The PPAR signaling pathway as a potential biomarker for the diagnosis of breast cancer

Hengyu Wang, Yuqin Yang, Qing Luo, Minfeng Liu, Can Luo, Jianyu Dong, Jingyun Guo, Changsheng Ye

Department of Breast Surgery, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong Province, P. R. China

Received August 8, 2018; Accepted February 12, 2019; Epub June 15, 2019; Published June 30, 2019

Abstract: Several previous studies have investigated the association between the peroxisome proliferator-activated receptor (PPAR) signal pathway and cancer risk; however, the results of these studies were inconsistent, and the role of the PPAR signaling pathway in cancer remains unclear. Therefore, the aim of this study was to further investigate the association between the PPAR signaling pathway and breast cancer risk. RNA-Seq expression data were derived from a breast cancer cohort of the Gene Expression Omnibus (GEO) dataset. A two-way hierarchical clustering analysis (HCA), a support vector machine (SVM) classifier, and a protein-protein interaction (PPI) network were built in the training dataset (GSE42568) using the twenty-one differentially expressed genes (DEGs) which were annotated into the PPAR signaling pathway. The accuracy of the candidate informative DEGs using the training dataset (GSE42568) in risk-stratifying samples was 72.73%, and the accuracy of the SVM classifier was 97.52%. The predictive ability of the validation datasets, GSE29431 and GSE21422, achieved reliable outcomes, with the risk-stratifying samples at 96.97% and 94.74%, respectively, from the two-way HCA. The accuracy yields were 92.42% and 94.74%, respectively, using the SVM classifier. A PPI analysis showed that the twenty-one DEGs formed a retinoid X receptor alpha (RXRA)-centric world with 14 nodes. Collectively, the twenty-one informative genes of the PPAR signaling pathway may represent the key genes associated with the occurrence of breast cancer. Our results provide the primary information and basic knowledge necessary to better understand the mechanisms of cancer pathogenesis.

Keywords: Breast cancer, gene expression omnibus datasets, peroxisome proliferator activated receptors signaling pathway, SVM classifier, two-way hierarchical clustering analysis

Introduction

Breast cancer (BC) is a common cancer among females in both developed and developing countries [1], with the risk of developing BC known to be influenced by both genetic and environmental factors. A total of 1.7 million new BC cases have been identified since 2012, with over 500,000 related deaths [2]. As BC cases continue to rise annually, corresponding to high mortality rates [3, 4], the association between potential environmental and genetic factors and BC risk needs to be further investigated. Recent studies have highlighted several genetic factors that appear to correlate with BC risk [5-7], including metabolism-related genetic variations [8].

Peroxisome proliferator-activated receptors (PPARs) comprise a cluster of nuclear transcription factors that are members of the nuclear hormone receptor super-family. They possess important functions in cellular differentiation and the regulation of carbohydrates and lipid metabolism [9]. Accordingly, polymorphisms in these receptors are assumed to affect the pathology of cancers and other diseases. Three PPAR subtypes, namely, PPAR alpha (PPARA), PPAR delta/PPAR beta (PPARD/PPARB), and PPAR gamma (PPARG), have been found to be dynamically regulated at multiple molecular levels. Endogenous PPARA ligands include palmitic acid, arachidonic acid, and stearic acid. In addition, other known ligands include compounds such as fenofibrate, bezafibrate, and non-steroidal anti-inflammatory drugs [10, 11]. Thus far, genetic variants of PPARA have been related to lipoprotein levels [12], cardiovascular disease [13], obesity [14], and type 2 diabetes [15]. These conditions arise through etiologic mechanisms that may also be relevant to breast carcinogenesis, including inflammation and
The PPAR signaling pathway in breast cancer

insulin resistance. Although the biology and epidemiology of PPARA suggest that this receptor may also play a role in BC, limited data exist on the possible link between PPARA and BC.

PPARG plays a pivotal role in regulating adipocyte differentiation, glucose and lipid homeostasis, and intracellular insulin-signaling events [16]. In addition, PPARG appears to have a contradictory role in tumorigenesis. Indeed, several studies have demonstrated the tumorigenic role of PPARG in a variety of cancers such as bladder tumors, renal pelvic tumors, hemangiooma, lipoma, skin fibrosarcoma, mammary adenocarcinoma, and hepatic tumors. The tumorigenicity of PPARs has not been fully recognized; however, recent studies have suggested various mechanisms for this reported effect. In contrast, the anti-tumorigenicity of PPARG agonists through other hypothesized mechanisms, as well as the downregulation of PPARG in some human cancers, has been reported [17-27]. Moreover, the activation of PPARG by its ligands can suppress the growth of tumor cells in liver, pancreatic, biliary, oral, esophageal, gastric, and colorectal tumors, suggesting that PPARG ligands may be a possible anticancer factor in PPARG-expressing tumors [28].

For complete activation, PPARs must heterodimerize with retinoid X receptors (RXR) to form a PPAR/RXR complex. This complex then binds to a specific DNA sequence, termed the PPAR-response element, in a given target gene [29]. RXRA is a nuclear receptor that regulates transcription, both as a homodimer and an obligate heterodimerization partner for 14 other nuclear receptors, including PPARA, PPARD, and PPARG [30].

Bioinformatics is an effective tool for collecting, classifying, and analyzing biological datasets, including gene expression microarray datasets [31, 32]. In fact, gene expression analysis by bioinformatics methods has been widely employed in genomics and biomedical research, broadening insights into the molecular mechanisms underlying human biology and disease [33]. Data mining of the available microarray datasets could assist scientists to bridge the research gap and to carry out more efficient experiments.

In this study, we analyzed public microarray data using a two-way hierarchical clustering analysis and a support vector machine (SVM) classifier in order to clarify potential associations between the PPAR signaling pathway and BC. Differentially expressed genes (DEGs) were first identified between the normal and tumor groups. From these DEGs, optimal informative genes were extracted using the DEGs annotated into the PPAR signal pathway. Candidate genes were then subjected to build a SVM classifier; the predictive capability of the candidate informative genes was then verified using two independent datasets. These informative genes were also utilized to construct protein-protein networks. Finally, we attempted to identify PPAR signal pathway-related genes to gain insight into the pathogenesis of BC.

Materials and methods

Data source

Two mRNA-seq expression datasets were accessed from the Gene Expression Omnibus data portal (https://www.ncbi.nlm.nih.gov/geo/). The GSE42568 dataset was used as a training set, which included 104 primary breast samples and 17 normal control samples. The GSE29431 and GSE21422 datasets were used as validation sets and consisted of 54 primary melanoma samples and 12 normal control samples and 14 primary melanoma samples and 5 normal control samples, respectively. The three mRNA expression datasets were assessed using GPL570 Platform. All microarray data were called using the GC robust multichip average method [34] and quantile normalized using the “affy” Bioconductor package by contributor.

Data preprocessing and differentially expressed genes (DEG) screening

Annotations to the probes were performed; probes that were not matched to the gene symbol were excluded. The average expression values were taken if different probes mapped to the same gene. DEGs in patients with BC versus those in healthy matched controls were analyzed using the DESeq package (version 3.10.3) of Bioconductor. A strict cutoff threshold was used and set to $P<0.05$ and fold change $\geq 2.0$.

Predictive capacity in proposed HCA and SVM classifier model

The DEGs that were annotated into the PPAR pathway were selected for further analysis.
The PPAR signaling pathway in breast cancer

Two-way HCA was performed on the expression values of genes that were significantly overlapping using the heatmap2 package (21) in R, and the distance was under default value.

A SVM classifier was constructed using the support vector classification function in sklearn in the svm package of Python (version 3), with the linear Kernel function (C=0.3) and a 3-fold cross-validation. In addition, the random seed was held at 100 to shuffle the training set. The capacity of classification were evaluated based on six metrics, including the accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating curve (AUC).

Verification of the classification model using other two independent dataset

Two-way HCA and the SVM classifier, based on the candidate informative genes, were conducted sequentially to further verify classification reliability by computing other two independent dataset as the test set.

Construction of mRNA-mRNA networks

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING; http://www. STRING database.org/) [35] is a gene or protein analysis tool designed to provide a critical assessment and integration of protein-protein interactions (PPIs). In this study, overlapping genes were mapped into the STRING database for PPI analysis and PPI scores >0.4 were selected as significant [36, 37]. A PPI network was then constructed using Cytoscape software (version 3.5.1) [38], and the degree was used for stating the role of the protein nodes in the network. Specifically, the greater the degree, the more important the nodes were in the network. In PPIs, genes closely associated with other genes were identified with degrees ≥10 [37, 39].

Results

HCA and SVM classifier for distinguishing disease status

To examine the PPAR signal pathway on BC risk, we used two independent methods, namely, HCA and SVM classifier. Using these methods, we identified shared DEGs in the PPAR signal pathway. Using HCA, the association between the expression pattern of candidate genes and the disease status of the samples were identified with the Euclidean method. SVM classifier discrimination between cancer patients and healthy samples was based on a hyperplane to maximize the distance between two samples on different sides of the plane, which were the closest samples to the plane in each category, respectively. The general workflow were showed in Figure 1.
Identification of selected informative genes

A total of 1833 DEGs were identified between normal and control samples using R software. Of these, 684 were upregulated, whereas 1149 were downregulated. Twenty-one candidate informative genes of the PPAR pathway were selected for further analysis. These included FABP4, LPL, ACADL, PCK1, SCD, ME1, ADIPOQ, PLIN4, ACADM, ACSL1, PPARG, CD36, PLIN1, PLIN2, NR1H3, ANGPTL4, EHHADH, FABP5, PLTP, RXRA, and SCP2. The distribution of the expression levels of the twenty-one candidate PPAR pathway-associated genes was shown in Figure 2A, with the detailed information provided in Table 1.

PPI network analysis

In view of the controversy between PPARs and cancer, the twenty-one candidate associated genes were selected to perform further analy-
The PPAR signaling pathway in breast cancer

sis. This PPI network was visualized using Cytoscape. Hub genes with a degree of interaction >10 were defined as those that strongly interacted with other candidate genes, including RXRA, LPL, FABP4, and PPARG (Figure 2B). These hub genes may represent key genes affected by the PPAR signal pathway that were associated with BC.

**HCA of candidate mRNAs**

A total of twenty-one candidate PPAR pathway-associated logarithmic expression values were subjected to HCA using the training set. As shown in Figure 2. As the result showed that all samples were distinctly subdivided into two clusters. The accuracy was 72.73% (88/121) and more specifically, 15 out of the 17 normal samples were incorporate into individual cluster and 86 out of the 104 tumor samples were classified into the other cluster (Figure 2C; Table 2).

**Assessment on the training dataset using the SVM classifier model**

To further confirm whether the candidate signature genes can discriminate between the two types of samples, an SVM classifier model was proposed on the basis of their expression values and achieved reliable result with the accuracy of 97.52%, a sensitivity of 99.05%, a specificity of 83.33%, a PPV of 97.25%, a NPV of 94.44%, and AUC of 97.62% (Figure 3B; Table 3).

**Validation of the classification model in GSE29431 cohort**

The performance of two-way HCA and the SVM classifier based on the twenty-one candidate signature genes was verified using the testing datasets. The results of two-way HCA indicated that all the samples in the validation dataset were stratified into two groups. The accuracy was thus determined to be 96.97% (64/66), only 2 normal samples incorrectly clustered into the tumor group and all tumor samples were extremely accurate classified into the corresponding group (Figure 3A; Table 2).

Likewise, the SVM model could correctly distinguish the tumor sample and normal samples attaining high accuracy (92.42%), and AUC, the sensitivity, specificity, PPV and NPV reaching 88.27%, 94.44%, 83.33%, 96.23%, 76.92%, respectively (Figure 3B; Table 3).

**Table 1.** Detail information of twenty-one candidate signature genes

<table>
<thead>
<tr>
<th>Status</th>
<th>Gene name</th>
<th>log_{2} FC</th>
<th>p-value</th>
<th>p-adjust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downregulated</td>
<td>FABP4</td>
<td>-4.54</td>
<td>2.32E-55</td>
<td>5.46E-51</td>
</tr>
<tr>
<td></td>
<td>LPL</td>
<td>-5.43</td>
<td>1.08E-30</td>
<td>1.02E-27</td>
</tr>
<tr>
<td></td>
<td>ACADL</td>
<td>-4.33</td>
<td>4.56E-15</td>
<td>5.42E-13</td>
</tr>
<tr>
<td></td>
<td>PCK1</td>
<td>-6.62</td>
<td>4.76E-27</td>
<td>2.67E-24</td>
</tr>
<tr>
<td></td>
<td>SCD</td>
<td>-3.5</td>
<td>1.29E-11</td>
<td>9.75E-10</td>
</tr>
<tr>
<td></td>
<td>ME1</td>
<td>-3.89</td>
<td>3.31E-24</td>
<td>1.26E-21</td>
</tr>
<tr>
<td></td>
<td>ADIPOQ</td>
<td>-4.3</td>
<td>4.59E-36</td>
<td>9.00E-33</td>
</tr>
<tr>
<td></td>
<td>PLIN4</td>
<td>-5.3</td>
<td>3.24E-28</td>
<td>2.24E-25</td>
</tr>
<tr>
<td></td>
<td>ACADM</td>
<td>-2.18</td>
<td>4.72E-12</td>
<td>3.86E-10</td>
</tr>
<tr>
<td></td>
<td>ACSL1</td>
<td>-3.78</td>
<td>6.27E-29</td>
<td>4.76E-26</td>
</tr>
<tr>
<td></td>
<td>PPARG</td>
<td>-4.98</td>
<td>4.24E-24</td>
<td>1.56E-21</td>
</tr>
<tr>
<td></td>
<td>CD36</td>
<td>-4.67</td>
<td>6.61E-26</td>
<td>3.11E-23</td>
</tr>
<tr>
<td></td>
<td>PLIN1</td>
<td>-5.99</td>
<td>1.01E-38</td>
<td>2.98E-35</td>
</tr>
<tr>
<td></td>
<td>PLIN2</td>
<td>-1.89</td>
<td>2.50E-05</td>
<td>5.97E-04</td>
</tr>
<tr>
<td></td>
<td>NR1H3</td>
<td>-1.89</td>
<td>3.64E-05</td>
<td>8.28E-04</td>
</tr>
<tr>
<td></td>
<td>ANGPT1L4</td>
<td>-2.09</td>
<td>4.21E-08</td>
<td>1.76E-06</td>
</tr>
<tr>
<td></td>
<td>EHHADH</td>
<td>-2.71</td>
<td>4.45E-09</td>
<td>2.29E-07</td>
</tr>
<tr>
<td></td>
<td>FABP5</td>
<td>-2.19</td>
<td>3.19E-04</td>
<td>5.66E-03</td>
</tr>
<tr>
<td></td>
<td>PLTP</td>
<td>-1.95</td>
<td>4.13E-05</td>
<td>9.20E-04</td>
</tr>
<tr>
<td></td>
<td>RXRA</td>
<td>-1.92</td>
<td>5.22E-04</td>
<td>8.60E-03</td>
</tr>
<tr>
<td></td>
<td>SCP2</td>
<td>-1.03</td>
<td>1.44E-04</td>
<td>2.84E-03</td>
</tr>
</tbody>
</table>

**Table 2.** Summary of clinical samples in the training and two test datasets

<table>
<thead>
<tr>
<th>Cancer</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>86</td>
<td>2</td>
</tr>
<tr>
<td>Healthy</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>17</td>
</tr>
</tbody>
</table>

**Note:** The row was the actual class, and the column was the predicted class (training dataset, GSE42568).

<table>
<thead>
<tr>
<th>Cancer</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>54</td>
<td>2</td>
</tr>
<tr>
<td>Healthy</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>54</td>
<td>12</td>
</tr>
</tbody>
</table>

**Note:** The row was the actual class, and the column was the predicted class (test dataset, GSE29431).

<table>
<thead>
<tr>
<th>Cancer</th>
<th>Healthy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Healthy</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>5</td>
</tr>
</tbody>
</table>

**Note:** The row was the actual class, and the column was the predicted class (test dataset, GSE21422).
Further validation in the GSE21422 cohort

Further validation of the predictive value of the candidate genes in the diagnosis of BC was conducted using another independent validation cohort. HCA of all samples \((n=19)\) revealed clear distinctions between BC patients and healthy samples. Samples were classified into cluster 1 and cluster 2, and achieved a prediction accuracy of 94.74\% with one normal sample incorrectly clustered into the tumor group (18/19) (Figure 3C). The SVM-based candidate gene risk classifier performed remarkably well. The accuracy was 94.74\%, with an AUC of 95.71\%, a sensitivity of 100.00\%, specificity of 80.00\%, PPV of 93.33\%, and a NPV of 100.00\% (Figure 3D; Table 3). In total, these results confirmed the twenty-one candidate signature genes can reliably discriminate the normal sample from controls in BC.

Discussion

Many studies have shown that BC is a metabolic disease [40, 41]. For example, the glutamate-to-glutamine ratio and aerobic glycolysis have been proposed as biomarkers of ER and Her2 status, respectively [42, 43]. PPARs are key transcriptional factors that catalyze and coordinate a variety of biochemical events in order to achieve energy homeostasis associated with many types of cancer, including hepatocellular carcinoma [44], lung adenocarcinoma [45], squamous cell carcinoma of the head and neck [46], bladder cancer [47], and skin carcinoma [48]. However, the association between the PPAR pathway and cancer risk remains controversial, and a prospective analysis is essential to determine any potential relationships.

Currently, several advanced biological techniques, including gene array and high-throughput sequencing, have been identified as ideal approaches for assessing the mechanisms of development and the immune responses to various diseases. In this study, a comprehensive bioinformatics analysis of several gene array datasets was applied to determine DEGs in BC and their associated pathways. A total of 1833 DEGs were identified to be associated with BC, including 684 upregulated DEGs and 1149 downregulated DEGs. Twenty-one candidate downregulated signature genes were then annotated into the PPAR signal pathway. Key selected downregulated DEGs were further used to investigate the PPI network analysis. Our results demonstrated that the RXRA and FABP4 genes presented with substantially more “weight” than the other genes in the interaction network.

Recent studies have shown that PPARG inhibits cellular proliferation and induces apoptosis through the upregulation of phosphatase and tensin homolog (PTEN), the downregulation of survivin, the downregulation of the X-linked inhibitor of apoptosis (XIAP), suppression of NF-κB and glycogen synthase kinase (GSK)-3β, upregulation of cyclin-dependent kinase (CDK) inhibitors, downregulation of CDK and cyclin D1, downregulation of prostaglandin-endoperoxide synthase 2 (PTGS2), upregulation of Kruppel-Like Factor 4 (KLF4), upregulation of Bax, downregulation of Bcl-2, and inhibition of telomerase activity and human telomerase reverse transcriptase (hTERT) expression through modulation of the Myc/Mad/Max network [49]. However, complete activation is dependent on heterodimerization with RXR, thus forming a PPARG/RXRA complex. We therefore speculated that the downregulated expression level of the PPARG/RXRA complex induces cellular proliferation and blocks apoptosis, resulting in BC.

To further investigate the association between the PPAR pathway and BC risk, the twenty-one candidate signature genes were selected for two-way HCA and to train the SVM classifier. 17 normal samples versus 104 tumor samples were used in the trial. The resulting outcomes showed that the accuracy of the informative genes was 72.73\%. The classification capability of the signature genes was further verified using two independent datasets (GSE29431 and GSE21422) that included 12 normal samples versus 54 tumor samples, and 5 normal samples versus 14 tumor samples, respective-

### Table 3. Performance of twenty-one candidate signature genes in SVM classifier

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Training dataset</th>
<th>Test datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GSE42568</td>
<td>GSE29431</td>
</tr>
<tr>
<td>Accuracy</td>
<td>97.52%</td>
<td>94.22%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>99.05%</td>
<td>94.44%</td>
</tr>
<tr>
<td>Specificity</td>
<td>83.33%</td>
<td>83.33%</td>
</tr>
<tr>
<td>PPV</td>
<td>97.25%</td>
<td>96.23%</td>
</tr>
<tr>
<td>NPV</td>
<td>94.44%</td>
<td>76.92%</td>
</tr>
<tr>
<td>AUC</td>
<td>97.62%</td>
<td>88.27%</td>
</tr>
</tbody>
</table>

Note: The tumor samples were set as positive group, while the normal sample as negative group.
The PPAR signaling pathway in breast cancer

Two-way HCA and SVM classifier analysis achieved consistent results, supporting our conclusion that these twenty-one candidate PPAR pathway-related genes exhibit a potential association between PPAR and BC risk. However, few reports exist regarding the involvement of these candidate genes in BC. Hence, further studies implicating the associations between the genes identified in this study and BC are warranted.

In conclusion, the present study identified twenty-one key DEGs in the progress of BC. Our results indicated that the four key genes, RXRA, LPL, FABP4, and PPARG, had a high degree of interaction, implying that they may co-function in the tumorigenesis of BC by participating in the regulation of the PPAR signaling pathway. However, further laboratory experiments are still required to confirm the exact association between these genes in order to clearly understand what correlation patterns exist among them. Collectively, the present study provided basic information, paving the road for future experimental research to investigate the mechanisms of BC development. Increasing knowl-

Figure 3. Performance of the two-way HCA and the support vector machine (SVM) classifier based on twenty-one candidate signature genes in the independent validation cohorts (GSE29431 and GSE21422). A. A heatmap of clustering analysis in the GSE29431 cohort. All samples were clustered into cluster 1 and cluster 2. B. ROC analysis of the SVM classifier in the GSE29431 cohort. All samples in the validation set were divided into the tumor group and normal group via the SVM classifier. C. The heatmap of HCA of all samples in the GSE21422 cohort. D. ROC analysis of the SVM-based twenty-one candidate signature genes in the GSE21422 cohort.
edge regarding the mechanisms of BC may lead to improved diagnostic efficacy, as well as the development of novel treatments.

Acknowledgements

This work was supported by the President’s Foundation of Nanfang Hospital, Southern Medical University (no. 2016L007).

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Changsheng Ye, Department of Breast Surgery, Nanfang Hospital, Southern Medical University, 1838 Tonghe Road, Guangzhou 510515, Guangdong Province, P. R. China. Tel: +86-20-37089917; Fax: +86-20-3708-9917; E-mail: yechsh2006@126.com

References


The PPAR signaling pathway in breast cancer


The PPAR signaling pathway in breast cancer


