Abstract. The aim of the present study was to identify the hub genes and provide insight into the tumorigenesis and development of breast cancer. To examine the hub genes in breast cancer, integrated bioinformatics analysis was performed. Gene expression profiles were obtained from the Gene Expression Omnibus (GEO) database and the differentially expressed genes (DEGs) were identified using the 'limma' package in R. Gene Ontology enrichment analysis and Kyoto Encyclopedia of Genes and Genomes pathway analysis was used to determine the functional annotations and potential pathways of the DEGs. Subsequently, a protein-protein interaction network analysis and weighted correlation network analysis (WGCNA) were conducted to identify hub genes. To confirm the reliability of the identified hub genes, RNA gene expression profiles were obtained from The Cancer Genome Atlas (TCGA)-breast cancer database, and WGCNA was used to screen for genes that were markedly correlated with breast cancer. By combining the results from the GEO and TCGA datasets, 15 hub genes were identified to be associated with breast cancer pathophysiology. Overall survival analysis was performed to examine the association between the expression of hub genes and the overall survival time of patients with breast cancer. Higher expression of all hub genes was associated with significantly shorter overall survival in patients with breast cancer compared with patients with lower levels of expression of the respective gene.

Introduction
Breast cancer is one of the most severe types of cancer worldwide and is the leading cause of cancer-associated mortality in women (1). The breast cancer incidence rate varies widely across regions, with rates ranging from 0.194% in East Africa to 0.897% in Western Europe, and is increasing gradually (2). A number of risk factors are associated with breast cancer, including long-term fertility, the use of hormonal contraception, physical inactivity and alcohol consumption; however, its etiology and pathogenesis are not fully understood (3).

Following the implementation of several large-cohort human tumor genome projects, including The Cancer Genome Atlas (TCGA) (4) and the International Cancer Genome Consortium, a large amount of genomic data was generated from tumor samples, which has aided cancer studies considerably (5,6). In addition, smaller-scale cancer projects led by individual institutions have made substantial contributions and provided large amounts of valuable data, which have been deposited into public databases including Gene Expression Omnibus (GEO) (7). The availability of cancer genome data has accelerated and may continue to accelerate a comprehensive understanding of the genetics underlying cancer pathophysiology, facilitating examination of the underlying molecular mechanisms involved in cancer initiation and progression, and improving diagnostic methods and preventive strategies.

Due to the limitations of experimental techniques (8,9), the development and application of microarray and sequencing technology brought cancer research into a new era. High-throughput techniques have been widely used for global gene expression profiling, which reflects the molecular basis of tumor phenotypes and has been used to classify tumors, identify pathogenic genes for various tumors, examine tumorigenesis, and distinguish between the occurrence and progression of tumors (10-12). The large numbers of gene microarray datasets in public databases have facilitated comprehensive analyses of gene expression in different types of cancer (13-16). Using bioinformatics and correlating the results with clinical data, new biomarkers for the diagnosis, therapy and prognosis of different types of cancer can be identified.

‘Weighted correlation network analysis’ (WGCNA) is an R package that has been used as a data mining method to identify...
co-expressed gene modules and examine the association between gene networks and phenotypes (17).

A number of studies have applied the aforementioned methods to screen and identify specific hub genes in different types of cancer. In an oral squamous cell carcinoma study, the authors revealed 10 relevant hub genes, and the findings were supported at the transcriptional and translational levels (16); and, in a previous study, 10 hub genes associated with pancreatic ductal adenocarcinoma (PDAC) were identified, and the cell cycle pathway was reported to serve an important role in PDAC (18).

In the present study, two microarray gene expression datasets, GSE10810 (19) and GSE65194 (20), were obtained from the GEO database. Strict calibration and filtering were used to obtain differentially expressed genes (DEGs), which were subsequently subjected to Kyoto Encyclopedia of Genes and Genomes (KEGG; www.genome.ad.jp/kegg/) pathway analysis and Gene Ontology (GO; www.geneontology.org) enrichment analysis. An integrated DEG protein-protein interaction (PPI) network was constructed together with a WGCNA co-expression network. Using a series of bioinformatics approaches, hub genes were identified, and an enrichment analysis was used to identify possible key pathways associated with breast cancer. Breast cancer RNA expression profiles were additionally obtained from TCGA and the same strategy was performed to verify the data. A series of Kaplan-Meier (KM) survival plots were constructed to identify the association between the expression of hub genes and the prognosis of breast cancer. At present, only a small number of previous studies used a similar approach to screen hub genes for breast cancer to the best of our knowledge. In the present study, the scale of the KM survival analysis was larger, which may improve support of the findings from the genetic screen. The present findings may provide further insight into the tumorigenesis and development of breast cancer at the molecular level, and provide precise and practically valuable markers for the diagnosis, therapy, monitoring and prognosis of breast cancer.

Materials and methods

Datasets. GEO is a public database of gene expression profiles and sequence-based data that is freely available for users. The size of datasets and the unity of the platform were evaluated, and two gene expression profile datasets (GSE10810 and GSE65194) were selected and downloaded from GEO. GSE10810 and GSE65194 were obtained from the GPL570 platform [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. GSE10810 contained 58 samples, including 31 breast cancer samples and 27 normal control samples, whereas GSE65194 contained 130 breast cancer samples and 11 normal control samples.

The gene expression profiles based on RNA-sequencing were additionally obtained from TCGA ‘GDC TCGA-BRCA’ cohort, University of California, Santa Cruz Xena (https://xena.ucsc.edu/), which contained 1,104 breast cancer samples and 113 normal samples.

Filtering of DEGs. The ‘limma’ R package (21,22) was applied to filter the DEGs between the group of patients with breast cancer and the normal group from the GSE10810 and GSE65194 datasets. The P-value of each DEG was calculated and then adjusted using the Bonferroni method (23). The threshold used to select genes that were significantly differentially expressed at values 2-fold greater than those of the control group was a log fold change (FC)≥2 and a Bonferroni P<0.01.

KEGG pathway and GO enrichment analyses of DEGs. KEGG was used to systematically analyze and annotate gene functions (24). The GO database classified the genes into three functional groups: ‘Molecular function’ (MF), ‘biological process’ (BP) or ‘cellular component’ (CC) (25). In the present study, KEGG pathway and GO enrichment analyses of the DEGs obtained in the previous step were conducted using the ‘clusterProfiler’ R package (26) with a cut-off P=0.05.

Integration of the PPI network and cluster analysis. The Search Tool for the Retrieval of Interacting Genes (STRING) (27) is a biological database for predicting pairs of PPIs. The interactions between DEGs were evaluated using STRING and genes with a combined score ≥0.9 were defined as key DEGs. Subsequently, Cytoscape (version 3.6.1; https://cytoscape.org/) (28) was used to develop the PPI network of the key DEGs that were identified. Molecular complex detection (MCODE) (29), a Cytoscape plugin, was used with the default parameters to identify the most important modules of the PPI network.

Construction of the co-expression network and analysis of clinically significant modules. The co-expression network was established using WGCNA, an R package designed for the construction of weighted gene co-expression networks (17). In the present study, an automatic one-step network construction and module detection method in WGCNA was performed with the default settings, including the calculation of Pearson's correlation coefficients, an unsigned type of topological overlap matrix, a merge cut height of 0.25 and a default minimal module size. The first principal component calculation module eigengene (ME) was used to quantify the similarity of the co-expression of entire modules. Pearson's correlation coefficients were calculated to assess the potential correlations between MEs and the phenotype.

Hub gene selection. Key genes were identified in the most significant module of the PPI network. The phenotype-associated modules in the WGCNA network were additionally identified, and the genes in those modules were extracted. Hub genes common to the two networks were selected as candidates for further analysis and validation.

Construction of a co-expression network from the TCGA dataset for further validation. To confirm the reliability of the identified DEGs from the aforementioned two GEO datasets, the TCGA-breast cancer data were analyzed using the same strategy to obtain DEGs in the TCGA database. A one-step function of WGCNA was used for TCGA DEG network construction and the detection of consensus modules. The correlation coefficients between MEs and phenotypes was calculated. The candidate genes that also appeared in
the TCGA co-expression network were considered the true hub genes.

KM survival analysis. KM plotter (kmplot.com/) contains and utilizes expression data from 5,143 breast cancer patients (30). The median expression level of each gene was used to divide patients into two groups, and overall survival analysis was performed to determine the association between the expression levels of hub genes and the overall survival time of patients with breast cancer. The hazard ratio was provided, and the P-value was calculated using log-rank tests.

Results

DEG filtering. With thresholds of $\log_{10}(P<0.01, 540$ and $2,509$ DEGs were extracted from the expression profiles in the GSE10180 and GSE65194 datasets, respectively. Scatter volcano plots were plotted to illustrate the distribution of each gene according to the logFC and -$\log(P$-value) values (Fig. 1A). Subsequent to performing the integrated bioinformatics analysis, 322 consistent DEGs were identified from the two datasets (Fig. 1B; Table S1). Among those DEGs, 69 genes were upregulated and 253 genes were downregulated. In addition, the gene expression pattern was consistent in the two datasets, as presented in the heat map (Fig. IC).

KEGG pathway analysis and GO enrichment analysis. KEGG pathway analysis was used to examine the functions of the DEGs. The top results for each functional group are presented in Fig. 2A and Table SII. ‘Oocyte meiosis’, ‘cell cycle’ or ‘progesterone-mediated oocyte maturation’ were the pathways in which upregulated genes were primarily enriched. The majority of the downregulated genes were enriched in ‘PPAR signaling pathway’, ‘5’ AMPK signaling pathway’, ‘regulation of lipolysis in adipocytes’ and ‘adipocytokine signaling pathway’.

The enriched GO terms are listed in Fig. 2B and Table SII. In the BP group, the majority of the upregulated genes were enriched in ‘nuclear division’, ‘mitotic nuclear division’, ‘organelle fission’ and ‘regulation of nuclear division’, and the downregulated genes were primarily enriched in ‘lipid localization’. In the CC group, the upregulated genes were primarily enriched in ‘spindle’, ‘spindle pole’ and ‘condensed chromosome, centromeric region’, while downregulated genes were mainly enriched in ‘lipid droplet’. In the MF group, the majority of the downregulated genes were enriched in ‘amido binding’, ‘growth factor binding’, and ‘peptide binding’.

Identification of key DEGs and significant clusters in the PPI network. The online STRING database was used to identify the interactions between DEGs. Genes with a combined score $>0.9$ were defined as key DEGs. A total of 95 key DEGs as network nodes and 244 edges were used to construct the PPI network (Fig. 3A). MCODE recognized three of the most significant clusters and identified 28 genes from the PPI network (Fig. 3B and Table I).

Construction of the weighted co-expression network and identification of key modules. A WGCNA was performed to classify the DEGs into different modules based on the similarity of their expression patterns using the method of average linkage clustering. In the present study, three modules (MEblue, MEturquoise and MEgrey) were identified and marked with different colors (blue, turquoise and gray, respectively) in Fig. 4A. Subsequently, all of the samples in the dataset were classified into a breast cancer group (case) and normal control group (normal) as two phenotypes. The blue module of the MEs demonstrated the most marked correlation with the breast cancer phenotype (Fig. 4B). All 35 genes in the blue module were identified (ANLN, ASPM, AURKA, BIRC5, BUB1B, CCNB1, CCNB2, CDC20, CDK1, CDKN3, CENPF, CENPU, CKS2, CXCL10, DTL, GINS1, HMGB3, HN1, IGF1, KIAA0101, MAD2L1, MELK, NUSAP1, PBK, PRC1, PTTG1, RRM2, TK1, TOP2A, TPX2, TYMS, UBE2C, UBE2T, UHRF1, ZWINT) and considered to be the most relevant genes for breast cancer.

Hub gene selection. According to the WGCNA results, 35 genes were correlated with the blue module. Of the 35 genes, 17 genes identified from the PPI network were consistent across the WGCNA network. Therefore, 17 common network genes [aurora kinase A (AURKA), baculoviral inhibitor of apoptosis repeat-containing protein 5 (BIRC5), mitotic checkpoint serine/threonine-protein kinase BUB1β (BUB1B), G2/mitotic-specific cyclin-B (CCNB1), CCNB2, cell division cycle 20 (CDC20), cyclin-dependent kinase (CDK)1, CDK inhibitor 3 (CDKN3), centrosome protein F (CENPF), insulin-like growth factor 1 (IGF1), mitotic spindle assembly checkpoint protein MAD2A (MAD2L1), protein regulator of cytokinesis 1 (PRC1), pituitary tumor-transforming gene 1 protein (PTTG1), DNA topoisomerase 2α (TOP2A), targeting protein for Xklp2 (TPX2), ubiquitin-conjugating enzyme E2 C (UBE2C) and ZW10 interacting (ZWINT)] were considered hub genes and subjected to further analysis and validation.

Construction of the co-expression network from the TCGA dataset for further validation. For further validation, TCGA-breast cancer expression dataset of 1,217 samples was downloaded (113 normal samples and 1,104 tumor samples), and the WGCNA was performed using the aforementioned method. A total of 4 modules were identified, MEturquoise, MEblue, MEbrown and MEGrey (Fig. 4C), and the MEblue and MEbrown modules exhibited more marked correlations with the breast cancer phenotype compared with the other modules (Fig. 4D). Upon integrating the 17 hub genes obtained from the PPI and WGCNA network, the MAD2L1 and IGF1 genes were not found in MEblue or MEbrown module and were excluded from the hub gene list. The remaining 15 genes (AURKA, BIRC5, BUB1B, CCNB1, CCNB2, CDC20, CDK1, CDKN3, CENPF, PRC1, PTTG1, TOP2A, TPX2, UBE2C and ZWINT) were present in MEblue. Based on the results obtained, 15 hub genes were identified in breast cancer.

KEGG pathway analysis and GO enrichment analysis of final 15 hub genes. According to the result mentioned above, the 15 hub genes were identified and performed with KEGG and GO enrichment analysis subsequently. The result of KEGG enrichment analysis showed that these hub genes were primarily
associated with 'cell cycle', 'oocyte meiosis' and 'p53 signaling pathway', and the GO enrichment analysis also demonstrated that they were significantly involved in cell cycle, cell division, nuclear division and chromosome segregation processes (Fig. S1 and Table SIII).

**KM survival analysis.** To further evaluate the prognostic importance of the hub genes in the present study, overall survival analysis was performed to examine the association between the expression of each gene and the overall survival time of patients with breast cancer (Fig. 5). The patients were separated into a high expression and a low expression group based on the median level of expression. All hub genes with higher expression levels were associated with a significantly shorter overall survival time among patients with breast cancer, suggesting that these hub genes are associated with the pathophysiology of breast cancer to varying extents and may serve as potential prognostic biomarkers to monitor the severity of breast cancer or predict the survival time of patients, or as therapeutic targets.
Discussion

Although the treatment of breast cancer has improved, it remains the most prevalent malignant tumor with the highest increase in prevalence among women worldwide (2,3). The identification of the molecular mechanisms of breast cancer is crucial to its diagnosis, therapy and prognosis. DNA microarray gene expression profiles are widely used to explore DEGs involved in tumorigenesis, which has provided valuable information for clinical applications (31).
In the present study, two gene expression profile datasets (GSE10810 and GSE65194) from the GEO database were retrieved and analyzed. A total of 322 consistent DEGs were first identified (69 upregulated DEGs and 253 downregulated DEGs) using the 'limma' R package. The common DEGs were filtered out and 17 hub genes were identified that were detected in the PPI and WGCNA co-expression networks, using an integrated bioinformatics analysis. To further validate these breast cancer hub genes, TCGA-breast cancer data were screened and extracted into the modules associated with a breast cancer phenotype using WGCNA. Of the 17 genes, 15 true hub genes (AURKA, BIRC5, BUB1B, CCNB1, CCNB2, CDC20, CDK1, CDKN3, CENPF, PRC1, PTTG1, TOP2A, TPX2, UBE2C and ZWINT) that are closely associated with breast cancer were identified. The present study may provide valuable information for treatment decisions and prognostic predictions for breast cancer.

The 15 hub genes were all commonly overexpressed among patients with breast cancer. According to the KEGG enrichment analysis, these hub genes were primarily associated with 'cell cycle', 'oocyte meiosis' and 'p53 signaling pathway', and the GO enrichment analysis also demonstrated that they were significantly involved in cell cycle, cell division, nuclear division and chromosome segregation processes. Based on the results of the KEGG and GO analyses, these hub genes were generally associated with chromosome instability, and may serve an important role in tumorigenesis and tumor proliferation. Furthermore, based on the results of a KM survival analysis of each hub genes, it was identified that higher expression of each gene was associated with a worse prognosis among patients with breast cancer. Therefore, the 15 hub genes may be closely associated with breast cancer pathophysiology and represent potential prognostic biomarkers.

An additional 3 of the 15 hub genes (BUB1B, TOP2A and AURKA) are frequently identified in the OncoKB cancer gene list (oncokb.org/#/cancerGenes), which collects a large number of genes that are associated with cancer, based on their inclusion in various different sequencing panels, the Sanger Cancer
Gene Census or the comprehensive study (32). \textit{BUB1B} encodes a kinase that is associated with spindle checkpoint function and controls proper chromosome segregation during cell division (33). The \textit{BUB1B} protein is localized to the kinetochore and is involved in the anaphase-promoting complex/cyclosome inhibition, which delays the onset of anaphase and ensures proper chromosome segregation. Therefore, \textit{BUB1B} serves important roles in tumor proliferation and progression among multiple cancer types (34). As a checkpoint-associated gene, \textit{BUB1B} overexpression may increase the risk of cancer (35).

\textit{TOP2A} encodes DNA topoisomerase, an enzyme that controls the topological states of DNA and cell progression (36). The \textit{TOP2A} protein is primarily associated with processes such as chromatid separation, chromosome condensation, and the relief of torsional stress that occurs during DNA transcription and replication. The upregulation of \textit{TOP2A} is associated with female breast cancer and other cancer types (37). As a negative regulator of p53, \textit{AURKA} promotes tumor growth and cell

Table I. Key differentially expressed genes identified from the protein-protein interaction network.

<table>
<thead>
<tr>
<th>A, MCODE cluster 1</th>
<th>Gene</th>
<th>MCODE score</th>
<th>Degree</th>
<th>Clustering coefficient</th>
<th>Topological coefficient</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CDK1</td>
<td>7.2</td>
<td>16</td>
<td>0.73333333</td>
<td>0.75</td>
<td>Upregulated</td>
</tr>
<tr>
<td></td>
<td>UBE2C</td>
<td>8.836.363.636</td>
<td>11</td>
<td>0.92727273</td>
<td>0.85795455</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ZWINT</td>
<td>9</td>
<td>9</td>
<td>1</td>
<td>0.86805556</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CENPF</td>
<td>9</td>
<td>10</td>
<td>0.93333333</td>
<td>0.85625</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BUB1B</td>
<td>7.813.186.813</td>
<td>14</td>
<td>0.8021978</td>
<td>0.79017857</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAD2L1</td>
<td>7.2</td>
<td>15</td>
<td>0.77142857</td>
<td>0.77083333</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRC1</td>
<td>7.822.222.222</td>
<td>9</td>
<td>0.97222222</td>
<td>0.88888889</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BIRC5</td>
<td>7.961.538.462</td>
<td>13</td>
<td>0.80769231</td>
<td>0.79326923</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCNB2</td>
<td>7.2</td>
<td>16</td>
<td>0.73333333</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCNB1</td>
<td>7.2</td>
<td>16</td>
<td>0.73333333</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CDC20</td>
<td>7.2</td>
<td>16</td>
<td>0.73333333</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MLF1IP</td>
<td>9</td>
<td>9</td>
<td>1</td>
<td>0.86805556</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TOP2A</td>
<td>9</td>
<td>13</td>
<td>0.80769231</td>
<td>0.7846154</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AURKA</td>
<td>8.192.307.692</td>
<td>13</td>
<td>0.84615385</td>
<td>0.80288462</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CDKN3</td>
<td>8</td>
<td>9</td>
<td>0.94444444</td>
<td>0.88194444</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TPX2</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>0.875</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B, MCODE cluster 2</th>
<th>Gene</th>
<th>MCODE score</th>
<th>Degree</th>
<th>Clustering coefficient</th>
<th>Topological coefficient</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IGF1</td>
<td>3.733.333.333</td>
<td>5</td>
<td>0.9</td>
<td>0.92</td>
<td>Downregulated</td>
</tr>
<tr>
<td></td>
<td>CLU</td>
<td>3.733.333.333</td>
<td>5</td>
<td>0.9</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VWF</td>
<td>3.733.333.333</td>
<td>5</td>
<td>0.9</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CLEC3B</td>
<td>3.733.333.333</td>
<td>5</td>
<td>0.9</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FIGF</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>CFD</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C, MCODE cluster 3</th>
<th>Gene</th>
<th>MCODE score</th>
<th>Degree</th>
<th>Clustering coefficient</th>
<th>Topological coefficient</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FABP4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>Downregulated</td>
</tr>
<tr>
<td></td>
<td>PPARG</td>
<td>2.4</td>
<td>4</td>
<td>0.83333333</td>
<td>0.875</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ADIPOQ</td>
<td>2.7</td>
<td>4</td>
<td>0.83333333</td>
<td>0.875</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LEP</td>
<td>3.0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EBF1</td>
<td>2.7</td>
<td>4</td>
<td>0.83333333</td>
<td>0.875</td>
<td></td>
</tr>
</tbody>
</table>

MCODE, molecular complex detection.

Gene Census or the comprehensive study (32). \textit{BUB1B} encodes a kinase that is associated with spindle checkpoint function and controls proper chromosome segregation during cell division (33). The \textit{BUB1B} protein is localized to the kinetochore and is involved in the anaphase-promoting complex/cyclosome inhibition, which delays the onset of anaphase and ensures proper chromosome segregation. Therefore, \textit{BUB1B} serves important roles in tumor proliferation and progression among multiple cancer types (34). As a checkpoint-associated gene, \textit{BUB1B} overexpression may increase the risk of cancer (35). \textit{TOP2A} encodes DNA topoisomerase, an enzyme that controls the topological states of DNA and cell progression (36). The \textit{TOP2A} protein is primarily associated with processes such as chromatid separation, chromosome condensation, and the relief of torsional stress that occurs during DNA transcription and replication. The upregulation of \textit{TOP2A} is associated with female breast cancer and other cancer types (37). As a negative regulator of p53, \textit{AURKA} promotes tumor growth and cell
survival (38). Myc proto-oncogene and AURKA regulate the expression of the other genes at the transcriptional level and contribute to the development of liver carcinoma (39).

The proteins encoded by the remaining 12 hub genes are associated with a number of tumor processes. PTTG1 prevents separin from promoting sister chromatid separation by encoding for securin proteins, and promotes tumor cell growth (40) and malignancy in breast cancer (41). CDK1 promotes cell cycle gene expression and is necessary for accurate cell division (42). Strategies targeting CDK1 inhibit the proliferation of liver cancer cells (43). Therefore, UBE2C may affect the progression of cancer to a certain extent (44,45). BIRC5 is a protein-coding gene from the inhibitor of apoptosis gene family. BIRC5 functions as a negative regulator of apoptosis (46,47). CCNB1 and CCNB2 are members of the cyclin family. As important components in cell cycle regulation, CCNB1 and CCNB2 appear to function as oncogenes and are associated with breast cancer, according to numerous studies (48-50). CDC20 serves as a regulatory protein during cell cycle progression, and performs certain functions in coordination with a series of other proteins, such as serving as an activator of the anaphase-promoting complex/cyclosome during the metaphase-anaphase transition, and the overexpression of CDC20 is associated with
Figure 5. Association between the expression of hub genes and the overall survival of patients with breast cancer. Increased expression of each hub gene above the median expression level was associated with a decreased overall survival time. HR, hazard ratio.
tumorogenesis and tumor progression (51-53). Furthermore, CDC20 is associated with reduced survival in patients with breast cancer (54). ZWINT is involved in kinetochore function and its overexpression affects the proliferation of breast cancer cells (55). CENPF is required for kinetochore function during cell division and is associated with the cell cycle, mitotic and cell proliferative pathways. CENPF, together with forkhead box protein M1, coordinate promote cancer malignancy (56,57). PRC1 encodes a protein that is involved in cytokinesis and is essential for cell cleavage (58). PRC1 overexpression was detected in p53-deficient cells, and the negative regulatory feedback mechanism was controlled by p53 (59). CDKN3 encodes a cyclin-dependent kinase inhibitor protein that is essential for normal mitosis and the G1/S transition (60). CDKN3 overexpression in cancer is typically associated with a poor survival outcome for patients (61). Therefore, it is a potential therapeutic target in cancer treatment studies (62). TPX2 encodes a spindle assembly factor required for the normal assembly of mitotic spindles and for the normal assembly of microtubules around the chromosomes during apoptosis (63). TPX2 may serve as a prognostic marker and promote the proliferation, progression, migration and invasion of breast cancer (64).

In conclusion, the present study identified 322 consistent candidate DEGs and demonstrated the presence of 15 hub genes using expression profiles from datasets containing multiple cohorts and a series of bioinformatics analyses. These hub genes were significantly enriched in ‘cell cycle’, ‘oocyte meiosis’ and ‘p53 signaling pathway’, in addition to cell division, nuclear division, chromosome segregation and other tumor-associated processes, which may prove their value in clinical applications designed to treat breast cancer. The present study may effectively improve our understanding of the innate causes of breast cancer, and the 15 hub genes may serve as biomarkers for the prediction, diagnosis, individualized prevention, treatment and prognosis of breast cancer.

Acknowledgements

The authors are grateful to the technical team at BGI-Shenzhen for their assistance and technical support. The authors would like to thank all the members of the BGI Cancer Genomics Institute research group for their enthusiastic involvement in the present study.

Funding

The present study was supported by funding from The Shenzhen Municipal Government of China (Shenzhen, China; grant no. JSGG20140702161347218).

Availability of data and materials

The datasets analyzed in the present study are available from the GEO repository at (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10810) and (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65194), and TCGA repository at (https:// xenabrowser.net/datasets/?cohort=GDC%20TCGA%20Breast%20Cancer%20). The data presented in the present study is available from the authors on reasonable request.

Authors’ contributions

HJ and XH analyzed the data and made major contributions to writing the manuscript. KS and GL analyzed the data and contributed to revising the article. JW and HY contributed to the design of the study and provided final approval of the manuscript. YH contributed to the design of the study and assisted with writing the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References


