86495_Auto_Edited.docx

WORD COUNT	3909	TIME SUBMITTED	25-AUG-2023 02:10PM
		PAPER ID	102239164

Name of Journal: *World Journal of Clinical Cases*

Manuscript NO: 86495

Manuscript Type: ORIGINAL ARTICLE

Clinical and Translational Research

Identification of potential diagnostic and prognostic biomarkers for breast cancer based on gene expression omnibus

Xiong Zhang, ZhiHui Mi

Abstract

BACKGROUND

Breast cancer is regarded as a highly malignant neoplasm in the female population, posing a significant risk to women's overall well-being. The prevalence of breast cancer has been observed to rise in China, accompanied by an earlier age of onset when compared to Western countries. Breast cancer continues to be a prominent contributor to cancer-related mortality and morbidity among women, primarily due to its limited responsiveness to conventional treatment modalities. The diagnostic process is challenging due to the presence of non-specific clinical manifestations and the suboptimal precision of conventional diagnostic tests. There is a prevailing uncertainty regarding the most effective screening method and target populations, as well as the specificities and execution of screening programs.

AIM

Consequently, it is of utmost importance to identify diagnostic and prognostic biomarkers for breast cancer.

METHODS

Overlapping differentially expressed genes were screened based on Gene Expression Omnibus (GSE36765, GSE10810, and GSE 20086) and The Cancer Genome Atlas datasets. A protein-protein interaction network was applied to excavate the hub genes among these differentially expressed genes. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway analyses, as well as gene set enrichment analyses, were conducted to examine the functions of these genes and their potential mechanisms in the development of breast cancer. For clarification of the diagnostic and prognostic roles of these genes, Kaplan–Meier and Cox proportional hazards analyses were conducted.

RESULTS

This study demonstrated that calreticulin (CALR), heat shock protein family B member 1 (HSPB1), insulin-like growth Factor 1 (IGF1), interleukin-1 receptor 1 (IL1R1), Krüppel-like factor 4 (KLF4), suppressor of cytokine signaling 3 (SOCS3), and triosephosphate isomerase 1 (TPI1) are potential diagnostic biomarkers of breast cancer as well as potential treatment targets with clinical implications.

CONCLUSION

The screening of biomarkers is of guiding significance for the diagnosis and prognosis of diseases.

INTRODUCTION

The global incidence of breast cancer worldwide has exceeded that of lung cancer, with 2.26 million new cases reported, making it the most prevalent cancer worldwide [5]. Invasive breast cancer, known for its high malignancy and unfavorable prognosis, represents the predominant form of breast cancer [32].

According to the latest report on breast cancer in China in 2020, the incidence and death rates of breast cancer in Chinese women accounted for 11.2% and 9.2% of the global incidence and death rates, respectively, ranking among the top in the world. The

diagnosis and management of breast cancer are currently experiencing a paradigm shift, transitioning from a standardized approach to personalized medicine. This shift encompasses a range of treatment options, including resection surgery [14], radiotherapy and targeted adjuvant therapy. However, progress in breast cancer treatment remains limited, and the prognosis is not promising. Various factors, such as genetics, lifestyle choices, obesity, and environmental influences, may all contribute to the onset and progression of breast cancer [4]. The study has documented that engaging in physical activity among individuals with breast cancer patients not only enhance their quality of life but also exerts an impact on their immune system [29]. In order to gain a deeper comprehension of the pathogenesis of breast cancer and enhance the precision of treatment, it is imperative to direct attention towards genetic research, tumor signaling pathways, and targeted therapies, which are progressively being implemented in clinic work. Furthermore, the adoption of molecular stratification of therapies and the utilization of biomarkers to inform prognosis and treatment choices are on the rise [4].

The occurrence, development, overall survival, recurrence, and non-recurrence of tumors are influenced by both the pathological type and clinical stage of tumors, as well as the expression and pathways of tumor genes. Extensive research has demonstrated a significant elevation in abnormal gene expression in breast cancer compared to normal tissues, and these genes playing a crucial role in proliferation, invasion, apoptosis, and overall survival [15, 27, 36]. The analysis of these abnormally expressed genes holds great clinical significance in terms of targeted therapy, prognosis, and predicting the risk of recurrence in breast cancer. Presently, numerous clinical investigations are being conducted on genes associated with tumor recurrence genes and signaling pathways. As a result, a predictive model has been developed to elucidate the role of genes in enhancing the conventional tumor classification and recurrence prediction. This model provides a greater wealth of genetic information and more precise predictive data [12, 20, 24] . For instance, Rodrigues-Ferreira *et al* (2019) [28] identified potential implications of EB1 and ATIP3 in the diagnosis and prognosis of breast cancer. Additionally, Ki67 has

proven to be an efficient prognostic indicator in the Chinese population following neoadjuvant chemotherapy [33].

In this study, we obtained breast cancer gene profiles (GSE36765, GSE20086, and GSE10810) from the Gene Expression Omnibus (GEO). By performing GEO2R online analysis, differentially expressed genes (DEGs) were identified in breast cancer tissues compared to non-cancerous tissues. Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment, co-expression, and protein-protein interaction (PPI) analyses were conducted on the DEGs. We then performed overall survival analyses with respect to the potential target genes. In addition, we employed the GEPIA and UALCAN web-based platforms to identify potential genes linked to pathological stage using The Cancer Genome Atlas (TCGA) data. Our findings indicate that CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1 exhibit potential as therapeutic targets with significant clinical implications.

MATERIALS AND METHODS

Microarray data

GEO datasets were selected based on the following criteria: studies with invasive breast cancer tissue samples, a description of the technology and platforms utilized, and the inclusion of adjacent normal tissues as controls. Three datasets (GSE36765 [14], GSE20086 [3], GSE10810 [26]) were downloaded from the GEO database, meeting these criteria. All three studies utilized the Affymetrix Human Genome U133 Plus 2.0 Array [HG-U133 plus 2] platform. GSE 36765 consisted of 30 breast cancer tumor tissue samples and 4 normal tissue samples; GSE20086 consisted of 6 breast cancer tumor tissue samples and 6 normal samples; GSE10810 consisted of 31 breast cancer samples and 27 normal samples. A total of 67 breast cancer tumor tissues samples and 37 normal tissue samples were included in this study. The detailed analysis process is shown in **Figure 1.**

Identification of DEGs

The identification of DEGs in selected datasets was conducted using the GEO2R tool, which integrates the GEO query and limma R software packages from the Bioconductor project. Through this analysis, the three GEO datasets were examined, and genes with a *P* value< 0.05 and a fold change >1.2 were determined as DEGs. Subsequently, a Venn diagram was constructed using the online tool Venny (v2.1.0) to identify the DEGs that were consistently present across all three datasets.

GO and KEGG pathway analysis

The GO terms (http://www.geneontology.org) and KEGG pathways (http://www.genome.jp) were identified and analyzed using the DAVID Functional Annotation Tool (v6.8) with the identifier parameter set as "official_gene_symbol" and the species parameter set as "Homo sapiens". A significant enrichment was defined as a *P* value< 0.05. The results obtained from DAVID were visualized using the ggplot2 package in R language (v3.6.3).

PPI network analysis

STRING (v11.0), which incorporates known and predicted interactions among over 932, 000, 000 proteins from various organisms, encompassing 24, 584, 628 proteins from 5090 organisms [30]. The DEGs were uploaded for modeling of multiple proteins, and the organism was set as "*Homo Sapiens*". The statistical significance of the network interaction relationship was determined based on an interaction score > 0.4, and disconnected nodes in the network were excluded. The high-confidence interaction relationships were imported into the Cytoscape software (v3.6.0) for visualization of gene interactions [11]. The cytoHubba plugin was utilized to identify the hub genes in the PPI network, which were subsequently chosen as candidate DEGs for subsequent experiments.

Expression and survival analyses for the candidate DEGs

GEPIA (http://gepia.cancer-pku.cn/) is an online database that encompasses comprehensive RNA sequencing expression data derived from 9736 tumors and 8537 normal samples sourced from the TCGA and GTEx projects [31]. GEPIA is structured around three primary functional modules: namely single gene analysis, cancer type

analysis, and multiple gene analysis. These modules enable the identification of differential gene expression between tumor and normal tissues, survival analysis, gene correlations, and other related analyses. In this study, we conducted Kaplan–Meier survival analysis to examine the association between the relative expression of candidate DEGs in patients with breast cancer and the overall survival time. Hazard ratios (HRs) and corresponding 95% confidence intervals were calculated to elucidate the relationship between patient survival and high/Low gene expression, thereby providing insights into the roles of genes in disease development. Multiple gene comparisons and principal component analysis were employed to visually assess the discriminatory ability of candidate genes in distinguishing between tumor and normal models.

The UALCAN (http://ualcan.path.uab.edu/index.html) is an online database that specializes in transcriptome and clinical data derived from TCGA data. It facilitates the examination of differential expression patterns between tumor and normal tissues, as well as the exploration of tumor stage, lymph node metastasis, and other pertinent related clinical parameters [6]. In this study, the UALCAN database was utilized to validate the expression of DEGs in both breast cancer and normal tissues.

RESULTS

Identification of DEGs

This study incorporated three gene sets, namely (GSE36765, GSE20086, and GSE10810), of which GSE36765 comprised of 30 tumor samples and 4 normal samples, GSE20086 consisted of 6 tumor samples and 6 normal samples, and GSE10810 encompassed 31 tumor samples and 27 normal samples. In comparison to the normal samples, a total of significant DEGs were identified across all datasets **(Figure 2),** including 87 upregulated genes, 120 downregulated genes, and 24 genes exhibiting both upregulation and downregulation.

Enrichment analysis of DEGs

Functional enrichment analysis of the obtained DEGs was performed using DAVID. The GO enrichment analysis primarily aimed to predict the functions of the target genes based on biological processes, cell components, and molecular functions. Utilizing DAVID, the enrichment analysis revealed that several biological processes, including signal transduction, negative regulation of apoptotic process, negative regulation of cell proliferation, protein stabilization, positive regulation of cell migration, positive regulation of fibroblast proliferation, and positive regulation of MAPK cascade, were significantly enriched among the DEGs **(Figure 3a).** In the category of cell components, the DEGs were found to be enriched in terms such as extracellular exosome, cell–cell junction, vesicle, plasma membrane, and secretory granule membrane, as depicted in **(Figure 3b).** Furthermore, the DEGs were associated with molecular protein binding, cadherin binding, MHC class II protein complex binding, signaling adaptor activity, and protein binding involved in heterotypic cell–cell adhesion, as shown in **Figure 3d**.

The following 16 KEGG pathways were enriched among the 231 DEGs: inflammatory bowel disease, central carbon metabolism in cancer, melanoma, prostate cancer, glioma, proteoglycans in cancer, viral carcinogenesis, transcriptional deregulation in cancer, Epsrein-Barr virus infection, human T-cell leukemia virus 1 infection, leishmaniasis, osteoclast differentiation, protein processing in endoplasmic reticulum, HIF-1 signaling pathway, EGFR tyrosine kinase inhibitor resistance, and inositol phosphate metabolism **(Figure 3c)**.

Identifying central genes in the PPI network

A PPI network was constructed based on the 231 significant DEGs, and the hub genes were identified using the STRING database and Cytoscape software. The PPI network **(Figure 4a)** encompassed a total of 169 nodes and 290 edges with any disconnected nodes being concealed. An interaction score exceeding 0.4 was deemed indicative of a high-confidence interaction relationship. The cytohubba module was employed to identify the genes with the highest degree of connectivity, as illustrated in **Figure 4b**. The set of genes that exhibited higher connectivity included CALR, CHD4, EEF1A1, EGFR, HSPB1, IGF1, IL1R1, KLF4, MANF, PKM, PTPRC, SEC11C, SOCS3, PIK3R1, and

TPI1. Gene expression profiles of the 15 central genes in breast cancer tumor *vs* normal samples identified using GEPIA are shown in **Figure 5**. Significantly differential was observed for CALR, EGFR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1 between the breast cancer tumor and normal groups. Thus, analyses of the correlations between these eight genes and the pathological stage of breast cancer were performed using GEPIA. The expression levels of CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1 were found to be significantly associated with the pathological stage of breast cancer (*P* value < 0.05), whereas EGFR exhibited no significant correlation (*P* value >0.05) **(Figure 6).** Therefore, this study focused on the seven genes CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1.

Functional enrichment analysis of DEGs in breast cancer

In order to enhance the understanding of the functions of the 231 DEGs derived from the three datasets, functional and pathway enrichment analyses were conducted using DAVID. Within the core BPs, particular attention was given to those that are closely associated with the development and advancement of cancer, including signal transduction, inactivation of MAPK activity, positive regulation of tumor necrosis factor production, regulation of cell migration, and regulation of fibroblast proliferation **(Figure 7)**, the diagrammatic figure based on the enrichment pathway analysis was provided in **Supplementary Figure 1.**

Overall survival and disease-free survival analyses

The overall survival and disease-free survival of patients with breast cancer were analyzed using the GEPIA database, with a focus on the expression of seven hub genes. The results indicated that there was no association between the expression of CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, or TPI1 and either overall survival or disease-free survival **(Figure 8)**. Only the expression of the reserve gene CDH4 was found to be significantly associated with overall survival in patients with breast cancer (*P* value< 0.05) **(Supplementary Figure 2**.

Correlation analysis based on GEPIA

Correlation analysis was conducted on the expression levels of seven candidate genes

in patients with breast cancer. Moderate positive correlations, with a threshold of R <0.5, were observed between CALR and TPI1, IGF1 and IL1R1, IGF1 and KLF4, IGF1 and SOCS3, and KLF4 and SOCS3 **(Figure 9)**.

Verification of the differential expression of CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1

UALCAN analysis revealed significant differences in the expression of these seven candidate genes between the normal and tumor groups, confirming their potential as marker genes **(Figure 10)**. Based on the reference group of normal samples, all seven genes exhibited significant deregulation in expression when considering age group, race, and lymph node metastasis status. In terms of age **(Supplementary Figure 3**, CALR, HSPB1, and TPI1 demonstrated high expression across all age groups. Notably, there were significant variations in CALR, HSPB1, and IGF1 expression between the age groups of 41–60 and 81–100 years, as well as in IGF1 and TPI1 expression between the age groups of 41–60 and 61–80 years. Additionally, IL1R1 expression exhibited significant differences between the age groups of 21–40 and 61–80 years. Regarding race **(Supplementary Figure 4),** our study specifically examined gene expression differences between Asian and non-Asian countries, given our focus on China. Notably, the expression levels of IGF1, SOCS3, and IL1R1 were found to be significantly reduced in Asian countries compared to non-Asian countries, whereas KLF4 expression was observed to be increased. Additionally, when considering the node metastasis status, significant variations in the expression of these genes were observed across the N0, N1, N2, and N3 stages **(Supplementary Figure 5)**.

Possibility of a seven-gene diagnostic biomarker for of breast cancer

Comparative analyses were conducted using GEPIA to evaluate seven potential biomarkers for breast cancer, based solely on tumor data. Among the seven genes, CALR exhibited the highest expression level among the seven genes, followed by HSPB1, TPI1, IL1R1, KLF4, SOCS3, and IGF1 **(Figure 11)**. Principle component analysis was performed on the seven genes using both breast cancer tumor data and normal tissue data. The results demonstrated that the seven genes cohort effectively

distinguished breast cancer samples from normal samples **(Figure 12)**, thereby supporting the utilization of this seven-gene biomarker for the diagnosis of breast cancer. To further validate the rationality and predictive capacity of biomarkers, we conducted an assessment of the expression of a gene set consisting of CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1. This evaluation aimed to differentiate between patients with breast cancer and controls using the receiver operating characteristic (ROC) curve. The area under the curve (AUC) value of 1 provided evidence that the selected biomarker could effectively distinguish the model.

DISCUSSION

Breast cancer exhibits the highest global incidence among all cancers and is the primary contributor to cancer-related morbidity and mortality among females [16]. A growing body of evidence suggests that various factors, including genetic and environmental factors, may contribute to the onset and progression of breast cancer. Some lifestyle factors, such as excessive nutrition, obesity, a high-fat diet, and excessive alcohol consumption, have been found to impact the occurrence of breast cancer [2, 17]. The symptoms of early-stage breast cancer may not be readily apparent, making it easy to overlook breast lumps, abnormalities in breast skin, and other symptoms [1]. Detecting symptoms during the mid- or advanced stages of breast cancer poses significant challenges for treatment. The recommended multidisciplinary treatment approach encompasses surgical intervention, radiotherapy, neoadjuvant therapy, and adjuvant therapy [13]. Targeted therapy, as an emerging modality, offers the advantages of specificity, notable efficacy, and reduced incidence of side effects. Currently, the molecules targeted in the treatment of breast cancer primarily consist of HER-2, VEGF, EGFR, PARP, PI3K/Akt/mTOR, and CDK4/6 [9, 10, 23, 25]. In patients with HER2-positive early-stage breast cancer who experienced invasive cancer post-neoadjuvant therapy, the T-DM1 adjuvant group exhibited a 50% lower risk of recurrence of death from invasive breast cancer compared to the trastuzumab group [22]. Nevertheless, the early detection, reduction of recurrence, and improvement of overall survival continue to

pose challenges in the clinical management treatment of breast cancer. Hence, the identification of novel biomarkers capable of predicting the recurrence of breast cancer and overall survival assumes significance, as it enables the classification of individuals into high and low risk groups based on these markers, thereby enhancing the effectiveness of subsequent treatment interventions. Numerous clinical studies pertaining to tumor recurrence are available in public databases. In this particular investigation, our attention was directed towards studies encompassing both tumor and adjacent normal tissue samples, with the aim of discerning genes linked to overall survival in patients afflicted with breast cancer.

The analysis incorporated three datasets (GSE36765, GSE20086, and GSE10810) sourced from the GEO database. Among these datasets, a total of 231 DEGs were identified as common, comprising 87 upregulated DEGs, 120 downregulated DEGs, and 24 DEGs exhibiting both up and downregulation. The chosen database exhibited an appropriate sample size, and the distribution of DEGs were deemed reasonable.

The present study reveals a significant association between the gene expression of CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, CHD4, and TPI1 and the prognosis of breast cancer. Specifically, only CHD4 expression demonstrated a correlation with overall survival in breast cancer patients, whereas the expression of CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3 and TPI1 did not exhibit any significant relationship with overall survival. Notably, a previous investigation documented the overexpression of CALR in breast cancer tissue compared to normal tissue, and further established its association with morality and the stemness index. Furthermore, the depletion of CALR results in the impairment of breast cancer stem cells, thereby impacting tumor initiation and metastasis, and augmenting the sensitivity to chemotherapy [21]. Our investigation demonstrated a significant association between aberrant CALR expression and lymph node metastasis. Previous research has also indicated that CALR exhibits potential as a biomarker of tumor status, exhibiting high expression in oral squamous cell carcinoma, pancreatic cancer, gastric cancer, and esophageal squamous cell carcinoma [7, 19, 35]. Clinical studies have demonstrated that the overexpression of HSPB1 and IGF1 in

breast cancer has an impact on disease outcome and the responsiveness of tumors to chemotherapy and radiotherapy [8]. Another study reported that the interference of the IL6 pathway through SOCS3 or IL6R inhibits tumor growth and metastasis in mouse xenograft models [18]. Increased TPI1 expression has been associated with a poor prognosis in patients with lung adenocarcinoma and has influenced the infiltration of immune cell [34]. In this study, we have discovered that these abnormally expressed genes not only contribute to the initiation and progression of tumors but also have a significant relationship with the pathological stage of tumors and lymph node metastasis. Our study revealed that the expression of reserve CDH4 was solely linked to the overall survival of breast cancer. Furthermore, the expression of CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1 exhibited significant associations with the pathological and lymph node metastasis stages of breast cancer.

CONCLUSION

+ADw-html+AD4APA-p+AD4-These gene serve as crucial indicators for evaluating the prognosis and guiding the treatment of breast cancer. Notably, the expression levels of these seven genes effectively differentiated breast cancer samples from normal samples in a principal component analysis. The integration of these seven genes into a biomarker panel holds potential for enhancing the accuracy of breast cancer diagnoses. By discerning the specific locations where aberrant gene expression triggers tumorigenesis, a theoretical foundation can be established for the advancement of gene-level targeted therapies with greater precision.+ADw-/p+AD4APA-/html+AD4-

ARTICLE HIGHLIGHTS

Research background

Breast cancer is widely recognized as a highly malignant neoplasm in women, posing a significant risk to their overall health. The diagnostic challenges of breast cancer arise from the heterogeneity of samples and the limitations of conventional techniques. Consequently, the identification of more stable biomarkers assumes paramount

importance in facilitating early breast cancer screening. In this context, bioinformatics methods have been employed to detect differentially expressed genes associated with proliferation, invasion, apoptosis, and overall survival in breast cancer.

Research motivation

To ascertain fundamental prognostic biomarkers in breast cancer, three databases were queried for genes associated with breast cancer as tumor markers.

Research objectives

Bioinformatics analysis of the molecular mechanism involved in breast cancer revealed that seven differentially expressed genes (DEGs) (CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1) play critical roles in the progression of breast cancer. Bioinformatics were used to identify hub genes and enrichment pathways in breast cancer, illustrating a biological relationship between the pathways and gene expression in breast cancer.

Research methods

Microarray data information, data processing of DEGs, PPI network and module analysis, and survival analysis were used to mine potential biomarkers. In addition, pathway enrichment analysis was also conducted to elaborate the pathogenesis of disease.

Research results

Three GEO datasets that included breast cancer tissues and normal tissues were analyzed; 231 DEGs were identified. 7 potential biomarkers (CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1) were closely related to the occurrence and progression of breast cancer. The discrimination of potential markers for the model is also relatively perfect, and the discrimination for cancer group and normal group is 100%.

Research conclusions

Through the utilization of bioinformatics analysis, the molecular mechanism of breast cancer was investigated, revealing the significant involvement of seven differentially expressed genes (CALR, HSPB1, IGF1, IL1R1, KLF4, SOCS3, and TPI1) in the advancement of breast cancer. These findings hold promise in enhancing our understanding of breast cancer pathogenesis, as well as in the identification of novel biomarkers and potential drug targets, thereby facilitating advancements in breast cancer diagnosis and therapeutics.

Research perspectives

Bioinformatics was employed to identify hub genes and significant pathways in breast cancer, thereby establishing a biological association between the pathways and gene expression potentially implicated in the advancement of breast cancer. The utilization of bioinformatics analysis revealed the relevant genes and cellular pathways implicated in the genesis and progression of breast cancer.

ACKNOWLEDGEMENTS

We thank the Gene Expression Omnibus database(http://www.ncbi.nlm.nih.gov/geo/) for providing us with analytical data.

86495_Auto_Edited.docx

ORIGINALITY REPORT

2%
SIMILARITY INDEX

PRIMARY SOURCES

1	f6publishing.blob.core.windows.net Internet	21 words — 1%
2	www.wjgnet.com Internet	18 words — < 1%
3	www.mdpi.com Internet	14 words — < 1%
4	Gonca G. Bural, Drew A. Torigian, Wichana Chamroonrat, Khaled Alkhawaldeh, Mohamed Houseni, Ghassan El-Haddad, Abass Alavi. "Quantitative assessment of the atherosclerotic burden of the aorta by combined FDG-PET and CT image analysis: a new concept", Nuclear Medicine and Biology, 2006 Crossref	13 words — < 1%
5	www.ncbi.nlm.nih.gov Internet	12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS