Identification of Potential Key Genes and Pathways for Inflammatory Breast Cancer Based on GEO and TCGA Databases

Qing Lv1,*, Yansong Liu2,*, Hu Huang1, Mingjie Zhu1, Junqiang Wu1, Dong Meng1

1Department of Breast Surgery, Affiliated Hospital of Jiangnan University, Wuxi, Jiangsu, People’s Republic of China; 2Department of Breast Surgery, Tumor Hospital of Mudanjiang City, Mudanjiang, Heilongjiang, People’s Republic of China

*These authors contributed equally to this work

Introduction: Inflammatory breast cancer (IBC) is a rare type of locally advanced breast cancer, accounting for about 2% to 4% of all breast cancer cases.1 IBC is a life-threatening disease characterized by diffuse erythema, edema, and peau d’orange of the skin.2 Despite its low prevalence, IBC causes 7% of breast cancer-associated deaths.3 According to the data from the Surveillance, Epidemiology, and End Results (SEER) program which was carried out by American National Cancer Institute, the diagnosis rate of IBC is increasing by the year, and it is a much higher rate of increase than the incidence of overall breast cancer diagnosis.4 Thus, it is necessary to further understand the unique biology and pathogenesis underlying this aggressive form of breast cancer.

To date, accumulating evidence has demonstrated that aberrant expression and mutation of genes are involved in the carcinogenesis and progression of IBC, and...
several markers have been identified. For instance, the tumors of IBC maintain the expression of E-cadherin which is strongly associated with tumor emboli formation.\textsuperscript{5,6} Another epithelial-mesenchymal transition (EMT) marker zinc-finger E-box-binding homeobox 1 (ZEB1) and the stem cell marker aldehyde dehydrogenase 1 (ALDH1) are also reported to contribute to IBC progression and invasion.\textsuperscript{7,8} However, the exact mechanism underlying IBC is yet to be fully elucidated.

During the last decades, due to the development of high-throughput technologies, including microarray and RNA-sequencing, the gene expression profiles have improved the understanding of IBC on its molecular mechanisms.\textsuperscript{9–12} However, due to the low incidence of IBC, such researches do not have satisfactory sample sizes, and analyses of samples from single centers may lead to biased results. A combined analysis of multiple datasets is urgently needed. Thus, we carried out this study, trying to identify differentially expressed genes (DEGs) in IBC distinct from those in non-IBC by analyzing datasets obtained from Gene Expression Omnibus (GEO) database, and explore the potential bio-functions by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. Furthermore, the protein–protein interaction (PPI) network was constructed, and five hub genes were screened out from it. The present study may provide new insight into the understanding of IBC pathogenesis and the identified hub genes may serve as potential targets for diagnosis and treatment.

**Materials and Methods**

**Microarray Dataset Selection and Processing**

To identify genes differentially expressed in IBC and non-IBC tissues, microarray datasets (GSE5847, GSE22597, GSE23720, and GSE45581) were downloaded from Genome Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo) database, and probe names in each expression matrix were annotated to Entrez ID according to the corresponding GEO platform files. All of the selected datasets contain transcriptome microarray data of IBC and non-IBC clinical samples, and the diagnosis of IBC was based on both clinical features and pathological results. Details of the four datasets are shown in Supplementary Table 1.

**Identification of DEGs**

DEGs between IBC and non-IBC samples from the four GEO datasets were screened using R scripts. Background correction was performed using the robust multi-array average (RMA) method. The R package, limma (https://www.bioconductor.org/packages/release/bioc/html/limma.html), was applied to analyze differential expression. Bayes tests were used to evaluate statistical significance. Adjusted $P$-value (Adj.$P$) < 0.05 and $\log_2$ fold-change (FC) > 1 were used as a threshold for DEG identifying. After respective differential expression analysis, the results were intersected to determine DEGs for all datasets.

**Functional Enrichment Analyses of DEGs**

To investigate the functional roles of the screened DEGs, Database for Annotation, Visualization and Integrated Discovery (DAVID; version 6.8, https://david.ncifcrf.gov/) were employed to perform the Gene Ontology (GO) enrichment analysis under three categories, biological process (BP), cellular component (CC) and molecular function (MF). KEGG pathway analysis was carried out using the KOBAS database (version 3.0, http://kobas.cbi.pku.edu.cn/). False discovery rate (FDR) < 0.05 was considered statistically significant.

**Protein–Protein Interaction (PPI) Network Construction and Hub Gene Identification**

Basing on the DEGs screened above, we performed a PPI analysis using the Search Tool for the Retrieval of Interacting Genes (STRING; version 11.0, https://string-db.org/). The combined score > 0.4 was used as the cut-off criteria. Subsequently, Cytoscape (version 3.7.1) software was utilized for PPI network analysis and visualization. The plug-in MCODE (version 1.5, http://apps.cytoscape.org/apps/mcode) was employed to identify significant modules of the constructed network. Moreover, to investigate the hub (key) genes, we used cytohubba (version 0.1, apps.cytoscape.org/apps/cytohubba), another plug-in of Cytoscape which could screen the important nodes from a given network by several topology algorithms. In the present study, the maximal clique centrality (MCC) method was selected to identify hub genes, and other parameters of cytohubba were set to default.

**The Cancer Genome Atlas (TCGA) Database Survival Analysis**

To further explore the effect of hub genes expression on patients’ clinical outcomes, we employed GEPIA2 (URL:
online tool to conduct a survival analysis of data from TCGA (URL: https://cancergenome.nih.gov/). Breast cancer samples were derived into hub gene high- and low-expression groups using the median transcript per million (TPM) value as a cutoff, respectively. A Log rank test was used for the hypothesis tests. To further confirm the biological functions of hub genes on tumorigenesis, Cox regression analysis was performed to evaluate the association between hub gene expression and prognosis in multiple malignancies, including breast cancer (BRCA), cervical squamous cell carcinoma (CESC), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), skin cutaneous melanoma (SKCM) and pancreatic adenocarcinoma (PAAD). P-value < 0.05 was considered to be statistically significant.

Clinical Tissues Collection
Eight IBC and twenty non-IBC clinical samples were collected by biopsy or surgery from patients who were primarily diagnosed between 2013 and 2019 at the Affiliated Hospital of Jiangnan University. The clinical and histological features of the specimens are shown in Supplementary Table 2. Immediately after removal, all specimens were transferred into liquid nitrogen and finally stored at −80°C before RNA extraction. Each patient enrolled in the present study has signed informed consent, and this study was approved by the ethical committee of the hospital.

RNA Extraction and Quantitative Reverse Transcriptional PCR (qRT-PCR)
Total RNA of clinical specimens was isolated using the E.Z.N.A Total RNA Kit I (Omega, USA) according to the protocol.
Complementary DNA (cDNA) was synthesized subsequently by using the Prime Script RT Reagent Kit (Takara Bio, Japan). Real-time PCR was performed on the CFX96 Touch Real-Time PCR Detection System (Bio-rad, USA) to measure the expression levels of hub genes using the comparative Ct method. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was considered as a normalization control. The sequences of primers are shown in Supplementary Table 3.

**Results**

**Identification of DEGs**

In the present study, we performed a multi-step analysis to explore hub genes of inflammatory breast cancer, and further investigated their potential biological functions and effects on the prognosis of breast cancer. “Hub genes” are defined as core genes whose expression products are hubs in the PPI network. Firstly, four IBC datasets (GSE5847, GSE22597,
GSE23720, and GSE45581) which contain a total of 131 IBC samples and 238 non-IBC samples, were downloaded from the GEO database. Differential expression analysis was performed for each dataset basing on the cut-off criteria (Figure 1A–D). Furthermore, the overlapping genes among the four datasets were considered as intersecting DEGs, including 53 upregulated and 61 downregulated genes (Figure 1E). The expression heatmap of the top 10 up- and down-regulated DEGs are shown in Figure 1F.

### Functional and Pathway Enrichment Analysis

GO and KEGG pathway enrichment analyses were performed to characterize the functional roles of selected DEGs. For category BP of GO analysis, the DEGs were significantly enriched in signal transduction, cell adhesion, inflammatory response, positive regulation of cell proliferation and transcription (Figure 2A). For category CC, the expression product of the DEGs were mainly located in the plasma membrane, exosome, extracellular region (Figure 2B). Moreover, for category MF, the main functions of proteins encoded by the DEGs were DNA binding, protein binding, protein homodimerization activity, etc (Figure 2C).

According to KEGG analysis, multiple pathways associated with tumorigenesis were enriched, including focal adhesion, Wnt, PI3K-AKT, VEGF, and Ras signaling pathways, etc. In addition, the DEGs were also enriched in cell adhesion and extracellular matrix (ECM) receptor interaction (Figure 2D).

### PPI Network Construction and Module Analysis

To further explore the interactions of DEGs in IBC, a PPI network involving 101 nodes and 471 edges (Figure 3A),
was constructed depending on the STRING database. Subsequently, Cytoscape plug-in MCODE was used for clustering analysis, and four functional modules were identified from the whole network (Figure 3B–E). Moreover, we performed pathway enrichment analyses for genes of the four modules to understand their potential bio-functions. According to the results, these modules were mainly associated with oncogenesis and cell adhesion, the enriched pathways included MAPK, ErbB, Wnt, VEGF signaling pathways, ECM-receptor interaction, and focal adhesion, etc (Table 1).

**Hub Genes Screening, Survival Analysis, and Expression Validation**
Plug-in cytohubba of Cytoscape was used to identify hub genes from the whole PPI network. According to the ranking by the MMC method, five hub genes (PTPRC, IL6, SELL, CD40, and SPN) were filtered out. The sub-network of these genes is demonstrated in Figure 4A.

To further explore the effect of hub genes expression on patients’ clinical outcomes, a total of 1069 breast cancer samples were derived into high- and low-expression groups according to the median expression level of each hub gene respectively, and survival analysis was performed by GEPIA2. In the five hub genes, SELL, and SPN were significantly associated with the overall survival of breast cancer (log-rank \( P_{\text{SELL}} = 0.016, P_{\text{SPN}} = 0.024; \) Figure 4B and C). However, PTPRC, IL6 and CD40 were not significantly correlated to breast cancer prognosis (log-rank \( P > 0.05; \) Supplementary Figure 1). Moreover, cox regression analysis was performed in multiple malignancies. The survival map showed that PTPRC, SELL, CD40, and SPN tended to reduce the risk of tumor diseases, and IL6 tended to be associated with poorer prognosis (Figure 4D).

In addition, we validated the expression of hub genes in clinical specimens. As shown in Figure 5, the relative expression level of PTPRC, SELL, CD40 and SPN in IBC samples were significantly lower than in non-IBC samples (\( P < 0.05 \)), and IL6 expression in IBC samples were significantly higher (\( P < 0.05 \)). This result was consistent with the bioinformatics results.

**Discussion**
In the present study, a total of 114 DEGs, including 53 upregulated and 61 downregulated genes, were identified from four GEO datasets of IBC. The results of GO and KEGG enrichment analyses indicated that the DEGs were strongly associated with multiple cancer-related functions and pathways, such as cell adhesion, cell proliferation, Wnt, PI3K-AKT, VEGF, and Ras signaling pathways, etc. The PPI network of selected DEGs was constructed basing on the STRING database, and module analysis was performed to further explore functional sub-networks. Finally, five hub genes were screened out from the whole PPI network, including PTPRC, IL6, SELL, CD40, and SPN. Survival analysis and expression validation of the five hub genes verified the robustness of the above results.

In recent years, an increasing number of studies working on DEGs in human cancers have been published. Due to the development of high-throughput technologies, including microarray and RNA-sequencing, current researches have been paying more attention to tumor genomes and transcriptomes instead of isolated genes. In a study basing on the original IBC dataset GSE5847 from GEO, 368 DEGs were identified. According to GO and

| Table 1 KEGG Enrichment Analysis of Top 4 Modules Identified from PPI Network |
|-----------------------------|-----------------|-----------------|
| **Module**                  | **KEGG Entry**  | **Description** |
| Module 1                    | hsa05200        | Pathways in cancer |
|                             | hsa04010        | MAPK signaling pathway |
|                             | hsa04012        | ErbB signaling pathway |
| Module 2                    | hsa04514        | Cell adhesion molecules (CAMs) |
|                             | hsa04810        | Regulation of actin cytoskeleton |
| Module 3                    | hsa04310        | Wnt signaling pathway |
|                             | hsa04370        | VEGF signaling pathway |
| Module 4                    | hsa04512        | ECM-receptor interaction |
|                             | hsa04510        | Focal adhesion |

**Abbreviations:** FDR, false discovery rate; ECM, extracellular matrix.
KEGG analyses, the DEGs were mainly enriched in cell adhesion, ECM-receptor interaction, angiogenesis, and focal adhesion, which were also involved in our results. However, for hub genes, there is no intersection between this study and ours. Compared with similar published researches on IBC, our study has a much larger sample size, consisting of 131 IBC and 238 non-IBC samples from up to four unique datasets. This ensures the robustness and high confidence in our results.

Based on the PPI network, the five hub genes were identified to be significantly associated with IBC pathogenesis and progression. Among these genes, IL-6 has been reported to be a candidate oncogene of IBC. IL-6 encodes a cytokine that functions in inflammation and the maturation of B cells, and is implicated in a wide range of inflammation-associated diseases, including almost all types of tumors. Molecular evidence has suggested that inflammatory signaling pathways are upregulated in IBC, especially NF-κB and IL-6. Moreover, aberrant IL-6 expression is proved to increase tumor stage, lymph node involvement, recurrence risk, and distant metastasis of breast cancer. In this study, IL-6 was identified to be a core gene of IBC and was strongly associated with poor prognosis of breast cancer. These findings are consistent with the previous ones.

Another inflammation associated gene, CD40, was also screened as a hub gene in our study. It is a member of the TNF-receptor superfamily, is essential for mediating a broad

Figure 4 Hub genes identification and survival analysis.
Notes: (A) PPI sub-network of the five hub genes. (B) Kaplan–Meier survival analysis of SELL. (C) Kaplan–Meier survival analysis of SPN. (D) Survival map of hub genes in multiple malignancies.
variety of immune and inflammatory responses. It is widely recognized that CD40 plays key roles in tumor immunity by inducting productive anti-tumor T cell immune responses, and its agonists have been applied to cancer immunotherapy in various malignancies. Interestingly, a study of gene expression in IBC reported that, compared with non-IBC samples, a series of NF-κB-regulated genes (including CD40) were observed to be upregulated in IBCs. Our study showed that CD40 is significantly downregulated in IBC, and is associated with favorable prognosis. In combination with the majority of previous studies, we tend to believe the tumor suppression effect of CD40 on IBC.

For hub gene PTPRC, SELL, and SPN, there have been few reports about their effect on IBC. PTPRC encodes a protein which is a member of protein tyrosine phosphatase (PTP) family, and has been recognized as an essential regulator of T- and B-cell antigen receptor signaling. SELL, which is also known as CD62L, is also related to inflammation and immunity. The protein it encodes is a cell surface adhesion molecule and takes part in selectin-induced immune modulation in cancer progression. SPN is also named CD43, it encodes a glycoprotein that functions in antigen-specific activation of T cells, and is reported to be a candidate biomarker of tumor diseases. In the present study, these genes are down-regulated in IBC and encode key proteins according to the PPI network. However, more molecular evidence is still required to verify the association between these genes and IBC.

To the best of our knowledge, our study is the first one to perform a combined bioinformatic analysis of multiple transcriptome microarray datasets on IBC. Compared with the small sample sized, single centered researches, the results of our study are less interfered by confounding factors and are more credible. On the other hand, our study has several limitations. Firstly, since TCGA database does not provide enough information for us to identify IBC samples from non-IBC ones, the survival analysis of the identified hub genes was performed using samples of all breast cancer subtypes. Therefore, although this result contributes to the understanding of the biological relevance of the hub genes, further study is still required to evaluate their prognostic values on IBC. Secondly, the association between the hub genes and clinicopathological features of IBC was not analyzed in this study due to the

Figure 5 Validation of hub genes expression in IBC and non-IBC clinical samples by qRT-PCR.
Notes: The relative expression level of the five hub genes PTPRC (A), IL6 (B), SELL (C), CD40 (D) and SPN (E) are shown in violin plot.
Abbreviation: qRT-PCR, quantitative reverse transcriptional polymerase chain reaction.
incomplete clinical information of GEO samples. We will perform these analyses when we collect enough specimens of IBC.

Take together, the identified five hub genes, PTPRC, IL6, SELL, CD40, and SPN, have been suggested to be potential key genes in IBC development and progression. These findings may provide a novel understanding of the pathogenesis of IBC, and these genes may be further considered as therapeutic targets. Moreover, high SELL and SPN expression were associated with a favorable prognosis of breast cancer, suggesting that the identified hub genes may have prognostic value. Further study is still needed to investigate the association between hub genes and survival in IBC samples and elucidate the underlying mechanisms of their effect on IBC.

Conclusion
In the present study, 114 DEGs between IBC and non-IBC were identified by bioinformatic analysis based on transcriptome microarray data from GEO database, and five hub genes (PTPRC, IL6, SELL, CD40, and SPN) were screened out. These genes are related to inflammation and immunity processes and may play key roles in IBC development and progression. Further study is needed to validate the results of our research.

Disclosure
The authors report no conflicts of interest in this work.

References


