Screening for key genes associated with invasive ductal carcinoma of the breast via microarray data analysis

J.L. Huan¹, X. Gao², L. Xing¹, X.J. Qin¹, H.X. Qian¹, Q. Zhou¹ and L. Zhu¹

¹Department of General Surgery, The Eighth People’s Hospital of Shanghai, Shanghai, China
²Department of Neurosurgery, Yidu Center Hospital of Weifang City, Weifang, China
³Department of Gynaecology and Obstetrics, The Eighth People’s Hospital of Shanghai, Shanghai, China

Corresponding author: J.L. Huan
E-mail: jinlianghuan@hotmail.com

Received July 31, 2013
Accepted August 7, 2014
Published September 29, 2014
DOI http://dx.doi.org/10.4238/2014.September.29.5

ABSTRACT. The aim of this study was to identify key genes related to invasive ductal carcinoma (IDC) of the breast by analyzing gene expression data with bioinformatic tools. Microarray data set GSE31138 was downloaded from Gene Expression Omnibus, including 3 breast cancer tissue samples and 3 normal controls. Differentially expressed genes (DEGs) between breast cancer and normal control were screened out (FDR < 0.05 and |logFC| > 2). Coexpression between genes was examined with String, and a network was then constructed. Relevant pathways and diseases were retrieved with KOBAS. A total of 56 DEGs were obtained in the IDC of the breast compared with normal controls. A gene coexpression network including 27 pairs of genes was constructed and all the genes in the network were upregulated. Further study indicated that most of the genes in the coexpression network were enriched in ECM-receptor interaction (COL4A2, FN1, and HMMR) and nucleotide excision repair
(CETN2 and PCNA) pathways, and that the most significantly related disease was autoimmune lymphoproliferative syndromes. A number of DEGs were acquired through comparative analysis of gene expression data. These findings are beneficial in promoting the understanding of the molecular mechanisms in breast cancer. More importantly, some key genes were revealed via gene coexpression network analysis, which could be potential biomarkers for IDC of the breast.

**Key words:** Breast cancer; Differentially expressed genes; Gene coexpression network analysis; Pathway analysis

**INTRODUCTION**

Breast cancer poses a great threat to human health, especially for women. It comprises 22.9% of all cancers (excluding non-melanoma skin cancers) in women (Buchholz, 2009) and caused 13.7% of cancer deaths in 2008 (Saracci, 2008). It can be divided into ductal carcinomas and lobular carcinomas (Florescu et al., 2011), where the former is more common.

Since early diagnosis and treatment are of great importance in reducing the mortality of breast cancer, people have been trying to identify more biomarkers (Weigel and Dowsett, 2010; Misek and Kim, 2011). The most important finding has been human epidermal growth factor receptor as a biomarker (Slamon et al., 1987), and now, targeted therapy has been successfully developed (Slamon et al., 2001; Vogel et al., 2002). Various new biomarkers have been reported, such as kallikrein gene 14 (Borgoño et al., 2003) and BAG-1 (Turner et al., 2001). Zehentner et al. (2004) indicated that mammaglobin can serve as a diagnostic tool for breast cancer. Dunning et al. (2003) found that a transforming growth factor β1 signal peptide variant was associated with increased incidence of invasive breast cancer.

Although considerable achievements have been made, more studies are needed to gain more knowledge about breast cancer and to discover effective biomarkers for diagnosis or treatment. Microarray technology is an effective tool to disclose the global changes in the incidence and development of cancer (DeRisi et al., 1996). Therefore, in the present study, we tried to identify potential biomarkers for invasive ductal carcinoma (IDC) of the breast via differential analysis and gene coexpression network analysis of gene expression data.

**MATERIAL AND METHODS**

**Microarray data**

The gene chip dataset GSE31138 was downloaded from Gene Expression Omnibus, containing 3 breast cancer samples and 3 normal controls. All samples were donated by the London BARTS Cancer Institute Laboratory. The platform was Affymetrix Human Genome U133 Plus 2.0 Array.

**Data pre-processing and differential analysis**

The original data were converted into recognizable expression data. The missing values were then filled in with the KNN method (Troyanskaya et al., 2001). The standardization
was performed with the median method (Fujita et al., 2006), followed by differential expression analysis with the package Limma (Kerr, 2003) of R. The Benjamini-Hochberg method (Benjamini et al., 2001) was used for multiple testing correction and the differentially expressed genes (DEGs) were selected out according to the criteria: \( P < 0.05, \) FDR \(< 0.05 \) and \( |\log_{10}FC| > 1.\)

**Gene coexpression network analysis**

String (Szklarczyk et al., 2011) was chosen for gene co-expression network analysis of the DEGs. The coexpression coefficients between DEGs were calculated on the basis of the characteristics of gene sequences and spatial structures. The coexpression pairs with coexpression >0.5 were selected out and used for gene coexpression network construction.

**Retrieval of relevant pathways and diseases**

Relevant pathways and diseases were retrieved for DEGs with KOBAS (KEGG Orthology Based Annotation System) (Xie et al., 2011). KOBAS was the first software to identify significantly enriched pathways using a hypergeometric test. It has been successfully used in pathway analysis for plants, animals and bacteria. Its purpose is to identify significantly enriched pathways and diseases for a set of genes or proteins, using pathway and disease information from multiple databases. In present study, \( P < 0.05 \) was set as the cut-off to screen out pathways related to IDC of the breast.

**RESULTS**

**Differentially expressed genes**

A good performance of data standardization was acquired. A total of 56 DEGs were then identified for breast cancer, 51 upregulated and 5 downregulated. It was obvious that most DEGs played roles in IDC via overexpression.

**Gene coexpression network**

The coexpression pairs with coexpression >0.5 were selected out and included in the network. A total of 27 pairs of coexpressed gene pairs were revealed (Figure 1), such as collagen 4A2 (COL4A2)-fibronectin 1 (FN1), COL4A2-TOP2A and neuroblastoma RAS viral (v-ras) oncogene homolog (NRAS)-WASL.

**Relevant pathways and diseases**

KOBAS (Xie et al., 2011) was chosen for pathway and disease analysis for DEGs and \( P < 0.05 \) was set as the cut-off. Finally, 2 relevant pathways [extracellular matrix (ECM)-receptor interaction and nucleotide excision repair] and 5 diseases [autoimmune lymphoproliferative syndromes (ALPS), malignant melanoma, adrenal carcinoma, oral cancer, and multiple myeloma] were obtained (Tables 1 and 2).
Figure 1. The coexpression network for differentially expressed genes.

Table 1. Relevant pathways of coexpressed genes.

<table>
<thead>
<tr>
<th>#Term</th>
<th>Database</th>
<th>Id</th>
<th>P value</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECM-receptor interaction</td>
<td>KEGG PATHWAY</td>
<td>hsa04512</td>
<td>0.017655</td>
<td>COL4A2, FN1, HMMR</td>
</tr>
<tr>
<td>Nucleotide excision repair</td>
<td>KEGG PATHWAY</td>
<td>hsa03420</td>
<td>0.035558</td>
<td>CETN2, PCNA</td>
</tr>
</tbody>
</table>

Table 2. Diseases related to the coexpressed genes in the network.

<table>
<thead>
<tr>
<th>#Term</th>
<th>Database</th>
<th>Id</th>
<th>P value</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autoimmune lymphoproliferative syndromes</td>
<td>KEGG DISEASE</td>
<td>H00108</td>
<td>0.025546096</td>
<td>NRAS</td>
</tr>
<tr>
<td>Malignant melanoma</td>
<td>KEGG DISEASE</td>
<td>H00038</td>
<td>0.035587324</td>
<td>NRAS</td>
</tr>
<tr>
<td>Adrenal carcinoma</td>
<td>KEGG DISEASE</td>
<td>H00033</td>
<td>0.035587324</td>
<td>NRAS</td>
</tr>
<tr>
<td>Oral cancer</td>
<td>KEGG DISEASE</td>
<td>H00016</td>
<td>0.040570449</td>
<td>NRAS</td>
</tr>
<tr>
<td>Multiple myeloma</td>
<td>KEGG DISEASE</td>
<td>H00010</td>
<td>0.045528742</td>
<td>NRAS</td>
</tr>
</tbody>
</table>

DISCUSSION

In the present study, 51 upregulated and 5 downregulated genes were obtained in the IDC of the breast. Gene coexpression network analysis was performed for all the DEGs, and relevant pathways and disease were then retrieved with KOBAS. Finally, 2 pathways (ECM-
screening for key genes by microarray data analysis

receptor interaction and nucleotide excision repair) and 5 diseases (ALPS, malignant melanoma, adrenal carcinoma, oral cancer and multiple myeloma) were revealed.

COL4A2, FN1 and HMMR (hyaluronan-mediated motility receptor) are components of ECM. The expression of ECM genes is associated with the prognosis of patients with lymph node-negative breast cancer as well as clinical benefit from endocrine treatment (Insalaco et al., 2012). FN1 expression is found to be upregulated during epithelial to mesenchymal transition, which is an early event in malignant transformation accompanied by a reduced adhesion of the tumor cells (Helleman et al., 2008). Several studies suggest that FN1 is related to tumor invasion and metastasis (Landstrom et al., 1992) by playing a key role in the tissue remodeling and cell migration events that occur during normal development. In particular, FN1 is a major constituent of the cell surface of many cultured cells, and it is either eliminated or reduced on the surface of oncogenically transformed cells (Horii et al., 2006). Many reports have suggested that there is a correlation between the loss of cell surface FN1 and the ability of a cell to metastasize (Caraglia et al., 2004). The changes in the cytoskeletal components such as production and organization of FN1, actin and collagen have been implicated in eliciting the transition from dormancy to metastatic growth (Calvo et al., 2008). This was in accordance with our finding since there was a coexpression between FN1 and COL4A2, which suggested that the interaction between the two proteins could be a good cut-in point to modulate the mobility of tumor cells. HMMR has also been reported to mediate migration, transformation, and metastatic spread of cancer cells (Du et al., 2011; Veiseh and Turley, 2011). Of course, further studies are needed to determine the use of these ECM genes in decisions regarding treatment and whether they can serve as targets for therapy (Vargas et al., 2012).

Centrin EF-hand protein 2 (CETN2) is a structural component of the centrosome, and proliferating cell nuclear antigen (PCNA) is a cofactor of DNA polymerase delta, both of which are involved in nucleotide excision repair. Nucleotide excision repair is linked with cancer risk (Lockett et al., 2005; Barry et al., 2012), and therefore, both proteins are worthy of further research to fully elucidate their roles in IDC.

According to Table 2, NRAS was associated with ALPS as well as several kinds of cancers, such as melanoma and oral cancer. ALPS is characterized by nonmalignant lymphadenopathy, splenomegaly, and autoimmune cytopenias (Krueger et al., 2002). Defective lymphocyte apoptosis secondary to mutations in the FAS gene is identified as a molecular basis for ALPS (Fisher et al., 1995). It was the first disease known to be caused by a primary defect in programmed cell death and was the first autoimmune disease with a defined genetic basis (John et al., 2008). Besides, previous studies have confirmed that NRAS is one of the specific and sensitive indices for breast cancer diagnosis and prognosis (Zhu et al., 2004). It was very interesting to find that breast cancer was associated with ALPS to a certain degree since it indicates a possibility to ameliorate both ALPS and breast cancer syndrome by targeted therapy on NRAS gene expression.

Several genes associated with breast cancer have been utilized as diagnostic or prognostic markers, or even therapeutic targets, such as NF-kb (Biswas et al., 2001, 2004) and HER-2 (Helms et al., 2010; Perez et al., 2010). However, currently available biomarkers for the early diagnosis and therapy of breast cancer are far from enough. Therefore, in the present study, we analyzed gene expression data to identify biomarkers for breast cancer. Since ECM-related COL4A2, FN1 and HMMR as well as CETN2 and PCNA are associated with breast cancer, they may be potential biomarkers. Meanwhile, NRAS, which showed close as-
sociations with several kinds of tumors and ALPS, could also be a promising biomarker for breast cancer.

REFERENCES


Screening for key genes by microarray data analysis


