Original Article
Bioinformatic analysis of differential expression and core GENEs in breast cancer

Hongchang Dong1, Shuai Zhang1,*, Yu Wei2,*, Chunyan Liu2, Na Wang1, Pan Zhang1, Jingling Zhu1, Jin Huang1

1The Key Laboratory of Xinjiang Endemic & Ethnic Diseases and Department of Biochemistry, Shihezi University School of Medicine, Shihezi, Xinjiang, China; 2The First Affiliated Hospital of Medical College of Shihezi University, Shihezi, Xinjiang, China. *Equal contributors.

Received November 27, 2017; Accepted December 22, 2017; Epub March 1, 2018; Published March 15, 2018

Abstract: Breast cancer (BRCA) is one of the most common malignancies in women. The gene expression profile of GSE103512 from the GEO database was downloaded in order to find key genes involved in the occurrence and development of BRCA. 75 samples, including 65 cancer and 10 normal samples, were included in this analysis. Differentially expressed genes (DEGs) between BRCA patients and health people were chosen using R tool. We next performed gene ontology (GO) analysis and Kyoto Encyclopedia of Gene and Genome (KEGG) pathway analysis using the Database for Annotation, Visualization and Integrated Discovery (DAVID). Moreover, Cytoscape with Search Tool for the Retrieval of Interacting Genes (STRING) was utilized to visualize protein-protein interaction (PPI) of these DEGs. The related genes and medicines specific to hub genes were predicted by CBioportal. We screened a total of 357 DEGs including 77 up-regulated and 280 down-regulated. A series of BRCA related GO terms and pathways were identified by analysis of these DEGs. Insulin-like growth factor 1 (IGF1); epidermal growth factor receptor (EGFR); v-jun avian sarcoma virus 17 oncogene homolog (JUN) and Estrogen Receptor 1 (ESR1) of the DEGs were screened by construction of the PPI network and the degree of connectivity. IGF1 and ESR1 were finally selected as potential hub genes and treatment targets of BRCA. In conclusion, this bioinformatics analysis demonstrated that DEGs and hub genes, such as IGF1, might regulate the development of gastric cancer. These DEGs could be used as new biomarkers for diagnosis and to guide the combination medicine of BRCA.

Keywords: Breast cancer, bioinformatics analysis, differential expression genes, biomarker, therapeutic

Introduction

Breast cancer is currently the main cause of death among women in the world [1]. Although the mortality of breast cancer has decreased, owing to the development of modern therapeutic approaches, the morbidity is still very high [2]. About 1,700,000 new cases of breast cancer and 521,900 deaths every year have been reported worldwide [3]. In China, especially in rural areas where people have lower income and less access to health care facilities, the morbidity of breast cancer has reached 24.20 per 100,000 [4]. In the United States, 231,840 women were diagnosed with and 40,290 died from breast cancer in 2015 [5]. Women of African descent show a lower incidence, but higher mortality rates and earlier age of onset [6]. Although China is a low-lying area relative to western countries, epidemiological data in recent years show that with the development of industrialization and the change of people's lifestyle, the overall incidence of breast cancer is increasing and the age of onset is decreasing [7, 8]. However, current diagnostic technologies, including imaging, molecular detection, and immunohistochemistry (IHC) have inherent limitations and may provide uncertain results [9, 10]. Thus, the early diagnosis and treatment of breast cancer are becoming more and more important.

Recently, many biomarkers and therapeutic target genes have emerged. Estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor-2 (HER2), and proliferation marker Ki-67 have been used to identify the subtypes and diagnose the breast cancer [2]. The most common treatment strategies today are endocrine therapy, chemothera-
Table 1. DEGs with \( P < 0.01 \) and |logFC| > 1

<table>
<thead>
<tr>
<th>EXPRESSION</th>
<th>GENE SYMBOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Down-regulated</td>
<td>HBA2///HBA1 HBA2///HBA1 LYVE1 ADH1B</td>
</tr>
<tr>
<td></td>
<td>HSPB6 HBB FABP4 PDK4 ADH1B CD36 CD36 HSPB6 HBA2///HBA1 CD36 PLIN1 CCL15-CCL14///CCL15///CCL14 HBA2///HBA1 FHL1 FHL1 LPL LYVE1 HBB GPX3 FOS CD36 FOSB FHL1 PDK4 NTRK2 PLA2G2A IGF1 SRPX IGSF10 IGF1 DPT TNXB//TNXB TXIST2 CCL18 GPX3N PIGR ADIPOQ EFEMP1 CXCL12 FHL1 FHL1 TFPI OGN FABP4 ALPL MMP19 PTGIS NTRK2 MT1M DST ENPP2 PDK4 CIDEC ABCA6 PIGR ADAMTS5 EGR1 PTGDS SVEP1 DPT MEG3 CD209 HBB IGFBP6 SYNMs CFD ALPL PTGIS CCL18 CAV1 C2orf40 MEG3 AOC3 SAA2///SAA1 MATN2 OGN SOCS3 PTGIS NTRK2 LINC01279 MYH11 ABCA8 VIM CCL28 SAA2///SAA1 CXCL2 ADAM33 EGR1 ANXA1 TGFBR3 TNXB///TNXA CCDC80 GALNT16 KIT CAV1 COL14A1 TFP1 ENPP2 DUSP1 FAM126A CRISP2L2 SLIT3 SCARA5 EBF1 IG1 MEG3 BTNL9 CHRD11 RHOJ ORS1 ITM2A IG1 FAP1 CAPN6 DPT///TNXB///TNXA CCL21 CEBPD MFAP4 CCL28 KLF4 PALMD CCDC80 COL14A1 SOD3 RGGC MYH11 TNXB ACACB SAMD5 SVEP1 ANK2 MYH11 MAO APDGFR A RRRN4CL G0S2 ANXA1 SYNP2 TSHZ2 TFPI SCARA5 CCDC80 PLTP ABCA9 SLC16A7 DUSP1 PTGDS MEST KLF4 EBF1 EMP1 ABI3BP PLPP3 COL14A1 DST PTGDS CXCL2 SOD2 PALMD RUNX1T1 JUN RHOJ PDGFD PLPP3 PDGFD TGFB2 RHOJ ACACB JUN ANK2 C7 EDNRE ACACB AOX1 F13A1 PCDH18 CD248 EBF1 DPT ITM2A ZFP36 GAS1 LPL ADAMTS11 LIFR PLPP3 SNORD114-3 RHOJ ADD3 TGFB3 AQP1 ITIH5 EGFRL COL14A1 CXorf36 GPAM CLEC3B C1R MYH11 CDR1 CCL13 MED AG SVEP1 ADGRA2 AOX1 FM02 ATF3 CFHR1///CFH SYNP2 TXSB FIGF ldb2 RND3 FLRT2 TTN ADAMTS4 CD209 EDNRE RHOU APCDD1 TNS1 CD163 GAS7 IGHD TBX18 MS4A4A KCNIP2 SYNP2 AQP1 TNS1 PLIN4 AGAR2 CD34 PRNP RHOJ TBX18 C1NFL1 DCLK1 ADAMTS5 ADD3 GNG11 FM02 PR0S1 EFEMP1 LHFP RSP03 ADAMTS5 KLF2 CAP26 LINR5B1 LINCO0341 KCNMB1 TF RCAN1 ANGPTL2 CDL11N ADD3 CD163 COL1SA1 CDC42EP5 PER1 MS4A4A GPM6B ISM1 Up-regulated</td>
</tr>
</tbody>
</table>
Core genes in breast cancer development

Int J Clin Exp Pathol 2018;11(3):1146-1156

py, and targeted therapy using antibodies recognizing cancer biomarkers; however, they are not always successful and may cause adverse effects and drug tolerance [11, 12]. Therefore, there is a constant need to investigate the development mechanism for breast cancer diagnosis and therapy.

High throughput sequencing is increasingly being widely used and it has been a very significant tool for life sciences, such as cancer grading, cancer early diagnosis, and prognosis prediction [12]. In this study, to further understand breast cancer, the GSE103512 was downloaded from GEO datasets and analyzed by R bioconductor for DEGs. Analysis of the biological process (BP), molecular function (MF), cellular component (CC), and KEGG pathways of the DEGs and three modules were performed. The regulatory network of these DEGs was constructed using string and the key genes with high degree of connectivity were selected by Cytoscape. Finally, the correlation between genes, survival, and disease was confirmed based on Kaplan Meier plotter online database (http://kmplot.com/analysis/) [13] and TCGA database.

Materials and methods

Data source and DEGs identification

The GSE103512 datasets including 65 breast cancer with 10 matched normal; 57 colorectal cancers with 12 matched normal; 60 non-small cell lung cancers with 9 matched normal; 60 prostate cancers with 7 matched normal, was downloaded from GEO. The breast cancer data were chosen to perform the differential expression genes analysis. The Bioconductor package of R was applied to detect the DEGs with cut-off criteria (adjust P value < 0.01 and |logFC| ≥ 1) [14]. A total of 357 genes with significant differential expression were selected after the analysis of GSE103512; of which, 77 were up-regulated genes and 280 were down-regulated (Table 1).

Gene ontology and KEGG pathway enrichment analysis

To identify characteristic biological alterations in high throughput genome and transcriptome data, gene ontology analysis (GO), an online tools containing numerous database, is a usual method for annotating genes and gene products [15]. Kyoto Encyclopedia of Genes and Genomes (KEGG) is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances [16]. DEGs from GSE were uploaded to DAVID to perform the BP (biological process), CC (cellular component), MF (molecular function) and pathway enrichment analysis. P < 0.05 was set as the cut-off criterion.

PPI network construction and core genes analysis

The STRING database (http://string-db.org) aims to provide a critical assessment and integration of protein-protein interactions, including direct (physical) as well as indirect (functional) associations [17]. DEGs from GSE103512 were uploaded to STRING for the known and predicted interactions analysis (confidence score ≥ 0.15, maximum number of interactors = 0). Next, the result was imported to Cytoscape for the core genes screening with degree cutoff = 25 [18].

Survival analysis of key genes

To assess the effect of 54,675 genes on survival using 10,461 cancer samples, 5,143 breast, 1,816 ovarian, 2,437 lung and 1,065 gastric cancer patients with a mean follow-up of 69/40/49/33 months [19] were analyzed. The primary purpose of the tool is as a meta-analysis based biomarker assessment [19]. In this study, key genes (ESR1, IGF1, EGFR, JUN) were assessed for their effect on survival with 95% confidence intervals and log rank P value < 0.01.

Correlation between hub genes and breast cancer

Study on the correlation between hub genes and breast cancer was performed by the online tools GEPIA (http://gepia.cancer-pku.cn/index.html). GEPIA is a web server for analyzing the RNA sequencing expression data of 9,736 tumors and 8,587 normal samples from the TCGA and the GTEx projects, using a standard processing pipeline [20].

Medicine and regulatory factors of hub genes prediction

CBioPortal (http://www.cbioportal.org/) website integrates the data of 126 tumor genome research, including TCGA and ICGC, covering
the data of twenty-eight thousand specimens. Here we used this website for prediction of medicine and regulation factors of hub genes based on such database including PhosphoSite, KEGG Drugs, pid, HumanCyc, Reactome, PANTHER and DrugBank [21].

Results

Screening of DEGs

To observe the alteration between breast cancer and normal tissues on the transcriptome level, 65 breast cancer and 10 matched normal mRNA expression values of GSE103512 datasets were analyzed by R conductor package. Cassette figures before (Figure 1A) and after (Figure 1B) data standardization are shown. We set the adjust P value < 0.01 and |logFC| ≥ 1 as cut-off criteria. The alignment of black dots on the same line indicates good standardization. Analysis of the mRNA expression fold-change of the DEGs showed a highly significant difference between BRCA and normal tissues (Figure 1C). These results revealed that the data can be directly used for further analysis. Finally, a total of 357 differential expressed genes were detected after the analysis of GSE103512, of which 77 were up-regulated genes and 280 were down-regulated (Table 1).

Identification of GO terms and pathways related to DEGs

The Database for Annotation, Visualization and Integrated Discovery is an online bioinformatics
Core genes in breast cancer development
resource that provides tools for the functional enrichment of large lists of genes or proteins [22]. To better understand breast cancer development, molecular function (Figure 2A), biological process (Figure 2B), cellular component (Figure 2C) and signaling pathway (Figure 2D) alteration in the disease were analyzed by DAVID.

**PPI network analysis reveals the hub genes among DEGs**

The potential main regulators may be the focal point for therapeutic and drug design. In this analysis, 77 up-regulated and 280 down-regulated genes were selected based on our criterion. To obtain the core genes during breast cancer development, we constructed the regulatory network by analysis of PPI and then IGF1 (degree = 36), EGFR (degree = 36), JUN (degree = 33), ESR1 (degree = 34) were selected as the hub genes (Figure 3).

**IGF1 and ESR1 alterations may play important role in breast cancer patient’s survival**

Survival analysis is widely used in clinical and epidemiological research, in randomized clinical trials for comparing the efficacy of treatments, and in observational (non-randomized) research to determine and test the existence of epidemiological association [22]. In our study, survival analysis was performed to assess the hub genes in patient’s survival. The results sug-
Core genes in breast cancer development

![Figure 4. Prognostic value of four genes (IGF1 (A), ESR1 (B), JUN (C), EGFR (D)) in BRCA patients. The desired Affymetrix IDs are valid: 209541_at (IGF1), 201464_x_at (ESR1), 202311_s_at (JUN), 1565483_at (EGFR). HR: hazard ratio, CI: confidence interval.]

gest that expression of IGF1 (HR = 0.75 (0.68-0.84) log rank $P = 4.2e-07$) (Figure 4A) and ESR1 (HR = 0.67 (0.6-0.74) log rank $P = 2.5e-13$) (Figure 4B) but not JUN (HR = 0.88 (0.79-0.98) log rank $P = 0.02$) (Figure 4C) and EGFR (HR = 0.87 (0.74-1.01) log rank $P = 0.074$) (Figure 4D) was associated with worse overall survival (OS) for breast cancer patients. To determine the correlation between hub genes and breast cancer, IGF1, and ESR1 expression was evaluated by GEPIA. Indeed, expression of IGF1 (Figure 5A) was down-regulated and ESR1 (Figure 5B) was up-regulated in breast cancer.

Various genes and medicines specific to IGF1 and ESR1 reveal IGF1 and ESR1 as potential targets for breast cancer treatment

To evaluate if IGF1 and ESR1 could be targets for breast cancer therapeutics, a website integrated PhosphoSite, KEGG Drugs, pid, HumanCyc, Reactome, PANTHER, and DrugBank databases, to analyze genes and drugs. In our results, 37 FDA approved drugs and 41 non-FDA approved drugs were selected for ESR1. Only Deoxy-Bigchap was selected for IGF1 (Figure 6). There were also many proteins that could regulate IGF1 and ESR in transport, phosphorylation, expression, and other potential ways (Figure 6).

Discussion

Approximately 232,340 new cases of invasive breast cancer and 39,620 breast cancer deaths are expected to occur among US women in 2013. Although exemestane is currently approved by the US Food and Drug Administration to prevent breast cancer recurrence, promising
results from clinical trials led the American Society of Clinical Oncology to include exemestane in their guidelines as a third option for chemoprevention [23]. Continued progress in the control of breast cancer will require sustained and increased efforts to provide high-quality screening, diagnosis, and treatment to all segments of the population.

Biomarker identification for any disease helps in the development of improved diagnostics and clinical treatment efficacy. Microarray data analysis has been widely used for disease biomarker discovery [23-26]. In this study, gene transcription alteration of breast cancer and normal in GSE103512 was analyzed. 77 up-regulated and 280 down-regulated DEGs were observed including an usual biomarker ESR1 in BRCA. These results indicate that these DEGs may be biomarkers of breast cancer. To have a better understanding of the biological contribution of these DEGs, MF, and BP analysis were executed. The response to cAMP, extracellular matrix organization, positive regulation of gene expression, cell chemotaxis, cell adhesion of BP and heparin binding, calcium ion binding, transmembrane receptor protein tyrosine kinase activity, transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding were identified, indicating the association of these GO term and breast cancer. In the cellular component analysis, most of these DEGs were enriched in GO terms such as extracellular space, extracellular region, proteinaceous extracellular matrix, extracellular exosome, extracellular matrix, cell surface, apical plasma membrane and membrane raft, etc. These GO terms are totally associated with membranes and suggest that secreted protein and membrane protein reactions such as some receptors and ligands interaction may play important roles in the breast cancer progression. Coincidentally, this speculation was assessed in the KEGG pathway analysis which contains the cytokine-cytokine receptor interaction term. Other pathways in our study may also associate with breast cancer development.

As the protein and protein interaction maybe essential for breast cancer, the PPI network of DEGs was constructed using STRING in our research. This data shows us a complicated regulation network in the breast cancer. But there should be some hub genes for breast cancer therapeutics, therefore, the core node with high connective degree was selected by Cytoscape. With this approach, 4 genes including IGF1, EGFR, ESR1 and JUN were selected for our analysis.

To assess the results of our analysis, survival analysis and expression comparison of the 4 genes was executed based on TCGA and other cancer databases. EGFR and JUN were without a significant effect on patient’s survival and were excluded in our study. Studies have shown that diabetes is a risk factor for breast cancer development and is associated with poor prognosis for breast cancer patients [27]. IGF1, a polypeptide protein similar to insulin molecular structure, its precursor could induce breast cancer cell proliferation via the IGF1 receptor [28]. However, there was no direct evidence for the effect of IGF1 on breast cancer diagnosis.
and therapeutics. Here, we demonstrate that the mRNA level of IGF1 was significantly decreased between BRCA tumor and normal tissues and seemed to be the core factor among these DEGs. There are many studies that have shown that mutation of ESR1 was essential for breast cancer and this gene has been applied as a breast cancer biomarker [2, 3, 29]. It also indicated the effectiveness of our analysis. In addition, the breast cancer patients in our study are ESR1 high expressers, but it is interesting that breast cancer patients with high ESR1 present higher survival than low ESR1 expression. We suspect this phenomenon is a self-protection mechanism against breast cancer and deserves experimental verification.

We also predict the potential drugs for breast cancer therapeutics according to our analysis, and this also highlights the potential of IGF1 and ESR1 as targets of breast cancer therapeutics. Screening of these drugs should be performed. The genes that regulate ESR1 transport, phosphorylation, expression, and other potential ways may also participate in the human self-protection against BRCA.

Our bioinformatics analysis identified DEGs that might play a central role in the occurrence, development, and prognosis of breast cancer. In this study, a total of 357 DEGs were selected, and EGFR, JUN, IGF1, and ESR1 might be the core genes of breast cancer. In order to get more accurate correlation results, we performed a series of verification experiments later to confirm the results of this prediction. Finally, IGF1 and ESR1 were selected as hub genes for diagnosis and treatment of breast cancer. Overall, this study provides some powerful evidence for future genomic individualized diagnosis and treatment of breast cancer.
Core genes in breast cancer development

Acknowledgements

We are grateful to Brouwer-Visser J, Cheng W, Bauer-Mehren A, Maisel D, Lechner K, Andersson E, Dudley JT, Milletti F who provided the GSE103580 dataset for this analysis, and it is our pleasure to acknowledge their contributions. This study was supported by the National Natural Science Foundation of China (No. 81771173 and 31060136).

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Jin Huang, The Key Laboratory of Xinjiang Endemic & Ethnic Diseases and Department of Biochemistry, Shihezi University School of Medicine, Shihezi 832002, Xinjiang, China. Tel: +86-(993) 205-7882; E-mail: huangjin623@163.com

References

Core genes in breast cancer development


