Am J Clin Exp Immunol 2025;14(2):45-67
www.ajcei.us /ISSN:2164-7712/AJCEI0159295

Original Article

Comprehensive analysis of autophagy-related prognostic genes in breast cancer using bulk and single-cell RNA sequencing

Yong Li[1,2], Chunmei Chen[1], Weiwen Li[1], Mingtao Shao[1], Yan Dong[1], Qunchen Zhang[1]

[1]*Department of Breast, Jiangmen Central Hospital, Jiangmen, Guangdong, P. R. China;* [2]*Department of General Surgery, The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong, P. R. China*

Received July 18, 2024; Accepted April 6, 2025; April 25, 2025; Published April 30, 2025

Abstract: Objective: This study aimed to utilize single-cell RNA sequencing (scRNA-seq) to elucidate the autophagic landscape in breast cancer and to develop a prognostic model for breast cancer patients based on traditional high-throughput RNA sequencing (bulk RNA-seq). Methods: We analyzed scRNA-seq data from the GSE75688 dataset to explore the expression patterns of autophagy-related genes (ARGs) across distinct cellular clusters. ARGs were retrieved from the GeneCards database, and bulk RNA-seq data were obtained from The Cancer Genome Atlas (TCGA). Cox proportional hazards regression was employed to construct a prognostic risk model based on ARGs. Patients were subsequently stratified into high-risk and low-risk groups according to their risk scores. For external validation, we used gene expression data from the GSE20685 and GSE48390 datasets. Receiver operating characteristic (ROC) curve analysis was performed to evaluate the performance of the 3-gene signature. Results: Using the FindClusters function in Seurat, all cells were grouped into four distinct clusters, highlighting the intratumoral heterogeneity within the samples. Significant differences in autophagy scores were observed among the clusters. Fifteen differentially expressed autophagy-related genes were identified, and a prognostic signature consisting of three autophagy-related genes - FEZ1, STX11, and ADAMTSL1 - was developed. Based on this model, patients were classified into high- and low-risk groups, with a statistically significant difference in survival between the two groups (log-rank test, $P = 0.0011$). The model demonstrated robust predictive performance with an AUC of 0.761 in the external validation dataset. A nomogram incorporating the 3-gene signature and clinical factors showed strong prognostic discrimination. Conclusion: This study uncovered significant variation in autophagy levels among different breast cancer cell clusters. Furthermore, we established a novel 3-gene autophagy-related prognostic model that effectively stratifies patient risk and provides a potential tool for personalized prognosis in breast cancer.

Keywords: Autophagy, breast cancer, prediction model, prognosis, single-cell RNA sequencing

Introduction

Breast cancer is one of the most prevalent malignant tumors among women worldwide, with a steadily increasing incidence rate [1, 2]. Timely detection and appropriate treatment significantly enhance the chances of survival, with the overall 5-year relative survival rate reaching approximately 90% [2]. The American Joint Committee on Cancer (AJCC) has traditionally utilized the tumor-node-metastasis (TNM) staging system for breast cancer classification. However, despite its prognostic utility, clinical outcomes can differ significantly among patients with the same TNM stage. This highlights the limitations of the TNM system in accurately predicting breast cancer prognosis.

In addition to TNM staging, various other factors - such as tumor pathological grade, molecular subtype, and the Ki67 labeling index - play critical roles in determining prognosis. Recent advances in cancer genomics, particularly in high-throughput sequencing technologies, have further refined prognostic evaluation. Gene expression profiling tools like the Oncotype DX 21-gene signature and the MammaPrint 70-gene signature have been employed to identify biomarkers that inform breast cancer prognosis and guide treatment decisions [3, 4].

https://doi.org/10.62347/XPCM9169

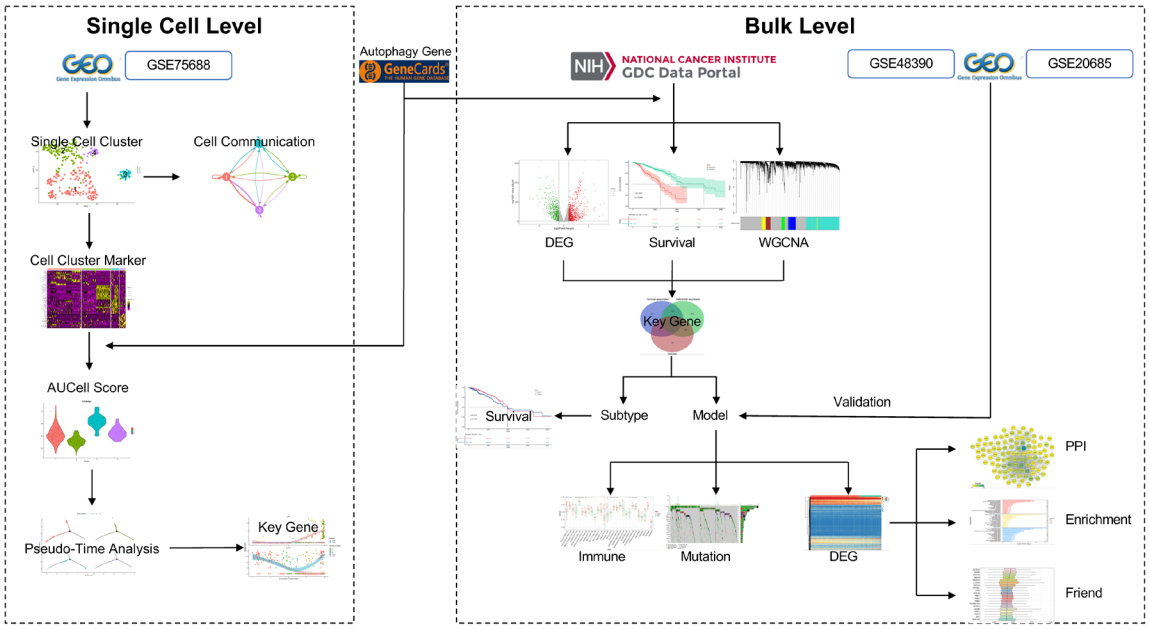

Figure 1. The flowchart of this study.

Autophagy is an intracellular degradation process that maintains cellular homeostasis by lysosomal degradation of macromolecules and damaged organelles. Dysfunction in autophagy is now recognized as a hallmark of cancer, though its role remains paradoxical [5, 6]. Multiple studies suggest that autophagy suppresses tumor initiation in early stages, while later supporting tumor progression and survival. Specifically, in breast cancer, autophagy has shown this dual nature. For instance, autophagic cell death induced by Bcl-2 in MCF-7 breast cancer cells has shown potential as a therapeutic strategy [7]. Conversely, inhibiting autophagy has been reported to enhance the efficacy of tamoxifen in ER-positive breast cancer [8], and to increase the cytotoxicity of epirubicin by promoting apoptosis [9]. Clearly, autophagy plays a complex and significant role in breast cancer progression.

Given this, autophagy-related genes (ARGs) may serve as potential prognostic markers for breast cancer - a disease known for its high heterogeneity. Unlike previous studies, we sourced the autophagy-related gene set from the GeneCards database and performed a comprehensive analysis to investigate the role of autophagy in prognosis. Using bioinformatics approaches and single-cell analysis, we explored tumor heterogeneity and autophagy status at the cellular level. Based on these findings, we constructed a prognostic model using ARGs to predict outcomes in breast cancer patients.

Methods

As shown in **Figure 1**, the study flowchart illustrates the overall research process. The single-cell sequencing dataset GSE75688, derived from breast cancer samples, was downloaded from the Gene Expression Omnibus (GEO) database [10]. This dataset includes a total of 326 breast cancer cells, as annotated by Chung et al., and was used for all single-cell analyses.

RNA sequencing data and corresponding clinical information were obtained from The Cancer Genome Atlas (TCGA) database (http://tcga-data.nci.nih.gov/tcga), comprising 1,104 breast cancer samples and 113 normal breast tissue samples. Additionally, two datasets - GSE20685 [11] and GSE48390 [12] - were retrieved from the GEO database for use as validation cohorts. Both datasets were derived from primary human breast cancer tissues and were generated using the same platform (GPL570).

To correct for batch effects across datasets, we applied the ComBat function from the SVA package in R. The sample sizes for GSE20685 and GSE48390 were 327 and 81, respectively. Furthermore, we obtained a list of 7,242 pyroptosis-related genes from the GeneCards database [13].

Single-cell analysis

We identified highly variable genes using the FindVariableFeatures function of the Seurat package [14]. Dimensionality reduction was then carried out using principal component analysis (PCA), and statistically significant principal components were determined using both the JackStraw and Elbow methods. Unsupervised clustering of cells was subsequently performed with Seurat's FindClusters function.

To identify significantly expressed genes within each cell cluster, we used the FindAllMarkers function with its default parameters, applying a cutoff of $|\log_2 FC| > 1$ and an adjusted *P*-value < 0.05. Differentially expressed genes (DEGs) were intersected with autophagy-related genes (ARGs) obtained from the GeneCards database to identify relevant autophagy-associated expression patterns.

To evaluate the autophagy activity across different cell clusters, we calculated AUCell scores using the AUCell package. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were then performed on the DEGs from the cluster exhibiting the highest autophagy level.

To explore potential relationships between the different cell clusters, we conducted pseudotime trajectory analysis using the Monocle package [15]. Furthermore, intercellular communication analysis was performed using CellPhoneDB [16], providing insights into ligand-receptor interactions among the cell populations.

Bulk RNA-seq analysis

Among the 7,242 autophagy-related genes (ARGs) retrieved from the GeneCards database (Table S1), 7,063 overlapped with the gene expression profiles from TCGA. Key genes were identified through the intersection of three criteria: (1) significant differential expression between normal and cancer tissue samples; (2) strong correlation of ARGs with overall survival; and (3) membership in gene modules highly correlated with the cancer phenotype.

Differentially expressed genes (DEGs) between normal and tumor samples were identified using the DESeq2 package in R, with the thresholds set at |fold change| > 1 and adjusted *p*-value < 0.05 [17]. Prognosis-related genes were identified through univariate Cox regression analysis ($P < 0.05$). To identify functionally relevant gene modules, we conducted weighted correlation network analysis (WGCNA) [18] using the WGCNA package [19], allowing us to detect ARG modules significantly associated with the cancer phenotype. Based on the expression patterns of the resulting key genes, we performed unsupervised clustering of cancer samples to verify their ability to distinguish autophagic states.

Next, we applied univariate Cox regression followed by least absolute shrinkage and selection operator (LASSO) regression to further narrow down prognostic ARGs. A multivariate Cox regression analysis was then used to construct the final prognostic model. The model's performance was validated using preprocessed external test datasets. Based on the model-derived risk scores, patients were stratified into high-risk and low-risk groups.

We further analyzed the DEGs between these two risk groups, including protein - protein interaction (PPI) network analysis, immune infiltration profiling, and mutational signature analysis. To evaluate the model's clinical independence, we integrated it with other clinical variables and constructed a nomogram using the rms package. The nomogram was validated through calibration curves and decision curve analysis to assess its predictive accuracy and clinical utility.

Statistical analysis

All statistical analyses were conducted using R Language for Statistical Computing (version 4.1). Wilcoxon rank-sum tests were applied when variables were non-normally distributed. *P* values were two-sided, and a p value < 0.05 was considered statistically significant.

Results

A total of 326 breast cancer cells were screened to identify highly variable genes, resulting in 1,270 high-variance genes selected from an initial pool of 55,823 genes for subsequent analysis. Based on the selected principal components, all cells were grouped into four distinct clusters using unsupervised clustering methods. Visualization through UMAP and t-SNE dimensionality reduction techniques revealed clear boundaries between the cell clusters (**Figure 2A**, **2B**).

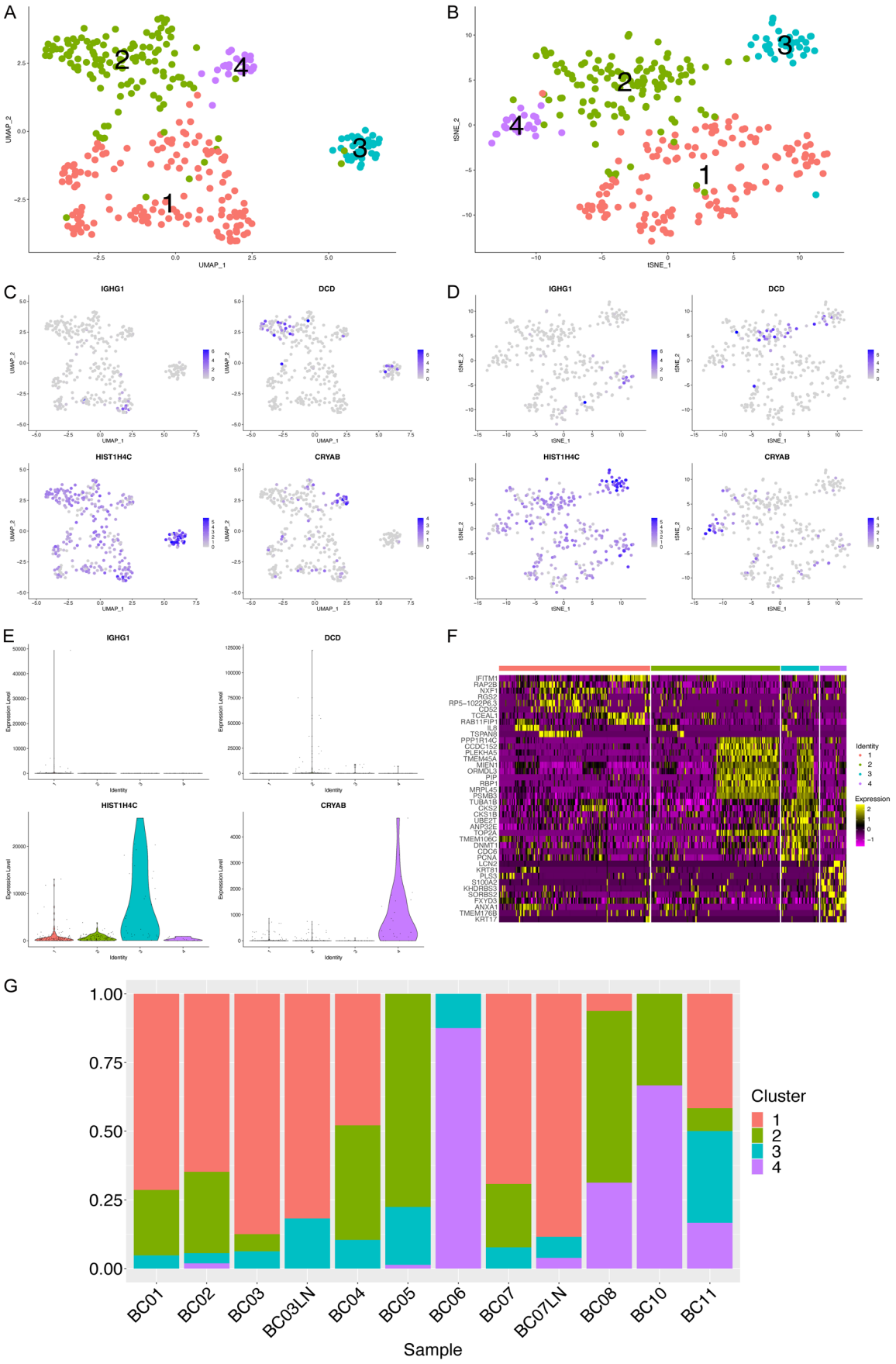
A
B
C
D
E
F
G
IGHG1
DCD
HIST1H4C
CRYAB
UMAP_1
UMAP_2
tSNE_1
tSNE_2
Identity
Expression Level
Expression
Cluster
Sample
BC01
BC02
BC03
BC03LN
BC04
BC05
BC06
BC07
BC07LN
BC08
BC10
BC11

Figure 2. Cell clusters. (A) UMAP dimensionality reduction illustrating clustering results. The X and Y axes represent the two UMAP dimensions. Each point denotes a cell, color-coded by cluster label. (B) t-SNE dimensionality reduction showing clustering results. The X and Y axes correspond to the two t-SNE dimensions. Cells are color-coded by cluster label, consistent with (A). (C) UMAP-based gene expression distribution. Axes represent UMAP dimensions. Cells are colored by gene expression levels, with darker shades indicating higher expression. Clusters correspond to those in (A). (D) t-SNE-based gene expression profile. Axes represent t-SNE dimensions. Cells are colored according to gene expression, with darker colors indicating higher levels. Clusters correspond to those in (B). (E) Cluster-specific gene expression. The X-axis indicates cell clusters, and the Y-axis shows expression levels. Only significantly differentially expressed genes with the highest $\log_2$ fold change ($\log_2$FC) in each cluster are shown. (F) Heatmap of cluster-specific gene expression. Displayed are the top $\log_2$FC genes per cluster. Upper bars represent individual clusters. (G) Cumulative histogram of cluster proportions across samples. The X-axis represents samples, and the Y-axis indicates the proportion of cells. Colors distinguish different clusters.

Following statistical screening, we identified 128 differentially expressed genes (DEGs) across the four clusters (Table S2), with 41 DEGs in cluster 1, 15 in cluster 2, 35 in cluster 3, and 37 in cluster 4. As shown in **Figure 2E**, the expression of specific genes within each cluster - such as *IGHG1*, *DCD*, *HIST1H4C*, and *CRYAB* - differed significantly compared to other clusters. These gene expression patterns were spatially confirmed through distribution plots in the reduced dimensional space (**Figure 2C**, **2D**), illustrating that the identified genes were predominantly expressed in their respective cell clusters.

To further explore DEG expression, we visualized the top 10 DEGs per cluster using a heatmap (**Figure 2F**), which demonstrated a strong correspondence between the DEGs and their respective clusters.

To evaluate the sample-specific distribution of cell clusters, we generated a histogram displaying the proportion of each cluster within each individual sample (**Figure 2G**). The results revealed substantial variability in cluster composition among different samples, with each sample containing at least two distinct cell clusters. This finding highlights the presence of notable intratumoral heterogeneity within the breast cancer samples.

Expression characteristics of autophagy genes in each cell cluster

A total of 7,242 autophagy-related genes were obtained from the GeneCards database. These genes were intersected with the differentially expressed genes (DEGs) identified from each cell cluster, resulting in 73 cluster-specific autophagy genes (Table S3). To visualize their expression patterns, violin plots and heatmaps were generated for the top 10 genes in each cluster - or all genes if fewer than 10 were available (**Figure 3A**, **3C**). Additionally, we assessed the expression correlation among the 73 autophagy-related genes across all cells (**Figure 3B**). Most genes showed either no correlation or positive correlations with one another, such as *PMAIP1* and *B2M*. However, a few genes demonstrated negative correlations, including *RPL23* and *RBP1*.

To quantify autophagy activity in each cluster, we used the AUCell package to calculate autophagy scores based on the expression of these 73 genes. As shown in **Figure 4A** and **4B**, autophagy scores varied significantly among the clusters. Cluster 3 had the highest autophagy score, indicating elevated autophagic activity, while cluster 2 exhibited the lowest score.

We then performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses using the 35 DEGs from cluster 3 (**Figure 4C**, **4D**). GO analysis revealed significant enrichment of 244 biological processes (Table S4), with the top five terms including *chromosome segregation, nuclear division*, *organelle fission*, *nuclear chromosome segregation*, and *sister chromatid segregation*. Similarly, KEGG analysis identified seven significantly enriched pathways (Table S5), with the top five being *cell cycle*, *oocyte meiosis*, *p53 signaling pathway*, *progestin-mediated oocyte maturation*, and *base excision repair*. These findings were consistent with the GO results and collectively underscored the strong association between cluster 3 and cell cycle - related processes.

Pseudotime trajectory analysis (**Figure 5A**, **5B**) revealed a non-linear differentiation trajectory among the clusters, forming an approximately triangular structure divided by a central branch point. Further analysis of this branch point (**Figure 5C**, **5D**) showed that 43 autophagy-related genes - two of which are highlighted in

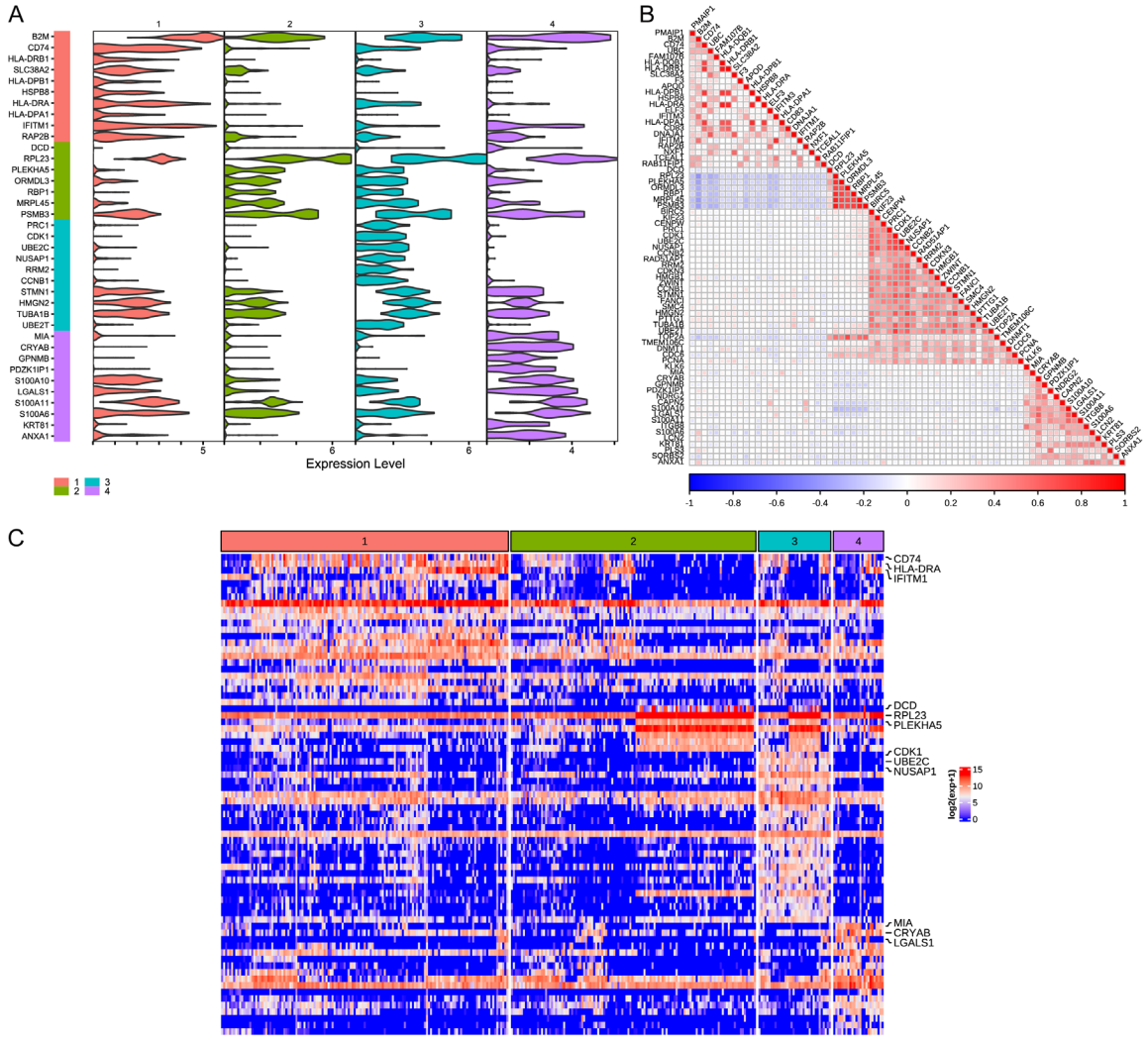

Figure 3. Autophagy gene expression. A. Expression of autophagy-related genes across cell clusters. The X-axis represents expression levels, and the Y-axis lists individual autophagy genes. Cell clusters are color-coded in alignment with **Figure 2A**. B. Correlation matrix of autophagy gene expression. Colors and square sizes represent the strength and direction of correlations: red indicates positive correlation, blue indicates negative correlation, and larger squares denote higher correlation magnitudes. C. Heatmap of autophagy gene expression across clusters. The top three most highly expressed autophagy genes in each cell cluster are labeled.

Panel D - exhibited bifurcated expression trends near this branching event, suggesting functional transitions between cellular states.

To investigate intercellular communication, we used CellPhoneDB to generate heatmaps (**Figure 6A**) and interaction network plots (**Figure 6B-F**). This analysis revealed that clusters 1 and 4 had the strongest intercluster communication, suggesting similar biological functions or coordinated activity. In contrast, clusters 2, 3, and 4 showed fewer receptor-ligand interactions, indicating relatively weaker communication among these clusters.

Differentially expressed ARGs in Bulk RNA-seq analysis

Following statistical threshold filtering, a total of 1,389 genes were retained for further analysis (**Figure 7A**). These genes demonstrated significant differential expression between normal and tumor tissue samples (**Figure 7B**). Integration with clinical survival data from TCGA identified 552 genes significantly associated with overall survival. Based on the expression profiles of these 552 genes, all breast cancer samples were stratified into high-risk and low-risk groups. Survival analysis using the log-rank

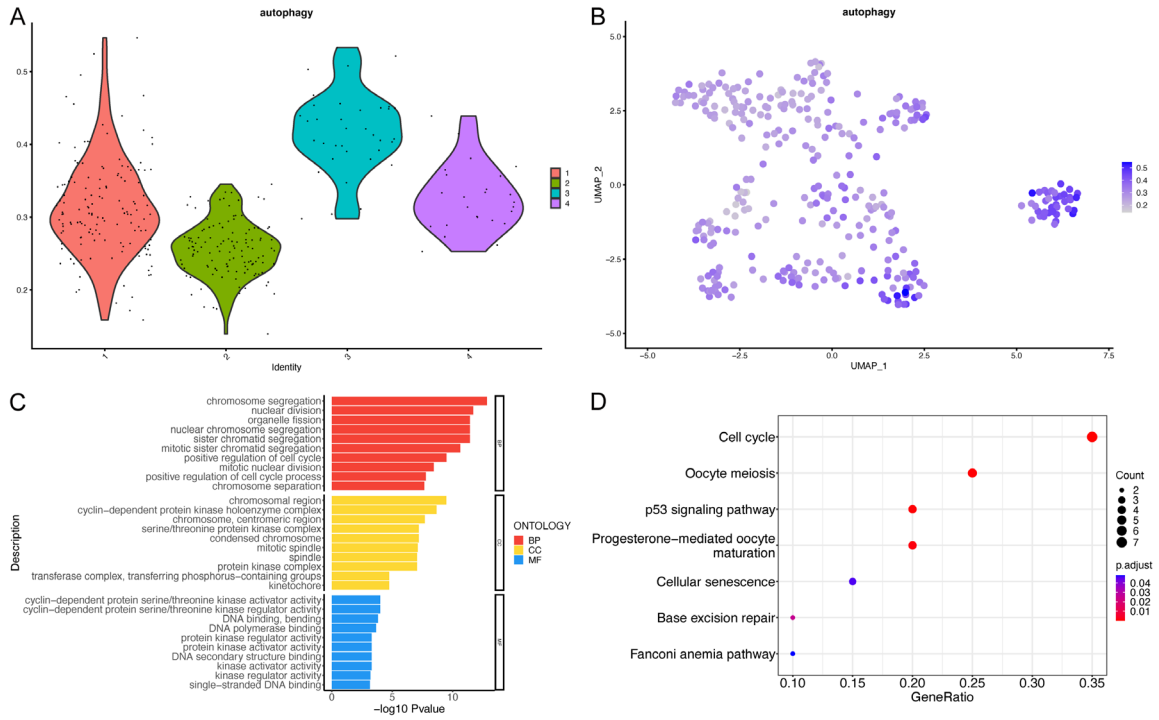

Figure 4. Autophagy score and functional enrichment analysis. A. Autophagy scores across cell clusters. The X-axis represents cell clusters, and the Y-axis shows autophagy scores. Data points are centered within the plot by cluster. B. UMAP plot of autophagy scores. Axes represent the two UMAP dimensions. Cells are color-coded based on autophagy scores, with darker colors indicating higher scores. Cell cluster identities correspond to those in **Figure 2A**. C. Bar chart of GO enrichment results. The X-axis displays the negative $\log_{10}$ of the adjusted *p*-value, and the Y-axis lists enriched Gene Ontology (GO) terms. The top 10 most significant terms from Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) categories are shown (if fewer than 10, all are displayed). D. Bubble plot of KEGG pathway enrichment. The X-axis indicates the proportion of genes enriched in each pathway (enriched genes/total differentially expressed genes), and the Y-axis shows pathway names. Dot size corresponds to the number of enriched genes, while color represents corrected *p*-values - redder colors indicate higher statistical significance.

test revealed a highly significant difference in survival outcomes between these groups ($P < 0.0001$; **Figure 7C**), confirming the strong prognostic value of these genes.

To further explore autophagy-related gene expression patterns, weighted gene co-expression network analysis (WGCNA) was applied to the 7,063 autophagy-related genes, resulting in the identification of five distinct gene modules (**Figure 7D**). Correlation analysis between these modules and the clinical phenotype (tumor vs. normal) revealed two modules with strong positive and negative associations with the disease state (**Figure 7E**). Genes from these significantly correlated modules were selected as candidates for further investigation.

By intersecting the genes identified through differential expression analysis, survival correlation, and WGCNA module membership, a final set of 15 key autophagy-related genes was obtained (**Figure 7F**). These genes included: *DCAF13*, *SLC35A2*, *FREM1*, *SLC7A5*, *CCND2*, *TUBA1C*, *MTHFD2*, *SQLE*, *NT5E*, *IL33*, *FEZ1*, *CDK5R1*, *STX11*, *ADAMTSL1*, and *FBLN5*.

Panorama of key genes

As shown in **Figure 8A**, the expression levels of the 15 key autophagy-related genes differed significantly between normal and tumor samples, with all comparisons yielding $P < 0.0001$ according to the Wilcoxon rank-sum test. Correlation analysis further revealed that most of these genes were strongly positively correlated with one another (**Figure 8B**). A mutation analysis using a waterfall plot (**Figure 8C**) indicated that among the 15 genes, *FREM1* exhibited the highest mutation frequency, followed by *ADAMTSL1*, both showing notably higher mutation rates than the remaining genes.

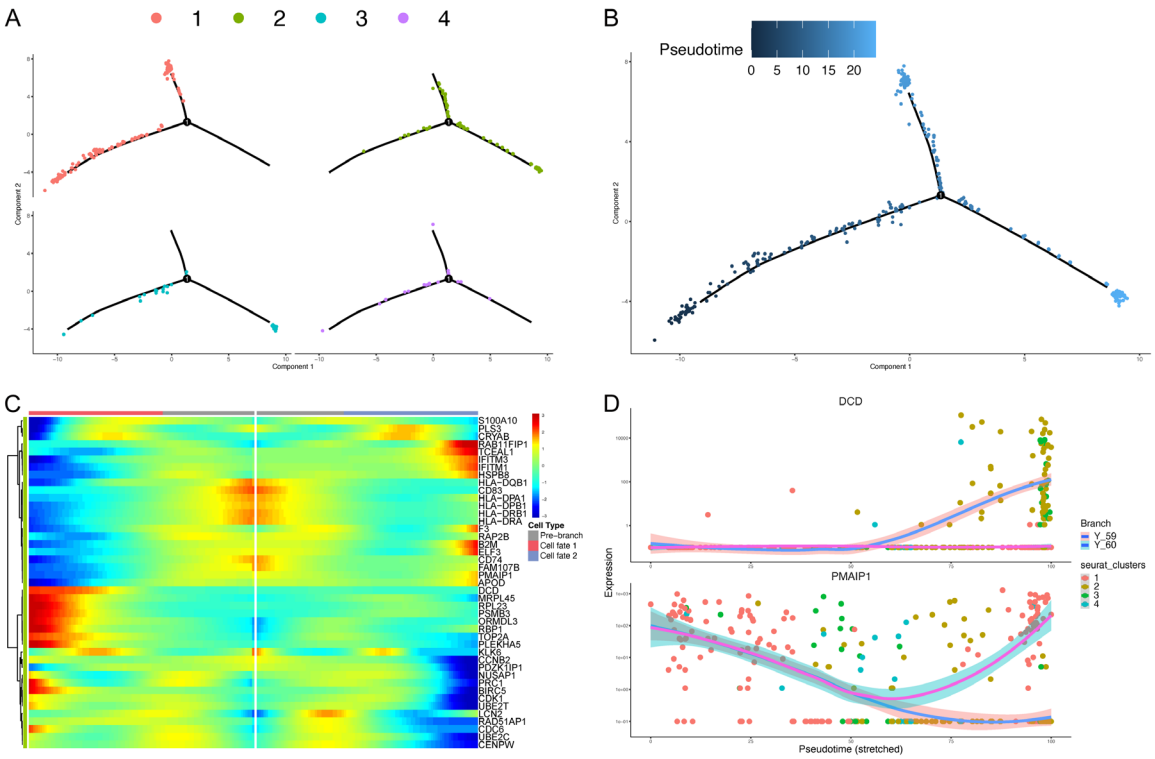

Figure 5. Cell trajectory. (A) Cell trajectory with cell clusters distinguished by different colors, corresponding to previous clustering results. (B) Pseudo-time values of cell loci. Pseudo-time values reflect the developmental trajectory, displayed using gradient colors, where darker shades represent earlier stages and lighter shades indicate later stages. (C) Heat map of gene expression at branch points, with "pre-branch" indicating stages before branching, and "cell fate 1" and "cell fate 2" denoting the two developmental paths following the branch, corresponding to the tracks in (A and B). (D) Gene expression trends at branch sites, where the X-axis represents pseudo-time values, the Y-axis represents expression values, points represent cells, cell clusters are differentiated by colors, and "branch" indicates the point at which gene expression trends diverge, influencing the development of cell trajectory in different directions.

To explore whether these key genes could classify tumors based on autophagy status, unsupervised clustering was performed using their expression profiles across all cancer samples. The optimal number of clusters was first determined (**Figure 8D**), and clustering was conducted using the K-means algorithm. Principal component analysis (PCA) was then used to visualize the clustering outcome, showing a clear separation of samples into two distinct groups (**Figure 8E**).

To validate the biological relevance of the clustering, survival analysis was performed. A log-rank test comparing overall survival between the two groups revealed a statistically significant difference ($P = 0.016$; **Figure 8F**), suggesting that the clustering based on autophagy-related gene expression may reflect underlying differences in autophagy status that are associated with patient prognosis.

Prognostic prediction model based on key genes

To develop a clinically applicable prognostic prediction model, univariate Cox regression and the Least Absolute Shrinkage and Selection Operator (LASSO) regression were applied to identify key prognostic genes. This screening process narrowed the list down to 10 candidate genes (**Figure 9A**, **9B**). Subsequently, multivariate Cox regression analysis was performed to further refine the model, ultimately identifying three key genes: *FEZ1*, *STX11*, and *ADAMTSL1* (**Figure 9C**).

Based on this three-gene model, risk scores were calculated for all samples, and patients were stratified into high- and low-risk groups. Kaplan-Meier survival analysis revealed a highly significant difference in survival outcomes between these two groups ($P = 0.0011$; **Figure**

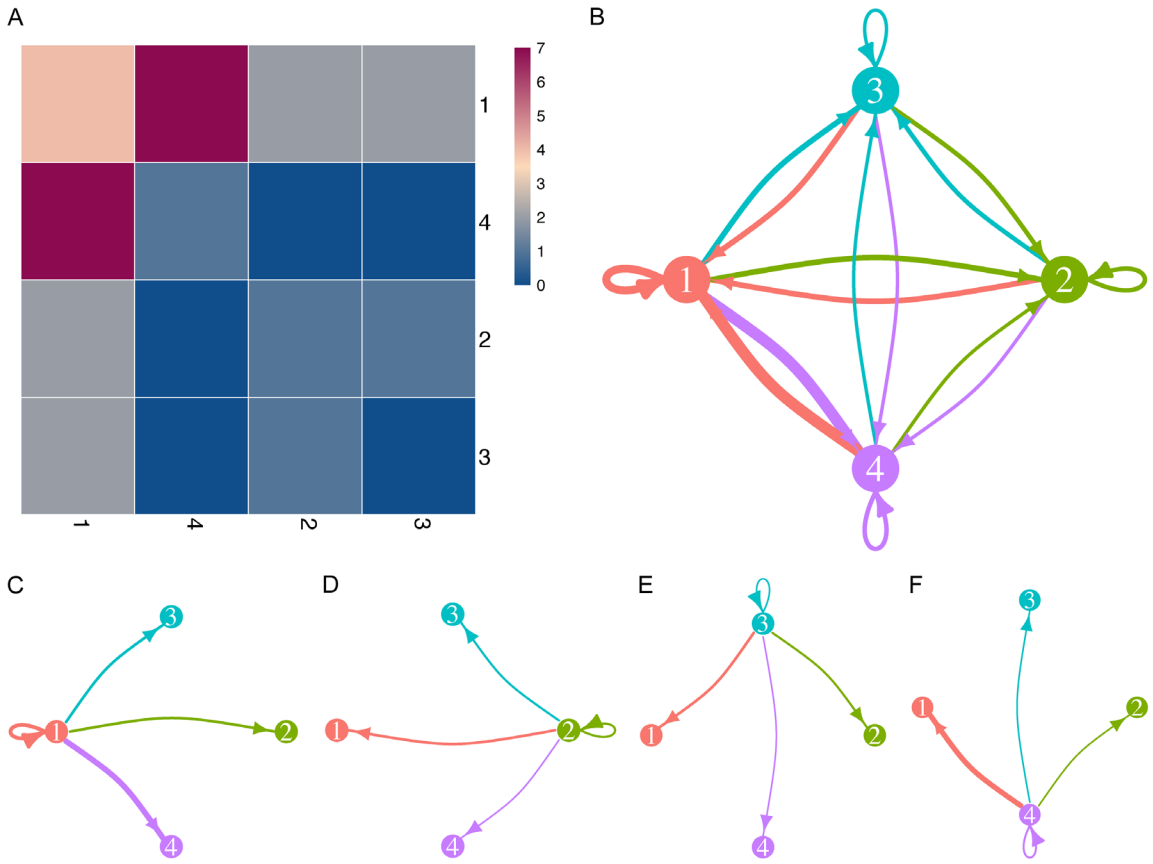

Figure 6. Cell communication. A. Heat map displaying cell communication, with colors representing the number of receptor-ligand pairs between clusters; larger numbers signify stronger interactions between clusters. B. Cell communication network diagram, wherein the edge thickness reflects the strength of interactions between cell clusters, with thicker edges indicating stronger interactions. C-F. Sectional diagrams of cell communication networks centered on each cell cluster.

9D), supporting the prognostic value of the model. Receiver operating characteristic (ROC) analysis for 3-year survival prediction yielded an area under the curve (AUC) of 0.585, indicating moderate predictive performance (**Figure 9E**).

A risk assessment plot (**Figure 9F**) illustrated the distribution of patient survival status across the risk spectrum, with clear distinctions between high- and low-risk groups. Notably, the expression patterns of the three model genes showed consistent trends with the calculated risk score: risk-associated genes were upregulated, while protective genes were downregulated as risk increased, aligning with their respective hazard ratios (HRs) in the model.

To validate the model externally, two datasets from the GEO database were integrated, and batch effects were corrected to ensure consistency across datasets (**Figure 10**). Survival analysis on the external validation cohort showed a significant difference in survival outcomes between the high-risk and low-risk groups, with a log-rank *p*-value of 0.0043 (**Figure 9G**). Furthermore, the area under the ROC curve (AUC) for the model in the external dataset was 0.761, significantly surpassing the threshold of 0.5 (**Figure 9H**), further confirming the model's accuracy and robustness in predicting patient prognosis.

Analysis of biological characteristics of the model

To further investigate the biological differences between the high-risk and low-risk groups, we performed a differential expression analysis of genes. After statistical screening, a total of 1,468 differentially expressed genes (DEGs) were identified, including 454 upregulated genes and 1,014 downregulated genes (**Figure 11A**, **11B**). These DEGs were subjected to Ge-

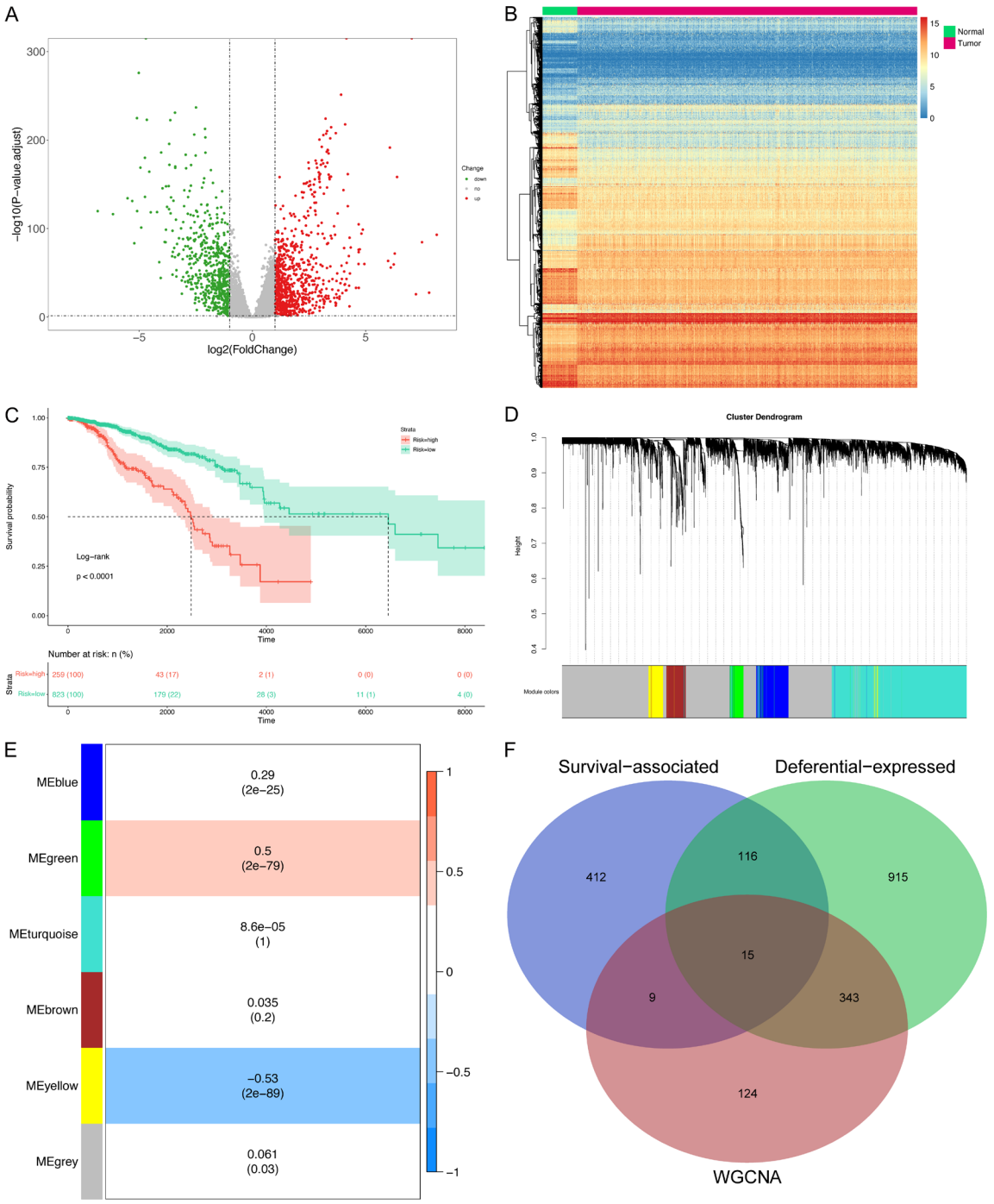

Figure 7. Key genes. A. Volcano map depicting differentially expressed genes, with the X-axis representing log2 (fold change) and the Y-axis representing -log10 (*p* value-adjust). Genes are represented by dots, green denotes down-regulated genes, red indicates up-regulated genes, and gray marks genes with no significant expression Changes. B. Heat map displaying differentially expressed genes. C. Survival curve illustrating the high-low risk group, with the top section presenting the survival curve (X-axis: survival time in days, Y-axis: survival rate), and the lower part containing the risk table (X-axis: survival time, Y-axis: group labels, color-coded to match the survival curve, and table data indicating the number of surviving samples and their percentage in the total sample count in each group). D. WGCNA hierarchical clustering, with the top section showing the hierarchical clustering tree and the bottom section displaying modules corresponding to genes, where each color represents a module, and the gray module is an ineffective module lacking significant co-expression characteristics. E. WGCNA phenotypic association heat map, revealing the association of each gene module with the disease, with correlation values within the figure and statistical significance *P*-values in parentheses. F. Venn diagram displaying the intersection of key genes, identifying a total of 15 genes for further study.

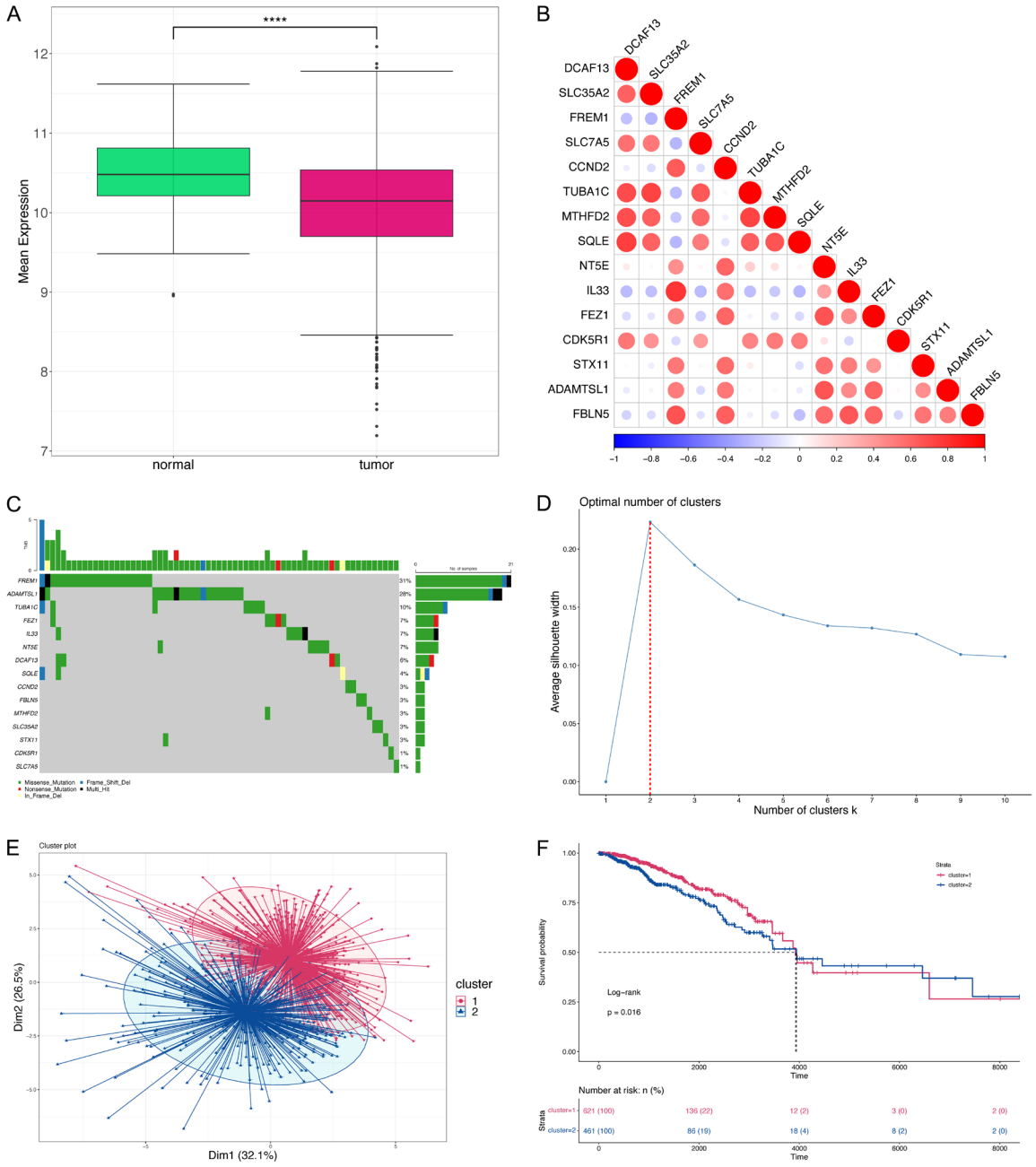

Figure 8. Landscape of key genes. A. The X-axis represents the sample type, and the Y-axis displays the average expression value of key genes. In the box plot, the median value is shown as the central line, the upper and lower quartiles are depicted by the upper and lower frame lines, and any outlier data points are indicated as individual dots. The statistical analysis employed the Wilcoxon rank sum test, with significance levels denoted by symbols (* for less than 0.05, ** for 0.01, *** for 0.001, **** for 0.0001), and no symbol indicates no significant difference. B. Bubble map illustrating the correlation among key genes. Red indicates a positive correlation, while blue represents a negative correlation. C. Waterfall plot displaying key gene mutations. D. Selection of the optimal number of clusters. The X-axis represents the number of clusters, and the Y-axis depicts the average silhouette score. The optimal number of clusters is identified as the value associated with the largest silhouette score, marked by the red dashed line. E. Dimensionality reduction plot of k-means clustering results. The X and Y-axes represent two dimensions, with each point representing a patient sample assigned to one of two groups, color-coded to distinguish between subgroups. F. Survival curves for inter-subgroup analysis.

ne Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. GO analysis revealed significant enrichment in 182 terms, with the top five terms in-

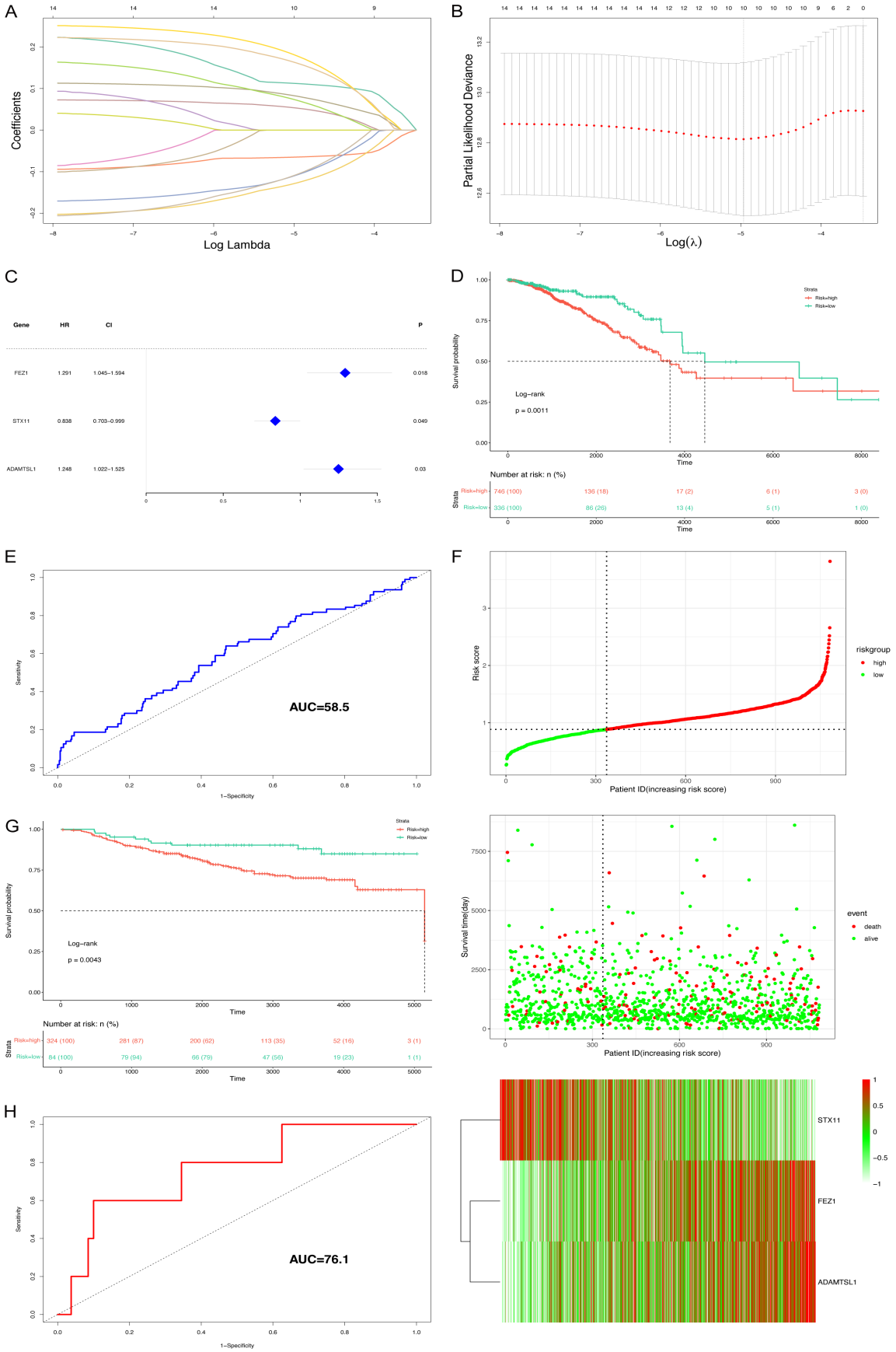
A
Coefficients
Log Lambda
B
Partial Likelihood Deviance
Log(λ)
C
Gene
HR
CI
P
FEZ1
1.291
1.045−1.594
0.018
STX11
0.838
0.703−0.999
0.049
ADAMTSL1
1.248
1.022−1.525
0.03
D
Survival probability
Log−rank
p = 0.0011
Time
Number at risk: n (%)
Risk=high
746 (100)
136 (18)
17 (2)
6 (1)
3 (0)
Risk=low
336 (100)
86 (26)
13 (4)
5 (1)
1 (0)
E
Sensitivity
1−Specificity
AUC=58.5
F
Risk score
Patient ID(increasing risk score)
riskgroup
high
low
G
Survival probability
Log−rank
p = 0.0043
Time
Number at risk: n (%)
Risk=high
324 (100)
281 (87)
200 (62)
113 (35)
52 (16)
3 (1)
Risk=low
84 (100)
79 (94)
66 (79)
47 (56)
19 (23)
1 (1)
Survival time(day)
Patient ID(increasing risk score)
event
death
alive
H
Sensitivity
1−Specificity
AUC=76.1
STX11
FEZ1
ADAMTSL1

Figure 9. Model. A. LASSO regression curve, depicting the convergence screening process of lasso regression for gene features. The X-axis shows the log lambda value, while the Y-axis represents the regression coefficient. Different features are depicted with lines of varying colors. B. Lambda value selection curve, used to determine the best lambda value for the regression model. Typically, the lowest point, indicated by the dotted line in the graph, is chosen as the best lambda value. C. Model forest plot, displaying genes comprising the model, hazard ratios (HR), 95% confidence intervals for HR, HR visualization, and statistical *P*-values. D. Survival curve for high-low risk groups, with the upper section presenting the survival curve (X-axis: survival time in days, Y-axis: survival rate) and the lower section displaying the risk table (X-axis: survival time, Y-axis: group labels). Colors in the table correspond to the survival curve, and table data includes the number of surviving samples and their percentage in the total sample count in each group. E. ROC curve for 3-year survival prediction. F. Risk triad. The upper figure is a scatter plot of risk groups, with the X-axis representing samples and the Y-axis representing the risk score. High and low-risk groups are differentiated by color. The middle figure is a scatter plot of risk outcomes, with the X-axis representing samples and the Y-axis representing survival time. Survival status is distinguished by color. The lower figure is a heat map of model gene expression. G. Validation of survival curves with external datasets. H. Validation of ROC curves with external datasets.

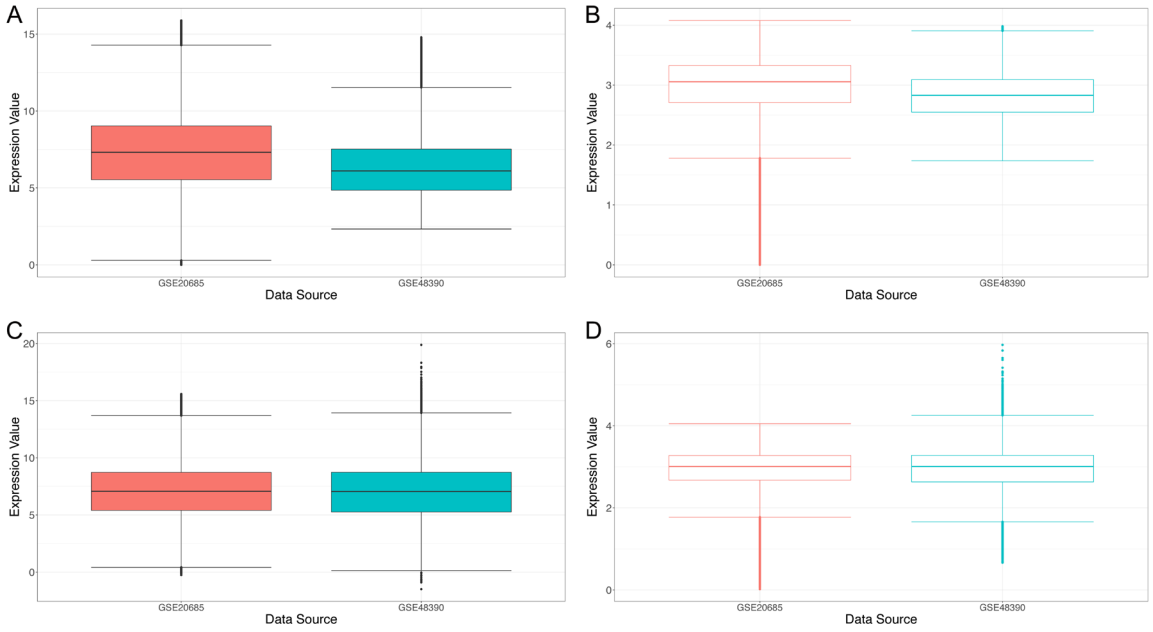

Figure 10. Validation data. The X-axis represents the data set, while the Y-axis represents gene expression values. In the box plots, the central line signifies the median value, the upper frame line represents the upper quartile, and the lower frame line indicates the lower quartile. A. Expression profiles without log standardization and without batch effect correction. B. Expression profiles normalized by log without batch effect correction. C. Expression profiles corrected for batch effect without log standardization. D. Expression profiles corrected for batch effect and standardized by log.

cluding extracellular matrix structural constituent, collagen-containing extracellular matrix, extracellular matrix organization, extracellular structure organization, and external encapsulating structure organization (**Figure 11C**). These processes are closely associated with extracellular matrix activities, which may be linked to cancer cell metastasis.

KEGG pathway analysis identified 15 significantly enriched pathways, with the top five including protein digestion and absorption, cytokine-cytokine receptor interaction, neuroactive ligand-receptor interaction, IL-17 signaling pathway, and ECM-receptor interaction (**Figure 11D**, **11E**). These pathways are predominantly involved in cell signaling, transduction, and protein activities, and may also implicate immune-related processes, particularly through cytokine- and chemokine-related pathways such as the IL-17 signaling pathway.

Further Gene Set Enrichment Analysis (GSEA) focused on autophagy-related pathways revealed that six pathways were significantly enriched among autophagy-related genes, collectively representing over 75% of the enrichment (**Figure 12**). These pathways included allograft

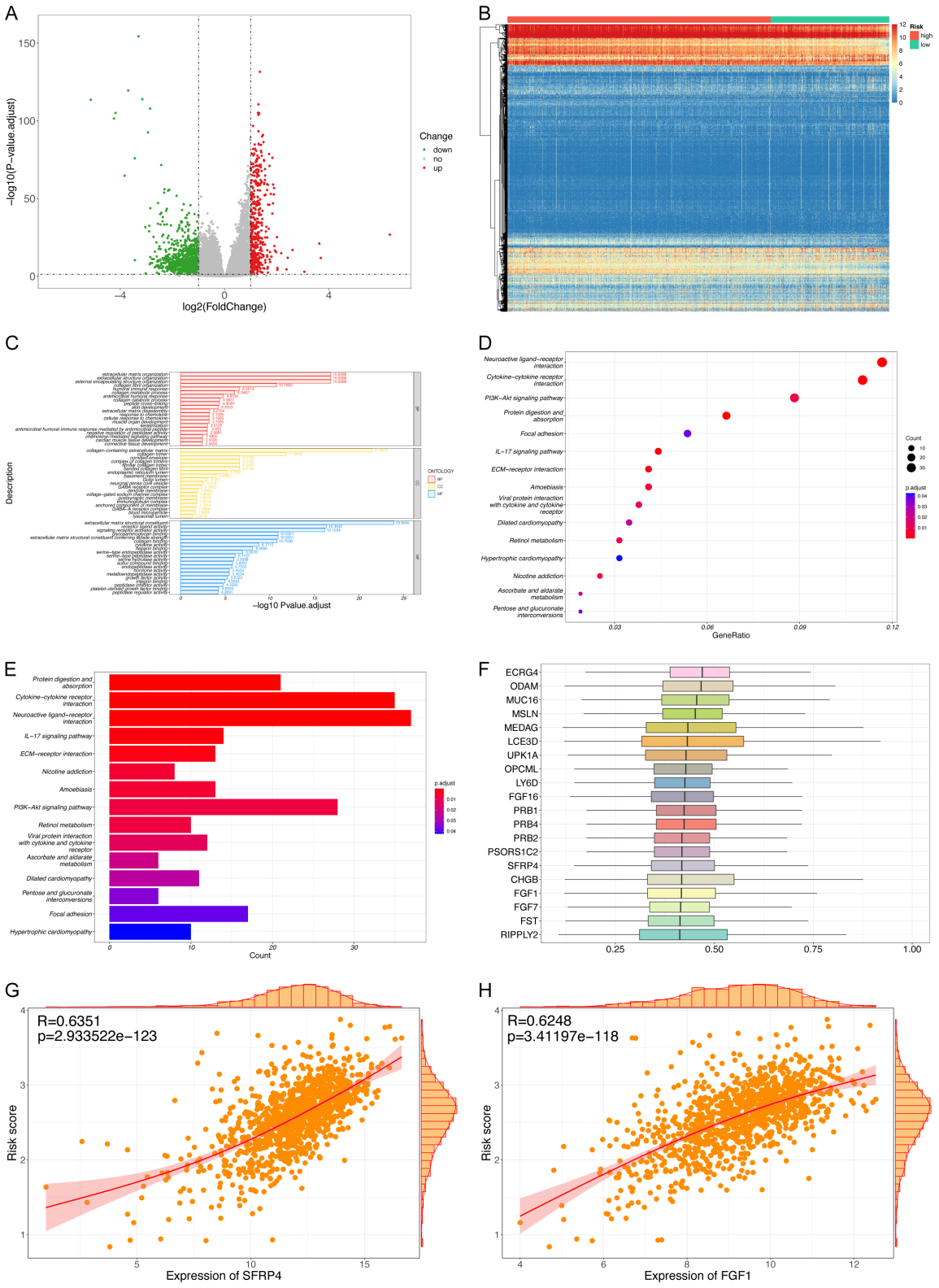

Figure 11. Biological characteristic of the model. (A) Volcano plot depicting differentially expressed genes in high and low-risk groups, with the x-axis representing log2 (fold change) and the y-axis representing -log10 (*p* value-adjust). Genes are represented by dots, with green indicating down-regulated genes, red for up-regulated genes, and gray for genes with no significant expression changes. (B) Heat maps displaying differentially expressed genes. (C) Bar chart of GO enrichment results, with the X-axis representing -log10 (*p* value-adjust) and the Y-axis listing

enriched GO terms. Only the top 20 most significant GO terms for BP, CC, and MF are displayed. (D) Bubble diagram illustrating KEGG enrichment results, with the X-axis representing the gene proportion (total number of genes enriched into a pathway/differentially expressed genes), the Y-axis listing pathway names, and dot size reflecting the number of genes enriched in each pathway. Color indicates the corrected *P*-value, with smaller *P*-values appearing redder, indicating higher significance. (E) Bar chart of KEGG enrichment results, with the X-axis indicating the number of genes and the Y-axis listing pathway names. Color represents the corrected *P*-value, with smaller *P*-values appearing redder, indicating higher significance. (F) Scatter plot of GO semantic similarity between genes and other genes, with the X-axis representing GO semantic similarity and the Y-axis showing the top 20 genes with the highest semantic similarity. (G, H) Scatter plot of the correlation between gene expression and risk score. The X-axis represents gene expression values, the Y-axis indicates the risk score, and the curve is the correlation fitting curve. The shaded area represents the confidence interval, while histograms and density curves are depicted outside the (G and H) correspond to SFRP4 and FGF1 genes, respectively.

rejection, cell cycle, EGFR tyrosine kinase inhibitor resistance, IL-17 signaling pathway, microRNAs in cancer, and proteoglycans in cancer, all of which align with the findings from the GO and KEGG analyses.

To assess the semantic similarity between GO terms, we calculated the semantic correlations among the identified genes. The results (**Figure 11F**) showed that genes such as *ECRG4*, *ODAM*, and *MUC16* exhibited a high degree of correlation with other genes. Additionally, we explored the relationship between these genes and the calculated risk score. Notably, genes like *SFRP4* and *FGF1* showed a highly significant positive correlation with the risk score (**Figure 11G**, **11H**), suggesting their potential role as risk factors.

Protein interaction network

The STRING database was utilized to construct a protein-protein interaction (PPI) network. Initially, the CytoHubba plug-in was employed to identify the top 100 hub genes based on node degrees, facilitating the construction of an interaction network for these hub genes. Visualization of the network was performed using Cytoscape software (**Figure 13A**). To gain a deeper understanding of the functional roles of these hub genes, we further investigated their associated miRNAs using the miRNet database, which provided insights into the genetic background and regulatory networks of the hub genes (**Figure 13B**). The results revealed that these hub genes were associated with both unique miRNAs and shared miRNA interactions, suggesting their involvement in similar regulatory processes and reflecting analogous biological characteristics.

Immune infiltration analysis

To further evaluate the extent of immune infiltration in the high-risk and low-risk groups, we applied the single-sample Gene Set Enrichment Analysis (ssGSEA) method to calculate immune cell scores for 28 immune cell types across all samples. Visualization of the results was achieved through heatmaps and box plots (**Figure 14A**, **14B**). The analysis revealed significant differences in the abundance of most immune cell types between the two groups. Notably, activated CD8 T cells, central memory CD8 T cells, activated CD4 T cells, central memory CD4 T cells, effector memory CD4 T cells, T follicular helper cells, gamma delta T cells, type 1 T helper cells, activated B cells, immature B cells, memory B cells, natural killer cells, CD56bright natural killer cells, activated dendritic cells, plasmacytoid dendritic cells, immature dendritic cells, and mast cells exhibited notable differences. Among these, key cell types such as activated CD8 T cells and activated CD4 T cells showed highly significant differences ($P < 0.0001$). These findings highlight substantial variations in the tumor microenvironment and immune landscape between the high-risk and low-risk groups.

Furthermore, we investigated the correlations between gene expression and immune cell infiltration levels, uncovering significant and strong associations between the *GZMB* gene, the *CXCL* family genes, and specific immune cell populations (**Figure 14C-F**). Notably, *GZMB* showed a strong positive correlation with both activated CD8 T cells and activated CD4 T cells. Similarly, *CXCL9* was positively correlated with activated CD8 T cells, while *CXCL10* exhibited a positive correlation with activated CD4 T cells. These findings suggest that these genes may play crucial roles in immune regulation within the tumor microenvironment.

Mutation characteristics analysis of risk groups

The gene mutation waterfall diagram (**Figure 15A**, **15B**) revealed differences in the frequency of mutated genes between the high-risk and

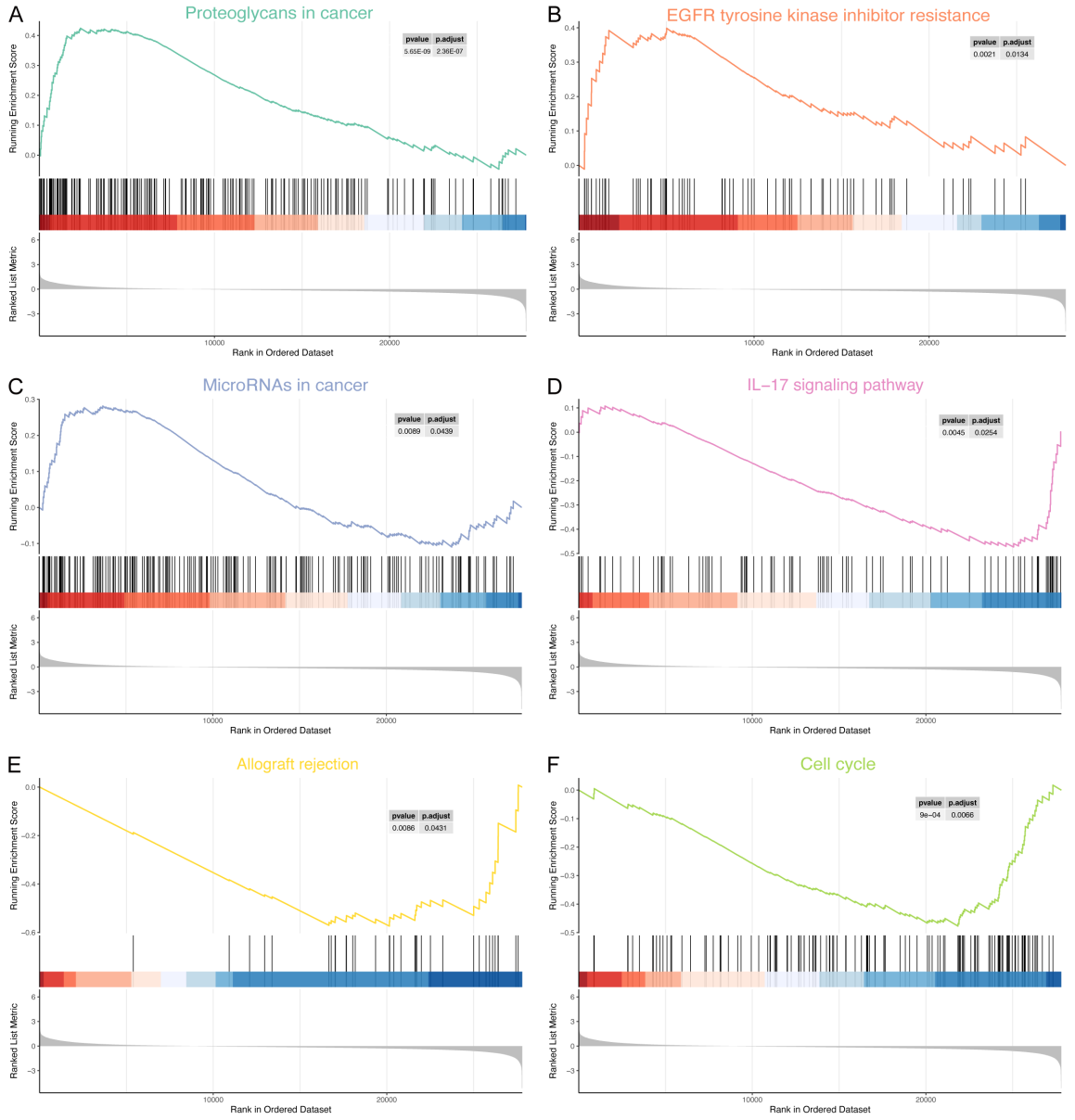

Figure 12. The axis of GSEA enrichment. The X-axis displays the rank of genes in the list of differentially expressed genes, where up-regulated genes are > 0 and down-regulated genes are < 0. The upper Y-axis represents the enrichment fraction, and the lower Y-axis depicts the logFC value. Six biological pathways highly correlated with autophagy are shown: Proteoglycans in cancer, EGFR tyrosine kinase inhibitor resistance, MicroRNAs in cancer, IL-17 signaling pathway, allograft rejection, and cell cycle.

low-risk groups. The high-risk group exhibited a higher frequency of *PIK3CA* mutations, whereas *TP53* mutations were more prevalent in the low-risk group. This discrepancy may reflect distinct genomic mutation profiles between the two groups. Copy number variation (CNV) analysis showed general concordance in the amplification and deletion sites across the two groups, although variations in the extent of amplification or deletion were observed (**Figure 15C**, **15D**).

To further link genomic mutation status to tumor characteristics, we calculated the tumor mutation burden (TMB) for each sample and visualized the results using a box plot (**Figure 15E**). The high-risk group demonstrated significantly higher TMB compared to the low-risk

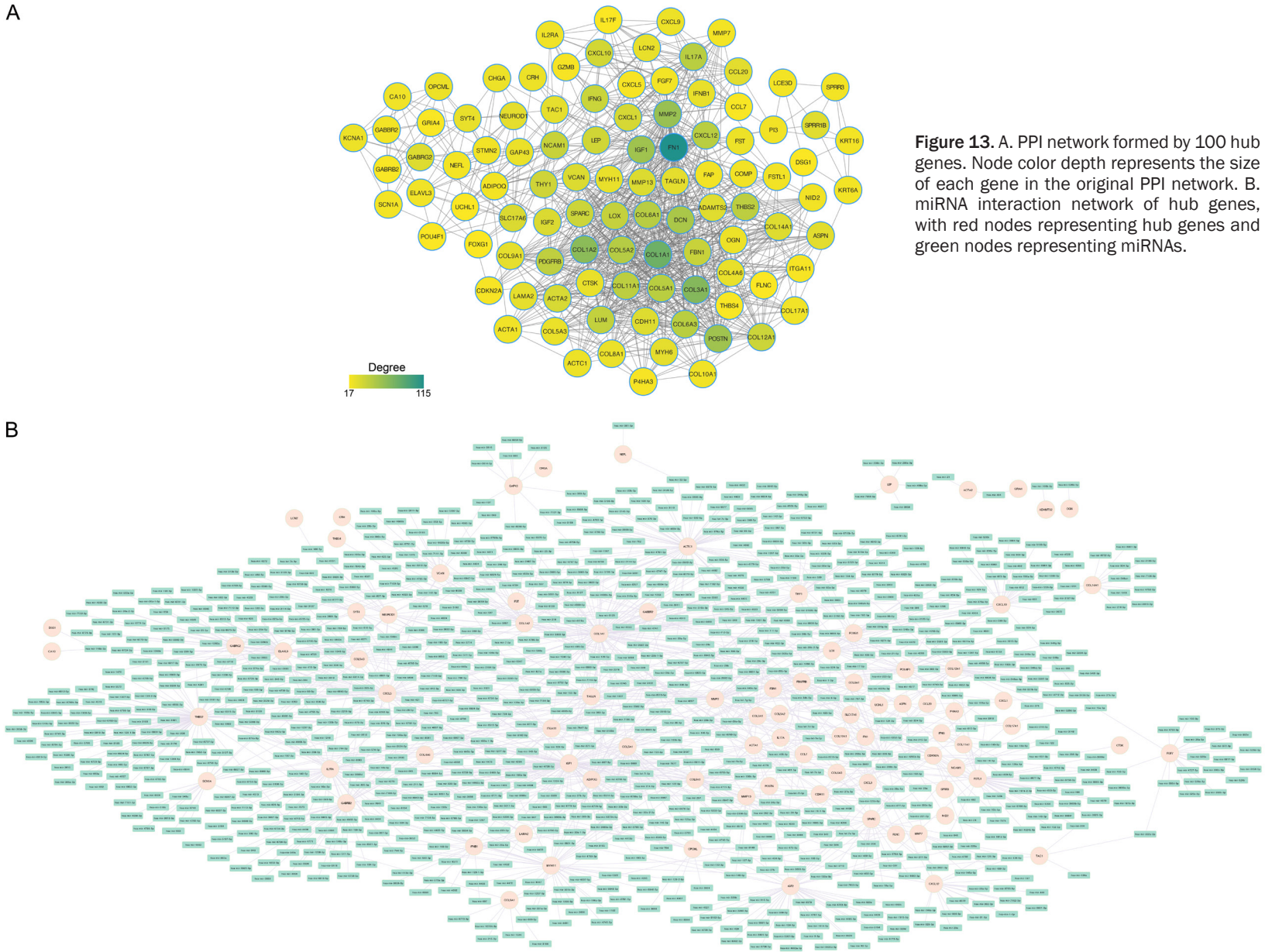

Figure 13. A. PPI network formed by 100 hub genes. Node color depth represents the size of each gene in the original PPI network. B. miRNA interaction network of hub genes, with red nodes representing hub genes and green nodes representing miRNAs.

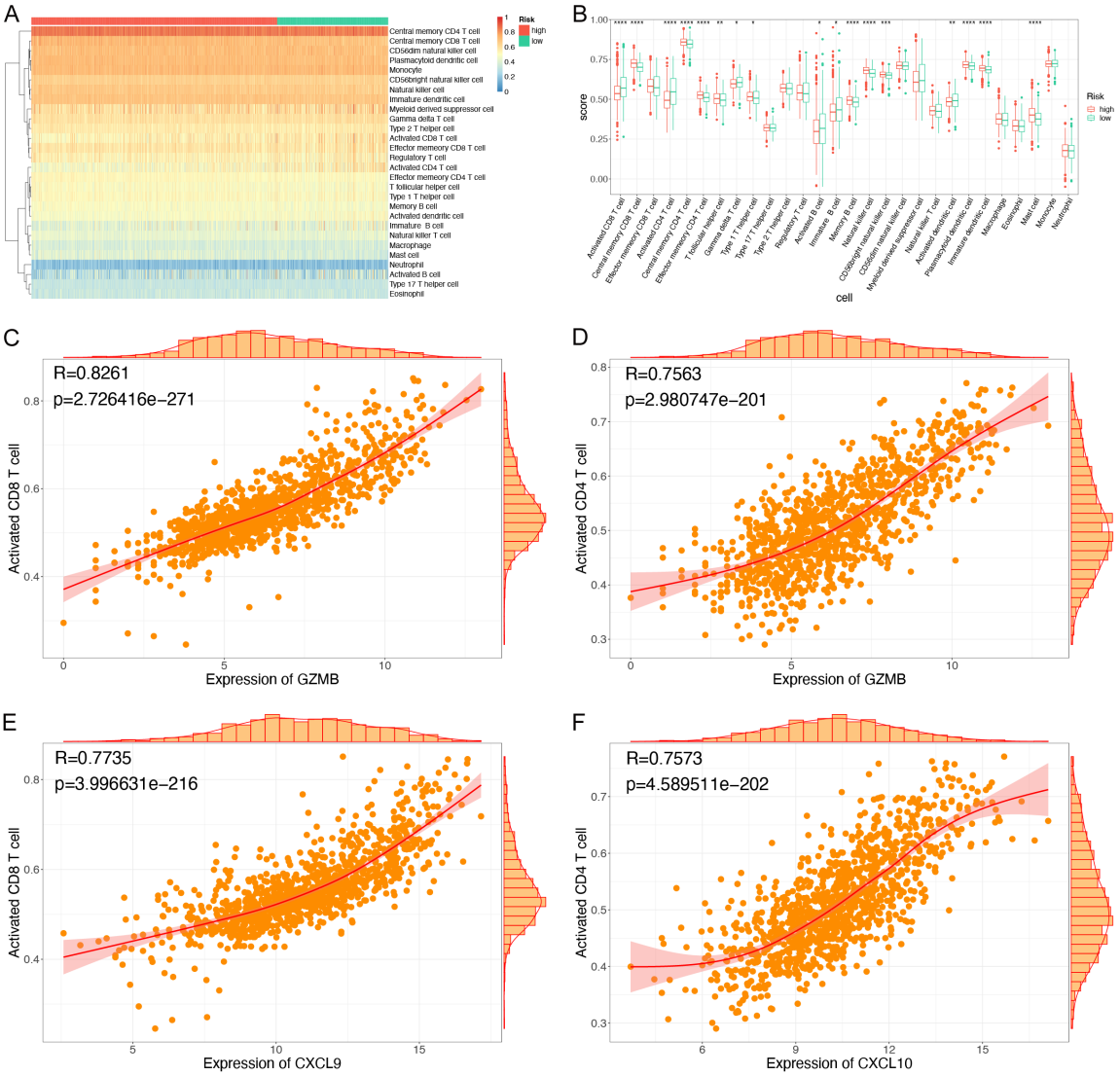

Figure 14. A. Heat map illustrating immune infiltration, with 28 types of immune cells listed as samples. B. Box plot showing immunization scores, with the X-axis representing 28 types of immune cells and the Y-axis indicating the level of immune infiltration. Each color represents a sample group. The statistical test employed the Wilcoxon rank sum test, with significance levels indicated by symbols (* for less than 0.05, ** for 0.01, *** for 0.001, **** for 0.0001), and no symbol denotes no significant difference. C-F. Scatter plots depicting the correlation between gene expression and immune cell infiltration. The X-axis represents gene expression values, while the Y-axis indicates the immune cell infiltration score. The curve represents the correlation fitting curve, and the shaded area represents the confidence interval. Histograms and density curves are displayed outside the figure.

group (P < 0.0001, Wilcoxon rank sum test), indicating a greater tumor burden and poorer prognostic outcomes, consistent with earlier findings. Similarly, microsatellite instability (MSI) analysis (**Figure 15F**) revealed a significantly higher proportion of MSI-high samples in the high-risk group, suggesting increased microsatellite instability in this group, which aligns with previous observations.

Finally, to evaluate potential differences in the response to immunotherapy, we performed TIDE analysis (**Figure 15G**, **15H**). The results showed no significant disparity in immunotherapy response between the high-risk and low-risk groups.

Establishment of a predictive nomogram

To integrate the risk model with other clinical factors, we constructed a nomogram that incorporated the risk score along with additional clinical variables. The results (**Figure 16A**) demonstrated strong alignment between the risk

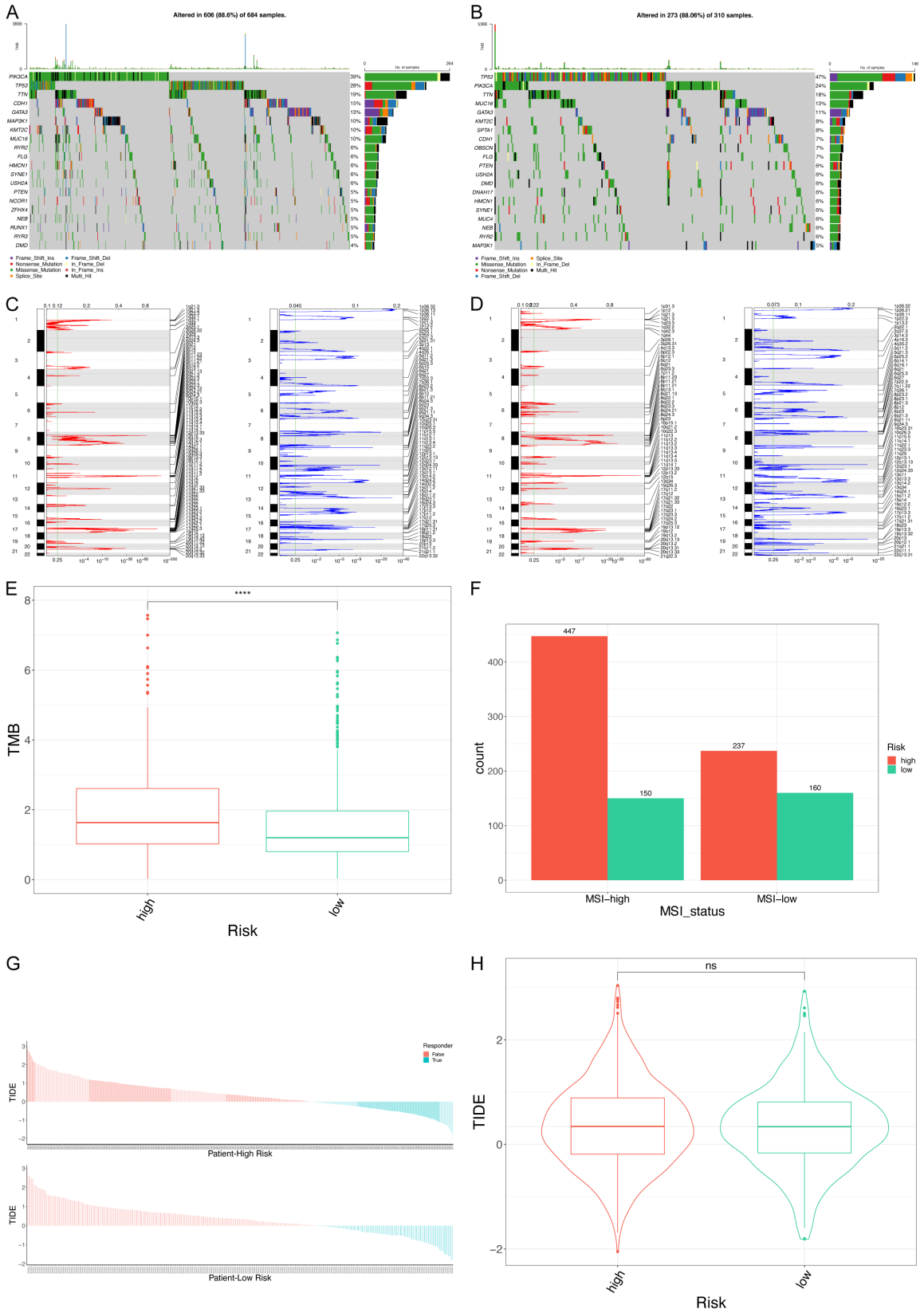

Figure 15. Mutation characteristic of the model. (A, B) Mutation characteristic of the model, (C, D) CNV peak plots, with the X-axis representing the variation level and the Y-axis indicating chromosome position. Red denotes amplification, while blue represents deletion. (C) corresponds to the high-risk group, and (D) represents the low-risk

group. (E) Box plot displaying TMB for the high-low risk groups. (F) Bar chart indicating MSI status for the high-low risk groups. The X-axis represents MSI status, the Y-axis displays the number of samples, and colors differentiate between high and low-risk groups. (G) TIDE scores for high and low-risk groups. The X-axis represents samples, the Y-axis shows the score, with scores > 0 indicating samples that do not respond to immunotherapy, and scores < 0 indicating a response. The high-risk group is displayed above, and the low-risk group is shown below. (H) Violin plot showing TIDE scores. "ns" signifies that the Wilcoxon rank sum test did not yield statistical significance.

score and clinical factors such as patient age, disease stage, and T, N, and M stages, accurately reflecting prognosis.

Next, we validated the nomogram using calibration curves and clinical decision curves. The calibration curve showed a close match to actual outcomes, with a predominantly diagonal distribution, indicating excellent predictive accuracy (**Figure 16B**). The decision curve (**Figure 16C**) revealed a higher net benefit for interventions based on the nomogram compared to simplistic approaches at 1, 3, and 5 years, underscoring the clinical relevance and utility of the model.

Discussion

Through single-cell analysis, we have uncovered the prevalence of autophagy in breast cancer cells, with varying intensities across different cell clusters. In this study, we identified 15 key autophagy-related genes associated with breast cancer survival. Unsupervised clustering based on these genes delineated two distinct groups, which were further used to develop a prognostic model. This model, consisting of three key genes - STX11, FEZ1, and ADAMTSL1 - was validated using external datasets. Additionally, we constructed a novel nomogram that integrates the risk score with clinical parameters, enhancing its clinical applicability.

Intra-tumor heterogeneity (ITH) presents a significant challenge in cancer treatment by promoting genetic variability that can drive tumor progression and the development of drug resistance [20]. Single-cell RNA sequencing (scRNA-seq) is instrumental in uncovering such cellular heterogeneity and revealing novel genetic traits associated with clinical outcomes [21]. Our study utilized scRNA-seq data to identify four distinct cell clusters within breast cancer, each exhibiting unique biological characteristics, highlighting the substantial heterogeneity of the tumor. The fact that each sample contains at least two distinct cell clusters emphasizes the complexity of intra-tumoral heterogeneity. Additionally, we observed fluctuations in autophagy levels across these clusters, with pseudo-temporal ordering identifying 43 autophagy-related genes with diverse expression patterns as cells differentiate. This dynamic analysis underscores the critical role of autophagy in the initiation and progression of breast cancer, as well as its potential involvement in therapeutic responses. Depending on the context, autophagy may either promote tumor cell survival or contribute to drug resistance and tumor progression [22-24].

Our findings indicate that among the three autophagy-related genes, STX11 acts as a protective gene, while FEZ1 and ADAMTSL1 are associated with increased risk. STX11, encoding Syntaxin 11, is enriched in immune cells such as natural killer cells, cytotoxic T cells, and monocytes/macrophages [25]. Silencing STX11 enhances phagocytosis of apoptotic cells and TNFα secretion, suggesting its antitumoral effect [26]. As a tumor suppressor gene, STX11 has been implicated in peripheral T-cell lymphomas [27]. FEZ1, a negative regulator of autophagy, forms a complex with the short coiled-coil protein (SCOC) and is involved in autophagy regulation. Knockdown of FEZ1 increases autophagic activity, promoting tumor progression [28, 29]. ADAMTSL1, an extracellular matrix component, is involved in cell-cell or cell-matrix interactions and exhibits significant differential methylation in breast cancer subtypes. It has shown high predictive value as a cancer biomarker [30], and polymorphisms in ADAMTSL1 have been linked to disease-free survival in breast cancer [31].

Despite the promising results, our study has some limitations. First, further evaluation of the prognostic effects of the risk model and nomogram in various molecular subtypes of breast cancer is needed. Second, due to the retrospective design of this study, certain clinical data, such as the specific chemoradiotherapy regimens used, were unavailable in the TCGA database, which could influence survival analy-

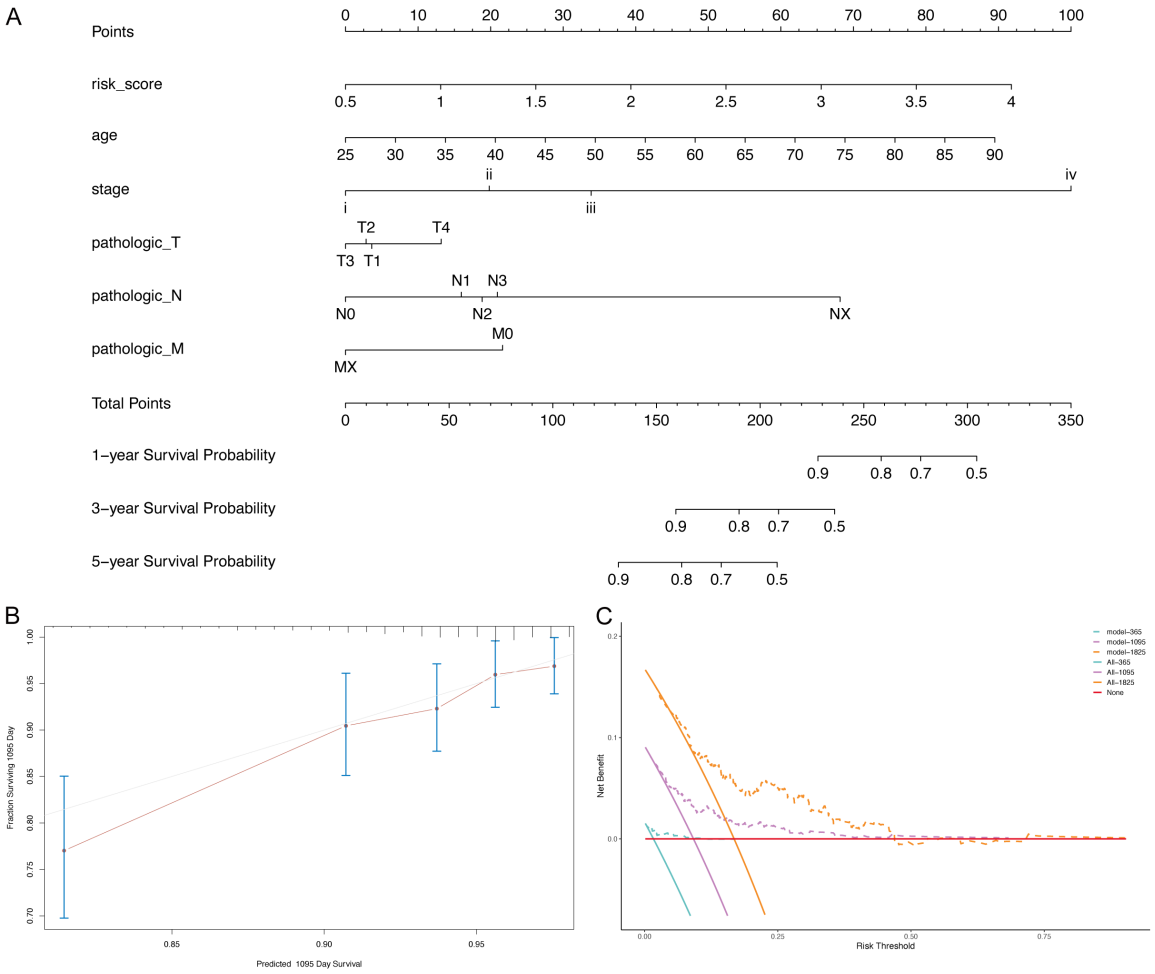

Figure 16. Nomogram and validation. A. Nomogram, displaying the indicator on the left side and the scoring scale on the right side. B. Nomogram calibration curve, with the X-axis representing the predicted probability and the Y-axis indicating the actual probability. The gray diagonal line serves as the ideal state reference. C. Clinical decision curve, with the X-axis representing the risk threshold, the Y-axis displaying the net benefit rate, and the "All" curve signifying that all samples received intervention without differences. The model curve (dashed line) represents intervention based on the prediction model, with time values indicated as 1, 3, and 5 years, respectively.

sis. Lastly, additional validation through functional experiments is necessary to confirm these findings.

In conclusion, our study reveals that autophagy levels vary across breast cancer cell clusters, suggesting that autophagy influences tumor cell differentiation. We identified a novel autophagy-related prognostic risk model comprising three key genes, providing a valuable tool for predicting patient prognosis and guiding clinical decision-making.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Yong Li, Department of Breast, Jiangmen Central Hospital, No. 23, Haibang St., Beijie, Jiangmen 529030, Guangdong, P. R. China. Tel: +86-750-3989522; E-mail: docleo-1985@sina.com

References

[1] Harbeck N and Gnant M. Breast cancer. Lancet 2017; 389: 1134-1150.

[2] Siegel RL, Miller KD, Fuchs HE and Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022; 72: 7-33.

[3] Weigelt B, Hu Z, He X, Livasy C, Carey LA, Ewend MG, Glas AM, Perou CM and Van't Veer LJ. Molecular portraits and 70-gene prognosis signature are preserved throughout the meta-

static process of breast cancer. Cancer Res 2005; 65: 9155-9158.
[4] Sparano JA and Paik S. Development of the 21-gene assay and its application in clinical practice and clinical trials. J Clin Oncol 2008; 26: 721-728.
[5] Ahmadi-Dehlaghi F, Mohammadi P, Valipour E, Pournaghi P, Kiani S and Mansouri K. Autophagy: a challengeable paradox in cancer treatment. Cancer Med 2023; 12: 11542-11569.
[6] Verma AK, Bharti PS, Rafat S, Bhatt D, Goyal Y, Pandey KK, Ranjan S, Almatroodi SA, Alsahli MA, Rahmani AH, Almatroudi A and Dev K. Autophagy paradox of cancer: role, regulation, and duality. Oxid Med Cell Longev 2021; 2021: 8832541.
[7] Akar U, Chaves-Reyez A, Barria M, Tari A, Sanguino A, Kondo Y, Kondo S, Arun B, Lopez-Berestein G and Ozpolat B. Silencing of Bcl-2 expression by small interfering RNA induces autophagic cell death in MCF-7 breast cancer cells. Autophagy 2008; 4: 669-679.
[8] Qadir MA, Kwok B, Dragowska WH, To KH, Le D, Bally MB and Gorski SM. Macroautophagy inhibition sensitizes tamoxifen-resistant breast cancer cells and enhances mitochondrial depolarization. Breast Cancer Res Treat 2008; 112: 389-403.
[9] Sun WL, Chen J, Wang YP and Zheng H. Autophagy protects breast cancer cells from epirubicin-induced apoptosis and facilitates epirubicin-resistance development. Autophagy 2011; 7: 1035-1044.
[10] Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, Ryu HS, Kim S, Lee JE, Park YH, Kan Z, Han W and Park WY. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun 2017; 8: 15081.
[11] Tomczak K, Czerwińska P and Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) 2015; 19: A68-77.
[12] Kao KJ, Chang KM, Hsu HC and Huang AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer 2011; 11: 143.
[13] Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A and Lancet D. GeneCards Version 3: the human gene integrator. Database (Oxford) 2010; 2010: baq020.
[14] Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P and Satija R. Integrated analysis of multimodal single-cell data. Cell 2021; 184: 3573-3587, e3529.
[15] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS and Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014; 32: 381-386.
[16] Efremova M, Vento-Tormo M, Teichmann SA and Vento-Tormo R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc 2020; 15: 1484-1506.
[17] Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014; 15: 550.
[18] Langfelder P and Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.
[19] Zhang B and Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005; 4: Article17.
[20] McGranahan N and Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell 2017; 168: 613-628.
[21] Potter SS. Single-cell RNA sequencing for the study of development, physiology and disease. Nat Rev Nephrol 2018; 14: 479-492.
[22] Yang ZJ, Chee CE, Huang S and Sinicrope FA. The role of autophagy in cancer: therapeutic implications. Mol Cancer Ther 2011; 10: 1533-1541.
[23] Chen HY and White E. Role of autophagy in cancer prevention. Cancer Prev Res (Phila) 2011; 4: 973-983.
[24] Kondo Y, Kanzawa T, Sawaya R and Kondo S. The role of autophagy in cancer development and response to therapy. Nat Rev Cancer 2005; 5: 726-734.
[25] Sutton RB, Fasshauer D, Jahn R and Brunger AT. Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. Nature 1998; 395: 347-353.
[26] Zhang S, Ma D, Wang X, Celkan T, Nordenskjöld M, Henter JI, Fadeel B and Zheng C. Syntaxin-11 is expressed in primary human monocytes/macrophages and acts as a negative regulator of macrophage engulfment of apoptotic cells and IgG-opsonized target cells. Br J Haematol 2008; 142: 469-479.
[27] Yoshida N, Tsuzuki S, Karube K, Takahara T, Suguro M, Miyoshi H, Nishikori M, Shimoyama M, Tsukasaki K, Ohshima K and Seto M. STX11 functions as a novel tumor suppressor gene in peripheral T-cell lymphomas. Cancer Sci 2015; 106: 1455-1462.

[28] Behrens C, Binotti B, Schmidt C, Robinson CV, Chua JJ and Kühnel K. Crystal structure of the human short coiled coil protein and insights into SCOC-FEZ1 complex formation. PLoS One 2013; 8: e76355.
[29] McKnight NC, Jefferies HB, Alemu EA, Saunders RE, Howell M, Johansen T and Tooze SA. Genome-wide siRNA screen reveals amino acid starvation-induced autophagy requires SCOC and WAC. EMBO J 2012; 31: 1931-1946.
[30] Li Z, Guo X, Wu Y, Li S, Yan J, Peng L, Xiao Z, Wang S, Deng Z, Dai L, Yi W, Xia K, Tang L and Wang J. Methylation profiling of 48 candidate genes in tumor and matched normal tissues from breast cancer patients. Breast Cancer Res Treat 2015; 149: 767-779.
[31] Kadalayil L, Khan S, Nevanlinna H, Fasching PA, Couch FJ, Hopper JL, Liu J, Maishman T, Durcan L, Gerty S, Blomqvist C, Rack B, Janni W, Collins A, Eccles D and Tapper W. Germline variation in ADAMTSL1 is associated with prognosis following breast cancer treatment in young women. Nat Commun 2017; 8: 1632.