13 https://www.mycancergenome.org/content/disease/lung-cancer/

Received May 2, 2017
Revised May 19, 2017
Accepted May 25, 2017
High-efficient Screening Method for Identification of Key Genes in Breast Cancer Through Microarray and Bioinformatics

ZIHAO LIU1*, GEHAO LIANG1*, LUYUAN TAN1, AN SU1, WENGUO JIANG2 and CHANG GONG1,2

1Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetic and Gene Regulation, Breast Tumor Center, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, P.R. China; 2Cardiff China Medical Research Collaborative, Cardiff University School of Medicine, Cardiff University, Cardiff, U.K.

Abstract. Background/Aim: The aim of the present study was to identify key pathways and genes in breast cancer and develop a new method for screening key genes with abnormal expression based on bioinformatics. Materials and Methods: Three microarray datasets GSE21422, GSE42568 and GSE45827 were downloaded from the Gene Expression Omnibus (GEO) database and differentially expressed genes (DEGs) were analyzed using GEO2R. The gene ontology (GO) and pathway enrichment analysis were established through DAVID database. The protein–protein interaction (PPI) network was performed through the Search Tool for the Retrieval of Interacting Genes (STRING) database and managed by Cytoscape. The overall survival (OS) analysis of the 4 genes including AURKA, CDH1, CDK1 and PPARG that had higher degrees in this network was uncovered Kaplan-Meier analysis. Results: A total of 811 DEGs were identified in breast cancer, which were enriched in biological processes, including cell cycle, mitosis, vessel development and lipid metabolism. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis revealed that the up-regulated DEGs were particularly involved in cell cycle, progesterone-mediated oocyte maturation and leukocyte transendothelial migration, while the down-regulated DEGs were mainly involved in regulation of lipolysis, fatty acid degradation and glycerolipid metabolism. Through PPI network analysis, 14 hub genes were identified. Among them, the high expression of AURKA, CDH1 and CDK1 were associated with worse OS of breast cancer patients; while the high expression of PPARG was linked with better OS. Conclusion: The present study identified key pathways and genes involved in breast cancer which are potential molecular targets for breast cancer treatment and diagnosis.

Breast cancer, originating from the epithelium of the mammary gland, has undergone an increase in incidence in China in the last few decades and remains the most frequently diagnosed cancer among women, with an estimated 272,400 new cases and 70,700 related deaths in 2015 (1, 2). Despite advances in endocrine and surgical therapy as well as chemotherapy, breast cancer-associated mortality in China still increases (2). This is attributed, at least partly, to the difficulties in breast cancer diagnosis and lack of treatments for advanced breast cancer patients.

Although immunohistochemistry analysis allows distribution of breast cancer tumors into three major subgroups: hormone receptor (HR)-positive, human epidermal receptor 2 (HER2)-positive and triple-negative breast cancer (3). Each subgroup shows specific prognosis and different responses to anticancer therapy, there is, however, accumulating evidence supporting the hypothesis that they share similar activated genes as well as signaling pathways (4, 5). These signaling pathways and multiple genes participate in the progression of breast cancer. Therefore, there is an urgent need to identify the key genes involved in breast cancer and explore highly efficient key-gene screening methods.

In addition, breast cancer, which is characterized by cumulative epigenetic and genetic aberrations as well as cancer cell heterogeneity, is certainly a complex disease. Understanding the pathological molecular alternations governing breast cancer initiation and progression is greatly important for prognosis and cancer prevention.
Since the introduction of bioinformatics, such as gene microarray technology and high-throughput sequence, which detect genetic alternations, we are able to analyze gene expression profiles during carcinogenesis and cancer progression. It can also uncover thousands of differentially expressed genes (DEGs) associated with different biological processes, different pathways, or molecular functions of cancer. Moreover, the bioinformatics methods make it possible to comprehensively analyze large amounts of data from the microarray output. Given the false-positives and heterogeneity of different microarray results, we processed three microarray datasets to obtain DEGs between normal and breast cancer tissues. Combined with bioinformatics, we identified key pathways and genes in breast cancer. Through survival analysis we found that **PPARG, AURKA, CDH1** and **CDK1** were critical DEGs in breast cancer.

**Materials and Methods**

**Microarray data and data processing.** Three gene expression profiles (GSE21422, GSE42568 and GSE45827) were obtained from public microarray data storage platforms (the Gene Expression Omnibus, GEO, [http://www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)). GSE21422 included 5 normal breast tissue samples and 14 breast cancer samples (6). GSE42568 contained 17 normal breast tissue samples and 104 breast cancer samples (7). GSE45827 consisted of 11 normal breast tissue samples and 130 breast cancer samples (8). GEO2R ([http://www.ncbi.nlm.nih.gov/geo/geo2r/](http://www.ncbi.nlm.nih.gov/geo/geo2r/)) was applied to screen DEGs between normal breast tissue and breast cancer in the three datasets. The *p*-value was adjusted for the correction of false positive outputs when using the Benjamini and Hochberg (False discovery rate) method and logFC represents the fold changes of down- or up-regulated genes of breast cancer against normal breast tissues. According to other studies, we set the adjusted *p*-value<0.01 and logFC>1 as significant DEGs (9, 10).

**Functional and pathway enrichment analysis of DEGs.** By GO enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis method, we mapped the potential relevant biological function annotation of the significant DEGs comprehensively. GO and KEGG was applied through DAVID database ([https://david.ncifcrf.gov/](https://david.ncifcrf.gov/)) which is an online tool for gene annotation, function visualization and large volume data integration (11). A *p*-value<0.05 was considered statistically significant.

**The protein-protein interaction (PPI) network and module selection.** To evaluate the PPI network information of the significant DEGs, we used an online database, Search Tool for the Retrieval of Interacting Genes (STRING, [http://www.string-db.org](http://www.string-db.org/)). Since we analyze more than 10,000 DEGs, we set confident interaction score ≥0.7 (high confident) as significant interaction in order to avoid uncertain/unidentified PPI. The PPI networks were visualized via Cytoscape software and modules of PPI network were screened using Molecular Complex Detection (MCODE) with the included criteria as follow: degree cutoff=2, node score cutoff=0.2, k-core=2, max. depth=100 (9, 10). The average degrees of MCODE score and nodes in modules were chosen as threshold, thus we set MCODE scores ≥9 and hub nodes ≥9 as criteria. The functional enrichment analysis was performed through DAVID in the modules.

**Survival analysis of hub genes in breast cancer.** According to the expression of hub genes, breast cancer patients were divided into two groups (high and low). The overall survival of the two groups was assessed by Kaplan-Meier plotter ([http://kmplot.com](http://kmplot.com)) while the log rank *p*-value and the hazard ratio (HR) (95% confidence intervals) were calculated.

**Results**

**Identification of DEGs.** A total of 875, 1,324 and 4,144 DEGs were up-regulated in GSE21422, GSE42568 and GSE45827 datasets, respectively (Figure 1A); while 1,145, 1,551 and 1,726 DEGs were down-regulated in GSE21422, GSE42568 and GSE45827 datasets (Figure 1B). In total, 811 DEGs presented the same expression tendency in all three datasets, including 344 up-regulated and 467 down-regulated in breast cancer tissues compared to normal breast tissues.

**Gene ontology (GO) and pathway enrichment analysis.** Through DAVID tool, we analyzed the GO and KEGG pathway enrichment of the identified DEGs (Table I). GO biological process (BP) analysis indicated the up-regulated DEGs were mainly involved in those regulating mitotic cell cycle and cell cycle; the down-regulated DEGs were mainly enriched in those relating to vessel development and lipid metabolic process. For GO cell component (CC), the up-regulated DEGs were mainly enriched in genes in microtubule cytoskeleton, chromosome, the centromeric region and spindle; the down-regulated DEGs were mainly enriched in those related to lipid particle, membrane raft and membrane microdomain. Concerning molecular function, the up-regulated DEGs were mainly enriched in microtubule binding, cytoskeletal protein binding and enzyme binding; the down-regulated DEGs were mainly enriched in glycosaminoglycan binding, hormone binding and RNA polymerase II core promoter proximal region sequence-specific binding. In addition, KEGG analysis showed that the up-regulated DEGs were mainly enriched in cell cycle, progesterone-mediated oocyte maturation and leukocyte transendothelial migration signaling pathways and the down-regulated DEGs were mainly enriched in regulation of lipolysis in adipocytes, fatty acid degradation and glycero-lipid metabolism (Table I).

**PPI network and module selection.** The PPI network consisted of 443 nodes and 1181 edges (Figure 2A). Based on the STRING output, there are 57 genes with their degree >10 and 192 genes with their module score >1, respectively. In order to select the most significant genes, we set criteria as module score >1 and degree >20. Twelve hub genes including: **IGF1, LEP, KIF11, PTEN, FOXO1, FGF2**, 4330
**Figure 1.** Identification of differentially expressed genes (DEGs) in mRNA datasets GSE21422, GSE42568 and GSE45827. A: A total of 344 presented the up-regulation tendency in all three datasets; B: A total of 467 show the down-regulation tendency in all three datasets.

**Figure 2.** Protein-protein interaction (PPI) network and selected modules with a higher degree. A: The constructed PPI network of differentially expressed genes; B: Module 1; C: Module 2.

CCNB1, PPARG, AURKA, IK3CA, CDH1 and CDK1 were selected. Through the plug-in MCODE, two significant modules were selected with MCODE score ≥9, nodes ≥9 and edges ≥9 with average MCODE score=4, nodes=7 and edges=3 (Figures 2B-2C). Functional enrichment analysis indicated that up-regulated genes in module one were significantly enriched in cell division and mitotic nuclear division, while the down-regulated genes in module two
Table I. Gene ontology and pathway enrichment analysis of differentially expressed genes in breast cancer.

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Gene function</th>
<th>Gene count</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up-regulated</td>
<td>G0000278</td>
<td>Mitotic cell cycle</td>
<td>79</td>
<td>1.20E-28</td>
</tr>
<tr>
<td></td>
<td>G0015630</td>
<td>Microtubule cytoskeleton</td>
<td>65</td>
<td>4.20E+14</td>
</tr>
<tr>
<td></td>
<td>G000775</td>
<td>Chromosome, centromeric region</td>
<td>25</td>
<td>1.50E-12</td>
</tr>
<tr>
<td></td>
<td>G0000819</td>
<td>Spindle</td>
<td>31</td>
<td>1.60E-12</td>
</tr>
<tr>
<td></td>
<td>G0008017</td>
<td>Microtubule binding</td>
<td>16</td>
<td>1.50E-05</td>
</tr>
<tr>
<td></td>
<td>G0019899</td>
<td>Enzyme binding</td>
<td>7</td>
<td>6.00E-05</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa04110</td>
<td>Cell cycle</td>
<td>15</td>
<td>8.70E-08</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa04914</td>
<td>Progesterone-mediated oocyte maturation</td>
<td>7</td>
<td>6.40E-03</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa04670</td>
<td>Leukocyte transendothelial migration</td>
<td>8</td>
<td>7.50E-03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Down-regulated</th>
<th>Term</th>
<th>Gene function</th>
<th>Gene count</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G001568</td>
<td>Lipid transport</td>
<td></td>
<td>52</td>
<td>1.60E-15</td>
</tr>
<tr>
<td>G001944</td>
<td>Vascular development</td>
<td></td>
<td>53</td>
<td>4.20E-15</td>
</tr>
<tr>
<td>G006629</td>
<td>Lipid metabolic process</td>
<td></td>
<td>82</td>
<td>1.40E-14</td>
</tr>
<tr>
<td>G0005811</td>
<td>Lipid particle</td>
<td></td>
<td>13</td>
<td>1.50E-07</td>
</tr>
<tr>
<td>G0045121</td>
<td>Membrane raft</td>
<td></td>
<td>25</td>
<td>4.20E-07</td>
</tr>
<tr>
<td>G0098857</td>
<td>Membrane microdomain</td>
<td></td>
<td>25</td>
<td>4.50E-07</td>
</tr>
<tr>
<td>G0005599</td>
<td>Glycosaminoglycan binding</td>
<td></td>
<td>20</td>
<td>1.10E-06</td>
</tr>
<tr>
<td>G0042562</td>
<td>Hormone binding</td>
<td></td>
<td>10</td>
<td>2.70E-05</td>
</tr>
<tr>
<td>G0005982</td>
<td>Transcription factor activity, RNA polymerase II core</td>
<td>23</td>
<td>4.90E-05</td>
<td></td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa04923</td>
<td>Regulation of lipolysis in adipocytes</td>
<td>14</td>
<td>6.20E-09</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa00071</td>
<td>Fatty acid degradation</td>
<td>9</td>
<td>3.80E-05</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa00561</td>
<td>Glycosylceramide metabolism</td>
<td>10</td>
<td>4.70E-05</td>
</tr>
</tbody>
</table>

If there were more than four terms enriched in this category, top four terms were selected according to p-value. Count: the number of enriched genes in each term.

were mainly enriched in positive regulation of wound healing and inflammatory response.

Hub genes expression analysis and Kaplan-Meier plotter. We picked out 3 significant hub genes from the 12 hub genes with their expression more than two-fold higher than normal tissue and 1 hub gene with its expression more than two-fold lower than normal tissue in all three datasets (GSE21422, GSE42568 and GSE45827). Through Kaplan–Meier plotter, we found high expression of PPAR γ (HR 0.66 [0.53-0.82], p<0.01) which was associated with better overall survival for breast cancer patients. On the other side, high expression of AURKA (HR 1.83 [1.47-2.28], p<0.01), CDH1 (HR 1.39 [1.12-1.72], p<0.01) and CDK1 (HR 1.55 [1.39-1.73], p<0.01) were associated with worse overall survival for breast cancer patients (Figure 3A-3D).

Discussion

In the present study, a total of 33 normal breast tissue samples and 248 breast cancer tissues were retrieved from the GEO database. A total of 811 DEGs including 344 up-regulated and 467 down-regulated genes were screened. In order to demonstrate the molecular function and the protein-protein interactions of these abnormal DEGs, we performed GO and KEGG pathway analysis. GO terms and KEGG analysis indicated that the up-regulated DEGs were mainly enriched in cell cycle, microtubule cytoskeleton and mitosis. In line with this, the abnormal function of cell cycle and mitosis would appear to be a major cause of cancer (12, 13). The down-regulated DEGs were mainly enriched in hormone binding, lipid metabolism, and glycosaminoglycan binding. In our study we found some lipid metabolism genes that were down-regulated. A recent work has also reported that some genes involved in lipid metabolism are down-regulated in cancer and potentially function as tumor suppressors (14). In addition, the down-regulated DEGs were mainly enriched in hormone binding. Strikingly, previous research has revealed that breast cancer cells can transform from HER2-positive to HER2-negative and become chemotherapy resistant (15). Thus, breast cancer potentially down-regulates its hormone-binding ability to become endocrine therapy resistant.
The combination of bioinformatics and microarray has enabled us to study a vast number of genes with abnormal expression and has improved the efficiency of analyzing complex genetic diseases such as cancer by more than five orders of magnitude since 2005. However, the occurrence of false-positive and false-negative results still exists (16). In order to avoid false-positive/negative and the heterogeneity between different sequencing platforms, we combined three microarray datasets together. We finally identified 4 hub genes which were found to be expressed more than two-fold higher/ lower in breast cancer tissues in all 3 GEO datasets. Survival analysis indicated high expression of PPARG was associated with better overall survival of breast cancer patients and the upregulation of
AURKA, CDH1 and CDK1 were significantly associated with worse overall survival of patients. Thus, molecular drugs targeting these 4 key genes might potentially be applicable in all breast cancer patients.

AURKA is a key regulator of mitosis, especially important in the passage from G2 to M (17). The forced expression of AURKA will result in gene copy number alternations and genetic mutations which are associated with the emergence of breast carcinoma (18). Further, AURKA overexpression correlates with high risk of breast cancer (19). Consistent with this, we found AURKA highly expressed in breast cancer patients and a predictor of poor prognosis. This indicates AURKA may function in breast tumor initiation and progression. The second hub gene, CDH1, encodes the E-cadherin protein. Concrete evidence, strikingly supports that E-cadherin functions as a tumor promoter. Clinical specimens have shown high expression of E-cadherin in 12% of lobular carcinomas and 70% of invasive ductal carcinomas (20, 21). Our study also found that high levels of CDH1 correlate with worse prognosis. Interestingly, the soluble form of E-cadherin contributes to tumor progression and metastasis via undermining cell junctions and activates onco-pathways (22). CDK1, belonging to the serine/threonine kinase family, is directly involved in regulation of cell cycle. It is essential to the initiation of mitosis and transition process of the cell cycle; the loss of function of CDK1 will lead to G2 phase cycle arrest (23, 24). PPARγ is a potential tumor suppressor and is downregulated in breast cancer tissue. Overexpression of PPARγ associates with better prognosis for patients. PPARγ is a member of the nuclear steroid receptor superfamily. In vitro, PPARγ activation will inhibit breast cancer cell proliferation, migration and invasion (25). Gene meta-analysis and clinical data suggest that a high level of PPARγ is a protective factor for breast cancer and colorectal cancer, which is consistent with our findings (26, 27).

In summary, the present study provided a comprehensive analysis of DEGs through bioinformatics to find potential genes that are involved in breast cancer progression and suggest useful targets for investigation or new biomarkers for cancer diagnosis.

Acknowledgements

This work was supported by grants from the National Key Research and Development Program of China (2017YFC1309100); the Natural Science Foundation of Guangdong Province (2014A03036003, 2014A030310378, 2014A020120509, 2015A030313172, 2015B050 501004, 2016A050502018 and 2016A030313237); China Scholarship Council (No. 201606385034); Cultivation for Major Projects and Emerging Interdisciplinary Funding Project of Sun Yat-sen University (17ykjc13) and Elite Young Scholars Program of Sun Yat-sen Memorial Hospital (Y201401). The Authors would also like to thank Cancer Research Wales and Wales National Life Science Research Network (NRN), Cardiff, Wales for their support.

References


Received May 2, 2017
Revised May 21, 2017
Accepted May 22, 2017