Abstract. Triple-negative breast cancer (TNBC) has attracted great attention due to its unique biology, poor prognosis, and aggressiveness. TNBC patients are more likely to suffer from metastasis. We screened and identified the TNBC-specific genes as potential biomarkers. A total of 167 breast cancer samples (45 TNBC and 122 non-TNBC) were used in the integrated analysis. Gene expression microarrays were used to screen the differentially expressed genes. We identified 65 core DEGs. According to the GO and KEGG analysis, the gene function enrichment in TNBC was revealed, such as basal cell carcinoma, prostate cancer, oocyte meiosis and choline metabolism in cancer pathways. Moreover, the PPI network reconstruction would benefit the screening of hubs. A RFS analysis of TNBC-specific genes was also conducted. RT-PCR was used to validate the expression pattern of hubs in TNBC. Finally, nine genes were identified and all of them were novel, specific and higher dysregulation expressed genes in TNBC. Such that, these genes will serve as potential biomarkers in TNBC and benefit further research in TNBC.

Introduction

Breast cancer is one of the most common malignancies in women of United States, which results in >40,000 deaths every year. These breast tumors consist of phenotypically diverse populations (1). Breast cancer can be divided into subtypes of two estrogen receptor (ER)-positive (luminal A and luminal B types), human epidermal growth factor receptor 2 (HER2) enriched, and triple-negative breast cancer (TNBC) (2). Triple-negative breast cancer is ER-negative, PgR-negative, and HER2-negative using clinical assays, which accounts for approximately 15% of all types of breast cancer (3,4). Furthermore, TNBC has attracted a tremendous amount of attention due to its unique biology, overall poor prognosis, aggressive, and pattern of metastases (5). Therefore, it is urgent to gain insight into the therapeutic targets when compared with endocrine-sensitive and HER2-positive breast cancer.

Gene expression profiling could categorize the characteristics of different subtypes and verify the genes as novel therapeutic targets (6). A limited number of studies have been conducted on the gene expression profile of TNBC. Yang et al discovered that FZD7 plays a critical role in cell proliferation in TNBC (7). Their finding have suggested that several Wnt pathway genes, such as FZD7, low density lipoprotein receptor-related protein 6 and TCF7 are overexpressed in TNBC (7). In another study, Mathe et al identified the novel genes associated with the lymph node metastasis in TNBC. According to an analysis of 33 TNBCs, 17 normal adjacent tissues and 15 lymph node metastases were identified (8). Wang et al identified the CDK7-dependent transcriptional addiction in triple-negative breast cancer (9). Furthermore, Abramovitz et al identified a 30-biomarker gene set that could distinguish the breast cancer into subtypes. This study also uses the subset genes for prognostication of OS and RFS (10).

He et al conducted an analysis to study the molecular characteristics of triple-negative breast cancer using microarray. Al-Ejeh et al conducted a meta-analysis for the gene expression profile in TNBC and clarified the genes for prognostication and therapy. In this integrated analysis, the combination of clinical samples with different types of chemotherapy from some databases would increase the heterogeneity (10,11).

Although the gene expression profiles convey significant findings in TNBC, there is a lack of sufficient conclusions to uncover the central mechanisms in TNBC. This situation requires the integration of different datasets. In the present study, we first conducted an integrated analysis of TNBC relative to the non-TNBC. The differentially expressed genes were identified by statistical analysis. Then, the core DEGs were selected for functional annotation. The PPI was also reconstructed. These analyses were conducted to reveal the biological pathway and molecular mechanisms regarding
TNBC. Finally, the selected hub genes would be the potential and specific biomarkers with prognostic value and used as the treatment targets in the future.

Materials and methods

Microarray data analysis. To identify the gene expression mode of TNBC, microarray data were collected from the Gene Expression Ominibus (GEO, https://www.ncbi.nlm.nih.gov/geo/), which is freely available for users. Three independent microarray databases (GSE27447, GSE61724, GSE18864) were downloaded. A total of 19 samples were studied in GSE27447 including 14 TNBC and 5 non-TNBC. GSE61724 consisted of 16 TNBC and 48 non-TNBC breast cancer samples. Moreover, 24 TNBC and 60 non-TNBC samples were involved in GSE18864.

Identification of differentially expressed genes (DEGs). Each dataset was analyzed independently. The raw data were normalized by R/Bioconductor software. The method of linear models for microarray data (LiMMA) was employed to screen normalized by R/Bioconductor software. The method of linear models for microarray data (LiMMA) was employed to screen the differentially expressed genes (DEGs). P-value <0.05 and |logFC| >1 were the cut-off criteria to identify the DEGs. The overall DEGs in the 4 datasets were involved in GSE18864.

The relative expression pattern of core DEGs in TNBC. The raw data was normalized. The relative expression equals the log2FC in each dataset. It was performed with the quotient of average log expression of TNBC and non-TNBC. The heat map was used to demonstrate and visualize the result using Package ‘gplots’ of R (13). A cluster analysis is also presented. According to the average of the relative expression, the expression pattern of core DEGs are presented.

Functional annotation of DEGs. The GO enrichment analysis was conducted to gain insight into the biological process of DEGs. GO included 3 groups: molecular function, biological process and cellular component. Kyoto Encyclopedia of Genes Genomes (KEGG, http://www.genome.jp/kegg/) is a knowledge database for systematic analysis of function annotation. In this study, the GO and KEGG were performed using web-based software KOBAS (http://kobas.cbi.pku.edu.cn/). P<0.05 was set as the threshold.

PPI network reconstruction. A protein-protein interaction (PPIs) analysis was conducted to visualize the functional relationships between the DEGs and other genes at a molecular level (4), which helped uncover the mechanisms in TNBC. The DEGs were used for protein-protein interaction (PPI) networks. These genes were submitted to the Biological General Repository for Interaction Datasets (BioGRID) (http://thebiogrid.org/) and retrieval of interactors. A topological analysis and visualization were conducted by CytoHubba plugin of Cytoscape (cytoHubba identifying hub) to screen the hub protein of the network.

Survival analysis of DEGs. Kaplan-Meier plotter (KM plotter, www.kmplot.com) is an online survival analysis tool to assess the effect of 22,277 genes on breast cancer prognosis from the microarrays of 1,809 patients (14). The patients with TNBC were split into two groups by the expression status of specific genes. The relapse-free survival (RFS) was depicted including the hazard ratio (HR) with 90% confidence intervals and the log rank P-value (15).

RT-PCR validation. To identify the markers, three TNBC lines (MDA-MB-231, MDA-MB-435, MDA-MB-468) and three non-TNBC lines (MCF-7, MDA-MB-453, SK-BR-3) were chosen (4). The total RNA was extracted with the TRIzol method. First-strand cDNA synthesis and RT-PCR validation were performed (4). The primers were designed with Primer Premier5. Reaction was set with 3 replicates, and run under the following conditions: 94˚C for 2 min, 35 cycles of 94˚C for 30 sec, 55˚C for 30 sec, 72˚C for 30 sec, and 72˚C for 10 min. The relative expression used the 2^{-ΔΔCt} calculation (16) with the human GAPDH gene as endogenous control for gene expression analysis. Nine genes were selected to perform real-time PCR in seven human cancer cell lines, in which PROM1 and KLK6 were upregulated genes and KRT18, GPR160, CMBL, AGR3, CREB3L4, CRIP1 and SDR16C5 were downregulated genes.

Results

Microarray analysis. In this study, the available gene expression datasets (GSE27447, GSE61724, and GSE18864) were used to gain insight into the molecular characteristics of TNBC. A total of 167 samples were analyzed (Table I). According to the statistical analysis, a total of 814 genes were contained in GSE27447 (442 downregulated and 372 upregulated genes) (Fig. 1A). Only 51 genes were identified in GSE61724 (37 downregulated genes and 14 upregulated genes) (Fig. 1B). In GSE18864, 159 genes were selected including 81 downregulated and 78 upregulated genes (Fig. 1C). The results are presented in Table I and Fig. 1.

Core DEGs of TNBC. In order to screen common specific genes in TNBC, the Veen diagram was processed. The overlap between the 3 datasets was investigated to identify how many common genes were involved in at least 2 datasets. These genes were named as the ‘core DEGs’. Twelve upregulated genes and 53 downregulated genes were identified. The upregulated genes are shown in Fig. 2A. In addition, the downregulated genes are shown in Fig. 2B. Only HORMAD1 in the upregulated genes and 5 genes (GPR160, NAT1, AGR2, AGR3 and ERBB4) were common in all of the three datasets (Fig. 2).

The relative expression pattern of core DEGs in TNBC. Among the core DEGs, three datasets revealed similar expression pattern. The cluster result of the sample revealed that GSE18864 was most consistent with the average expression, followed by GSE61724 and GSE27447. HORMAD1, EGF5, KLK6 and GABRP represented higher expression than other genes. AGR2, AGR3, ANKR3D0A, NME5 and CYP4Z3P had lower expression (Fig. 3A and B). These significant genes would be potential biomarkers to characterize TNBC.

KEGG and GO results. To gain insight into the biological pathways in TNBC, a KEGG pathway analysis was conducted (Fig. 4A and B). A plot was made with R package clusterProfiler with a P-value cut-off of <0.05 (17). The result showed
that ‘nicotine addiction’ (GABRP) and ‘basal cell carcinoma’ (FZD7) pathways were involved in the ‘upregulated genes. In addition, ‘ubiquinone and other terpenoid-quinone biosynthesis pathways’ (NQO1), ‘prostate cancer’ (CREB3L4, AR), ‘oocyte meiosis’ (AR, PGR), ‘nitrogen metabolism’ (CA12), ‘caffeine metabolism’ (NAT1), ‘choline metabolism in cancer’ (SLC44A5, SLC44A4), ‘chemical carcinogenesis’ (GSTM3, NAT1) were included in downregulated genes.

Based the GO function annotation, the upregulated genes were categorized into 39 groups (Fig. 4C). The top five terms were: ectodermal cell differentiation, ectoderm development, cell fate commitment involved in formation of primary germ layer, tissue regeneration and negative regulation of cysteine-type endopeptidase activity involved in apoptotic process. Furthermore, 32 GO terms were involved in the downregulated genes. The top five terms were: xenobiotic

<table>
<thead>
<tr>
<th>GEO series</th>
<th>No. of TNBC</th>
<th>No. of non-TNBC</th>
<th>Downregulated genes</th>
<th>Upregulated genes</th>
<th>Total DEGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE27447</td>
<td>5</td>
<td>14</td>
<td>442</td>
<td>372</td>
<td>814</td>
</tr>
<tr>
<td>GSE61724</td>
<td>16</td>
<td>48</td>
<td>37</td>
<td>14</td>
<td>51</td>
</tr>
<tr>
<td>GSE18864</td>
<td>24</td>
<td>60</td>
<td>81</td>
<td>78</td>
<td>159</td>
</tr>
</tbody>
</table>

Table I. Microarray database and DEGs in the present study.

Figure 1. DEGs revealed in 3 selected microarrays by volcano plot. The log2 of the fold change is shown on the x-axis. The negative log of P-value is shown on the y-axis. (A) The presentation of GSE27447, (B) the presentation of GSE61724, (C) the presentation of GSE18864.

Figure 2. The Venn plot depicting the overlap of the DEGs. (A) Upregulated genes, (B) downregulated genes.
metabolic process, cellular response to xenobiotic stimulus, choline transmembrane transporter activity, response to xenobiotic stimulus and dystroglycan binding (Fig. 4C). P-value <0.01 was the cut-off criterion.

Construction of PPI network in TNBC. The reconstruction of the PPI network in TNBC was based on the 65 core DEGs with their interactors. As a result, the PPI network consisted of 133 nodes and 188 edges in the upregulated interactions (Fig. 5A), and in order to screen the hub, the top 50 groups were selected from the network to construct a subnetwork (Fig. 5B). Furthermore, 872 nodes and 995 edges in the downregulated interactions (Fig. 5C). The top 50 groups were also selected from the network to construct a subnetwork (Fig. 5D).

A local-based metric (i.e., degree), and a global-based metric (i.e., betweenness centrality) were adopted to determine...
the main genes (18). By these indexes, a set of genes were identified, including 5 upregulated genes (CRYAB, LAMP3, CSF2RB, KLK6, FZD7) and 5 downregulated genes (AR, KRT18, WWP1, PGR, ERBB4) (Table II). More attention should be paid to these genes in further research since they were chosen as the potential candidate biomarkers of TNBC.

The RFS analysis of TNBC-specific genes. The prognostic value of TNBC-specific genes was also implemented in www.kmplot.com. The relapse-free survival (RFS) for patients with TNBC was assessed by the low and high expression of each gene. By the log-rank P-value, the top 6 significant genes (LAMP3, C16orf54, CSF2RB, CA12, ERBB4 and GPR160) were listed (Fig. 6). The hR of LAMP3 was 0.5 and log-rank P-value of it was 0.0019 (Fig. 6A). The hR of C16orf54 was 0.45 and log-rank P-value was 0.0051 (Fig. 6B). The hR of CSF2RB was 0.55 and log rank P-value was 0.0067 (Fig. 6C). The hR of CA12 was 1.77 and log rank P-value was 0.0087 (Fig. 6D). The hR of ERBB4 was 0.52 and log rank P-value was 0.01 (Fig. 6E). The HR of GPR160 was 0.53 and logrank P-value was 0.026 (Fig. 6F). The result showed that the

Table II. Topological characters of the core genes.

<table>
<thead>
<tr>
<th>GEO series</th>
<th>Protein symbol</th>
<th>Degree</th>
<th>Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upregulated hub</td>
<td>CRYAB</td>
<td>64</td>
<td>1957.5</td>
</tr>
<tr>
<td></td>
<td>LAMP3</td>
<td>64</td>
<td>1957.5</td>
</tr>
<tr>
<td></td>
<td>CSF2RB</td>
<td>14</td>
<td>1105</td>
</tr>
<tr>
<td></td>
<td>KLK6</td>
<td>12</td>
<td>946</td>
</tr>
<tr>
<td></td>
<td>FZD7</td>
<td>10</td>
<td>117</td>
</tr>
<tr>
<td>Downregulated hub</td>
<td>AR</td>
<td>238</td>
<td>210081.7991</td>
</tr>
<tr>
<td></td>
<td>KRT18</td>
<td>90</td>
<td>94918.36244</td>
</tr>
<tr>
<td></td>
<td>WWP1</td>
<td>84</td>
<td>79340.42971</td>
</tr>
<tr>
<td></td>
<td>PGR</td>
<td>63</td>
<td>31428.61489</td>
</tr>
<tr>
<td></td>
<td>ERBB4</td>
<td>51</td>
<td>38581.13683</td>
</tr>
</tbody>
</table>
high expression of CA12 and the low expression of LAMP3, C1orf54, CSF2RB, ERBB4 and GPR160 exhibited associations with unfavorable relapse-free survival.

**RT-PCR validation.** To validate the DEGs in the integrated analysis, 9 TNBC-specific genes were chosen from DEGs by combining the PPI and marker selection result. PROM1 and KLK6 were selected as the upregulated genes. In addition, KRT18, GPR160, CMBL, AGR3, CREB3L4, CRIP1 and SDR16C5 were chosen as the downregulated genes. Primers are listed in Table III. In general, the expression patterns in TNBC and non-TNBC lines were consistent with that in the integrated analysis (Fig. 7). The KLK6 and PROM1 genes were highly expressed in TNBC cell lines (Fig. 7A and B). In addition, the SDR16C5 was lowly expressed in TNBC cell lines (Fig. 7C). The KRT18 was highly expressed in TNBC cell lines (Fig. 7D). The GPR160 was lowly expressed in TNBC cell lines (Fig. 7E). The CMBL was lowly expressed in TNBC cell lines (Fig. 7F). The AGR3 was lowly expressed in TNBC cell lines (Fig. 7G). The CREB3L4 was also lowly expressed in TNBC cell lines (Fig. 7H). However, the expression of CRIP1 was slightly higher in TNBC lines compared with that in non-TNBC lines (Fig. 7I). Finally, the RT-PCR result could be also used to identify the reliability of the integrated analysis.

**Discussion**

Identification of valid biomarkers in diagnosis, treatment and prognosis in TNBC is a challenging task. Although individual datasets have produced hundreds of genes (Table I), only six common DEGs are achieved. In consideration of the sample limitation and heterogeneity from patients and array platforms,
Figure 6. Prognostic value of 6 genes of TNBC patients. The related Affymetrix IDs were acquired. (A) 205569_at (LAMP3), (B) 1559584_a_at (C16orf54), (C) 205159_at (CSF2RB), (D) 215867_x_at (CA12), (E) 214053_at (ERBB4), (F) 223423_at (GPR160). HR, hazard ratio; CI, confidence interval.

Figure 7. RT-PCR validation of ten genes in six breast cancer cell lines. (A) The gene expression of KLK6, (B) the gene expression of PROM1, (C) the gene expression of SDR16C5, (D) the gene expression of KRT18, (E) the gene expression of GPR160, (F) the gene expression of CMBL, (G) the gene expression of AGR3, (H) the gene expression of CREB3L4, (I) the gene expression of CRIP1.
individual analyses were only confined to uncover the central mechanisms behind TNBC (18). The integrated analysis from different gene expression profiles would be efficient for insight into TNBC.

In our study, we achieved 65 DEGs from at least 2 datasets (Fig. 2). Furthermore, KEGG enrichment result is related to a range of cancer pathways, such as basal cell carcinoma (FZD7), prostate cancer (CREB3L4, AR), choline metabolism in cancer (SLC44A5, SLC44A4) and chemical carcinogenesis (GSTM3, NAT1), which provide more evidence to uncover the mechanisms of TNBC (Fig. 4). The GO annotation was mainly enriched in ectodermal cell differentiation, tissue regeneration, xenobiotic metabolic process, and prostate gland development. It is consistent with KEGG annotation. Then, the PPI network was reconstructed.

We chose a global and local metric to investigate the interaction between the DEGs and other intimate genes in TNBC. The main genes in the network would benefit the screening of hubs. These hub genes deserve more attention and further study since they are potential candidate biomarkers for TNBC.

It has been reported that FZD7 is overexpressed in 67% of TNBC (19). The FZD7-involved canonical Wnt pathway is the basis for the formation of TNBC. It could be regarded as the biomarker and potential therapeutic target of TNBC (7). LAMP3 is a protein included in cell proliferation term of GO. The activation of the IFN pathway is an early event following AC chemotherapy in 9 TNBC (20). Nagelkerke et al reported the role of UPR-induced LAMP3 in hypoxia-mediated migration of breast cancer cells (21).

FAM171A1 has been identified as an overexpressed gene in a large meta-analysis of seven patients with ER-negative BrCa tumor, which is also revealed as a hub gene (22). CSF2RB is related to the molecular profiling and computational network of TAZ-mediated mammary tumorigenesis (23). The overexpression of IRX1 is associated with growth arrest in gastric cancer, which can inhibit peritoneal spreading and metastasis (24). However, it has not been reported in the TNBC. In the present study, it was found that IRX1 serves as a common gene in GSE27447 and GSE18864, which is a highly-expressed gene with average lgFC of 1.14 in TNBC compared with non-TNBC (Figs. 2 and 3).

C16orf54 is reported in immune/inflammation-related genes in the stromal gene set of breast cancer in multivariate Cox proportional hazard models with HR of 0.507. There is no research on C16orf54 in TNBC (25).

The role of UPR-induced LAMP3 in hypoxia-mediated migration of breast cancer cells (21).

Table III. The primers used in the RT-PCR validation.

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene full name</th>
<th>Sequence primers from 5’ to 3’</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAPDH</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>F: GCACCGTCAAGGGCTGAGAAC  R: GGATCTCCTCCCTGGAAGATG</td>
</tr>
<tr>
<td>AGR3</td>
<td>Anterior gradient 3, protein disulphide isomerase family member</td>
<td>F: CCAACCTTGGCATGCTACAAT  R: TCATCTCCTCCCTGAGATAG</td>
</tr>
<tr>
<td>CMBL</td>
<td>Carboxymethylenobutenolide homolog</td>
<td>F: CCCTCTGGCGACTGGTCTATC  R: GCACTGATCTCTCATGATCTCTG</td>
</tr>
<tr>
<td>CREB3L4</td>
<td>cAMP responsive element binding protein 3 like 4</td>
<td>F: CCCAGCTTACATGCATTCACA  R: AAGTCATCTCCGTGAGGTCGTA</td>
</tr>
<tr>
<td>CRIP1</td>
<td>Cysteine rich protein 1</td>
<td>F: GCTGAGCAGCAAGGCACCAAC  R: CAAACATGCTGCTAGCA</td>
</tr>
<tr>
<td>GPR160</td>
<td>G protein-coupled receptor 160</td>
<td>F: TTTCACTCTGGTATGATTTGGA  R: CATTCCTGTCCAGTGTTTTT</td>
</tr>
<tr>
<td>KLK6</td>
<td>Kallikrein related peptidase 6</td>
<td>F: GCCTACCTGCGCAGATCA  R: ATCCACTCCTGGCAGAATCC</td>
</tr>
<tr>
<td>KRT18</td>
<td>Keratin 18</td>
<td>F: CTCCGCAAGTCATTGATGA  R: TACCTCCTGCTGTTCCTTCTCT</td>
</tr>
<tr>
<td>PROM1</td>
<td>Prominin 1</td>
<td>F: TCCGAGAAGATACCTGAGAGA  R: CATACAAAAAGATAATACCCACAGAG</td>
</tr>
<tr>
<td>SDR16C5</td>
<td>Short chain dehydrogenase/reductase family 16C member 5</td>
<td>F: AGTAGCCGACCAGGTTAAAAAGA  R: TGGTACGATTCGCGTACCTT</td>
</tr>
</tbody>
</table>

F, forward; R, reverse.
Komit et al also reported that HORMAD1 is involved in the carcinogenesis of TNBC (27). In the present study, HORMAD1 possesses the IgFC 1.71, which could be a potential biomarker in TNBC. Chakrabarti et al reported that the ELF5 regulates the mammary gland stem cell fate by influencing the notch signaling (28). CRYAB, ELF5, and GABRP have been reported as the molecular targeted therapies (29). KLK6 may be a promising biomarker in epithelial-to-mesenchymal transition (EMT) due to the fact that it belongs to a family of serine proteases, which involves the clinical biomarker KLK3 in prostate cancer (30). However, there is no research on KLK6 in TNBC.

Among the downregulated core DEGs, the present analysis showed that some novel genes were not investigated in TNBC patients previously, such as KRT18, GPR160, CMBL, AGR3, CREB3L4, CRIPI and SDR16C5. Several downregulated genes have been reported in previous research including STC2, FOXA1, ATP8B1, SLC39A6, DUSP4, KRT18, TSPAN1, ERBB4, GATA3 (18). FOXA1, ANKR3D0A, CMBL, GPR160 and AGR2 were downregulated genes in TNBC compared with non-TNBC (4). AR has been used to indicate the prognosis and AGR2 were downregulated genes in TNBC compared to the clinical biomarker KLK3 in prostate cancer previously, such as KRT18, GPR160, CMBL, AGR3, proteases, which involves the clinical biomarker KLK3 in prostate cancer (30). However, there is no research on KLK6 in TNBC.

In conclusion, the TNBC-specific gene expression profiles have been identified and the RFS analyses of these TNBC-specific genes have been performed. Furthermore, by integrated analysis, we achieved a set of core DEGs in TNBC compared with non-TNBC. The function annotation and PPI network reconstruction would be conducive to understand the underlying mechanisms in TNBC. According to the combination of the relative expression pattern, GO and KEGG annotation and PPI network reconstruction, a set of hub genes have been identified, including HORMAD1, PROM1 and KLK6 from upregulated genes, KRT18, GPR160, CMBL, AGR3, CREB3L4, CRIPI and SDR16C5 from the downregulated genes, which are closely associated with TNBC. These hub genes are novel and specific genes that have not been investigated much. They also have higher expression of dysregulation. In this way, these hub genes will act as potential biomarkers in TNBC and contribute to the study of TNBC.

Acknowledgements
This study was supported by the National Natural Science Foundation of China (nos. 81373573 and 81460655) and the Mongolian Medicine Systems Biology Science and Technology Innovation Team Plan of Inner Mongolia.

References


