Identification of Potential Target Genes in HER-2 Positive Breast Cancer by Bioinformatics Analysis.

Song Wang
Southwest Medical University

Yi Quan (✉️ 415974602@qq.com )
southwest medical university https://orcid.org/0000-0003-1724-6894

Hongying Lyu
Southwest Medical University

Jian Deng
Southwest Medical University

Research

**Keywords:** HER-2 positive breast cancer, Differentially expressed genes, Bioinformatics analysis, Target genes, Protein-Protein interaction

**DOI:** https://doi.org/10.21203/rs.3.rs-34161/v1

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

**Background:** HER-2 positive breast cancer has a high risk of for relapse, metastasis and drug resistance, and is correlated with a poor prognosis. Thus, the study objective was to reveal target genes and key pathways in HER-2 subtype breast cancer.

**Methods:** The gene expression dataset (GSE29431) was downloaded from the Gene Expression Omnibus database (GEO), and the differentially expressed genes (DEGs) were determined using LIMMA package in R software. Subsequently, Functional enrichment analysis were performed in ClusterProfiler package of R platform. The Search Tool for the Retrieval of Interacting Genes (STRING) database was used to construct a Protein-Protein Interaction (PPI) network of DEGs. Module analysis and target genes were identified by Cytoscape software. Further more, The influence of target genes on overall survival (OS) was assessed using the Kaplan-Meier plotter database.

**Results:** The differential expression analysis revealed 96 genes were up-regulated while 407 genes were down-regulated in HER-2 positive breast cancer tissue compared to normal breast tissue. Functional enrichment analysis showed that the DEGs were mainly involved in regulation of lipid metabolic process, PPAR signaling pathway and PI3K-Akt signaling pathway. PPI network construction revealed a total of 199 nodes and 560 edges, and 12 target genes were identified by the highest value of degree. In addition, target genes were associated with worse overall prognosis, including NUSAP1, PTTG1, CEP55, TOP2A, CCNB1, CENPF, MELK, AURKA, UBE2C, BUB1B, KIF20A and RRM2.

**Conclusion:** The present study identified 12 target genes associated with the development of HER-2 subtype breast cancer, which may help to provide new biomarkers and therapeutic targets.

Background

Breast cancer is the most common cancer in women all over the world [1], and has become a major cause of morbidity [2]. In china, the incidence rate of breast cancer has increased sharply in recent years, severely threatening chinese womens health and life quality [3]. As is known to all, Breast cancer is an intrinsically heterogeneous disease and has different molecular subtypes and biological characteristics. HER2-positive breast cancer accounts for about 20% of all breast cancer and is closely associated with highly aggressive clinical behavior and poor prognosis, due to HER-2 gene amplification or protein overexpression [4, 5]. Trastuzumab plus chemotherapy have become the first-line systemic treatment, which significantly prolongs overall survival (OS) in breast cancer patients [6, 7]. However, due to the development of primary or secondary resistance to targeted drugs, some patients still experienced disease recurrence or metastasis after targeted therapy [8]. Therefore, it is necessary to explore the underlying molecular mechanism and identify target genes in HER-2 subtype breast cancer, which contributes to search for novel therapeutic targets for the treatment.

The Gene Expression Omnibus (GEO) is a public repository for microarray expression data supported by the National Center for Biotechnology Information (NCBI), containing a great deal of patients across all
major tumor types.

In the present study, gene expression dataset (GSE29431) was obtained in the Gene Expression Omnibus (GEO). Subsequently, We performed differential gene expression analysis by LIMMA package between the HER-2 positive breast cancer and normal breast tissue. With these selected DEGs, Gene Ontology (GO) functional and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed, and the protein-protein interaction (PPI) network was constructed. Cytoscape software was applied to conduct module analysis and select target genes with a high degree of connectivity. Finally, the expression level and prognostic value of target genes were further assessed using UALCAN and Kaplan-Meier plotter database.

**Methods**

**Microarray Expression Data Source**

Microarray dataset (ID Number: GSE29431) based on GPL570 platform of Affymetrix Human Genome U133 Plus 2.0 Array [HG-U133_Plus_2] was obtained from GEO database (https://www.ncbi.nlm.nih.gov/geo/), including 28 HER-2 positive breast cancer samples and 12 normal breast tissue samples. The gene expression data was normalized using RMA algorithm [9].

**Identification of DEGs**

The identification of DEGs was performed using LIMMA package [10], with a criteria of adj P-value (Benjamin-corrected P-value) < 0.01 along with at least two-fold change. False discovery rate was controlled by correcting the P-values according to the Benjamini–Hochberg procedure. A volcano plot was produced using R package to show results of differential gene expression analysis.

**GO and KEGG pathway enrichment analysis of DEGs**

To obtain the biofunctions of the DEGs, we performed Gene Ontology (GO) annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis using the clusterProfiler package of R software [11]. With Benjamin-corrected p value < 0.05, GO terms and KEGG pathways were considered to be significantly enriched, and GO annotation contained biological process, molecular function and cellular component.

**PPI network construction and target gene identification**

The protein–protein interaction (PPI) analysis was carried out using the Search Tool for the Retrieval of Interacting Genes (STRING) database [12] (https://string-db.org/cgi/input.pl), with a combined score = 0.7 as the threshold. Subsequently, the PPI network was visualized by Cytoscape software [13]
We respectively used the Molecular Complex Detection (MCODE) [14] tool and CytoHubba [15], two plugin software of cytoscape, to performed module analysis of the PPI network and screen the target genes. What's more, a node degree of 15 was selected as the threshold.

**Expression analysis of target genes**

UALCAN [16] (http://ualcan.path.uab.edu) is a publicly interactive web resource used to validate the reliability of the identified target genes from the dataset (GSE29431). We utilized UALCAN to analyze the key genes transcript expression in the HER-2 positive breast cancer samples which is derived from the TCGA project.

**Survival analysis of target genes**

The Kaplan-Meier plotter [17](http://kmplot.com/analysis/) is a comprehensive online platform that can assess the effect of a great number of genes on survival based on TCGA database. To assess the association of target genes with survival in breast cancer, we conducted survival analysis using the TCGA-BRCA database. P-values were determined by log-rank test.

**Results**

**Identification of DEGs**

With a cutoff adj p value < 0.01 and fold change ≥ 2, We identified a total of 503 DEGs, including 96 up-regulated genes and 407 down-regulated genes. A volcano plot of the DEGs is shown in Figure. 1.

**GO and KEGG pathway enrichment analysis of DEGs**

In order to identify the important biological function in the specified gene set, we conducted functional enrichment analysis using clusterProfiler package. For GO analysis, the results showed that regulation of lipid metabolic process was mainly enriched for biological processes (BP), extracellular matrix was primarily enriched for cellular components (CC) and amide binding was chiefly enriched for molecular functions (MF). Further more, KEGG pathway analysis indicated that these DEGs were mainly enriched in PPAR signaling pathway and PI3K-Akt signaling pathway. The results of functional enrichment analysis were shown in Figure. 2.

**PPI network construction and target genes identification**

To identify potential interactions between DEGs, The PPI network was constructed by STRING database and visualized by Cytoscape software. With a combined score > 0.7, a PPI network(Figure. 3) containing
199 nodes and 560 edges was constructed. As was shown in the Figure. 4, the most significant module totally consisted of upregulated genes was identified using the MCODE plugin. Subsequently, with the highest connectivity degree = 15, 12 genes selected from this module were regarded as target genes, including NUSAP1, PTTG1, CEP55, TOP2A, CCNB1, CENPF, MELK, AURKA, UBE2C, BUB1B, KIF20A and RRM2. These genes were shown in the Table 1.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Gene Description</th>
<th>Expression</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUSAP1</td>
<td>Nucleolar and spindle associated protein 1</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>PTTG1</td>
<td>Regulator Of Sister Chromatid Separation</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>CEP55</td>
<td>Centrosomal protein 55</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>TOP2A</td>
<td>DNA topoisomerase 2-alpha</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>CCNB1</td>
<td>Cyclin B1</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>CENPF</td>
<td>Centromere Protein F</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>MELK</td>
<td>Maternal Embryonic Leucine Zipper Kinase</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>AURKA</td>
<td>Aurora Kinase A</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>UBE2C</td>
<td>Ubiquitin Conjugating Enzyme E2 C</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>BUB1B</td>
<td>BUB1 mitotic checkpoint serine/threonine kinase B</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>KIF20A</td>
<td>Kinesin family member 20A</td>
<td>UP</td>
<td>15</td>
</tr>
<tr>
<td>RRM2</td>
<td>Ribonucleotide reductase regulatory subunit M2</td>
<td>UP</td>
<td>15</td>
</tr>
</tbody>
</table>

**Expression analysis of target genes in the UALCAN database**

Based on TCGA RNA-seq data available from UALCAN database, expression analysis demonstrated that the mean expression values of target genes were significantly higher in HER-2 positive breast cancer
tissues compared with adjacent normal tissues (P < 0.05). So, these results (Fig. 5.) further validate reliability of the identified target genes from the GSE29431.

**Survival analysis of target genes**

We analyzed whether target genes was associated with survival in breast cancer patients using the Kaplan-Meier plotter database. Finally, We identified CCNB1, CENPF, NUSAP1, CEP55, PTTG1, TOP2A, MELK, AURKA, UBE2C, BUB1B, CENPF, AURKA, UBE2C, BUB1B, KIF20A and RRM2, which were associated with worse overall survival situation (Figure 6.), and the difference was statistically significant (logrank p < 0.05).

**Discussion**

Breast cancer is a highly heterogeneous malignant tumor, whose biological behavior is closely related to her-2 status. Her-2 subtype breast cancer often presents with disease relapse and metastasis, and the survival rate is low. In recent years, targeted drugs therapies have improved overall survival for HER-2 subtype breast cancer patients [18]. Despite this standardized initial therapy, with the development primary and secondary resistance to targeted drugs, some patients could rapidly developed disease progression [19]. Thus, further studies will be needed to explore the pathogenesis and search new therapeutic targets.

In this study, we identified 503 DEGs in the GSE29431 using LIMMA package. Among 503 DEGs, 96 genes were up-regulated and 407 genes were down-regulated The GO enrichment analysis showed that DEGs were mainly related to regulation of lipid metabolic process (BP term), while the KEGG pathway enrichment analysis showed that DEGs were mainly enriched PPAR signaling pathway and PI3K-Akt signaling pathway. Using STRING database and Cytoscape software, we constructed PPI network and identified 12 target genes (CCNB1, CENPF, NUSAP1, CEP55, PTTG1, TOP2A, MELK, AURKA, UBE2C, BUB1B, CENPF, AURKA, UBE2C, BUB1B, KIF20A and RRM2). And the reliability of 12 target genes was further validated by UVALCAN database. Finally, using Kaplan-Meier plotter, we found that over-expression of key genes were associated with poor prognosis.

The lipid metabolism is closely associated with breast cancer[20,21]. It has been reported the use of statins reduced the chance of recurrence. in breast cancer [22]. Activated glutamine metabolism in HER-2 positive breast cancer had clinical implication as a potential therapeutic target [23]. PPAR pathway plays a major role in regulating cancer cell proliferation. Ham et al [24] found that the PPAR with c-Myc governed the tumorigenicity of breast cancer. However, it is unclear whether PPAR signaling pathway is involved in targeted drug resistance. Further studies are needed to validate this. The PI3K/AKT pathway, as an important intracellular signaling path, participated in the development of breast cancer [25]. Studies have suggested that inhibition of PI3K/AKT/mTOR pathway can improve resistance to anti-HER2 therapies. According to BOLERO-3 trial [26], the addition of everolimus, a mTOR inhibitor, significantly
prolonged PFS in trastuzumab resistance in HER2-overexpressing breast cancer (7.0 vs 5.8 months, p = 0.0067).

NUSAP1 is a microtubule-associated protein, which take a important part in the spindle assembly and cell proliferation. NUSAP1 has been shown to be closely related to the development of breast cancer and renal cell carcinoma [27,28]. Knockdown of NUSAP1 inhibited the migration, proliferation and invasion of IBC cells [29]. Pituitary tumor-transforming gene 1 (PTTG1) serves as a proto-oncogene and regulates sister chromatid separation [30]. Additionally, PTTG1 overexpression, commonly observed in many malignant tumor, promoted the invasive ability of tumor cells [31-32]. According to the expression pattern of PTTG1 in primary and metastatic breast cancer, it was expected to be an important molecular target in breast cancer treatment [33]. Studies showed silencing PTTG1 gene was an effective method for the treatment of liver cancer [34]. CEP55 is a key regulator of cytokinesis, and was associated with genomic instability [35]. Notably, the proliferative potential and invasive ability of the breast cancer cells were markedly reduced when CEP55 was down-regulated [36]. TOP2A gene is located on human chromosome 17q21, and is closely related to invasive and proliferative potential. Clinical studies have showed that TOP2A played a predictive role for anthracycline-based chemotherapy [37]. CCNB1, known as cyclin B1, is a highly conserved cyclin. Testing cyclin B1 in patients treated with T-DM1 was beneficial to identify early the patients more likely to benefit from this drug [38]. CENPF is a cell cycle-regulated protein associated with kinetochores, and is overexpressed in breast cancer, hepatocellular carcinoma and other tumors [39-40]. CENPF was reported to be a novel target protein contributing to the anti-tumour effects of zoledronic acid [41].

MELK is a serine/threonine kinase in the snf1/AMPK family, and is associated with undifferentiated phenotype, poor prognosis, chemoresistance, and Radioresistance [42]. Many studies showed MELK was a significantly therapeutic target for breast cancer [43]. AURKA, a human Aurora kinase, can promote cell cycle and suppress cell apoptosis. By stimulating the PI3K/AKT/mTOR pathway, overexpressed AURKA activated the development of chemotherapeutic resistance [44,45]. UBE2C, as a member of the ubiquitin-conjugating enzyme family (E2), was associated to the progression of cancers [46]. Mo et al [47] reported UBE2C expression was positively correlated with HER2 expression (P<0.05). Additionally, the suppression of UBE2C could inhibit growth of BRCA cells and sensitized breast cancer cells to doxorubicin [48]. BUB1 mitotic checkpoint serine/threonine kinase B (BUB1B) encodes a kinase participating in the spindle checkpoint, and is in correlation with distant metastasis and progression of breast carcinoma [49].

Kinesins (also known as KIFs) are a superfamily of molecular motors engaged in mitosis and migration. Khongkow et al identified KIF20A was involved in paclitaxel action and resistance [50]. Ribonucleotide reductase M2 subunit (RRM2), a rate-limiting enzyme for DNA synthesis and repair, was associated with cell proliferation, invasiveness and migration [51]. Previous study suggested that targeting RRM2 may be a novel strategy for breast cancer treatment. Small molecular antagonist of RRM2 gene significantly reduced tamoxifen-resistant cell proliferation and decreased tumor growth [52]. Suppression of RRM2 synthesis could enhance the chemosensitivity to adriamycin [53].
In present study, using bioinformatics analysis, we identified 12 potential targeted genes in HER-2 positive breast cancer, and these findings provided new targets for diagnosis and treatment. However, animal experiments and population-based studies are necessary to confirm the results in this study.

**Abbreviations**

BP: Biological process; CC: Cellular component; DEGs: Differentially expressed genes; GEO: Gene Expression Omnibus; GO: Gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; MF: Molecular function; NCBI: National Center for Biotechnology Information; OS: Overall survivals; PPI: Protein–protein interaction;

**Declarations**

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Availability of data and materials**

The dataset (GSE29431) analyzed in the present study was downloaded from the GEO database(http://www.ncbi.nlm.nih.gov/geo/).

**Competing interests**

The authors declare that they have no competing interests.

**Funding**

No funding

**Authors' contributions**

SW came up with the idea and performed the statistical analysis. HL and JD wrote the manuscript. YQ reviewed and revised the paper. All authors read and approved the final manuscript.
Acknowledgements

Not applicable.

References


Figures

Cutoff for logFC is 2
The number of up gene is 96
The number of down gene is 407

Figure 1

Volcano plot of 503 DEGs in GSE29431 Each dot represents an individual gene (Red dots presented up-regulated genes with logFC>2 and adj p value < 0.01, while blue dots presented down-regulated genes with logFC<2 and adj p value < 0.01).
Figure 2

GO and KEGG pathway functional enrichment analysis of DEGs. The results in the enrichment analysis were classified into four groups: (A) Cellular component, (B) biological processes, (C) molecular function and (D) KEGG pathway.
Figure 3

Protein-protein interaction network of DEGs The Red circle presented up-regulated genes, while the black square presented down-regulated genes.
Figure 4

The most significant module from the PPI network. The Red circle presented upregulated genes.
The expression analysis of target genes in the UALCAN database. The Blue boxplot represented normal breast tissue samples, while green boxplot represented HER-2 positive breast cancer samples. These target genes were significantly overexpressed in the HER-2 positive tumor tissues, including (A) CCNB1 (P=7.17E-12); (B) CENPF (P=6.26E-08); (C) NUSAP1 (P=1.77E-12); (D) CEP55 (P=4.11E-09); (E) PTTG1.
(P=5.29E-13) ; (F) TOP2A (P=1.04E-04) ; (G) MELK (P=1.56E-09) ; (H) AURKA (P=1.54E-09) ; (I) UBE2C (P=3.46E-10) ; (J) BUB1B (P=7.15E-12) ; (K) KIF20A(P=8.02E-12) ; (L) RRM2 (P=1.37E-12).
**Figure 6**

Overall survival analysis of target genes in the breast cancer patients. The prognosis value of 12 target genes: (A) CCNB1; (B) CENPF; (C) NUSAP1; (D) CEP55; (E) PTTG1; (F) TOP2A; (G) MELK; (H) AURKA; (I) UBE2C; (J) BUB1B, also known as SSK1; (K) KIF20A; (L) RRM2.