infiltration in human breast cancer. *Cell.* 2022;185(7):1189–1207.e25. doi:10.1016/j.cell.2022.02.021
10. Network CGAR, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.* 2013;45(10):1113–1120. doi:10.1038/ng.2764

11. Yan Y, Sun D, Hu J, et al. Multi-omic profiling highlights factors associated with resistance to immuno-chemotherapy in non-small-cell lung cancer. *Nat Genet.* doi:10.1038/s41588-024-01998-y
12. Khaliq AM, Rajamohan M, Saeed O, et al. Spatial transcriptomic analysis of primary and metastatic pancreatic cancers highlights tumor microenvironmental heterogeneity. *Nat Genet.* 2024;56(11):2455–2465. doi:10.1038/s41588-024-01914-4
13. Wagner J, Rapsomaniki MA, Chevrier S, et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell.* 2019;177(5):1330–1345.e18. doi:10.1016/j.cell.2019.03.005
14. Chiu YC, Zheng S, Wang LJ, et al. Predicting and characterizing a cancer dependency map of tumors with deep learning. *Sci Adv.* 2021;7(34):eabh1275. doi:10.1126/sciadv.abh1275
15. Jubran J, Slutsky R, Rozenblum N, Rokach L, Ben-David U, Yeager-Lotem E. Machine-learning analysis reveals an important role for negative selection in shaping cancer aneuploidy landscapes. *Genome Biol.* 2024;25(1):95. doi:10.1186/s13059-024-03225-7
16. Wu L, Wen Y, Leng D, et al. Machine learning methods, databases and tools for drug combination prediction. *Brief Bioinform.* 2022;23(1):bbab355. doi:10.1093/bib/bbab355
17. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* 2021;13(1):112. doi:10.1186/s13073-021-00930-x
18. Wang Z, Gu Y, Huang L, et al. Construction of machine learning diagnostic models for cardiovascular pan-disease based on blood routine and biochemical detection data. *Cardiovasc Diabetol.* 2024;23(1):351. doi:10.1186/s12933-024-02439-0
19. Shi M, Yang A, Lau ESH, et al. A novel electronic health record-based, machine-learning model to predict severe hypoglycemia leading to hospitalizations in older adults with diabetes: a territory-wide cohort and modeling study. *PLoS Med.* 2024;21(4):e1004369. doi:10.1371/journal.pmed.1004369
20. Guo Y, Li L, Zheng K, et al. Development and validation of a survival prediction model for patients with advanced non-small cell lung cancer based on LASSO regression. *Front Immunol.* 2024;15:1431150. doi:10.3389/fimmu.2024.1431150
21. Guan S, Xu Z, Yang T, et al. Identifying potential targets for preventing cancer progression through the PLA2G1B recombinant protein using bioinformatics and machine learning methods. *Int J Biol Macromol.* 2024;276(Pt 1):133918. doi:10.1016/j.ijbiomac.2024.133918
22. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet.* 2016;48(5):481–487. doi:10.1038/ng.3538
23. Gleeher P, Cox N, Huang RS. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PLoS One.* 2014;9(9):e107468. doi:10.1371/journal.pone.0107468
24. He J, Wei Q, Jiang R, et al. The core-targeted RRM2 gene of berberine hydrochloride promotes breast cancer cell migration and invasion via the epithelial-mesenchymal transition. *Pharmaceuticals.* 2022;16(1):42. doi:10.3390/ph16010042
25. Fischer M, Müller GA. Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. *Crit Rev Biochem Mol Biol.* 2017;52(6):638–662. doi:10.1080/10409238.2017.1360836
26. Boichuk S, Parry JA, Makielski KR, et al. The DREAM complex mediates GIST cell quiescence and is a novel therapeutic target to enhance imatinib-induced apoptosis. *Cancer Res.* 2013;73(16):5120–5129. doi:10.1158/0008-5472.CAN-13-0579
27. Dzikiewicz-Krawczyk A, Kok K, Slezak-Prochazka I, et al. ZDHHC11 and ZDHHC11B are critical novel components of the oncogenic MYC-miR-150-MYB network in Burkitt lymphoma. *Leukemia.* 2017;31(6):1470–1473. doi:10.1038/leu.2017.94
28. Liu Y, Zhao X, Seitz A, et al. Circular ZDHHC11 supports Burkitt lymphoma growth independent of its miR-150 binding capacity. *Sci Rep.* 2024;14(1):8730. doi:10.1038/s41598-024-59443-3
29. Graser S, Stierhof YD, Nigg EA. Cep68 and Cep215 (Cdk5rap2) are required for centrosome cohesion. *J Cell Sci.* 2007;120(Pt 24):4321–4331. doi:10.1242/jcs.020248
30. Kim DH, Lee KE. Discovering breast cancer biomarkers candidates through mRNA expression analysis based on the cancer genome atlas database. *J Pers Med.* 2022;12(10):1753. doi:10.3390/jpm12101753
31. Yang S, Tang D, Zhao YC, et al. Novel genetic variants in KIF16B and NEDD4L in the endosome-related genes are associated with nonsmall cell lung cancer survival. *Int J Cancer.* 2020;147(2):392–403. doi:10.1002/ijc.32739
32. Hey S, Wiesner C, Barcelona B, Linder S. KIF16B drives MT1-MMP recycling in macrophages and promotes co-invasion of cancer cells. *Life Sci Alliance.* 2023;6(11):e202302158. doi:10.26508/lsa.202302158
33. Cai ZR, Zheng YQ, Hu Y, et al. Construction of exosome non-coding RNA feature for non-invasive, early detection of gastric cancer patients by machine learning: a multi-cohort study. *Gut.* 2025;gutjnl–2024–333522. doi:10.1136/gutjnl-2024-333522
34. Xiang R, Ma L, Yang M, et al. Increased expression of peptides from non-coding genes in cancer proteomics datasets suggests potential tumor neoantigens. *Commun Biol.* 2021;4(1):496. doi:10.1038/s42003-021-02007-2
35. Li H, Yang P, Wang J, et al. HLF regulates ferroptosis, development and chemoresistance of triple-negative breast cancer by activating tumor cell-macrophage crosstalk. *J Hematol Oncol.* 2022;15(1):2. doi:10.1186/s13045-021-01223-x
36. Gunaydin G. CAFs interacting with TAMs in tumor microenvironment to enhance tumorigenesis and immune evasion. *Front Oncol.* 2021;11:668349. doi:10.3389/fonc.2021.668349
37. Kundu M, Butti R, Panda VK, et al. Modulation of the tumor microenvironment and mechanism of immunotherapy-based drug resistance in breast cancer. *Mol Cancer.* 2024;23(1):92. doi:10.1186/s12943-024-01990-4
38. Hartman ML, Czyz M. BCL-G: 20 years of research on a non-typical protein from the BCL-2 family. *Cell Death Differ.* 2023;30(6):1437–1446. doi:10.1038/s41418-023-01158-5
39. Lee S, Hu Y, Loo SK, et al. Landscape analysis of adjacent gene rearrangements reveals BCL2L14-ETV6 gene fusions in more aggressive triple-negative breast cancer. *Proc Natl Acad Sci U S A.* 2020;117(18):9912–9921. doi:10.1073/pnas.1921333117
40. Romero P, Benhamo V, Deniziaut G, et al. Medullary breast carcinoma, a triple-negative breast cancer associated with BCLG Overexpression. *Am J Pathol.* 2018;188(10):2378–2391. doi:10.1016/j.ajpath.2018.06.021
41. Zynda ER, Schott B, Babagana M, et al. An RNA interference screen identifies new avenues for nephroprotection. *Cell Death Differ.* 2016;23(4):608–615. doi:10.1038/cdd.2015.128
42. Nguyen PM, Dagley LF, Preaudet A, et al. Loss of Bcl-G, a Bcl-2 family member, augments the development of inflammation-associated colorectal cancer. *Cell Death Differ.* 2020;27(2):742–757. doi:10.1038/s41418-019-0383-9

43. Qiu Q, Zhang P, Zhang N, Shen Y, Lou S, Deng J. Development of a prognostic nomogram for acute myeloid leukemia on ighd gene family. *Int J Gen Med.* 2021;14:4303–4316. doi:10.2147/IJGM.S317528
44. Khass M, Vale AM, Burrows PD, Hw S. The sequences encoded by immunoglobulin diversity (DH) gene segments play key roles in controlling B-cell development, antigen-binding site diversity, and antibody production. *Immunol Rev.* 2018;284(1):106–119. doi:10.1111/immr.12669
45. Wang Z, Liu H, Gong Y, Cheng Y. Establishment and validation of an aging-related risk signature associated with prognosis and tumor immune microenvironment in breast cancer. *Eur J Med Res.* 2022;27(1):317. doi:10.1186/s40001-022-00924-4
46. Li B, Liu D, Xu J. Overexpression of lncRNA MAPT-AS1 exacerbates cell proliferation and metastasis in breast cancer. *Transl Cancer Res.* 2022;11(4):835–847. doi:10.21037/tcr-22-719
47. Wang D, Li J, Cai F, et al. Overexpression of MAPT-AS1 is associated with better patient survival in breast cancer. *Biochem Cell Biol.* 2019;97(2):158–164. doi:10.1139/bcb-2018-0039
48. Cheng Y, Cheng Y, Yang F, Yao Z, Wang O, Wang O. Knockdown of lncRNA MAPT-AS1 inhibits proliferation and migration and sensitizes cancer cells to paclitaxel by regulating MAPT expression in ER-negative breast cancers. *Cell Biosci.* 8:7. doi:10.1186/s13578-018-0207-5
49. Singgih EL, van der Voet M, Schimmel-Naber M, Brinkmann EL, Schenck A, Franke B. Investigating cytosolic 5'-nucleotidase II family genes as candidates for neuropsychiatric disorders in Drosophila (114/150 chr). *Transl Psychiatry.* 2021;11(1):55. doi:10.1038/s41398-020-01149-x
50. Yu X, Sun R, Yang X, He X, Guo H, Ou C. The NT5DC family: expression profile and prognostic value in pancreatic adenocarcinoma. *J Cancer.* 2023;14(12):2274–2288. doi:10.7150/jca.85811
51. Yu XW, She PW, Chen FC, et al. Metabolic subtypes and immune landscapes in esophageal squamous cell carcinoma: prognostic implications and potential for personalized therapies. *BMC Cancer.* 2024;24(1):230. doi:10.1186/s12885-024-11890-x
52. Staeger H, Brauchlin A, Schoedon G, Schaffner A. Two novel genes FIND and LIND differentially expressed in deactivated and Listeria-infected human macrophages. *Immunogenetics.* 2001;53(2):105–113. doi:10.1007/s002510100306
53. Zhou J, Hu M, He M. TNFAIP3 Interacting Protein 3 Is an Activator of Hippo-YAP Signaling Protecting Against Hepatic Ischemia/Reperfusion Injury. *Hepatology.* 2021;74(4):2133–2153. doi:10.1002/hep.32015
54. Yang H, Shen X, Wang H, Shuai W. TNIP3 overexpression protects against arrhythmogenic remodeling in the heart failure mice. *Europace.* doi:10.1093/europace/eaaf002

Journal of Inflammation Research

Publish your work in this journal

The Journal of Inflammation Research is an international, peer-reviewed open-access journal that welcomes laboratory and clinical findings on the molecular basis, cell biology and pharmacology of inflammation including original research, reviews, symposium reports, hypothesis formation and commentaries on: acute/chronic inflammation; mediators of inflammation; cellular processes; molecular mechanisms; pharmacology and novel anti-inflammatory drugs; clinical conditions involving inflammation. The manuscript management system is completely online and includes a very quick and fair peer-review system. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/journal-of-inflammation-research-journal>

Dovepress
Taylor & Francis Group