Identification of an eight-lncRNA prognostic model for breast cancer using WGCNA network analysis and a Cox-proportional hazards model based on L1-penalized estimation

ZHENBIN LIU, MENGHU LI, QI HUA, YANFANG LI and GANG WANG

Department of Ulcer and Vascular Surgery, First Teaching Hospital of Tianjin University
of Traditional Chinese Medicine, Tianjin 300193, P.R. China

Received February 14, 2019; Accepted May 24, 2019

DOI: 10.3892/ijmm.2019.4303

Abstract. An ever-increasing number of long noncoding (lnc) RNAs has been identified in breast cancer. The present study aimed to establish an lncRNA signature for predicting survival in breast cancer. RNA expression profiling was performed using microarray gene expression data from the National Center for Biotechnology Information Gene Expression Omnibus, followed by the identification of breast cancer-related preserved modules using weighted gene co-expression network (WGCNA) network analysis. From the lncRNAs identified in these preserved modules, prognostic lncRNAs were selected using univariate Cox regression analysis in combination with the L1-penalized (LASSO) Cox-proportional Hazards (Cox-PH) model. A risk score based on these prognostic lncRNAs was calculated and used for risk stratification. Differentially expressed RNAs (DERs) in breast cancer were identified using MetaDE. Gene Set Enrichment Analysis pathway enrichment analysis was conducted for these prognostic lncRNAs and the DERs related to the lncRNAs in the preserved modules. A total of five preserved modules comprising 73 lncRNAs were mined. An eight-lncRNA signature (IGHA1, IGHGP, IGKV2-28, IGLL3P, IGLV3-10, AZGP1P1, LINC00472 and SLC16A6P1) was identified using the LASSO Cox-PH model. Risk score based on these eight lncRNAs could classify breast cancer patients into two groups with significantly different survival times. The eight-lncRNA signature was validated using three independent cohorts. These prognostic lncRNAs were significantly associated with the cell adhesion molecules pathway, JAK-signal transducer and activator of transcription 5A pathway, and erbb pathway and are potentially involved in regulating angiotensin II receptor type 1, neuropeptide Y receptor Y1, KISS1 receptor, and C-C motif chemokine ligand 5. The developed eight-lncRNA signature may have clinical implications for predicting prognosis in breast cancer. Overall, this study provided possible molecular targets for the development of novel therapies against breast cancer.

Introduction

Breast cancer is the most common cancer in females worldwide and affects ~12% of women (1). Treatments for cancer usually include surgery, chemotherapy and radiation therapy (2). In addition, hormone-blocking therapy is preferred by hormone receptor-positive cancers (2). The prognosis of breast cancer varies according to a number of factors, such as stage of cancer, grade of cancer and age (3). Better prognosis prediction can potentially improve patient survival and guide tailored therapy of patients. Therefore, identifying prognostic biomarkers has become an urgent and highly active area of research.

Long non-coding RNAs (lncRNAs) are non-coding transcripts that are longer than 200 bp. Although previously believed to be junk DNA, studies have increasingly recognized the critical roles of lncRNAs in regulating cell machinery, cell cycle, differentiation and apoptosis (4,5). Emerging evidence have proved that lncRNAs are involved in the regulation of diverse genes and pathways involved in the development of breast cancer and endocrine resistance (6,7). LncRNA HOX transcript antisense intergenic RNA (HOTAIR) is not only an independent biomarker for metastasis in estrogen receptor (ER)-positive breast cancer patients (8), but is also known to strengthen ER signaling and facilitate tamoxifen resistance (9). Recently, Li *et al* (10) reported that lncRNA Angelman syndrome chromosome region represses the invasion and

Correspondence to: Dr Gang Wang, Department of Ulcer and Vascular Surgery, First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, 314 Anshanxi Road, Tianjin 300193, P.R. China
E-mail: tjtcmliuzhenbin@sina.com

Abbreviations: lncRNAs, Long noncoding RNAs; DERs, Differentially expressed RNAs; NCBI, National Center for Biotechnology Information; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas; BRCA METABRIC, Breast Cancer Molecular Taxonomy of Breast Cancer International Consortium; WGCNA, weighted gene co-expression network; Cox-PH, Cox-proportional Hazards; LASSO, L1-penalized; MetaDE, Meta-analysis to detect differentially expressed genes; GSEA, Gene Set Enrichment Analysis; CAMS, cell adhesion molecules

Key words: long noncoding RNA, weighted gene co-expression network, module, pathway, prognosis

metastatic capability of breast cancer cells by modulating the degradation of Enhancer of Zeste Homolog 2. Furthermore, Tracy *et al* (11) identified a list of lncRNAs that participate in breast cancer progression based on transcriptome-wide sequencing of breast cancer cell lines. Although a few lncRNAs have been implicated in the biology of breast cancer, the prognostic potential of lncRNAs in breast cancer has not been fully elucidated.

Meng *et al* (12) developed a four-lncRNA signature that is predictive for breast cancer prognosis. However, their results were only based on lncRNA expression profiling in 887 breast cancer patients from Gene Expression Omnibus (GEO). Likewise, Sun *et al* (13) identified an eight-lncRNA signature which could serve as an independent biomarker for prediction of overall survival of breast cancer using case and control datasets only downloaded from The Cancer Genome Atlas (TCGA) database. In the present study, an integrated analysis was performed using publicly available microarray expression profiles of breast cancer patients from the GEO, TCGA and Breast Cancer Molecular Taxonomy of Breast Cancer International Consortium (BRCA METABRIC) repositories to identify prognostic lncRNAs through weighted gene co-expression network (WGCNA) network analysis, univariate Cox regression analysis, and Cox-proportional Hazards (PH) model based on L1-penalized (LASSO) estimation. Moreover, the biological significance of these lncRNAs in breast cancer was explored by constructing lncRNA-mRNA networks and performing pathway enrichment analysis.

Materials and methods

Data sources. The microarray expression profiles of at least 100 human breast cancer tissue samples were searched in the National Center for Biotechnology Information GEO (http://www.ncbi.nlm.nih.gov/geo/) based on the Affymetrix-GPL570 platform. The search retrieved the datasets GSE21653 (n=266), GSE76124 (n=198), GSE5460 (n=129) and GSE58812 (n=107). The four datasets were used for WGCNA network analysis.

GSE21653 was also used as the training set for the survival analysis of this study. The GSE20685 (Affymetrix-GPL570 platform) dataset comprising 327 breast cancer samples was downloaded from the GEO database. RNA-seq expression data of 1,063 breast cancer samples were obtained from the TCGA database (https://portal.gdc.cancer.gov/projects/TCGA-BRCA) and RNA-seq expression data of 1,904 breast cancer samples were acquired from BRCA METABRIC (Illumina High-seq 2000 platform). The GSE20685, TCGA and BRCA METABRIC datasets were used as the validation sets. The clinical characteristics of patients in the four datasets are shown in Table I.

In addition, microarray expression data was retrieved from NCBI GEO based on the following criteria: Both human breast cancer tissue samples and paired normal tissues samples were included; the total number of samples was >100; the platform used was Affymetrix-GPL570 platform. Three datasets, including GSE65194 (14) (153 breast cancer samples and 11 normal samples), GSE29044 (15) (73 breast cancer samples and 36 normal samples), and GSE42568 (16) (104 breast cancer samples and 17 normal samples), were retrieved for the identification of consensus differentially expressed RNAs (DERs) between breast cancer and normal samples using MetaDE analysis.

Data preprocessing. Raw CEL profiles of the datasets generated using the Affymetrix-GPL570 platform were subjected to median normalization, background normalization and quantile normalization using the oligo (17) package (version1.41.1, http://www.bioconductor.org/packages/release/bioc/html/oligo.html) in R (version 3.4.1).

Fragments per kilobase of exon per million reads mapped (FPKM) expression values of the datasets downloaded from the BRCA METABRI and TCGA repositories were subjected to quantile normalization using the preprocessCore package (18) (version1.40.0, http://bioconductor.org/packages/release/bioc/html/preprocessCore.html) in R (version 3.4.1).

For all datasets used in the study, the probe sets with RefSeq IDs were identified according to Affymetrix-GPL570 annotation files. From all probe sets with RefSeq transcript IDs, the probe sets were selected that were annotated as non-coding RNAs in the Refseq database (19). Moreover, the sequencing reads provided by Affymetrix-GPL570 were mapped to the GRCh38 human genome assembly using Clustal 2 (20) (http://www.clustal.org/clustal2/). The resulting lncRNAs combined with the annotated lncRNAs in the Refseq database were used for subsequent analysis.

WGCNA network analysis. Using GSE21653 as the training set, GSE76124, GSE5460 and GSE58812 as validation sets, a WGCNA network was constructed as previously described (21,22) using the WGCNA package (23) (version 1.61, https://cran.r-project.org/web/packages/WGCNA/index.html) to identify breast cancer-related modules. Briefly, the four datasets were compared and analyzed by correlation analysis. The soft threshold power of β was calculated using scale free topology criterion, followed by generation of the weighted adjacency matrix. Modules with >80 RNAs at the minimum cut height of 0.99 were selected using the dynamic tree cut method. Among these selected modules, the preserved modules were then determined using the WGCNA package. In addition, functional annotation analysis of the preserved modules was performed using the userListEnchment function of the WGCNA package. In addition, the correlations of these modules with clinical factors of patients in GSE21653 were investigated using the WGCNA package.

Survival analysis. For survival analysis, GSE21653 was used as a training set, whereas GSE20685, the TCGA and BRCA METABRIC datasets were used as the test sets. Based on the survival information time in GSE21653, univariate Cox regression analysis was performed to analyze the associations of the lncRNAs included in the preserved WGCNA modules with prognosis with the survival package in R. The lncRNAs with log-rank P<0.05 were identified as prognosis-related lncRNAs.

Based on these prognosis-related lncRNAs, a Cox-PH model was applied based on LASSO estimation to select the optimal panel of prognostic lncRNAs as previously described (24,25). The optimal lambda was determined after running 1,000

Table I. Summary of clinical characteristics in GSE21653, GSE20685, TCGA and BRCA METABRIC datasets.

Clinical characteristics	GSE21653 (N=266)	GSE20685 (N=327)	TCGA (N=1,063)	BRCA-METABRIC (N=1,904)
Age (years)	54.48±14.06	47.89±10.69	58.24±13.19	61.09±12.98
Molecular subtype				
Basal	75	-	141	199
ERBB2	24	-	64	220
LuminalA	89	-	413	679
LuminalB	49	-	190	461
-	29	-	255	345
Histology				
IDC	213	-	771	1,727
ILC	22	-	190	141
TUB	6	-	0	0
Others	21	-	100	36
-	4	-	2	0
pN				
Stage 0	120	-	-	-
Stage 1	140	-	-	-
-	6	-	-	-
pT				
Stage 1	59	-	-	-
Stage 2	126	-	-	-
Stage 3	68	-	-	-
-	13	-	-	-
SBR grade				
Grade 1	45	-	-	-
Grade 2	89	-	-	-
Grade 3	125	-	-	-
-	7	-	-	-
ER status				
Positive	150	-	784	1,445
Negative	113	-	236	429
-	3	-	43	30
erbB2 status				
Positive	29	-	-	188
Negative	216	-	-	1,512
-	21	-	-	204
Ki67 status				
Positive	144	-	-	-
Negative	58	-	-	-
-	64	-	-	-
p53 status				
Positive	69	-	-	
Negative	125	-	-	
-	72	-	-	-
PR status				
Positive	136	-	682	-
Negative	127	-	337	-
-	3	-	44	-

Table I. Continued.

Clinical characteristics	GSE21653 (N=266)	GSE20685 (N=327)	TCGA (N=1,063)	BRCA-METABRIC (N=1,904)
Death				
Dead	83	83	139	1,103
Alive	169	244	930	801
-	14	0	0	0
Overall survival time (months)	60.03±41.38	94.71±38.45	36.27±34.78	125.03±76.33

Age and over survival time were expressed as the mean ± standard deviation; ERBB2, epidermal growth factor receptor 2; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; TUB, tubular carcinoma; pN, category: regional lymph nodes; pT, pathologic tumor size; SBR, Scarff-Bloom-Richardson; ER, estrogen-receptor; PR, progesterone receptor; -, information unavailable; TCGA, The Cancer Genome Atlas; BRCA METABRIC, Breast Cancer Molecular Taxonomy of Breast Cancer International Consortium.

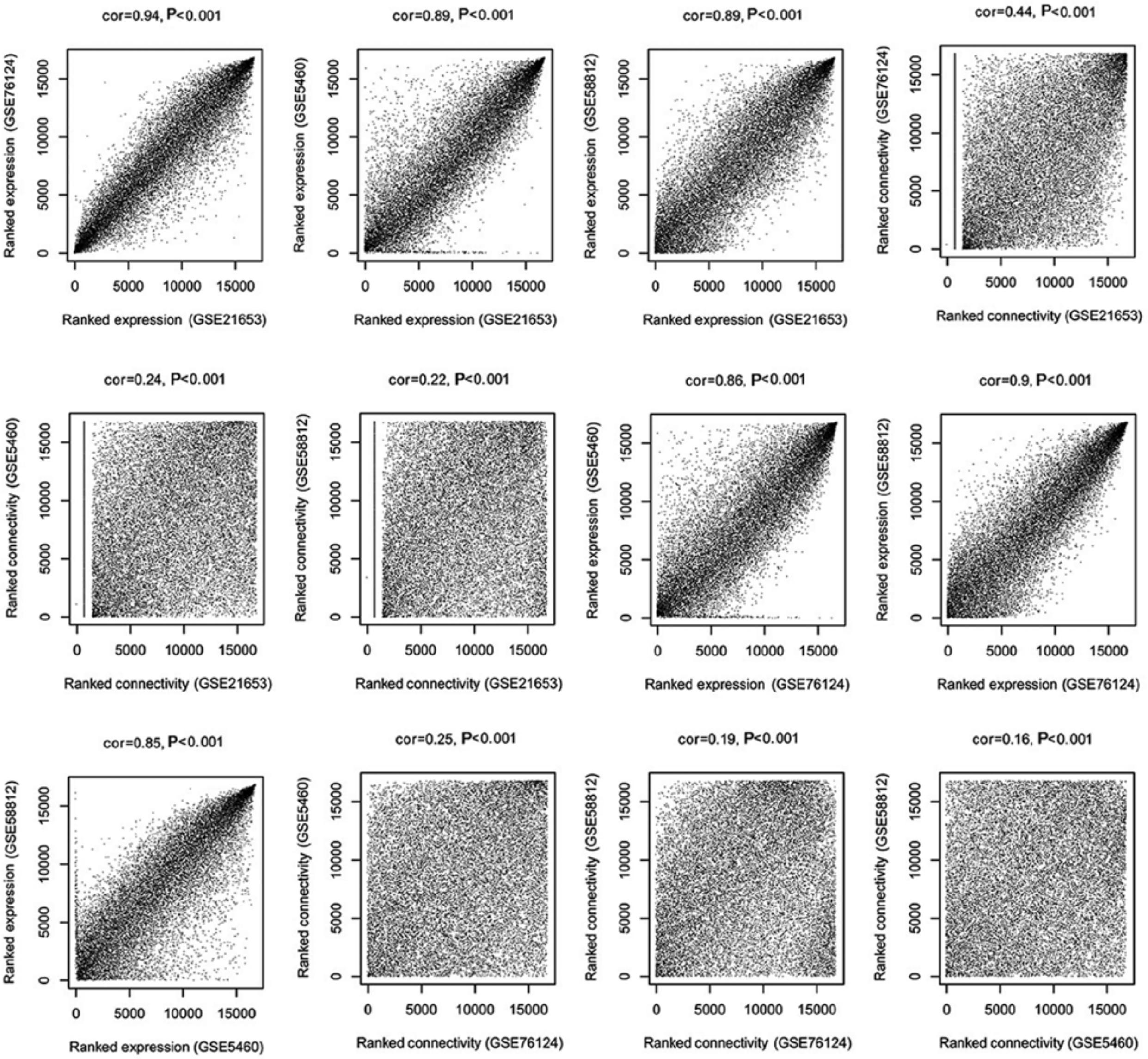

Figure 1. Correlation analysis between the GSE21653, GSE76124, GSE5460, and GSE58812 datasets.

stimulations through cross-validation likelihood. The risk score is the logarithm of hazard ratio from the fitted Cox-PH model to dichotomize the samples (24,25). Using the Cox-PH coefficients and the optimal group of these prognostic lncRNAs, a risk scoring model was generated for prognosis prediction as follows: Risk-score = Σ (βlncRNAn x exprlncRNAn).

βlncRNAn denotes Cox-PH coefficient of lncRNAn while exprlncRNA denotes the lncRNAn expression levels.

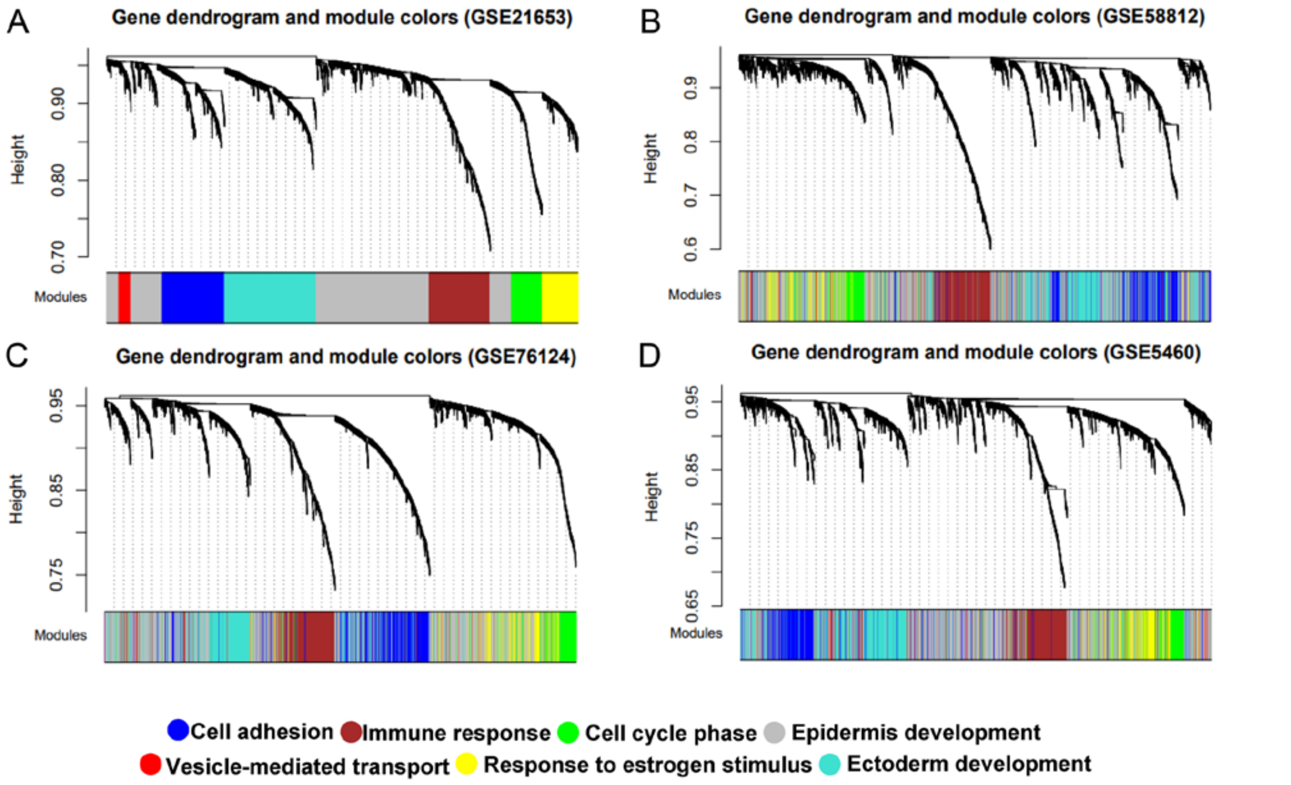

Figure 2. Identification of weighted gene co-expression network modules in the datasets. (A) GSE21653, (B) GSE58812, (C) GSE76124 and (D) GSE5460. The RNAs are organized into various modules that are marked with different colors. Blue, Cell adhesion; Brown, Immune response; Green, Cell cycle phase; Grey, Epidermis development; Red, Vesicle-mediated transport; Yellow, Response to estrogen stimulus; Turquoise, Ectoderm development.

Using this risk scoring model, the risk score was calculated for each sample in GSE21653 (training set). All patients in this set were divided into the high and low risk groups based on the median risk score. Moreover, the robustness of this prognostic model was evaluated in the GSE20685, TCGA and BRCA METABRIC datasets (validation sets). For this purpose, all samples were dichotomized in each set into two different risk groups with the median risk score as cutoff. The two risk groups for survival were compared using a Kaplan-Meier curve with Wilcoxon log rank test.

Selection of consensus DERs. As mentioned above, the GSE22866, GSE50161 and GSE4290 comprised both breast cancer and normal tissue samples. Consensus DERs between breast cancer and normal samples across the three sets were screened using the MetaDE package (26) (https://cran.r-project.org/web/packages/MetaDE/). The strict threshold was tau2=0, Q pval>0.05, P<0.05 and FDR<0.05.

Pathway enrichment analysis. The DERs associated with the prognostic lncRNAs were focused on in the preserved modules identified from the WGCNA network. lncRNA-mRNA networks were constructed using these DERs and the prognostic lncRNAs. To explore the potential biological roles of these prognostic lncRNAs selected by the LASSO Cox-PH model in breast cancer, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was performed for these lncRNA-mRNA networks using gene set enrichment analysis (27) (http://software.broadinstitute.org/gsea/index.jsp) software. Nominal P<0.05 was chosen as the cutoff value.

Results

WGCNA network module mining. The GSE21653, GSE76124, GSE5460 and GSE58812 datasets were used for WGCNA network analysis. After data preprocessing, 15,998 mRNAs were identified to be shared among the GSE21653, GSE76124, GSE5460 and GSE58812 datasets, out of which 851 were lncRNAs. Correlation analysis showed good correlation among all pairs of the four datasets based on the expression levels of the shared RNAs (correlation coefficients >0.9, $P<1x10^{-200}$; Fig. 1).

A WGCNA network was constructed using GSE21653 (training set). With WGCNA applying scale free topology criterion, the soft threshold power of β was 5 when scale-free topology model fit R^2 was maximized (0.9) and the mean connectivity for the network was 5. A total of seven modules were identified (module size ≥80 and cut height ≥0.99) in the network (Blue, Brown, Green, Turquoise, Yellow, Red, and Grey; Fig. 2A). In addition, module mining was separately conducted for GSE76124, GSE5460 and GSE58812 (validation sets; Fig. 2B-D). For the three sets, genes were processed in the same manner as for GSE21653. A multi-dimensional scaling plot was generated to analyze the expression of genes in the seven modules of GSE21653. Results revealed that the genes within the same module cluster together (Fig. 3A). Hierarchical clustering analysis of the seven modules was independently performed for each of the four datasets. The modules on the same branches showed similar gene expression patterns (Fig. 3B).

Table II. Analysis of weighted gene co-expression network modules.

Module	Color	Module size	Number of mRNAs	Number of lncRNAs	Preservation Z-score	Module annotation
MEblue	Blue	423	417	6	21.8833	Cell adhesion
MEbrown	Brown	413	374	39	39.0982	Immune response
MEgreen	Green	209	205	4	20.3943	Cell cycle phase
MEgrey	Grey	1,215	1,171	44	2.6106	Epidermis development
MEred	Red	82	80	2	0.8947	Vesicle-mediated transport
MEturquoise	Turquoise	623	601	22	15.1238	Response to estrogen stimulus
MEyellow	Yellow	242	240	2	10.1586	Ectoderm development

Module size, number of mRNAs and lncRNAs; preservation Z-score >10, modules with highly preservation. Lnc, long noncoding.

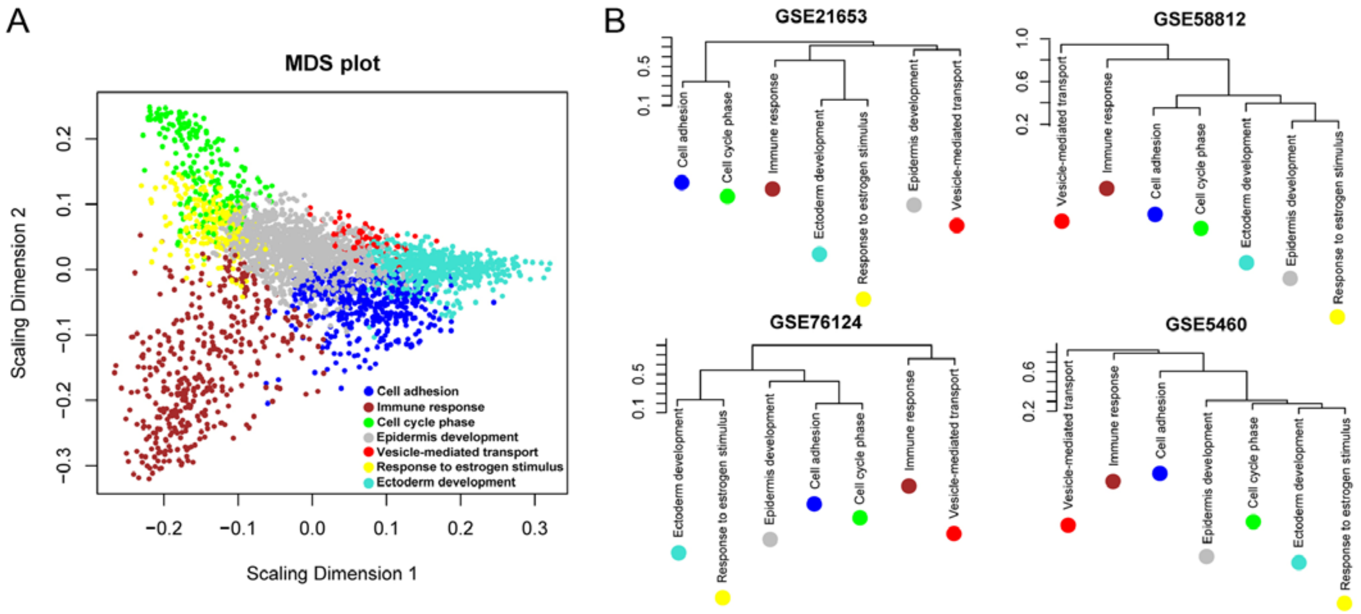

Figure 3. Analysis of WGCNA modules. (A) A multi-dimensional scaling plot showing the expression of genes in the seven modules of GSE21653. Dimension 1 and 2, respectively represents the first and second principal component. (B) Hierarchical clustering analysis of WGCNA modules for GSE21653, GSE58812, GSE76124, and GSE5460. Blue, Cell adhesion; Brown, Immune response; Green, Cell cycle phase; Grey, Epidermis development; Red, Vesicle-mediated transport; Yellow, Response to estrogen stimulus; Turquoise, Ectoderm development. WGCNA, weighted gene co-expression network.

According to the results of module preservation analysis, the Blue, Brown, Green, Turquoise and Yellow modules were highly preserved (preservation Z-score >10, Table II), thereby indicating that the five modules are breast cancer-related modules. According to results of module annotation, each of the five modules were associated with cell adhesion, immune response, cell cycle phase, response to estrogen stimulus and ectoderm development (Table II).

Identification of a prognostic signature of lncRNAs and development of prognostic scoring model. The five preserved modules included 73 lncRNAs. Of these, 39 lncRNAs were found to be related to prognosis in GSE21653 (training set) based on univariate cox regression analysis (P<0.05; Table III). With expression of the 39 prognosis-related lncRNAs as input, Cox-PH model based on LASSO penalization identified eight prognostic lncRNAs, including IGHA1, IGHGP, IGKV2-28, IGLL3P, IGLV3-10, AZGP1P1, LINC00472 and SLC16A6P1 (Table IV). Based on cox-PH coefficients and expression values of the eight lncRNAs, the risk score for each patient was calculated as follows: Risk-score = (-0.7451) x ExpIGHA1 + (-0.2334) x ExpIGHGP + (-0.0358) x ExpIGKV2-28 + (0.1031) x ExpIGLL3P + (-0.9832) x ExpIGLV3-10 + (-0.3707) x ExpAZGP1P1 + -1.6801) x ExpLINC00472 + (-1.6795) x Exp SLC16A6P1.

Based on the median risk score, all patients in GSE21653 were classified into a high risk group (n=126; >median risk score) and a low-risk group (n=126; <median risk score). Overall survival time was significantly increased in the low risk group compared with the high risk group (logRank P=0.0002; Fig. 4A). The risk stratification capability of the eight-lncRNA signature was validated in three independent datasets (GSE20685, TCGA and BRCA METABRIC datasets). Similarly, risk scores were determined for each

Table III. Prognosis-related lncRNAs identified by univariate Cox regression analysis.

LncRNAs	Module-color	P-value
SNHG14	Blue	0.0047
IGHA1	Brown	0.0003
IGLL3P	Brown	0.0005
IGKV1OR2-108	Brown	0.0010
IGLV3-10	Brown	0.0010
IGKV1-39	Brown	0.0013
IGKV3-20	Brown	0.0015
IGKV1-37	Brown	0.0019
IGKV2-28	Brown	0.0019
IGKV4-1	Brown	0.0021
IGHGP	Brown	0.0023
TRBC2	Brown	0.0029
IGHV3-72	Brown	0.0032
TRDC	Brown	0.0037
IGKV1OR2-2	Brown	0.0038
IGKV1-13	Brown	0.0042
IGHV3-20	Brown	0.0048
IGHV3-7	Brown	0.0048
HCP5	Brown	0.0062
IGKV1-17	Brown	0.0074
IGHV3-23	Brown	0.0078
IGHV4-61	Brown	0.0080
GBP1P1	Brown	0.0083
BNIP3P1	Green	0.0004
CKS1BP2	Green	0.0022
LINC00472	Turquoise	0.0005
SLC16A6P1	Turquoise	0.0010
AZGP1P1	Turquoise	0.0062
GSTT2	Turquoise	0.0098
RNU1-123P	Turquoise	0.0130
RNU4-46P	Turquoise	0.0150
RNU6-564P	Turquoise	0.0170
RASA4CP	Turquoise	0.0200
RN7SL494P	Turquoise	0.0260
CYP2B7P	Turquoise	0.0270
CYP21A1P	Turquoise	0.0280
GOLGA2P5	Turquoise	0.0300
FABP5P2	Yellow	0.0140
RNU6-146P	Yellow	0.0170

Prognosis-related lncRNAs were identified by univariate Cox and LncRNAs with P<0.05 were retained. Lnc, long noncoding.

Table IV. Information of optimal panel of prognostic lncRNAs .

LncRNAs	Coef	Hazard ratio	P-value
IGHA1	-0.7451	0.8667	0.0030
IGHGP	-0.2334	0.8775	0.0220
IGKV2-28	-0.0358	0.8779	0.0183
IGLL3P	-0.1031	0.7999	0.0051
IGLV3-10	-0.9832	0.8309	0.0095
AZGP1P1	-0.3706	0.8855	0.0462
LINC00472	-1.6801	0.6592	0.0044
SLC16A6P1	-1.6795	0.8025	0.0101

Optimal panel of prognostic lncRNAs were screened using Cox-proportional Hazards model based on L1-penalized estimation. Coef, Cox-PH coefficient; HR, Hazard Ratio; Lnc, long noncoding.

set. Samples with the risk scores larger than the median risk score were classified under the high-risk group, while samples with risk score smaller than median risk score were classified under the low risk group. As shown in Fig. 4B-D, all patients in each dataset were split into two risk groups with significantly different survival times (GSE20685, logRank P=7.97x10^{-05}; TCGA, logRank P=2.07x10^{-04}; BRCA METABRIC, logRank P=1.87x10^{-08}). These observations highlighted the prognostic value of the eight lncRNAs in breast cancer.

Construction of lncRNA-mRNA networks and pathway enrichment analysis. All eight prognostic lncRNAs were included in the Brown and Turquoise modules. As mentioned in Methods section, consensus DERs between breast cancer and normal samples were screened across GSE22866, GSE50161, and GSE4290. Consequently, 1,372 consensus DERs (tau2=0, Qpval>0.05, P <0.05 and FDR<0.05) were obtained, including 55 lncRNAs and 1,317 coding RNAs. As shown in Fig. 5, lncRNA-mRNA networks were constructed. These prognostic lncRNAs and the DERs related to these prognostic DERs are shown in the Brown and Turquoise modules.

KEGG pathway enrichment analysis was performed using the generated lncRNA-mRNA networks. Enrichment analysis revealed that genes in modules related to AZGP1P1, IGLL3P, IGHA1, IGLV3-10, IGHGP, LINC00472, IGKV2-28 and SLC16A6P1 were associated with several pathways, such as cell adhesion molecules (CAMS) pathway, T cell receptor pathway, JAK-signal transducer and activator of transcription pathway, and erbb pathway (Fig. 6). Moreover, a number of genes enriched in these pathways, such as angiotensin II receptor type (AGTR)1, neuropeptide Y receptor Y1 (NPY1R), KISS1 receptor (KISS1R) and C-C motif chemokine ligand (CCL) 5 were identified.

Discussion

Although lncRNAs are well recognized as playing important roles in the biology of tumorigenesis (28), additional studies focusing on the involvement of lncRNAs in breast cancer should be conducted. Based on the analysis of 1,222 breast cancer cases and control datasets downloaded from the TCGA database, Sun *et al* (13) reported that an eight-lncRNA signature consisted of AC007731.1, AL513123.1, C10orf126, WT1-AS, ADAMTS9-AS1, SRGAP3-AS2, TLR8-AS1, and HOTAIR was an independent prognostic factor associated with overall survival by WGCNA analysis and multivariate

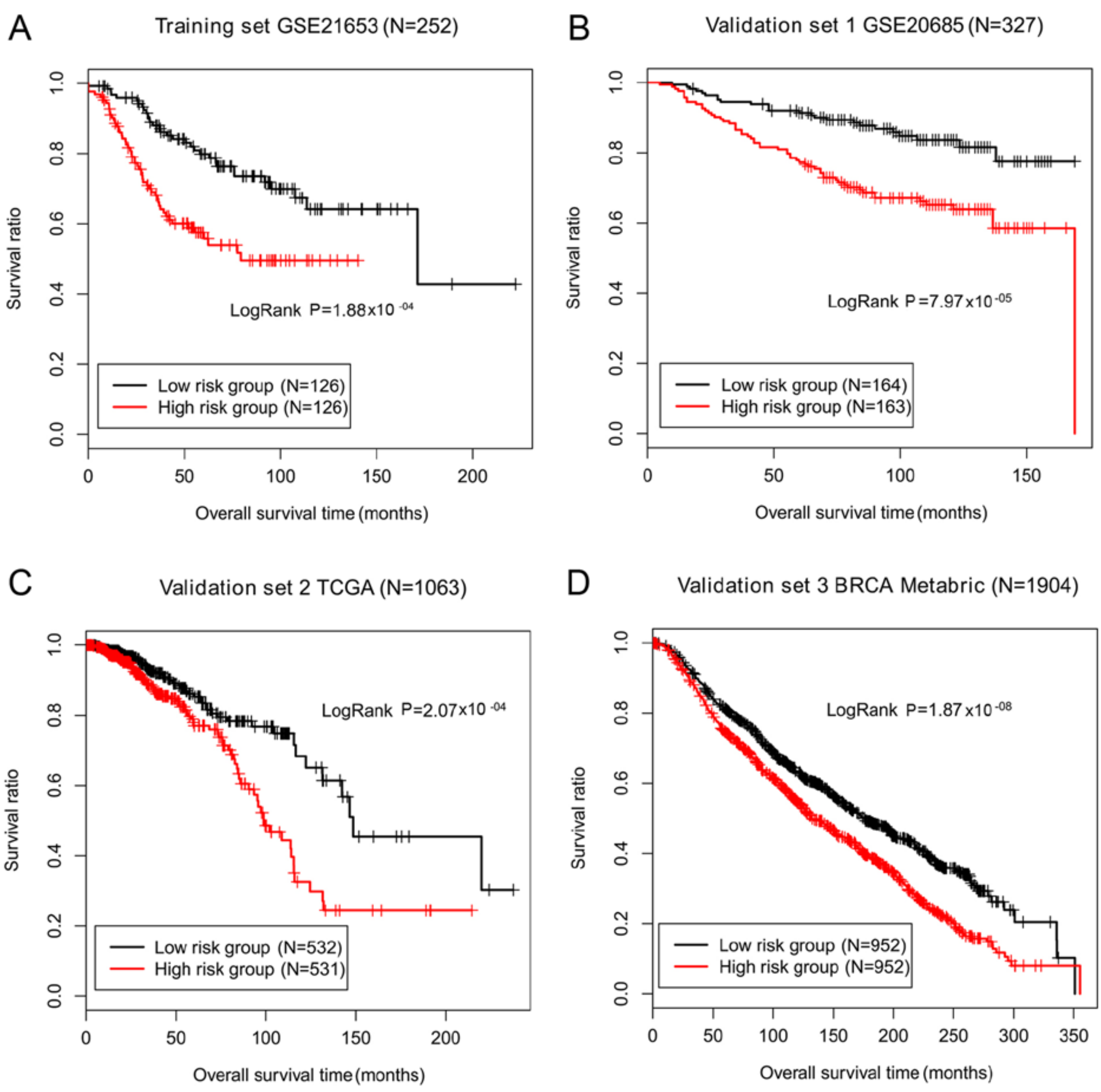

Figure 4. Kaplan-Meier survival curves for low-risk and high-risk groups in the datasets. (A) GSE21653 set, (B) GSE20685 set, (C) TCGA set and (D) BRCA METABRIC set. Black and red curves are for low risk and high risk groups, respectively. TCGA, The Cancer Genome Atlas; BRCA METABRIC, Breast Cancer Molecular Taxonomy of Breast Cancer International Consortium.

Cox hazard model. By mining the microarray gene expression data in the GEO, TCGA and BRCA METABRIC repositories, an eight-lncRNA signature (IGHA1, IGHGP, IGKV2-28, IGLL3P, IGLV3-10, AZGP1P1, LINC00472, SLC16A6P1) was determined for prognosis prediction using WGCNA network analysis, univariate Cox regression analysis, and a LASSO PH model. The eight-lncRNA signature found in the present study is different from the study of Sun *et al* (13). Moreover, risk scores were calculated based on the eight-lncRNA signature, which could dichotomize patients into two risk groups with different survival times. The predictive capability of this eight-lncRNA signature was successfully confirmed in three independent datasets. The above findings suggested that the eight lncRNAs identified could serve as prognostic biomarkers for breast cancer.

To the best of our knowledge, all eight prognostic lncRNAs, except for LINC00472, have not been reported in breast cancer to date. Shen *et al* (29) provided evidence that LINC00472 plays a tumor suppressive role in breast cancer and could thus serve as a prognostic biomarker. Similarly, Shen *et al* (30) revealed that LINC00472 is significantly linked to disease-free survival in patients with grade 2 breast cancer. In addition, Lu *et al* (31) showed that the inhibition of LINC00472 in breast cancer progression is mediated by miR-141 and mRNA programmed cell death 4.

Another highlight of the present study was that the lncRNA-mRNA networks were constructed with these prognostic lncRNAs and the DERs associated with these prognostic lncRNAs in breast cancer. Pathway enrichment analysis was performed for these lncRNAs and DERs to elucidate the underlying mechanisms. The results of the present study showed that these lncRNAs were significantly associated with the CAMS pathway, JAK-STAT pathway and erbb pathway. CAMS have been established to participate in breast cancer cell angiogenesis, migration, invasion and metastasis (32,33). The JAK-STAT pathway plays an important role in inflammation

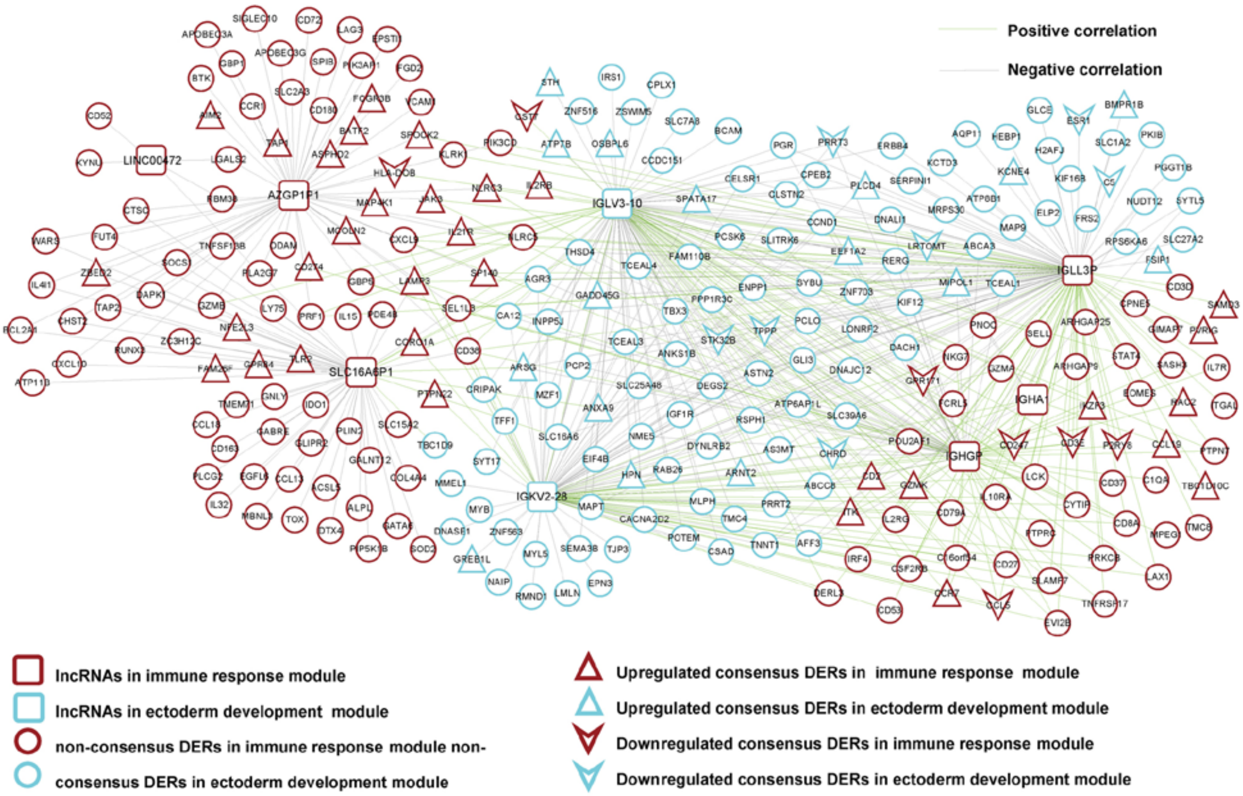
Positive correlation
Negative correlation
AZGP1P1
LINC00472
SLC16A6P1
IGLV3-10
IGKV2-28
IGLL3P
IGHA1
IGHGP
lncRNAs in immune response module
lncRNAs in ectoderm development module
non-consensus DERs in immune response module non-
consensus DERs in ectoderm development module
Upregulated consensus DERs in immune response module
Upregulated consensus DERs in ectoderm development module
Downregulated consensus DERs in immune response module
Downregulated consensus DERs in ectoderm development module

Figure 5. LncRNA-mRNA networks. RNAs from Turquoise or Brown modules are colored in turquoise or brown, respectively. Square nodes represent lncRNAs. Regular or inverted triangles indicate upregulated or downregulated consensus DERs identified by the MetaDE method, respectively. Brown, Immune response; Turquoise, Ectoderm development. Round nodes represent non-consensus DERs. Black or green link denotes positive or negative correlation between two nodes.

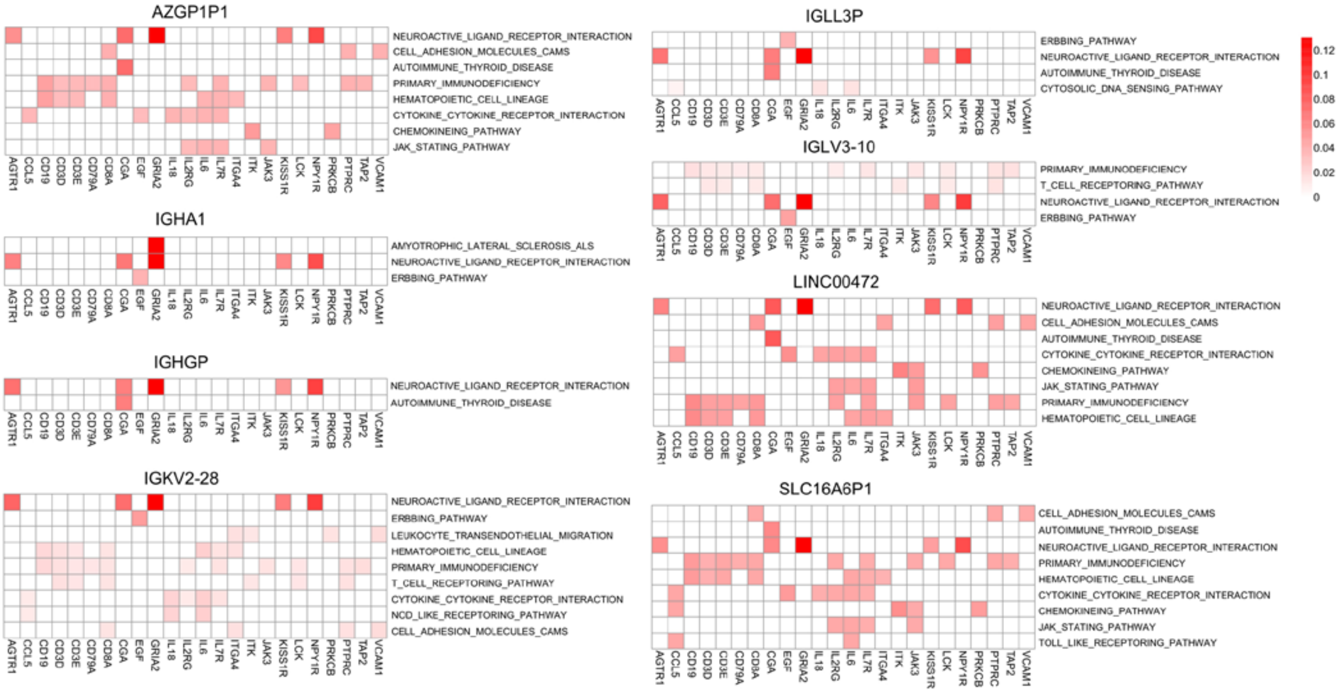

Figure 6. Results of pathway enrichment analysis on the lncRNA-mRNA networks. Horizontal axis indicates the genes, and vertical axis indicates the pathways. A red diamond indicates that a gene is significantly enriched in a pathway. Deeper red symbols indicate stronger correlations. Blank diamond indicates that a gene is not significantly enriched in a pathway.

and carcinogenesis (34) and has been regarded as a novel therapeutic target in breast cancer (35). The ErbB receptor tyrosine kinase family is comprised of ErbB1, ErbB2, ErbB3 and ErbB4. A large body of evidence demonstrated that the ErbB2 gene and protein serve as biomarkers for prognosis and therapy response (36,37). Furthermore, HER kinase domain mutations enhance tumor progression could potentially serve as prognostic markers (38). In addition, De Cola *et al* (39) showed that downregulation of ErbB receptors contributes to targeted therapy resistance. The above findings confirmed that the CAMS, JAK-STAT and erbb pathways are important for breast cancer and indicated that these prognostic lncRNAs participate in the regulation of these pathways in breast cancer.

In the present study, a list of genes that were enriched in these significant pathways for these prognostic lncRNAs, such as AGTR1, NPY1R, KISS1R and CCL5 were also identified. An *in vitro* study by Oh *et al* (40) showed that AGTR1 promotes tumor growth and angiogenesis. There is evidence that NPY1R in peripheral blood is significantly linked to tumor metastasis and prognosis of breast cancer patients (41). KISS1R, a G protein coupled receptor, promotes invadopodia formation and invasion in breast cancer cells (42). CCL5 facilitates cell proliferation and survival of breast cancer cells by regulating metabolism (43). These genes are potentially involved in the mechanisms underlying the predictive value of this eight-lncRNA signature and could serve as molecular biomarkers for breast cancer.

The present study has certain limitations. Although the eight-lncRNA prognostic signature was validated in three independent datasets, further testing on clinical data is warranted. In addition, the present study focused solely on microarray expression datasets. Therefore, experimental studies are required to verify the findings of the present study. Additionally, further studies should be conducted to elucidate the mechanisms underlying the actions of these prognostic lncRNAs.

In conclusion, the present study recommends an eight-lncRNA signature for survival prediction in breast cancer. These prognostic lncRNAs could affect cancer development partly by regulating the CAMS, JAK-STAT, erbb pathways, as well as AGTR1, NPY1R, KISS1R and CCL5. These findings hold promise for the identification of promising therapeutic targets for breast cancer.

Acknowledgements

Not applicable.

Funding

No funding was received.

Availability of data and materials

The datasets used during the current study are available from the corresponding author on reasonable request.

Authors' contributions

ZL performed data analyses and wrote the manuscript. ML, QH and YL contributed significantly to the data analyses and manuscript revision. GW conceived and designed the study. 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

1. McGuire A, Brown JA, Malone C, Mclaughlin R and Kerin MJ: Effects of age on the detection and management of breast cancer. Cancers 7: 908-929, 2015.
2. Radice D and Redaelli A: Breast cancer management: Quality-of-life and cost considerations. Pharmacoeconomics 21: 383-396, 2003.
3. Jiralerspong S and Goodwin PJ: Obesity and breast cancer prognosis: Evidence, challenges, and opportunities. J Clin Oncol 34: 4203-4216, 2016.
4. Kitagawa M, Kitagawa K, Kotake Y, Niida H and Ohhata T: Cell cycle regulation by long non-coding RNAs. Cell Mol Life Sci 70: 4785-4794, 2013.
5. Liz J and Esteller M: lncRNAs and microRNAs with a role in cancer development. Biochim Biophys Acta 1859: 169-176, 2016.
6. Hayes EL and Lewis-Wambi JS: Mechanisms of endocrine resistance in breast cancer: An overview of the proposed roles of noncoding RNA. Breast Cancer Res 17: 40, 2015.
7. Godinho MF, Sieuwerts AM, Look MP, Meijer D, Foekens JA, Dorssers LC and van Agthoven T: Relevance of BCAR4 in tamoxifen resistance and tumour aggressiveness of human breast cancer. Br J Cancer 103: 1284-1291, 2010.
8. Sørensen KP, Thomassen M, Tan Q, Bak M, Cold S, Burton M, Larsen MJ and Kruse TA: Long non-coding RNA HOTAIR is an independent prognostic marker of metastasis in estrogen receptor-positive primary breast cancer. Breast Cancer Res Treat 142: 529-536, 2013.
9. Xue X, Yang YA, Zhang A, Fong KW, Kim J, Song B, Li S, Zhao JC and Yu J: LncRNA HOTAIR enhances ER signaling and confers tamoxifen resistance in breast cancer. Oncogene 35: 2746-2755, 2016.
10. Li Z, Hou P, Fan D, Dong M, Ma M, Li H, Yao R, Li Y, Wang G and Geng P: The degradation of EZH2 mediated by lncRNA ANCR attenuated the invasion and metastasis of breast cancer. Cell Death Differ 24: 59-71, 2017.
11. Tracy KM, Tye CE, Page NA, Fritz AJ, Stein JL, Lian JB and Stein GS: Selective expression of long non-coding RNAs in a breast cancer cell progression model. J Cell Physiol 233: 1291-1299, 2017.
12. Meng J, Li P, Zhang Q, Yang Z and Fu S: A four-long non-coding RNA signature in predicting breast cancer survival. J Exp Clin Cancer Res 33: 84, 2014.
13. Sun M, Wu D, Zhou K, Li H, Gong X, Wei Q, Du M, Lei P, Zha J and Zhu H: An eight-lncRNA signature predicts survival of breast cancer patients: A comprehensive study based on weighted gene co-expression network analysis and competing endogenous RNA network. Breast Cancer Res Treat 175: 59-75, 2019.
14. Maubant S, Tesson B, Maire V, Ye M, Rigaill G, Gentien D, Cruzalegui F, Tucker GC, Roman-Roman S and Dubois T: Transcriptome analysis of Wnt3a-treated triple-negative breast cancer cells. PLoS One 10: e0122333, 2015.
15. Colak D, Nofal A, Albakheet A, Nirmal M, Jeprel H, Eldali A, Al-Tweigeri T, Tulbah A, Ajarim D, Malik OA, *et al*: Age-specific gene expression signatures for breast tumors and cross-species conserved potential cancer progression markers in young women. PLoS One 8: e63204, 2013.
16. Clarke C, Madden SF, Doolan P, Aherne ST, Joyce H, O'Driscoll L, Gallagher WM, Hennessy BT, Moriarty M, Crown J, *et al*: Correlating transcriptional networks to breast cancer survival: A large-scale coexpression analysis. Carcinogenesis 34: 2300-2308, 2013.
17. Carvalho B: An Introduction to the Oligo Package, 2009.
18. Bolstad BM, Irizarry RA, Astrand M and Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193, 2003.

19. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, *et al*: Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44: D733-D745, 2016.
20. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, *et al*: Clustal W and clustal X version 2.0. Bioinformatics 23: 2947-2948, 2007.
21. Lou Y, Tian GY, Song Y, Liu YL, Chen YD, Shi JP and Yang J: Characterization of transcriptional modules related to fibrosing-NAFLD progression. Sci Rep 7: 4748, 2017.
22. Perumal D, Leshchenko VV, Kuo PY, Jiang Z, Readhead B, Eden C, Athaluri Divakar SK, Zhang W, Cho HJ, Chari A, *et al*: Weighted gene co-expression network analysis (WGCNA) identifies highly proliferative myeloma subgroup responsive to CDK4/ARK5 inhibition. Blood 124: 3445, 2014.
23. Horvath S and Langfelder P: Tutorials for the WGCNA package for R: WGCNA Background and glossary, 2011.
24. Tibshirani R: The lasso method for variable selection in the Cox model. Stat Med 16: 385-395, 1997.
25. Huang S, Yee C, Ching T, Yu H and Garmire LX: A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. PLoS Comput Biol 10: e1003851, 2014.
26. Chong QI, Hong L, Cheng Z and Yin Q: Identification of metastasis-associated genes in colorectal cancer using metaDE and survival analysis. Oncol Lett 11: 568-574, 2016.
27. Walker MG and Shi J: Gene set enrichment analysis (GSEA) for interpreting gene expression profiles. Current Bioinformatics 2: 133-137, 2007.
28. Evans JR, Feng FY and Chinnaiyan AM: The bright side of dark matter: lncRNAs in cancer. J Clin Invest 126: 2775, 2016.
29. Shen Y, Katsaros D, Loo LW, Hernandez BY, Chong C, Canuto EM, Biglia N, Lu L, Risch H, Chu WM and Yu H: Prognostic and predictive values of long non-coding RNA LINC00472 in breast cancer. Oncotarget 6: 8579-8592, 2015.
30. Shen Y, Wang Z, Loo LW, Ni Y, Jia W, Fei P, Risch HA, Katsaros D and Yu H: LINC00472 expression is regulated by promoter methylation and associated with disease-free survival in patients with grade 2 breast cancer. Breast Cancer Res Treat 154: 473-482, 2015.
31. Lu P, Yang X, Yang Y, Wang F, Li L and Gu Y: Linc00472 suppresses breast cancer progression and enhances doxorubicin sensitivity through regulation of miR-141 and programmed cell death 4. Rsc Advances 8: 8455-8468, 2018.
32. Mcsherry EA, Brennan K, Hudson L, Hill AD and Hopkins AM: Breast cancer cell migration is regulated through junctional adhesion molecule-A-mediated activation of rap1 GTPase. Breast Cancer Res 13: R31, 2011.
33. Li DM and Feng YM: Signaling mechanism of cell adhesion molecules in breast cancer metastasis: Potential therapeutic targets. Breast Cancer Res Treat 128: 7-21, 2011.
34. Christy J and Priyadharshini L: Differential expression analysis of JAK/STAT pathway related genes in breast cancer. Meta Gene 16, 2018.
35. Hosford SR and Miller TW: Clinical potential of novel therapeutic targets in breast cancer: CDK4/6, Src, JAK/STAT, PARP, HDAC, and PI3K/AKT/mTOR pathways. Pharmacogenomics Pers Med 6: 203-215, 2014.
36. Ross JS, Fletcher JA, Linette GP, Stec J, Clark E, Ayers M, Symmans WF, Pusztai L and Bloom KJ: The Her-2/neu gene and protein in breast cancer 2003: Biomarker and target of therapy. Oncologist 8: 307-325, 2003.
37. Ross JS, Slodkowska EA, Symmans WF, Pusztai L, Ravdin PM and Hortobagyi GN: The HER-2 receptor and breast cancer: Ten years of targeted anti-HER-2 therapy and personalized medicine. Oncologist 14: 320-368, 2009.
38. Boulbes DR, Arold ST, Chauhan GB, Blachno KV, Deng N, Chang WC, Jin Q, Huang TH, Hsu JM, Brady SW, *et al*: HER family kinase domain mutations promote tumor progression and can predict response to treatment in human breast cancer. Mol Oncol 9: 586-600, 2015.
39. De Cola A, Volpe S, Budani MC, Ferracin M, Lattanzio R, Turdo A, D'Agostino D, Capone E, Stassi G, Todaro M, *et al*: MiR-205-5p-mediated downregulation of ErbB/HER receptors in breast cancer stem cells results in targeted therapy resistance. Cell Death Dis 6: e1823, 2015.
40. Oh E, Kim JY, Cho Y, An H, Lee N, Jo H, Ban C and Seo JH: Overexpression of angiotensin II type 1 receptor in breast cancer cells induces epithelial-mesenchymal transition and promotes tumor growth and angiogenesis. Biochim Biophys Acta 1863: 1071-1081, 2016.
41. Liu L, Xu Q, Cheng L, Ma C, Xiao L, Xu D, Gao Y, Wang J and Song H: NPY1R is a novel peripheral blood marker predictive of metastasis and prognosis in breast cancer patients. Oncol Lett 9: 891-896, 2015.
42. Goertzen CG, Dragan M, Turley E, Babwah AV and Bhattacharya M: KISS1R signaling promotes invadopodia formation in human breast cancer cell via β-arrestin2/ERK. Cell Signal 28: 165-176, 2016.
43. Gao D, Rahbar R and Fish EN: CCL5 activation of CCR5 regulates cell metabolism to enhance proliferation of breast cancer cells. Open Biol 6: 160122, 2016.