Bioinformatics Analysis of Pivotal Module and Biomarkers Related to the Prognosis of Breast Cancer Based on Single-cell Transcriptome Data

Rui Liu

The Third Affiliated Hospital of Kunming Medical University: Yunnan Cancer Hospital

Xin Yang

Yunnan Provincial People's Hospital: First People's Hospital of Yunnan

Yuhang Quan

The Third Affiliated Hospital of Kunming Medical University: Yunnan Cancer Hospital

Yiyin Tang

The Third Affiliated Hospital of Kunming Medical University: Yunnan Cancer Hospital

Yafang Lai

Guiyang Railway Health Supervision Institute of China Railway Chengdu Bureau Group Limited Company

Chunxiang Li

The Third Affiliated Hospital of Kunming Medical University: Yunnan Cancer Hospital

Dechun Yang

The Third Affiliated Hospital of Kunming Medical University: Yunnan Cancer Hospital

Maohua Wang

The Third Affiliated Hospital of Kunming Medical University: Yunnan Cancer Hospital

Anhao Wu (✉ 529691000@qq.com)

The Third Affiliated Hospital of Kunming Medical University: Yunnan Cancer Hospital

Research Article

Keywords: breast cancer, tumor heterogeneity, single cell sequencing, disease-related modules, prognostic markers

Posted Date: April 20th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-424509/v1>

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

[Read Full License](#)

Abstract

Background: The effect of breast cancer heterogeneity on prognosis of patients is still unclear, especially the role of immune cells in prognosis of breast cancer. Therefore, the discovery of new markers to assess the effect of breast cancer heterogeneity on patient prognosis is crucial to improve prognosis and survival of patients.

Methods: Single cell transcriptome sequencing data of breast cancer were downloaded from GEO database. PCA and UMAP were used for dimensionality reduction analysis and cell clustering. Find All Markers function was used to calculate differential genes in each cluster, and Do Heatmap function was used to plot the distribution of differential genes in each clusters. CellPhoneDB was used to analyze ligand-receptor interactions. TRRUST database combined with Cytoscape were used to construct a receptor-ligand-transcription factor interaction network. WGCNA is used to analyze pivotal modules associated with breast cancer prognosis. Univariate regression analysis and KM survival analysis were used to identify prognostic genes in prognostic modules. Multivariate regression analysis combined with risk scoring were used to construct a breast prognosis model, which was verified by TCGA and ICGC sample data.

Result: In this study, 14 cell clusters were identified in two single-cell datasets (GSE75688 and G118389). The results of ligand receptor interaction network revealed that macrophages and DC cells were the most frequently interacting cells with other cells in breast cancer. The results of WGCNA analysis suggested that the MEblue module is most relevant to the overall survival time of triple-negative breast cancer. Twenty-four prognostic genes in the blue module were identified by univariate Cox regression analysis and KM survival analysis. Multivariate regression analysis combined with risk analysis was used to analyze 24 prognostic genes to construct a prognostic model. The verification result of our prognostic model showed that there were significant differences in the expression of PCDH12, SLIT3, ACVRL1, and DLL4 genes between the high-risk group and the low-risk group.

Conclusion: PCDH12, SLIT3, ACVRL1 and DLL4 are prognostic biomarkers and relate to the type and proportion of immune cells in breast cancer.

Introduction

Breast cancer ranks first in the incidence of female cancer, and shows an increasing trend year by year[1]. Globally, about 2.1 million cases of breast cancer were newly diagnosed in women in 2018, accounting for nearly a quarter of cancer cases in women [2]. Due to advances in early diagnosis and comprehensive treatment strategies, the prognosis of breast cancer patients has been greatly improved in recent years, but about 30% of breast cancer patients still develop metastases after diagnosis and treatment. The 5-year overall survival rate for patients with non-metastatic breast cancer was greater than 80%; however, the survival rate for patients with metastatic breast cancer was less than 30%[3, 4].

Although the expression of estrogen receptor (ER), progesterone receptor (PR) and ERBB2 receptor (HER2) laid the foundation for the classification of breast cancer, and breast cancer is divided into at least five molecular subtypes (ie Luminal A, Luminal B, Her2-enriched, Basal-like and Normal-like) based on gene expression. However, as research progresses, the genomic/transcriptome level of breast cancer typing continues to increase[5]. These studies confirm the heterogeneity of breast cancer. More importantly, heterogeneity is also reported to be one of the leading causes of breast cancer treatment failure, recurrence, and patient death[6]. Tumor heterogeneity not only leads to differences in survival and prognosis of different patients, but also causes different biologic characteristics of cancer cells and different responses to chemotherapy drugs [7].

Although the heterogeneity of breast cancer has been found and confirmed, the implementation of individualized treatment of breast cancer also needs to take into account the existence of different molecular subtypes of breast cancer and different cell subsets within the same tumor tissue[8]. However, the biological relationships between different clonal subsets and between clones and microenvironment in breast cancer tissues are still unclear. Traditional gene sequencing methods can only detect population cells and cannot reflect genetic characteristics at the single-cell level. Single-cell sequencing technology is helpful to study tumor heterogeneity from the differences at the single-cell level and facilitate the comparison of the differences between different subtypes of the same tumor[9]. In this study, single-cell sequencing data were used to identify the inter-tumor and intra-tumoral heterogeneity of breast cancer samples, and a multi-factor interaction network of receptor-ligand-transcription factors in breast cancer was constructed by identifying the characteristic genes of immune cell subtypes and combining with known immune cell marker genes. WGCNA was used to identify prognostic signature, and a prognostic model was constructed for evaluation and verification.

Materials And Methods

1.1 Data sources and processing

Two sets of data, GSE75688 and GSE118389, were downloaded from GEO, among which GSE75688 contained single cell sequencing data (sRNA-seq) of primary breast cancer and metastatic breast cancer, and GSE118389 was sRNA-seq data of triple negative breast cancer. The Create Seurat Object function is used to process the Seurat object. After the two sets of data were analyzed by PCA, the Find Integration Anchors function is used to integrate the two sets of data in the S4 object. The data was finally divided into three groups: primary breast cancer, metastatic breast cancer, and triple negative breast cancer.

1.2 PCA dimension reduction, cell clustering and annotation

The Seurat package in R is used to preprocess the integrated data. After PCA dimension reduction, JackstrawPlot and ElbowPlot are used to show the overall situation of the data. According to experience and debugging, 0.2 is selected as the threshold value for cell clustering, and 14 cell clusters are obtained. Subsequently, marker genes in Cell marker and Panglao DB databases and genes reported in the literature were used to annotate the cell clusters[10, 11]. The Find All Markers function in the Seurat package was

used for differential analysis of single-cell data. According to the expression levels of top5 genes and marker genes of immune cell reported in the literature, dotplot and violin plots were used to display marker genes in each cluster. In addition, dotplot were used to show the proportion of immune cells in different groups according to the frequencies of individual cells in the Primary BC group, Metastatic BC group and TNBC group.

1.3 Construction of ligand-receptor network and joint analysis of transcription factors

In this section, the cellphoneDB software (<https://github.com/Teichlab/cellphonedb>) is adopted. After downloading the software to build a good environment, according to the code cellphonedb method `statistical_analysis meta.txt counts.txt`, the interaction between ligands and receptors in each group of data is analyzed[12]. Among them, the meta.txt file is the barcode and corresponding annotated cells; counts.txt is the barcode and gene expression matrix. In the results, P values of cell types and enriched interaction ligands and receptors were shown. With $P < 0.05$ as the threshold, cellphonedb plot `heatmap_plot meta.txt p values`. In order to further analyze the transcription factor regulation of ligands and receptors, the database TRRUST (<https://www.grnpedia.org/trrust/>) was applied, and the hypergeometric test method are used to trace the transcription factors of target ligand and receptor genes, and Cytoscape 3.7.2 is used in the visual display of results.

1.4 Weighted gene co-expression network analysis

WGCNA is a systems biology approach to construct scale-free networks using gene expression data. WGCNA package of R was used to construct a weighted co-expression network based on the expression profile data of the multifactorial network genes[13]. The prognostic modules were screened according to the characteristics of overall survival time and overall survival status. GO enrichment analysis was performed on each module by Metascope to analyze the biological functions of each module. On the other hand, the website KOBAS is used to perform KEGG Pathway enrichment analysis of co-expressed genes in modules related to the prognosis of breast cancer.

1.5 Univariate regression analysis, KM-survival prognostic analysis and multivariate regression analysis

Univariate Cox regression analysis was used to screen for genes significantly expressed in key modules. Then, the genes with significant differences in the univariate regression analysis were analyzed by Lasso dimensionality reduction, and the output results were used as candidate genes. On the other hand, K-M survival analysis was used to identify prognostic related genes, and $P < 0.05$ was used as a threshold to screen genes with significant prognostic effects. The candidate genes and prognostic genes were intersected and visualized by Venn diagram. In this study, 24 genes associated with prognosis of breast cancer were identified by univariate Cox regression analysis. Forest maps were used to show the results of univariate cox regression for 24 genes. The survminer and survival packages are used to perform multivariate regression analysis. Age, lymphatic node metastatic status (N0 vs NN (excluding NX)), T stage, radiotherapy, race, breast cancer stage, and overall survival were combined with 24 genes for

multivariate regression analysis to identify pivotal genes that were significantly associated with breast cancer prognosis.

The breast cancer samples were divided into high-risk and low-risk groups based on the median expression value of the screened key genes associated with breast cancer progression. Time-dependent ROC results showed that the AUC of the combined Signature 3-year model group was 0.801 for the analysis of immune infiltration levels in the low-risk group. The time-dependent ROC is used to reflect the accuracy and precision of the prediction model. It is generally believed that the model with $AUC \geq 0.7$ can be used to predict the prognostic outcome at a specific time.

Results

2.1 quality control, dimensionality reduction, cell clustering and annotation of breast cancer single cell sequencing data

In this study, two sets of data (GSE75688 and GSE118389) were downloaded. GSE75688 contains single-cell RNA sequencing data of primary and metastatic breast cancer, a total of 563 cells. Among them, 12 samples were applied bulk sequencing. Except for non-single cell sequencing data, 549 single cell data were retained. Among the 549 data, 441 cells are primary breast cancer and 108 cells are metastatic breast cancer. GSE118389 is the scRNA sequencing data of triple-negative breast cancer, which contains 1534 cells. In the quality control of single-cell transcriptome data, the number of feature genes greater than 200 and less than 2500, the proportion of mitochondrial (percent.mt) less than 10% are used as a threshold for data screening and filtering. The first threshold is set to eliminate the empty oil droplets. To avoid the small number of RNA, data less than 200 is eliminated. The second threshold is set to eliminate more than two cells into one oil droplet. Subsequently, the data is analyzed by PCA and UMAP dimensionality reduction, and the results are visualized in the form of heat map, JackstrawPlot, and ElbowPlot (Fig. 1).

Next, the Find Integration Anchors function is used to integrate two single-cell transcriptome sequencing data from two data sets (non-merge, because merge is only A data merge, which cannot remove batch effect), which minimizes the error caused by different batches of experiments, and is used to construct the final S4 object. Subsequently, the Scale Data was used for data centralization and standardization, Seurat was continued to be used for PCA and UMAP dimensionality reduction analysis. According to references and debugging effects, the clustering analysis were performed with a threshold of resolution = 0.2, and 14 clusters were obtained. The cell clusters were annotated according to the marker genes in the Cellmarker, PanglaoDB database and references. The results showed that 14 clusters were endothelial cells, DCs, basal cells, acinar cells, T cells, NK cells, B cells, macrophage cells, Fibroblast cells, neutrophils, epithelial cells, neurons cells, HPCs, ductal cells (Fig. 2A). The Find Variable Features function is used to find the genes with the largest differences among the difference cells clusters. The results show that HP, KCNJ, SCGB2, CPB1 and SPP are the first five significantly different genes (Fig. 2B). Meanwhile,

expression of common cell marker CD3D□MS4A1□AIF1□LUM□S100A8□KRT14□TFF3□CD34□CLDN1 in 14 clusters was analyzed, Violin map is used to show the results of marker genes in each cluster (Fig. 2C).

2.2 differential gene expression and multi factor interaction analysis of immune cells in breast cancer

In order to analyze the differentially expressed genes in each cluster, the Find All Markers function is used to calculate the expression of differential genes in each cluster, and the Do Heatmap function is used to plot the distribution of differential genes in different cell types (Fig. 3A). According to the conventional genes of cell annotation in the literature, the expression of marker genes of 6 types of immune cells including NK, DCs, macrophages, B cells, T cells and neutrophils was analyzed (Fig. 3B). Subsequently, the frequency of cells in each cluster is counted and used to explore the enrichment ratio of immune cell populations. The results show that NK cells are significantly aggregated in triple-negative breast cancer, the proportion of macrophages is significantly increased in primary breast cancer, and B cells, T cells and neutrophils may play important role in metastatic breast cancer (Fig. 3C-D). Excitingly, it has been reported that T cells and neutrophils are involved in metastasis of breast cancer[14]. Next, the MSigDB database is used to perform functional annotation analysis of cell types, which is conducive to revealing the functional status of immune cells. The analysis results showed that the functions of specifically expressed genes in T cells, Macrophage cells, B cells, DC cells, and Neutrophils were significantly enriched in 20, 10, 7, 6, 3, and 3 terms, respectively (Fig. 3E).

On the basis of clarifying the effect of differential gene expression in breast cancer samples on immune cells. The cellphoneDB software is used to analyze the ligand-receptor relationship between cells. In the output of the results of the ligand receptor (Supplementary Table 1 result ligand receptor), the heatmap plot function is used to analyze the interaction between immune cells. The results showed that macrophages and DCs cells were significantly active and interacted with a variety of cells (Fig. 4A). In order to further analyze the interaction between ligand and receptor, the TRRUST database was applied, and the hypergeometric test method was used to analyze the interaction between differential genes, immune cell marker genes, and ligand-receptor. The multi-factor interaction network between immune cells was constructed and visualized with Cytoscape 3.7.2 (Fig. 4B).

2.3 WGCNA Analysis reveals key modules associated with breast cancer progression and patient survival

Based on the genes in the immune cell multi-factor interaction network, the grouping information of single-cell data (primary, metastatic, and triple-negative breast cancer) and expression matrix, the co-expression network was constructed and WGCNA analysis was performed. The average-linkage hierarchical clustering method is used for gene cluster analysis. According to the standard of hybrid dynamic shearing tree, the minimum number of genes (soft threshold) of each gene network module is set. The results indicate that power = 18 is used as the threshold for subsequent analysis (Fig. 5A). After the eigengenes are calculated, the modules are subjected to cluster analysis, and finally 6 modules are obtained (Fig. 5B). According to different groups, age, gender, race, N, stage, T, radiotherapy, overall survival status, and overall survival time were used as indicators to screen the prognostic module with the highest correlation with breast cancer survival. The Pearson correlation coefficient between the ME of

each module and the sample feature is calculated (the higher the module, the more important it is). The results showed that the ME blue module was most correlated with the overall survival time of triple-negative breast cancer and the difference was the most significant ($R = 0.13$, $P = 3e-05$) (Fig. 5C). The contained genes are the main components that representing the function and characteristics of the module. The MEblue module contains 144 genes (Fig. 5D). These results indicate that ME blue may be a prognostic-related module in triple-negative breast cancer, and plays an important role in predicting the development of the disease and the overall survival of the patient. The GO enrichment analysis of the modular genes on the Metascape website shows that the ME blue module is mainly enriched in terms related to cell differentiation, movement, and proliferation, such as developmental process, locomotion, cell proliferation, multicellular organismal process, etc. (Fig. 5E). In order to further understand the mechanism of the ME blue module involved in the occurrence and progression of triple-negative breast cancer, the KOBAS website was used to perform KEGG enrichment analysis on the genes in the blue module. The results showed that the disordered genes in the ME blue module are mainly involved in pathway in cancer, transcriptional mis-regulation in cancer, PI3K-AKT signaling pathway, Ras signaling pathway, MAPK signaling pathway, cytokine-cytokine receptor interaction, AMPK signaling pathway (Fig. 5F).

2.4 Univariate regression analysis combined with KM survival analysis to identify prognostic genes in ME-blue modules

The coxph function in the survival package was used to analyze the relationship between the genes in the blue module and the overall survival (OS) of 1194 samples in the TCGA data. 144 genes in the blue module were analyzed by univariate regression, and 130 genes with significant regression differences were obtained with $P < 0.05$ as the threshold (Supplementary Table 2 Univariate Cox) were obtained (Fig. 6A). Next, these 130 genes were subjected to lasso dimensionality reduction analysis, and the number of output genes was still 130 (Fig. 6B). On the other hand, the Kaplan-Meier method was used to perform an overall survival analysis of 130 genes in the blue module, and 24 genes were found to be associated with prognosis of breast cancer ($P < 0.05$, Supplementary Table 3 KM-survival. gene.0.05). This means that in the key module, there are 24 genes that are significantly different in regression analysis and have significant prognostic properties in survival analysis (Fig. 6A). Two genes (ABCC9, NPR1) were randomly selected for survival curve display (Fig. 6C). Next, the correlation coefficients and the univariate regression analysis results of 24 genes were visualized (Fig. 6D and E).

2.5 Multivariate regression analysis constructed a risk model of breast cancer prognosis-related genes

In order to identify prognostic markers of breast cancer, the 24 prognostic-related genes in the blue module combined with the clinical factors of TCGA (age, lymph node metastasis status (N0 vs Nn (excluding Nx)), T stage, radiotherapy or not, race, breast Cancer stage) were subjected to multivariate regression analysis. The results suggest that the expression of ECM2, PCDH12, EPAS1, CD93, DLL4, and ARHGEF15 are significantly different in age, N (lymph node metastasis status), and radiotherapy or not (Fig. 7A). Subsequently, the sample data were scored and grouped according to the median value of

prognosis related gene expression. The Kaplan-Meier survival analysis of high risk and low risk group showed that the prognosis of high-risk group was poor (Fig. 7B, $P < 0.001$). The results of time-dependent ROC (receiver operating characteristic) analysis found that the above-mentioned prognostic genes showed good predictive effects on the 1-year, 3-year, and 5-year survival of breast cancer (AUC were all greater than 0.75) (Fig. 7C). Among them, the AUC of 3-year survival model was 0.801, which confirmed good accuracy of the prediction model (Generally, $AUC \geq 0.7$ is considered to be a good predictor).

2.6 Validation of a risk model constructed by genes related to the prognosis of breast cancer

In this part of the analysis, breast cancer data from ICGC was used to verify the model constructed by the prognostic-related genes in the ME blue module. 1542 sample data with overall survival time and overall survival status are used to verify the accuracy and validity of the above model. The results indicate that the prognosis of high-risk group in the model is poor ($P < 0.001$, Fig. 8A). The time-dependent ROC results show that the AUC values for 1, 3 and 5 years are 0.614 (Fig. 8B), 0.634 (Fig. 8C), and 0.632 (Fig. 8D), respectively. In addition, according to our prognostic model, age, N lymph node metastasis and radiotherapy showed significant differences in breast cancer samples with different risk scores (high-risk and low-risk) (Fig. 8E-G). Moreover, the verification results of ROC also confirmed that the AUC value of age, N lymph node metastasis was high, which indicated good accuracy of ROC (Fig. 8H-J). These results indicate that our model has practical application value in the prognosis and survival of breast cancer.

Discussion

The heterogeneity of breast cancer is the main reason for treatment failure and recurrence. In recent years, the development of single-cell sequencing technology has given us a deeper understanding of the heterogeneity of breast cancer. Cell types and specific gene expression characteristics in tumor tissues can be accurately distinguished by single-cell transcriptome. In fact, not only breast cancer cells show significant heterogeneity, but also non cancer cells are the main content of heterogeneity in breast cancer[15, 16]. Non-cancer cells in breast cancer include fibroblasts, adipocytes, endothelial cells and various immune cells[17]. Among non-cancer cells, the role of immune cells is particularly significant. The progression of breast cancer is characterized by increased immune cell infiltration in tumor parenchyma and stroma, including $CD4^+$ and $CD8^+$ granzyme B^+ cytotoxic T cells, B cells, macrophages and dendritic cells[18]. In addition, tumor-infiltrating lymphocytes have been reported as a prognostic indicator of breast cancer chemotherapy response and patient survival[19]. In this study, NK cells were found to be significantly aggregated in triple-negative breast cancer, and the number of macrophages was significantly increased in primary breast cancer, while the proportion of B cells, T cells, and neutrophils was raised in metastatic breast cancer. Although the high total number of NK cells reflects a good survival rate, the infiltration and activation of NK cells vary greatly with different types of breast cancer. The heterogeneity of NK cells and their actual role in the microenvironment of breast cancer need to be further elucidated[20]. Tumor-associated macrophages (TAM) are the main component of breast cancer microenvironment[21]. The increased density of macrophages in breast cancer tissues is related to the

poor prognosis of patients. Macrophages are not only involved in the immune escape of breast cancer, but also involved in the angiogenesis of tractable tumors[22]. In addition, B cells, T cells and neutrophils have all been reported to participate in the immune escape and metastasis of breast cancer[14, 23, 24]. These results indicate that our analysis results are correct and reasonable in terms of cell clusters and immune cell infiltration.

After the cells are clustered and annotated, a multi-factor interaction network of ligand-receptor combined with transcription factors is constructed and used to discover modules that are significantly related to the prognosis of breast cancer. The results showed that the blue module had the highest correlation with the overall survival time of breast cancer ($P = 3e-05$). The functions of the blue module are mainly enriched in terms related to cell developmental, locomotion, and proliferation. The decrease of cell development is related to the poor differentiation level and cell stem characteristics of breast cancer; the enhancement of cancer cell motility is related to tumor invasion and metastasis, the disorder of cell proliferation is the basis of tumor tumorigenesis and progression. The enrichment results of KEGG showed that the main signaling pathways for differential gene enrichment in the blue module include pathway in cancer, transcriptional mis-regulation in cancer, PI3K-AKT signaling pathway, Ras signaling pathway, MAPK signaling pathway, cytokine-cytokine receptor interaction, MAPK signaling pathway, etc. PI3K-AKT is over-activated in most breast cancers and promotes the excessive proliferation of cancer cells through the mTOR complex[25]. For example, the loss of expression of the negative regulatory proteins PTEN and INPP4B (tumor suppressor genes) of the PI3K-AKT pathway is associated with the occurrence and progression of triple-negative breast cancer, and the loss of PTEN expression is found in more than half of TNBC patients[26]. In the Ras signaling pathway, activated Ras promotes cell cycle and cell proliferation by recruiting Raf1 protein to initiate a kinase cascade to activate MAPK (ERK1/2) and transcription factors Fos and c-Jun[27, 28]. In addition, activation of the Ras-MAPK pathway has been reported to promote TNBC immune escape[29]. p38MAPK signal was found to promote the invasion and metastasis of breast cancer by enhancing the epithelial-mesenchymal transition of cancer cells[30]. AMPK and its downstream mTOR are involved in regulating the material and energy metabolism of cancer cells 30903363. For example, AMPK-mediated lipid metabolism reprogramming promotes breast cancer cell proliferation and migration[31]. These results indicate that the cellular functions and signal pathways enriched in the blue module play a key role in the occurrence and progression of breast cancer. The validation results of the breast cancer prognosis model constructed by multivariate regression risk analysis showed that PCDH12, SLIT3, ACVRL1 and DLL4 genes are significantly different in high-risk group and low-risk breast cancer, which can be used as risk factors for breast cancer prognosis.

In recent reports, the high expression of PCDH12 was found to be associated with the high pathological grade of papillary renal cell carcinoma[32]. As a new type of tumor suppressor gene, SLIT3 has been reported to play a role in breast, liver, lung, and colon cancer, and the promoter methylation of SLIT3 has been reported to be associated with tumor occurrence and progression[33, 34]. ACVRL1 (activin receptor like protein 1) encodes ALK1, which is a member of transforming growth factor - β receptor family and is associated with angiogenesis[35]. ACVRL1 expression can be used as a prognostic marker for metastatic colorectal cancer patients receiving chemotherapy and bevacizumab[36]. DLL4, as one of the main

components of the Notch pathway, is reported to be highly expressed in breast cancer and associated with advanced stage and distant metastasis of the patient[37]. these studies confirm the correctness of PCDH12–SLIT3–ACVRL1 and DLL4 genes as risk factors for breast cancer prognosis.

Conclusion

MeBlue is a prognostic module in triple negative breast cancer. The expressions of PCDH12–SLIT3–ACVRL1 and DLL4 are not only related to the type and proportion of immune cells, but also contribute to the prognosis of breast cancer.

Declarations

Ethics approval and consent to participate

Clinical and animal experiments are not involved in this study. Therefore, ethical and participatory consent elements are not included.

Consent for publication

This study did not involve the collection of patient samples and data.

Availability of data and materials

The datas that support our findings of this study are openly available in GEO public database. The single-cell transcriptome data used in this study were obtained from GSE75688 and GSE118389 datasets. The URL of the GSE75688 dataset is: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75688>

The URL for the GSE118389 dataset is <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118389>

Competing interests

The authors declare that they have no competing interests.

Funding

The project was funded by National Natural Science Foundation of China (No. 8196110236).

Authors' contributions

RL, XYand YHQ jointly covered the research protocol and performed the main bioinformatics analysis on the single cell sequencing data.

YYT and YFL participated in data analysis and manuscript compilation.

LCX and YDC completed the statistical analysis and visualization of the results.

WAH and WMH discussed the results, revised the manuscript, and provided financial support for this study.

Acknowledgements

Clinical and animal experiments are not involved in this study. Therefore, ethical and participatory consent elements are not included, and this study did not involve the collection of patient samples and data. All authors read and approved the final manuscript. Grateful acknowledgement is made to all colleagues and friends for their contributions to this paper in various ways.

References

1. Anastasiadi Z, Lianos GD, Ignatiadou E, Harissis HV, Mitsis M: **Breast cancer in young women: an overview.** *Updates Surg* 2017, **69**:313-317.
2. Mokhatri-Hesari P, Montazeri A: **Health-related quality of life in breast cancer patients: review of reviews from 2008 to 2018.** *Health Qual Life Outcomes* 2020, **18**:338.
3. Escala-Garcia M, Morra A, Canisius S, Chang-Claude J, Kar S, Zheng W, Bojesen SE, Easton D, Pharoah PDP, Schmidt MK: **Breast cancer risk factors and their effects on survival: a Mendelian randomisation study.** *BMC Med* 2020, **18**:327.
4. Ellington TD, Henley SJ, Wilson RJ, Miller JW: **Breast Cancer Survival Among Males by Race, Ethnicity, Age, Geographic Region, and Stage - United States, 2007-2016.** *MMWR Morb Mortal Wkly Rep* 2020, **69**:1481-1484.
5. Yeo SK, Guan JL: **Breast Cancer: Multiple Subtypes within a Tumor?** *Trends Cancer* 2017, **3**:753-760.
6. Zeng X, Liu C, Yao J, Wan H, Wan G, Li Y, Chen N: **Breast cancer stem cells, heterogeneity, targeting therapies and therapeutic implications.** *Pharmacol Res* 2021, **163**:105320.
7. Tuasha N, Petros B: **Heterogeneity of Tumors in Breast Cancer: Implications and Prospects for Prognosis and Therapeutics.** *Scientifica (Cairo)* 2020, **2020**:4736091.
8. Ding S, Chen X, Shen K: **Single-cell RNA sequencing in breast cancer: Understanding tumor heterogeneity and paving roads to individualized therapy.** *Cancer Commun (Lond)* 2020, **40**:329-344.
9. Kinker GS, Greenwald AC, Tal R, Orlova Z, Cuoco MS, McFarland JM, Warren A, Rodman C, Roth JA, Bender SA, et al: **Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity.** *Nat Genet* 2020, **52**:1208-1218.
10. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, Liu L, Huang D, Jiang J, Cui GS, et al: **Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma.** *Cell Res* 2019, **29**:725-738.
11. Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F, Ellisen LW: **Unravelling subclonal heterogeneity and aggressive disease states in TNBC**

- through single-cell RNA-seq. *Nat Commun* 2018, **9**:3588.
12. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R: **CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes.** *Nat Protoc* 2020, **15**:1484-1506.
 13. Guo W, Feng W, Fan X, Huang J, Ou C, Chen M: **Osteomodulin is a Potential Genetic Target for Hypertrophic Cardiomyopathy.** *Biochem Genet* 2021.
 14. Coffelt SB, Kersten K, Doornebal CW, Weiden J, Vrijland K, Hau CS, Verstegen NJM, Ciampricotti M, Hawinkels L, Jonkers J, de Visser KE: **IL-17-producing $\gamma\delta$ T cells and neutrophils conspire to promote breast cancer metastasis.** *Nature* 2015, **522**:345-348.
 15. Montemagno C, Pagès G: **Metastatic Heterogeneity of Breast Cancer: Companion and Theranostic Approach in Nuclear Medicine.** *Cancers (Basel)* 2020, **12**.
 16. Liu J, Xu T, Jin Y, Huang B, Zhang Y: **Progress and Clinical Application of Single-Cell Transcriptional Sequencing Technology in Cancer Research.** *Front Oncol* 2020, **10**:593085.
 17. Avagliano A, Fiume G, Ruocco MR, Martucci N, Vecchio E, Insabato L, Russo D, Accurso A, Masone S, Montagnani S, Arcucci A: **Influence of Fibroblasts on Mammary Gland Development, Breast Cancer Microenvironment Remodeling, and Cancer Cell Dissemination.** *Cancers (Basel)* 2020, **12**.
 18. Goff SL, Danforth DN: **The Role of Immune Cells in Breast Tissue and Immunotherapy for the Treatment of Breast Cancer.** *Clin Breast Cancer* 2021, **21**:e63-e73.
 19. Sui S, An X, Xu C, Li Z, Hua Y, Huang G, Sui S, Long Q, Sui Y, Xiong Y, et al: **An immune cell infiltration-based immune score model predicts prognosis and chemotherapy effects in breast cancer.** *Theranostics* 2020, **10**:11938-11949.
 20. Wu SY, Fu T, Jiang YZ, Shao ZM: **Natural killer cells in cancer biology and therapy.** *Mol Cancer* 2020, **19**:120.
 21. Larionova I, Tuguzbaeva G, Ponomaryova A, Stakheyeva M, Cherdyntseva N, Pavlov V, Choinzonov E, Kzhyshkowska J: **Tumor-Associated Macrophages in Human Breast, Colorectal, Lung, Ovarian and Prostate Cancers.** *Front Oncol* 2020, **10**:566511.
 22. Choi J, Gyamfi J, Jang H, Koo JS: **The role of tumor-associated macrophage in breast cancer biology.** *Histol Histopathol* 2018, **33**:133-145.
 23. Gu Y, Liu Y, Fu L, Zhai L, Zhu J, Han Y, Jiang Y, Zhang Y, Zhang P, Jiang Z, et al: **Tumor-educated B cells selectively promote breast cancer lymph node metastasis by HSPA4-targeting IgG.** *Nat Med* 2019, **25**:312-322.
 24. Kresovich JK, O'Brien KM, Xu Z, Weinberg CR, Sandler DP, Taylor JA: **Prediagnostic Immune Cell Profiles and Breast Cancer.** *JAMA Netw Open* 2020, **3**:e1919536.
 25. Miricescu D, Totan A, Stanescu S, Il, Badoiu SC, Stefani C, Greabu M: **PI3K/AKT/mTOR Signaling Pathway in Breast Cancer: From Molecular Landscape to Clinical Aspects.** *Int J Mol Sci* 2020, **22**.
 26. Li J, Ho WY, Tsang JYS, Ni YB, Chan SK, Tse GM: **Expression of biomarkers in the AKT pathway correlates with malignancy and recurrence in phyllodes tumours of the breast.** *Histopathology* 2019,

27. Santos E, Crespo P: **The RAS-ERK pathway: A route for couples.** *Sci Signal* 2018, **11**.
28. Meng L, Liu S, Liu F, Sang M, Ju Y, Fan X, Gu L, Li Z, Geng C, Sang M: **ZEB1-Mediated Transcriptional Upregulation of circWWC3 Promotes Breast Cancer Progression through Activating Ras Signaling Pathway.** *Mol Ther Nucleic Acids* 2020, **22**:124-137.
29. Loi S, Dushyanthen S, Beavis PA, Salgado R, Denkert C, Savas P, Combs S, Rimm DL, Giltane JM, Estrada MV, et al: **RAS/MAPK Activation Is Associated with Reduced Tumor-Infiltrating Lymphocytes in Triple-Negative Breast Cancer: Therapeutic Cooperation Between MEK and PD-1/PD-L1 Immune Checkpoint Inhibitors.** *Clin Cancer Res* 2016, **22**:1499-1509.
30. Wen S, Hou Y, Fu L, Xi L, Yang D, Zhao M, Qin Y, Sun K, Teng Y, Liu M: **Cancer-associated fibroblast (CAF)-derived IL32 promotes breast cancer cell invasion and metastasis via integrin β 3-p38 MAPK signalling.** *Cancer Lett* 2019, **442**:320-332.
31. Zhang ZG, Zhang HS, Sun HL, Liu HY, Liu MY, Zhou Z: **KDM5B promotes breast cancer cell proliferation and migration via AMPK-mediated lipid metabolism reprogramming.** *Exp Cell Res* 2019, **379**:182-190.
32. Feng X, Zhang M, Meng J, Wang Y, Liu Y, Liang C, Fan S: **Correlating Transcriptional Networks to Papillary Renal Cell Carcinoma Survival: A Large-Scale Coexpression Analysis and Clinical Validation.** *Oncol Res* 2020, **28**:285-297.
33. Wang J, Yu XF, Ouyang N, Zhao SY, Guan XF, Yao HP, Chen R, Chen T, Li JX: **[Expression and prognosis effect of methylation-regulated SLIT3 and SPARCL1 genes in smoking-related lung adenocarcinoma].** *Zhonghua Yi Xue Za Zhi* 2019, **99**:1553-1557.
34. Dickinson RE, Dallol A, Bieche I, Krex D, Morton D, Maher ER, Latif F: **Epigenetic inactivation of SLIT3 and SLIT1 genes in human cancers.** *Br J Cancer* 2004, **91**:2071-2078.
35. Capasso TL, Li B, Volek HJ, Khalid W, Rochon ER, Anbalagan A, Herdman C, Yost HJ, Villanueva FS, Kim K, Roman BL: **BMP10-mediated ALK1 signaling is continuously required for vascular development and maintenance.** *Angiogenesis* 2020, **23**:203-220.
36. Hanna DL, Loupakis F, Yang D, Cremolini C, Schirripa M, Li M, Matsusaka S, Berger MD, Miyamoto Y, Zhang W, et al: **Prognostic Value of ACVRL1 Expression in Metastatic Colorectal Cancer Patients Receiving First-line Chemotherapy With Bevacizumab: Results From the Triplet Plus Bevacizumab (TRIBE) Study.** *Clin Colorectal Cancer* 2018, **17**:e471-e488.
37. Zohny SF, Zamzami MA, Al-Malki AL, Trabulsi NH: **Highly Expressed DLL4 and JAG1: Their Role in Incidence of Breast Cancer Metastasis.** *Arch Med Res* 2020, **51**:145-152.

Figures

Figure 1

Quality control and PCA analysis of two single-cell transcriptome sequencing data sets from the GEO database A: The DimHeatmap function is used to analyze the main source of heterogeneity in the data set, and to determine the PC data for further downstream analysis. B: The Violin distribution form shows the distribution and proportion of the number of genes (nFeature), the number of UMIs (nCount), and the content of mitochondrial genes (percent.mito) in the cell. C: The JackStrawPlot function is used to compare the p-value distribution and uniform distribution of each PC and determine the important PC (powerful and rich PC with low p-value characteristics). D: The ElbowPlot function sorts the principal components according to the percentage of variance.

Figure 2

Single-cell clusters of breast cancer and the expression of marker genes in each cluster. A: Visualization of cell UMAP clustering results. B: Visualization of the genes with the highest differential change and non-differential genes. C: The violin chart shows the expression of common cell marker genes in the corresponding cell clusters.

A

B

Figure 4

CellphoneDB analysis of cell-cell interaction A: Heat map of cell -cell interaction based on ligand -receptor. B: Multi-factor interaction network of ligand-receptor combined transcription factor (Blue represents the transcription factor TF, green represents the immune marker gene, red represents the ligand gene ligand, and light blue represents the receptor gene receptor).

Figure 5

WGCNA analysis was used to identify modules which significantly associated with breast cancer prognosis. A: Analysis of network topology for various soft-thresholding powers. B: Module clustering analysis based on eigengenes. C: Correlation analysis of each module and its traits. C: Frequency statistical analysis of genes in each module. D: Functional enrichment analysis of breast cancer prognosis-related modules. E: KEGG Enrichment Analysis of dis regulated genes in MEblue module.

Figure 6

Univariate regression analysis and KM survival analysis of the BLUE module related to breast cancer prognosis. A: Winn plots showing the intersection of significantly different genes in univariate regression analysis and KM population survival analysis. B: KM survival curves of ABCC9 and NPR1. C: Lasso dimensionality reduction analysis of 130 genes with significant differences in univariate analysis. D:

Correlation coefficient distribution of 24 intersection genes by univariate regression analysis. E: Visualization of univariate regression analysis of 24 intersection genes.

Figure 7

Efficiency evaluation of multivariate regression analysis and risk scoring. A: The results of the multi-factor analysis are displayed by forest map. B: Kaplan-Meier survival curves of high and low risk groups. C: Time-dependent ROC was applied to evaluate the accuracy of the model in predicting 1 -, 3 -, and 5-year survival

Figure 8

The accuracy of breast cancer risk model constructed by prognostic related genes in predicting patient survival, clinicopathological indicators and prognosis A: KM survival curve of high and low risk group in ICGC data. B: Accuracy of the time-dependent ROC assessment model in ICGC data for 1-year survival of breast cancer patients. C: Accuracy of the time-dependent ROC assessment model in ICGC data for 3-year survival of breast cancer patients. D: Accuracy of the time-dependent ROC assessment model in ICGC data for 5-year survival of breast cancer patients. E-G: Correlation between risk score and clinical factors (AGE <60 vs AGE >60; N0 vs N1; Non-radiation vs Radiation). H-J: Prognostic accuracy of risk scoring.

Supplementary Files

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

- [SupplementaryTable1resultligandreceptor.xlsx](#)

- [SupplementaryTable2UnivariateCox.xlsx](#)
- [SupplementaryTable3KMsurvival.gene.0.05.xlsx](#)