Construction and Evaluation of a Prognostic Model Based on Metastasis-Associated Genes in Breast Cancer

Zhixiao Liao
The First Clinical Medical College of Guangzhou University of Traditional Chinese Medicine

Yueyang Deng (✉ dyylovelzx@163.com)
Tianjin Cancer Hospital Airport Hospital

Research Article

Keywords: Breast cancer, Distant metastasis, Overall survival, Nomogram, TCGA database

Posted Date: August 21st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3265001/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Objective: The aim of the study was to investigate the gene expression profile features in distant metastatic breast cancer (BC) patients, identify the metastasis-associated genes correlated with prognosis, and construct a survival rate nomogram.

Methods: Transcriptome data of BC patients were downloaded from The Cancer Genome Atlas (TCGA) database, and divided into metastatic and non-metastatic groups. Differentially expressed genes (DEGs) were analyzed between the two groups, and Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was performed to explore the potential functions of DEGs. Univariate COX, LASSO regression, and multivariate Cox regression models were applied to screen prognostic-related genes, and a prediction model was established.

Results: A total of 215 DEGs were identified. FAM9C, CRISP2, TFPI2, TUBA3E, IL12Rβ2, BP1 and CSN3 were independent influencing factors for overall survival (OS) rate. Area under the curve (AUC) values outweighed 0.6, and calibration curves did not deviate from the reference line.

Conclusion: The metastasis-related genes prognostic nomogram for BC patients established in this study had favourable predictive power that could provide a theoretical reference for subsequent studies.

1 Introduction

Breast cancer (BC) incidence has surpassed lung cancer as the most common malignancy, with an estimated 2.3 million new cases, accounting for 11.7% of all cancers diagnosed worldwide in 2020(1). Distant metastasis is the major cause of death of BC patients. Approximately 3%-8% of patients have metastasized when first diagnosed, and about 30%-40% of patients with localized disease remain at risk of recurrence and metastasis even after radical treatment(2). Metastatic BC patients have a poor prognosis with a 5-year survival rate of around 20%(3).

BC is highly heterogeneous at the clinical and molecular level. BC metastasis is a multi-step, multi-stage, polygene process, and tumors with the same clinical, pathological, and hormone receptor status may have different metastatic phenotypes. And the research on metastasis-mediated genes at the molecular level has become a focus of current studies.

Previous studies have indicated metastasis-associated genes have great potential for cancer prognostic assessment(4–8). Therefore, this study was conducted to screen the differentially expressed genes (DEGs) of metastatic BC by analyzing the transcriptome data in The Cancer Genome Atlas (TCGA) database, afterwards utilize Cox regression and Lasso regression analysis to identify the independent prognostic factors, and then establish a predictive model.

2 Materials and Methods
2.1 Data Collection

Gene expression data of BC were downloaded from TCGA database. The patients that met the following criteria were enrolled: (1) patients had histologically confirmed diagnosis of BC; (2) women were aged over 18 years; (3) samples were annotated with distant metastasis information; (4) survival information was complete; and (5) follow-up time was more than 1 month.

All eligible samples were classified into the metastasis group (patients at M1 stage) and the non-metastasis group (patients at M0 stage). Overall survival (OS), defined as the time interval between confirmed diagnosis and all-cause death or last follow-up, was considered as the primary outcome index.

2.2 Identification and Enrichment Analysis of DEGs

SPSS 26.0 and R 4.1.2 were used for statistical analyses. Significance level was set at two-sided $P<0.05$. Differential analyses were performed between metastatic and non-metastatic groups, and then DEGs volcano plot was drawn. False discovery rate (FDR) $\leq 0.05$ and $|\log_2 \text{Fold Change}| (|\log_2 \text{FC}|) \geq 1$ were set as the thresholds for DEGs. DEGs were then subjected to Gene Ontology (GO) functional enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, in order to examine the potential functions and pathways.

2.3 Construction of DEGs Prognostic Model

Firstly, univariate Cox regression analysis was conducted on DEGs to screen a set of candidate prognostic genes for OS. Then, genes with $P$-value $<0.05$ were included in LASSO regression analysis to avoid overfitting. Finally, multivariate Cox regression analysis was performed to ascertain independent prognostic factors.

2.4 Validation of DEGs Prognostic Model

Receiver operating characteristic (ROC) curves and calibration curves were used to evaluate the discriminatory ability and predictive accuracy, and bootstrap method was calculated from 1000 iterations.

Based on the median risk score of the nomogram, BC patients were further stratified as high- or low-risk groups. The log-rank test was used to determine survival differences between groups in Kaplan-Meier (KM) analysis. The correlation between risk scores and TNM stage was assessed.

Immune score, stromal score, ESTIMATE score and tumor purity were calculated by the ESTIMATE algorithm, and compared between groups. The proportions of 22 human immune cell subsets in each sample was calculated by the CIBERSORT algorithm, and compared between groups.

The differential analysis of all genes between high- and low-risk groups was generated, and these genes were ranked by log2 FC value, and then Gene Set Enrichment Analysis (GSEA) was used to investigate the biological mechanisms differences.
3 Results

3.1 Identification of DEGs

A total of 462 BC samples were enrolled, among which 16 had distant metastasis and 446 had no distant metastasis. Totally 215 DEGs, including 199 upregulated genes and 16 downregulated genes, were identified in the metastasis cohort compared with the non-metastasis cohort (Fig. 1A).

3.2 Enrichment Analysis of DEGs

In biological process category of Go analysis, the top-3 GO terms were response to xenobiotic stimulus (GO: 0009410, 8.1%), axon development (GO:0061564, 7.6%), and regulation of peptidase activity (GO: 0052547, 7.6%). In cellular component category, the top-3 GO terms were collagen – containing extracellular matrix (ECM) (GO: 0062023, 7.6%), specific granule lumen (GO: 0035580, 3.0%), and cornified envelope (GO: 0001533, 2.5%). In molecular function category, the top-3 GO terms were receptor ligand activity (GO: 0048018, 8.3%), endopeptidase activity (GO: 0004175, 6.2%), and enzyme inhibitor activity (GO: 0004857, 5.7%) (Fig. 1B).

KEGG analysis revealed that there were 5 enriched signaling pathways, of which were arachidonic acid metabolism, salivary secretion, metabolism of xenobiotics by cytochrome P450, regulation of lipolysis in adipocytes, and chemical carcinogenesis-DNA adducts (Fig. 1C).

3.2 Construction of DEGs Prognostic Model

In the univariate analysis, a total of 10 DEGs were significantly correlated with OS ($P<0.05$) (Fig. 2A). Next, the above 10 OS-related genes were confirmed by LASSO regression (Fig. 2B-C). The further multivariate analysis demonstrated that FAM9C, CRISP2, TFPI2, TUBA3E, IL12RB2, BPI, and CSN3 were independent prognostic factors of OS ($P<0.05$) (Fig. 2D). A survival risk model of OS was constructed based on the above genes (Fig. 2E), and the formula for risk score calculation was as follows: risk score = $(0.832\times FAM9C) + (0.455\times CRISP2) + (-0.208\times TFPI2) + (-0.750\times TUBA3E) + (-1.084\times IL12R\beta2) + (-0.682\times BPI) + (0.208\times CSN3)$.

3.3 Verification of DEGs Prognostic Model

The AUC values predicting the 1-, 3- and 5-year OS were 0.643, 0.732 and 0.772 (Fig. 3A), indicating a good discriminatory capacity of the model. The calibration curves were close to the 45-degree diagonal line (Fig. 3B-D), suggesting optimal agreement between nomogram predictions and actual observations.

The prognosis of the low-risk group was significantly better than that of the high-risk group ($P<0.001$) (Fig. 3E), indicating that this model was able to accurately distinguish high-risk patients from low-risk patients. The risk scores of different T stage, M stage, N stage, and stage all exhibited significant differences ($P<0.05$) (Fig. 3H-K), indicating that high-risk group had more aggressive behaviors.
The high-risk group represents a lower immune \( (P<0.05) \) and ESTIMATE score, and higher tumor purity than the low-risk group (Fig. 3F). The proportion of CD8+ T cells, follicular helper T cells, CD4+ memory activated T cells, M1 macrophages, and resting dendritic cells were significantly lower in the high-risk group \( (P<0.05) \) (Fig. 4A). In addition, patients in the high-risk group had remarkably lower expression of immune checkpoint genes, such as CD70, TNFSF4, CTLA4, PDCD1, TNFRSF8, TNFRSF9, TIGIT, ICOS, CD28, TNFSF14, CD80, and TNFSF15 \( (P<0.05) \) (Fig. 4B). The above results demonstrated that high-risk group had lower immune infiltration level than low-risk group.

GSEA displayed that androgen response, angiogenesis, and epithelial-mesenchymal transition (EMT) pathways were considerably enriched in the high-risk group (Fig. 4C). Hallmarks like allograft rejection, complement, E2F targets, IL2-STAT5 signalling, IL6-JAK-STAT3 signalling, inflammatory response, interferon alpha and gamma response, KRAS signalling, spermatogenesis, and TNFA signaling via NFKB were enriched in the low-risk group (Fig. 4D).

### 4 Discussion

The present study analyzed the BC dataset from TCGA database through bioinformatics analysis. Firstly, 215 DEGs related to metastasis were identified through differential analysis. Following univariate Cox regression analysis, LASSO regression analysis, and multivariate Cox regression analysis were used to identify 7 OS prognostic signature genes, including FAM9C, CRISP2, TFPI2, TUBA3E, IL12Rβ2, BP1 and CSN3. Next, the OS nomogram showed relatively good predictive performance after validation.

In this study, GO functional analysis found that the DEGs of metastatic BC were mainly enriched in the regulation of peptidase activity \( (BP/MF) \) and ECM \( (CC) \). Peptidases are proteolytic enzymes that hydrolyze peptide bonds. ECM represents a vital component of the tumor microenvironment. Studies have revealed that tumor cells with metastatic potential would invade the extracellular matrix, and secrete specific proteolytic enzymes that act on collagen to degrade the basement membrane, which can promote cancer cell migration, invasion, and metastasis(9).

KEGG pathway analysis found that arachidonic acid metabolism pathways were significantly enriched. Arachidonic acid can participates in various metabolic pathways and coproduces multiple metabolites. Among these metabolic products, prostaglandin E2 (PGE2), epoxyeicosatrienoic acids (EETs), and hydroxy-eicosatetraenoic acid (HETE) have been proven to promote the proliferation, invasion and metastasis of tumour cells, inhibit apoptosis, and stimulate angiogenesis(10).

The results of this study identified 7 metastasis-related genes as independent factors associated with OS in BC patients, and these genes have also been reported to be associated with the occurrence and development of tumors. Of these, β2 components encoded by the IL-12Rβ2 gene belongs to the IL-12 receptor (IL-12R) heterodimer(11). It can combine with IL-12 and enhance the secretion of interferon-γ (IFN-γ), which exhibits anti-proliferative and pro-apoptotic effects against tumor cells. Moreover, it can promotes the proliferation and differentiation, as well as tumor cell-killing effect of NK cells, cytotoxic T lymphocytes, and macrophages(12). The TUBA3E gene belongs to a gene family encoding several
microtubule cytoskeleton α- and β-tubulin protein isotypes. Studies have found that a diet high in milk fat and exposure to bisphenol A are accompanied by changes in multiple genes such as TUBA3E, which is therefore plausibly involved in BC development(13). BP1 belongs to the homeobox gene family. The up-regulation of BP1 in BC was significantly linked to tumor grade, pathological stage, and estrogen receptor status(14). As a common serine protease inhibitors, TFPI2 plays an important role in angiogenesis, tumorigenesis and metastasis. It was reported that TFPI2 gene expression is downregulated in various tumors and is closely related to tumor progression and unfavorable outcome(15–17). In addition, studies have shown that overexpression of FAM9C gene in liver cancer can promote the proliferation, immune evasion, invasion and metastasis of tumor cells(18). However, other studies have found that the FAM9C gene is a type of cervical cancer tumor suppressor gene, and HPV infection can induce decreasing or silencing of this gene by abnormal promoter methylation(19). CRISP2 is a member of the cysteine-rich secretory protein (CRISP) family, and its specifically low expression in cervical cancer and precancerous lesions is helpful in increasing the diagnostic accuracy(20). CSN3, a pivotal subunit of the constitutive COP9 signalosome (CSN), has been implicated in development and progression of myelomas and liver cancer, and the knockdown of CSN3 can inhibit tumor cell proliferation and induce apoptosis(21, 22). The above results were similar to the conclusions of the present study, suggesting that these genes may be a potential target for anti-tumor treatment, and further studies to elucidate the specific function and molecular mechanism in BC are needed.

In this research, different methods were used to analyze the immune infiltration between groups, and the degree of partial immune infiltration in the high-risk group was significantly lower than that in the low-risk group, indicating that patients in the high-risk group are in an immunosuppressive status and might not benefit from immunotherapy as a consequence. GSEA analysis further observed that the high-risk group were concentrated on the hallmarks including angiogenesis and EMT. Tumor angiogenesis assists the formation of aberrant vascular networks and is related to tumor progression, metastasis, and recurrence(23). EMT process can allow cancer cells lose the polarities and connections with the basement membrane, and as a result, tumor cells acquire higher migration and invasion capacities(24).

Several limitations of this study should be acknowledged. TCGA database lacks several vital information, such as clinical treatment and molecular subtypes, and the number of metastasis-associated samples was limited. In the future, relevant experiments are still needed to further verify and explore these results in this paper, by expanding the scope of data collection, supplementing biological experiments, and collecting clinical samples.

5 Conclusion

In conclusion, we used metastasis-associated genes in TCGA database to identify independent prognostic factors for BC patients, and then established a OS nomogram, which can assist in providing more accurate reference for prognosis prediction and treatment decisions.

Declarations
Data Availability Statement

Publicly available datasets were analyzed in this study. These data were derived from TCGA database.

Author Contributions

Conception and design: ZL. Administrative support: YD. Collection and assembly of data: ZL. Data analysis and interpretation: ZL. Manuscript writing: ZL and YD. Final approval of manuscript: ZL and YD.

Funding

The Tianjin Municipal Education Commission Scientific Research Program (Natural Science) of China (No.2019KJ048) funded this study.

Acknowledgments

The authors thank the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) for providing the TCGA dataset.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References


Figures

![Figure 1](image-url)
A Volcano plots of DEGs between metastatic and non-metastatic groups. B Go analysis of DEGs between metastatic and non-metastatic groups. C KEGG analysis of DEGs between metastatic and non-metastatic groups.

Figure 2

Figure 3

A ROC curves of 1-, 3-, and 5-year OS prediction. B-D Calibration plots for predicting 1-(B), 3-(C), and 5-(D) year OS. E KM survival curves of high- and low-risk groups. F Correlation analysis between the risk score
of the prognostic risk model and Estimate scores. HK Correlation analysis between risk score and TNM stage.

**Figure 4**

A Correlation analysis of immune cell infiltration and risk score. B Correlations analysis of immune checkpoint genes expression levels and risk score. C-D GSEA plots of the high-risk group(C) and the low-risk group(D).