



# Identification of key genes relevant to the prognosis of ER-positive and ER-negative breast cancer based on a prognostic prediction system

Bin Xiao<sup>1</sup> · Jianfeng Hang<sup>1</sup> · Ting Lei<sup>1</sup> · Yongyin He<sup>1</sup> · Zhenzhan Kuang<sup>1</sup> · Li Wang<sup>2</sup> · Lidan Chen<sup>1</sup> ·  
Jia He<sup>1</sup> · Weiyun Zhang<sup>1</sup> · Yang Liao<sup>1</sup> · Zhaohui Sun<sup>1</sup> · Linhai Li<sup>1</sup>

Received: 4 November 2018 / Accepted: 29 January 2019 / Published online: 19 March 2019

© The Author(s) 2019

## Abstract

Few prognostic indicators with differential expression have been reported among the differing ER statuses. We aimed to screen important breast cancer prognostic genes related to ER status and to construct an efficient prognostic prediction system. mRNA expression profiles were downloaded from TCGA and GSE70947 dataset. Two hundred seventy-one overlapping differentially expressed genes (DEGs) between the ER– and ER+ breast cancer samples were identified. Among the 271 DEGs, 109 prognostically relevant mRNAs were screened. mRNAs such as *RASEF*, *ITM2C*, *CPEB2*, *ESR1*, *ANXA9*, and *VASN* correlated strongly with breast cancer prognosis. Three modules, which contained 28, 9 and 8 enriched DEGs, were obtained from the network, and the DEGs in these modules were enriched in response to hormone stimulus, epithelial cell development, and host cell entry. Using bayes discriminant analysis, 48 signature genes were screened. We constructed a prognostic prediction system using the 48 signature genes and validated this system as relatively accurate and reliable. The DEGs might be closely associated with the prognosis in patients with breast cancer. We validated the effectiveness of our prognostic prediction system by GEO database. Therefore, this system might be a useful tool for preliminary screening and validation of potential prognosis indicators for ER+ breast cancer derived from mechanistic research.

**Keywords** Breast cancer · Estrogen receptor · Differentially expressed genes · Prognosis · Prognostic prediction system

---

Bin Xiao and Jianfeng Hang contributed equally to this work.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11033-019-04663-4>) contains supplementary material, which is available to authorized users.

✉ Zhaohui Sun  
zhaohui3@126.com

✉ Linhai Li  
xyuf9002@smu.edu.cn

<sup>1</sup> Department of Laboratory Medicine, General Hospital of Southern Theatre Command of PLA, No. 111, Liuhua Road, Yuexiu district, Guangzhou 510010, Guangdong Province, China

<sup>2</sup> Department of Medical Research, General Hospital of Southern Theatre Command of PLA, Guangzhou, Guangdong Province, China

## Introduction

Breast cancer is the most common malignancy affecting women, with approximately 1,400,000 new cases having been reported worldwide in 2012 [1]. The incidence and mortality rates of breast cancer are high and represent a major health burden among women especially in China [2]. Several risk factors have been identified as significantly related to breast cancer, including a positive family history, early menstruation, late-stage pregnancy or menopause, overweight after menopause, a fatty diet, alcohol consumption, and an older age [3]. In addition, high levels of estrogen metabolites have been closely linked with the development of breast cancer in women [4], and 70–80% of patients with breast cancer are thought to harbor estrogen receptor (ER)-positive disease [5].

Estrogen signaling pathways play important roles in breast cell proliferation and apoptosis via the activation of estrogen and ER [6]. ER-related target genes also serve as important markers for breast cancer treatment and prognosis. Reportedly,

Forkhead box protein A1 (FOXA1) is a significant independent predictor of a favorable prognosis among patients with breast cancer, as this marker positively correlates with markers associated with ER positivity and negatively correlates with the tumor size, tumor grade, tumor status, and Erb-B2 receptor tyrosine kinase 2 expression [7]. Forkhead box P1 (FOXP1), a forkhead box transcription factor subfamily P member, was reported to play a crucial role in breast cancer cell proliferation through the regulation of estrogen signaling; in addition, FOXP1 expression may indicate a good prognosis in patients treated with tamoxifen [6]. Ali et al. [8] found that *AURKA*, which encodes aurora kinase A, is a stronger prognostic factor than other markers, such as the proliferation marker Ki-67 and polo-like kinase 1, among patients with ER-positive breast cancer. DPP3 is overexpressed in breast cancer and elevated levels of *DPP3* mRNA correlate with increased NRF2 downstream gene expression and poor prognosis, particularly for ER-positive breast cancer [9]. PIK3CA mutations and AKT activation are often detected in breast cancer, are mostly associated with better or insignificant outcomes in estrogen receptor-positive (ER+) early stage breast cancer, and tend to accompany worse prognoses in ER– disease [10]. Inhibition of both p110 $\alpha$  and p110 $\beta$  isoforms of PI3K may be considered short-term cytotoxic agents and long-term cytostatic agents [11]. Additionally, ER-beta (ER $\beta$ ), which correlates with the PTEN/PI3K/pAKT pathway, acts as a favorable prognostic marker for triple-negative breast cancer [12]. In contrast, cyclinD1 overexpression indicates a poor prognosis and shorter overall survival among patients with ER-negative breast cancer. However, no correlation between cyclinD1 upregulation and clinical outcome has been identified in ER-positive samples [13].

Although several mRNAs related to breast cancer prognosis have been reported, several new prognostic indicators that may provide useful clinical guidance need to be explored, and few prognostic indicators with differential expression among different ER statuses have been reported. Therefore, in the present study we utilized mRNA expression profiles from The Cancer Genome Atlas (TCGA) database and the GSE70947 dataset in the Gene Expression Omnibus (GEO) database to screen for prognostically relevant differentially expressed genes (DEGs) between ER-negative and ER-positive breast cancers. Subsequently, we constructed a prognostic prediction system with the intent to provide a useful tool for clinical breast cancer prognosis prediction.

## Materials and methods

### Data source and data preprocessing

The breast cancer-related mRNA expression profile and corresponding clinical information were downloaded from the TCGA database (<https://gdc-portal.nci.nih.gov/>). A total

of 780 breast cancer samples, including 179 ER-negative and 601 ER-positive samples, were available. An Illumina HiSeq 2000 RNA Sequencing platform (Illumina, Inc., San Diego, CA, USA) was used to generate mRNA-seq data. An additional GSE70947 dataset was acquired from the GEO database (<http://www.ncbi.nlm.nih.gov/geo/>); this dataset was generated using the GPL13607 Agilent-028004 SurePrint G3 Human GE 8 $\times$ 60K Microarray (Agilent Technologies, Santa Clara, CA, USA). This dataset was composed of 148 breast cancer samples, including 33 ER-negative and 115 ER-positive samples, and 148 matched adjacent normal breast tissue samples. The breast cancer samples were selected as the analysis object.

We downloaded the files with the suffix “cel” for the GSE70947 dataset and obtained the raw gene expression values via gene annotation for each probe. If one gene symbol was matched to multiple probe IDs, the mean expression value was used as the gene expression level. Data preprocessing was performed via background correction and quantile normalization with the robust multi-array average algorithm.

### Differentially expressed gene screening and hierarchical clustering analysis

The mRNA expression profile data were initially identified in the TCGA database according to the HUGO (Human Genome Organisation) Gene Nomenclature Committee (HGNC, <http://www.genenames.org/>), which includes 19,034 protein coding gene annotations. Next, the R limma package (Version 3.01; R Project for Statistical Computing, Vienna, Austria) was used to screen DEGs in both the TCGA database and the GSE70947 dataset, and the multtest R package was used to calculate the false discovery rate (FDR) of each gene. Finally, a FDR < 0.05 and  $\log_2$ FC (fold change) > 0.585 (IFCI > 1.5) were selected as threshold values for DEG selection, and overlapping DEGs in both datasets were identified through Venn diagram analysis. A coupled two-way clustering analysis of the top 25 upregulated genes and top 25 downregulated genes was conducted using the pheatmap package (R software).

### Screening for prognostically relevant mRNAs

A set of 692 tumor samples was acquired from the TCGA database, along with information about the survival duration and status. We performed Cox regression analyses, using the R survival package, to identify the mRNAs with significant prognostic relevance among the DEGs and calculated P values using the log-rank test. mRNAs with P values < 0.05 were considered to be significantly DEGs. Next, we ranked the P values of those mRNAs in descending order depending

on the logRank test and subjected the top 6 mRNAs to a Kaplan–Meier survival curve analysis.

### Construction of a co-expression network of mRNAs relevant to breast cancer prognosis

After screening for significantly prognostically relevant mRNAs, we extracted the expression values of those mRNAs from the TCGA database. A Pearson correlation coefficient was calculated to assess the relevance between these different mRNAs using the *Cor* function in the R-language. Co-expression mRNA pairs with correlation coefficient *rl* values  $\geq 0.6$  were retained. Subsequently, a co-expression network was analyzed using the *Cor* function in the R-language and visualized using Cytoscape 2.8.0 [17].

### Functional annotation and submodule analysis of co-expressed genes

In R, the clusterProfiler package is an open-source program for biological function analyses such as Gene Ontology (GO) function annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses. In the present study, we utilized the clusterProfiler package to conduct the GO and KEGG functional enrichment analyses based on Fisher's exact test to investigate the biological functions of genes in the co-expression network. The Fisher's exact test algorithm was as follows:

	DEGs	Non-DEGs	Summation
Pathway/GO genes	$n_{11}$	$n_{12}$	M
Non-pathway/GO genes	$n_{21}$	$n_{22}$	N–M
Summation	K	N–K	N

$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i} \binom{N-M}{K-i}}{\binom{N}{K}}$$

Here, N represents the number of genes in the whole genome; M represents the number of genes in one pathway gene set; and K indicates the number of DEGs. The Fisher's test score is indicated by a p value that represents the probability that at least x genes in K belonged to the pathway genes.

Cytoscape 2.8.0 (National Resource for Network Biology, Bethesda, MD, USA) plugins Mcode (which clusters a given network based on topology to find dense connections) and Bingo (a Java-based tool used to determine which GO

categories are statistically overrepresented in a set of genes) were used to perform the function module analysis and module annotation based on the hypergeometric distribution of the co-expression network. The Mcode analysis parameters were a Degree cutoff of 2, K-core of 2, and Node score cutoff of 0.2.

### Construction of the discriminant prognosis prediction system

A total of 693 tumor samples with survival information were downloaded from the TCGA database and used as training data for the discriminant prognostic prediction system. The clinical information of these samples were listed in Supplemental Table 1. We used a survival cutoff of 15 months to divide the samples into poor (survival time < 15 months) and good prognosis groups (survival time  $\geq 15$  months) [14]. Prior probability was assumed by using a Bayesian Discriminant. First, we ranked the previously identified prognostically relevant mRNAs according to descending logRank (p value) order. Next, a Bayes discriminant analysis was implemented using the Bayesian discriminant function in the e1071R package. After incrementally adding each gene and deleting genes that might affect the discrimination accuracy, the discriminant analysis was continued until the highest discrimination accuracy was achieved. Next, we defined the discriminant coefficient of each sample as the prognostic score and the gene combination as the signature gene, and defined the discriminant system as a prognostic prediction system. Through this methodology, 48 signature genes were obtained.

### Validation of the prognostic prediction system

To evaluate the effectiveness of the constructed prognostic prediction system, a Kaplan–Meier survival curve analysis was performed to validate the relevance of the classification results from the discriminant system to the actual survival time, using the sample type identified via the discriminant system and the survival information. In addition, a GSE61304 dataset of 52 breast cancer samples with survival and ER status information was downloaded from GEO for validation of our discriminant prediction system. The expression values of previously screened signature genes were extracted from the GSE61304 dataset, and prognostic scores for these signature genes were calculated using our prognostic prediction system. Then, using the same standard dataset, we divided the 52 samples according to the prognostic score into poor and good prognostic groups. Finally, we performed a Kaplan–Meier survival curve analysis to validate the relevance of the classification results from the discriminant system to the actual survival time.

## Results

### Differentially expressed gene screening and hierarchy cluster analysis

After identifying a total of 17,137 protein coding mRNA expression levels that corresponded to protein coding mRNAs in the HGNC database, mRNAs with expression values  $< 5$  in the TCGA database were deleted. Subsequently, 12,524 mRNAs were screened. After removing genes with lower expression values, the expression level density peak increased significantly (Supplemental Figure 1). In addition, 668 and 1115 DEGs were identified in the TCGA and GEO databases, respectively, and, of these, 271 genes overlapped in the two databases (Supplemental Table 2).

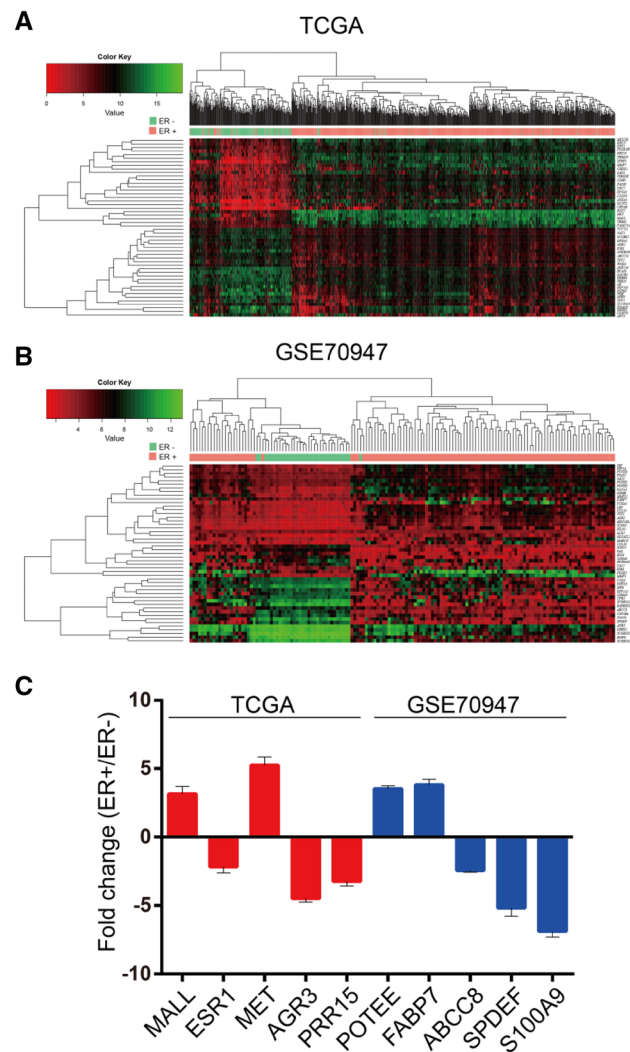
The hierarchy cluster analysis showed that the top 50 DEGs in the TCGA database and GSE70947 dataset could correctly distinguish ER-negative and ER-positive breast cancer samples with correlated expression profiles (Fig. 1a, b). The expression of some DEGs in the ER-negative and ER-positive samples was also confirmed by qPCR (Fig. 1c).

### Screening for prognostically relevant mRNA

Using a threshold value of  $P < 0.05$ , 109 mRNAs were identified as significantly related to the overall survival of breast cancer patients (Supplemental Table 3). Among these, the top six prognostically relevant mRNAs were *RAS* and EF-hand domain containing *RASEF* ( $P = 1.22E-06$ ), *CMYA5* ( $P = 1.57E-06$ ), *ITM2C* ( $P = 4.19E-06$ ), *NBEA* ( $P = 2.73E-05$ ), *CPEB2* ( $P = 3.43E-05$ ), and *VASN* ( $P = 3.54E-05$ ) (Fig. 2) (Supplemental Table 4).

### Co-expression network and submodule analysis of prognostically relevant mRNAs

A co-expression network was constructed to analyze interactions between the proteins encoded by prognostically relevant mRNAs. A total of 93 genes with 1267 corresponding relationships were integrated into this network (Fig. 3). Three modules were obtained from the network using the Mcode plugin.; Module 1, the red module, had nine enriched DEGs (e.g., *CPEB2* and *LONRF2*); module 2, the blue module, had nine enriched DEGs (e.g., *RASEF*) and module 3, or the green module, had 28 enriched DEGs (e.g., *ANXA9* and *ESR1* which encoded ER $\alpha$ ) (Fig. 3). In addition, green and red lines were used to indicate negative and positive correlations of the DEGs, respectively.

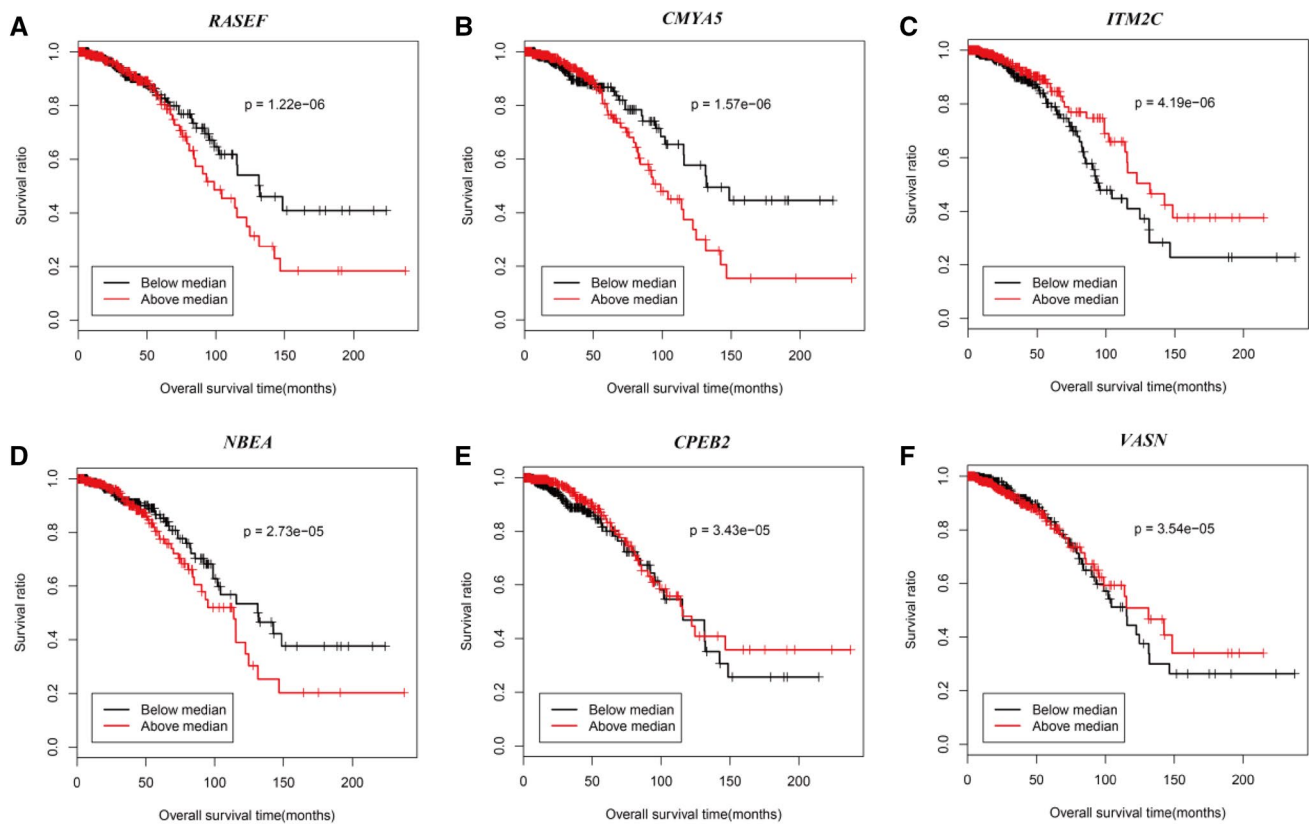


**Fig. 1** A hierarchical cluster map of the top 50 DEGs. **a** A hierarchical cluster map of the top 50 DEGs identified from the TCGA database. **b** A hierarchical cluster map of the top 50 DEGs identified from the GSE70947 dataset. ER-positive and -negative samples are denoted as red and green, respectively, on the sample strip. ER estrogen receptor, TCGA The Cancer Genome Atlas, DEGs differentially expressed genes. **c** qPCR assay indicating the expression of some DEGs between ER-negative and ER-positive breast cancer samples. Error bar indicated SD. \*\*\* $P < 0.001$ . (Color figure online)

### Functional annotation

Using the Bingo plugin, the DEGs were significantly enriched in a total of 21 GO terms and 10 KEGGs (Fig. 4). Most DEGs were significantly associated with GO terms such as intracellular signaling cascade and response to organic substance, and associated with pathways such as MAPK signaling and endocytosis. In addition, the DEGs in module 1 were enriched in response to epithelial cell development, those in module 2 were closely linked to





**Fig. 2** Kaplan–Meier survival curves for the top six genes related to prognosis, ranked according to logRank P values. **a** Kaplan–Meier survival curves for *RASEF*. **b** Kaplan–Meier survival curves for *CMYA5*. **c** Kaplan–Meier survival curves for *ITM2C*. **d** Kaplan–Meier survival

curves for *NBEA*. **e** Kaplan–Meier survival curves for *CPEB2*. **f** Kaplan–Meier survival curves for *VASN*. The P value indicates the logRank significance between the groups. The red and black lines indicate high and low expression samples, respectively. (Color figure online)

host cell entry, and those in module 3 were associated with hormone stimulus (Fig. 3).

### Construction of the discriminant prognostic prediction system

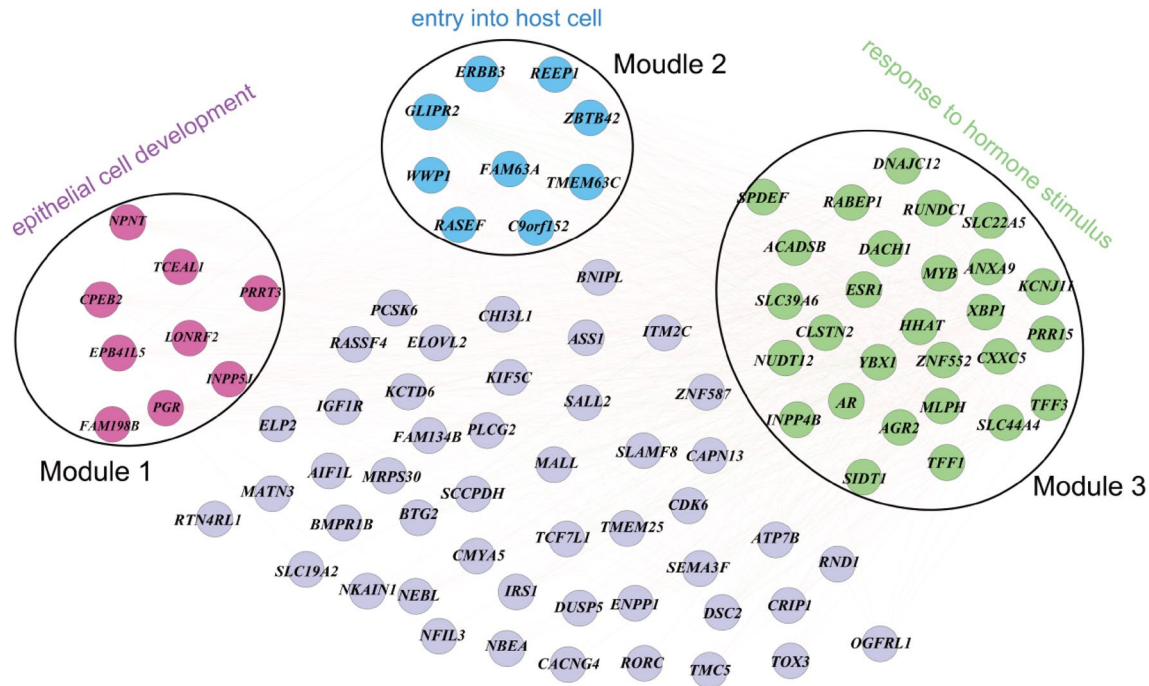
The 692 tumor samples in the TCGA database were classified as having poor ( $n=287$ , survival time  $< 15$  months) or good prognoses ( $n=405$ , survival time  $\geq 15$  months) (Supplemental Table 5). A Bayesian discriminant analysis was used to construct a prognostic prediction system comprising a set of 48 signature genes, such as *ITM2C*, *CPEB2*, *NBEA*, *RASEF*, *LONRF2*, and *ANXA9* (Supplemental Table 6). First, some of these genes showed their connections by the protein–protein network analysis (Supplemental Figure 2). Moreover, the correlations between different stages with good or poor prognosis were analyzed based on the discriminant analysis comprising the 48 signature genes (Supplemental Figure 3). The discriminant scores ranged from  $-3$  to  $3$ ; the discriminant scores for poor prognosis ranged from  $0$  to  $3$ , whereas those for good prognosis ranged from  $-3$  to  $0$ . The discriminant prognostic prediction scoring system was

constructed using the discriminant score for each sample in the discriminant system (Fig. 5), as follows:

$$\begin{aligned} \text{prognostic score} &= \hat{\alpha}_{i=1}^{48} \text{Bayes discriminia analysis} \\ &= \left\{ \begin{array}{l} [0, 3] \sim \text{bad} \\ [-3, 0] \sim \text{good} \end{array} \right\} \end{aligned}$$

### Validation of the prognostic prediction system

A Kaplan–Meier survival curve analysis demonstrated that the identified signature genes all indicated different prognoses in ER-negative ( $P=0.007$ ), ER-positive ( $P=0.00093$ ), and all samples ( $P=1.6E-05$ ), and the survival rate in the good prognosis group was significantly higher than that in the poor prognosis group (logRank  $p < 0.05$ ). In addition, the expression values of the 48 signature genes in the GSE61304 dataset and the prognostic prediction system were used to divide 52 tumor samples from the GSE61304 dataset into 34 samples with good prognoses and 18 with poor prognoses as a validation dataset. Notably, the good prognosis group had a



**Fig. 3** Co-expression network of prognostically relevant mRNAs. The green lines indicate the correlation coefficient is negative. The red lines indicate the correlation coefficient is positive. Three modules,

which contain enriched DEGs with different biological functions, were obtained from the network. The rest purple DEGs could not be assigned into any functional modules. (Color figure online)

remarkably higher disease-free survival (DFS) rate than did the ER-negative ( $P=0.015$ ), ER-positive ( $P=0.00854$ ), and total samples ( $P=0.00086$ ), indicating that our prognostic prediction system was relatively accurate and reflective of the actual estimated prognoses of the samples (Fig. 5d–f).

## Discussion

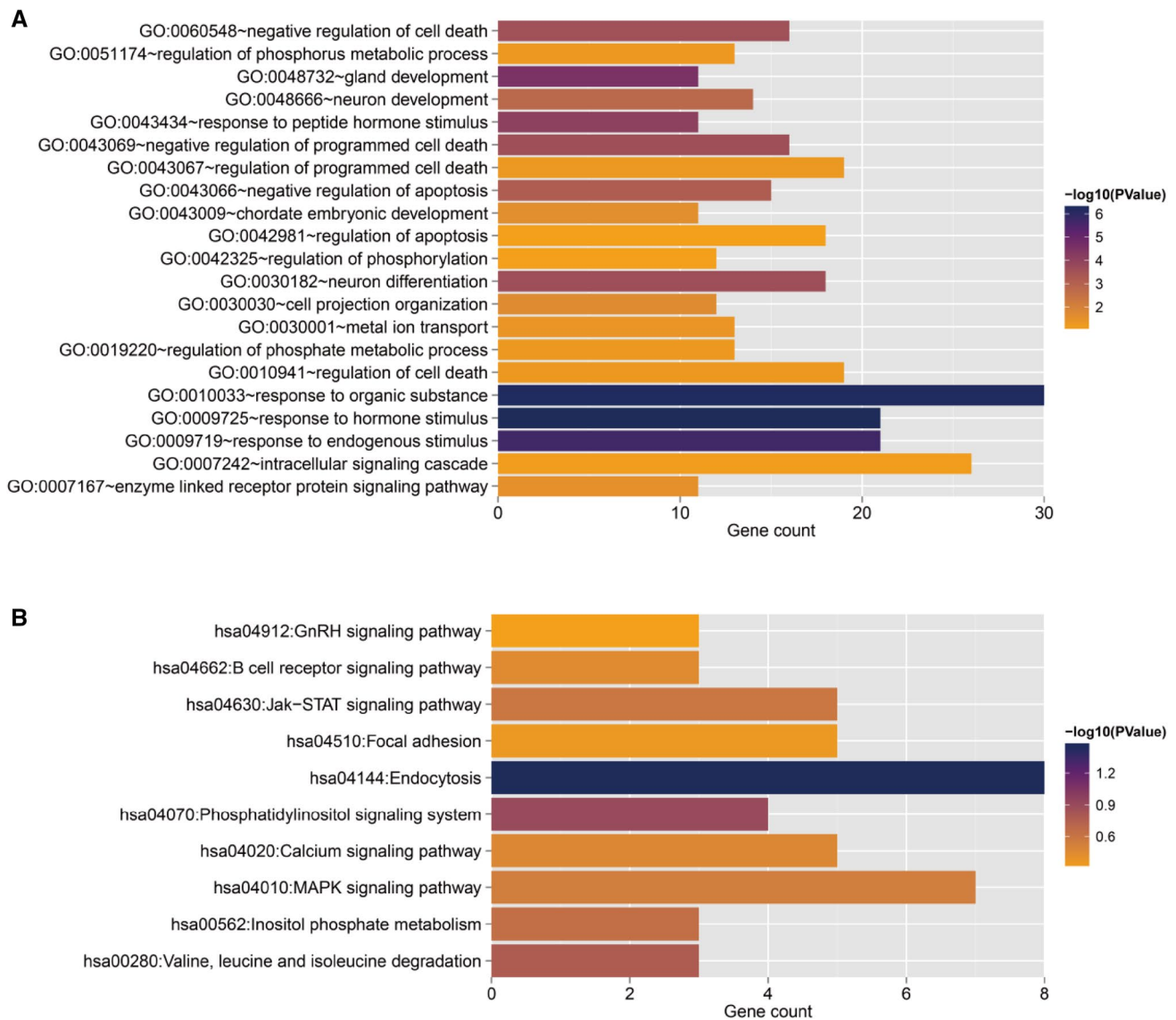
In the present study, we used data from the TCGA and GEO database to identify prognostically relevant DEGs for breast cancer according to the ER status, and analyzed the functions of these DEGs and their inter-relationships via corresponding mRNA correlation coefficients. A total of 109 mRNAs were found to correlate significantly with prognosis. Of these, *RASEF*, *ITM2C*, *CPEB2*, and *VASN* exhibited strong correlation with the breast cancer prognosis.

The expression of *RASEF*, which encodes a member of the Rab family of GTPases, was found to be higher in breast tumors than in cutaneous malignant melanoma lesions [15]. In addition, Oshita et al. suggested that *RASEF* overexpression might promote lung cancer cell proliferation by enhancing extracellular signal-regulated kinase (ERK) 1/2 signaling and that *RASEF* might be a poor prognostic marker in lung cancer [16].

*ITM2C*, which encodes integral membrane protein 2C, participates in tumor necrosis factor-induced cell death [17].

Supiot et al. [18] suggested an association of *ITM2C* with rectal cancer development via a cell apoptosis-dependent process. In addition, Alvarez et al. [19] identified *ITM2C* deletion in hereditary *BRCA1*-deleted breast cancers. Notably, *BRCA1* mutations are predominantly observed in triple-negative breast cancers [20]. These findings were consistent with our predicted association between downregulated *ITM2C* and poor prognosis in patients with breast cancer, especially with ER-negative status.

The overexpression of *CPEB2* (cytoplasmic polyadenylation element binding protein 2) has been reported to correlate with anoikis resistance, thereby contributing to the metastasis of triple-negative breast cancer [21]. Functional annotation identified an association of *CPEB2* with epithelial cell development, and *CPEB2* was shown to promote differentiation and inhibit the epithelial-to-mesenchymal transition in mammary epithelial cells [22]. Breast carcinogenesis occurs via the oncogenic transformation of mammary epithelial cells [23]. We also predicted that *CPEB2* upregulation would be a prognostic factor for breast cancer, and identified *CPEB2* as most relevant to *ESR1* in the co-expression network. Notably, in our study, *ESR1* (estrogen receptor 1) was a top 7 gene related to prognosis and was enriched for hormone stimulus in the GO analysis. Approximately 70% of breast cancers express ER, and most exhibit the ER+ phenotype, which is sensitive to ER inhibition [24]. Ramos et al. reported that hypermethylation of *ESR1* could



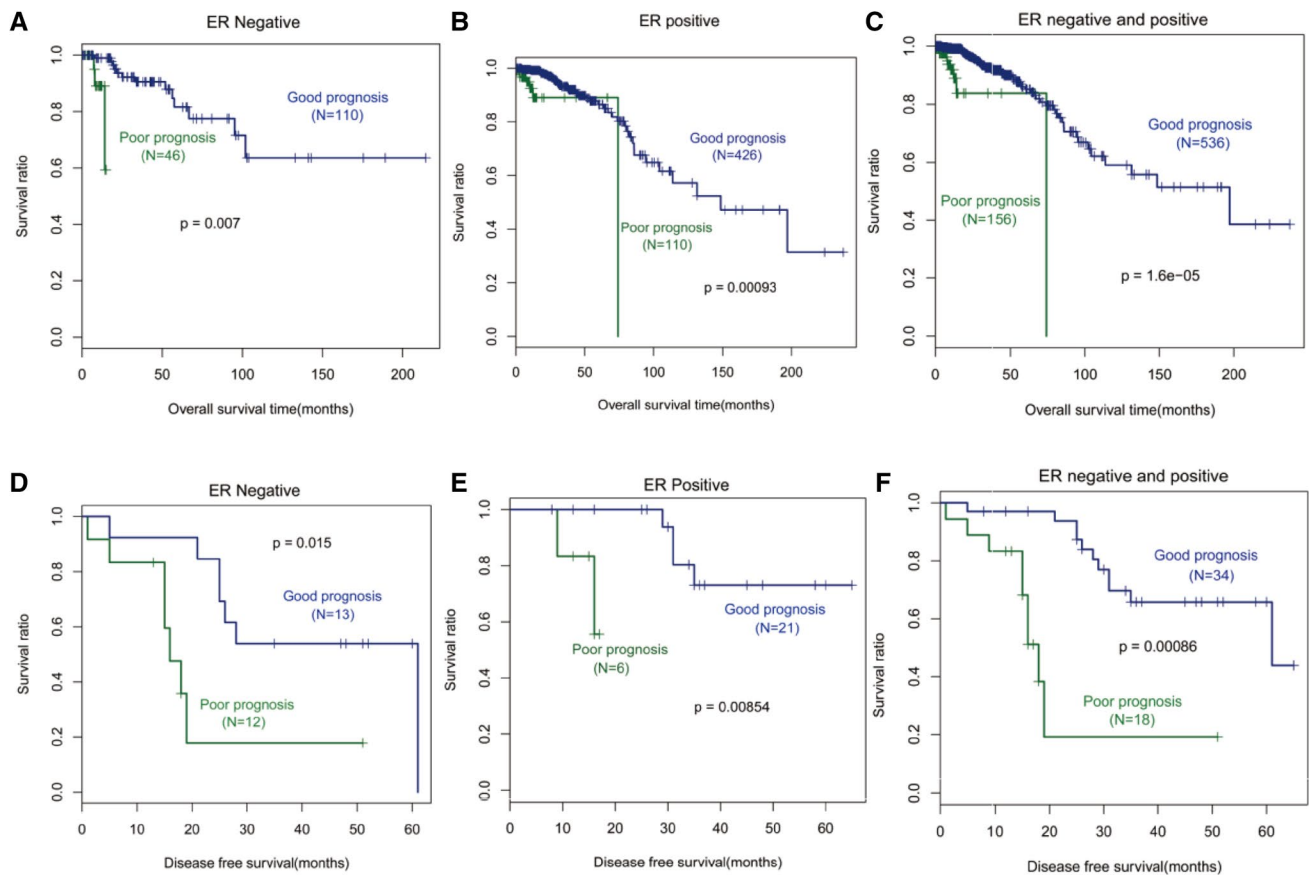
**Fig. 4** Results of a functional enrichment analysis. **a** The results of a Gene Ontology analysis of genes in the co-expression network. **b** The results of a KEGG pathway analysis of genes in the co-expression network. *KEGG* Kyoto Encyclopedia of Genes and Genomes

be used as an indicator of poor prognosis in sporadic breast cancer [25]. Similarly, we found that *ESR1* might be a prognostically relevant gene, particularly for ER+ breast cancer.

In the present study, another predicted prognostically relevant gene, *ANXA9*, was found to correlate significantly with *ESR1*. Strong expression of *ANXA9*, which encodes the calcium-dependent phospholipid-binding protein family member annexin A9, was found to correlate with the metastasis of breast cancer to the bone [26]. Miyoshi et al. demonstrated reduced survival among colorectal cancer patients with high *ANXA9* expression relative to those with low *ANXA9* expression, suggesting that *ANXA9* expression is a marker of poor prognosis in this population [27]. *VASN*, which encodes the ADAM Metallopeptidase Domain 17 (ADAM 17) substrate

vasorin, acts as a transforming growth factor-beta (TGF- $\beta$ ) trap. Potentially, the downregulation of *VASN* might inhibit ADAM17 and consequently activate TGF- $\beta$  signaling [28]. TGF- $\beta$  signaling plays an important role in breast cancer progression, and expression of the TGF- $\beta$  receptor may strongly reduce overall survival among patients with ER-negative tumors [29]. Although *ANXA9* and *VASN* have not been confirmed as prognostically relevant for breast cancer, the above information and our results led us to speculate that *ANXA9* and *VASN* might act as prognostic indicators.

We additionally attempted to construct a prognostic prediction system using a set of 48 signature genes that could be used to indicate good or poor prognosis and yield high prognostic scores. Moreover, we extracted the expression



**Fig. 5** The construction and validation a prognostic prediction system according to a Kaplan–Meier survival analysis. **a–c** Kaplan–Meier survival curves demonstrating correlations between the prognostic prediction system and actual survival information. **d–f** Kaplan–Meier

survival curves showing correlations between the GSE61304 samples and actual survival information. The blue and green lines indicate the survival curves for good and poor prognosis, respectively. (Color figure online)

values of those 48 signature genes from another GSE61304 dataset to validate our prognostic prediction system. Accordingly, our system yielded prognostic scores that could be used to classify the samples as having either a good prognosis or poor prognosis. Interestingly, the Kaplan–Meier survival curve analysis revealed that the DFS was considerably higher among the good prognosis group than among the poor prognosis group, regardless of the ER status. This finding demonstrates the accuracy and reliability of our prognostic prediction system.

There are limitations in our manuscript. The gene signature is derived from the segregation of patients with different ER status based on Bayes discriminant analysis, which could cause bias of data analysis. The 48 gene signature was screened based on bioinformatics analysis and this study may just provide clues for future study of patients with ER positive breast cancer. Prospective studies such as randomized controlled trials and cohort studies have not been performed to validate the reliability of prognostic prediction system outcomes. The future focus

of our work is to collect more samples and improve our prognostic prediction system experimentally. It is also necessary to select some key genes relevant to the prognosis of ER-positive and ER-negative breast cancer for randomized controlled trials. We can extract serum from clinical breast cancer patients and up-regulate or down-regulate key genes. Meanwhile, we need to set up a blank control and experimental control, and verify the validity of key genes by comparing the prognosis of each group to further verify the prognosis prediction system.

In summary, a total of 109 prognostically relevant mRNAs were identified with regard to ER status. DEGs such as *RASEF*, *ITM2C*, *CPEB2*, *ESR1*, *ANXA9*, and *VASN*, which correlate strongly with other mRNAs, might be closely associated with prognosis among patients with breast cancer. In addition, we constructed a prognostic prediction system comprising 48 signature genes, and our validation of this system demonstrated its efficacy and consistency. Therefore, our system will contribute to exploring novel prognostic factors both from ER-positive



and from ER-negative breast cancers based on mechanistic studies.

**Funding** This work was supported by The Military logistics Research Project (Number CWH17C017 to Linhai Li), Science and Technology Program of Guangzhou, China (Number 201804010186 to Bin Xiao) and the National Natural Science Foundation of China No. 81402409 (Li Wang).

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Ghoncheh M, Mirzaei M, Salehiniya H (2015) Incidence and mortality of breast cancer and their relationship with the Human Development Index (HDI) in the world in 2012. *Asian Pac J Cancer Prev* 16(18):8439–8443
- Zuo TT, Zheng RS, Zeng HM et al (2017) Female breast cancer incidence and mortality in China. *Thorac Cancer* 8(3):214–218
- Wang K, Yee C, Tam S et al (2018) Prevalence of pain in patients with breast cancer post-treatment: a systematic review. *Breast* 42:113–127
- Im A, Vogel VG, Ahrendt G et al (2009) Urinary estrogen metabolites in women at high risk for breast cancer. *Carcinogenesis* 30(9):1532–1535
- Higa GM, Fell RG (2013) Sex hormone receptor repertoire in breast cancer. *Int J Breast Cancer* 2013:284036
- Shigekawa T, Ijichi N, Ikeda K et al (2011) FOXP1, an estrogen-inducible transcription factor, modulates cell proliferation in breast cancer cells and 5-year recurrence-free survival of patients with tamoxifen-treated breast cancer. *Horm Cancer* 2(5):286–297
- Mehta RJ, Jain RK, Leung S et al (2012) FOXA1 is an independent prognostic marker for ER-positive breast cancer. *Breast Cancer Res Treat* 131(3):881–890
- Ali HR, Dawson SJ, Blows FM et al (2012) Aurora kinase A outperforms Ki67 as a prognostic marker in ER-positive breast cancer. *Br J Cancer* 106(11):1798–1806
- Lu K, Alcivar AL, Ma J et al (2017) NRF2 Induction supporting breast cancer cell survival is enabled by oxidative stress-induced DPP3-KEAP1 interaction. *Cancer Res* 77(11):2881–2892
- Yang SX, Polley E, Lipkowitz S (2016) New insights on PI3K/AKT pathway alterations and clinical outcomes in breast cancer. *Cancer Treat Rev* 45:87–96
- Hosford SR, Dillon LM, Bouley SJ et al (2017) Combined inhibition of both p110alpha and p110beta isoforms of phosphatidylinositol 3-Kinase is required for sustained therapeutic effect in PTEN-deficient, ER+ breast cancer. *Clin Cancer Res* 23(11):2795–2805
- Wang J, Zhang C, Chen K et al (2015) ERbeta1 inversely correlates with PTEN/PI3K/AKT pathway and predicts a favorable prognosis in triple-negative breast cancer. *Breast Cancer Res Treat* 152(2):255–269
- Umekita Y, Ohi Y, Sagara Y et al (2002) Overexpression of cyclinD1 predicts for poor prognosis in estrogen receptor-negative breast cancer patients. *Int J Cancer* 98(3):415–418
- Wang P, Wang Y, Hang B et al (2016) A novel gene expression-based prognostic scoring system to predict survival in gastric cancer. *Oncotarget* 7(34):55343–55351
- Jonsson G, Bendahl PO, Sandberg T et al (2005) Mapping of a novel ocular and cutaneous malignant melanoma susceptibility locus to chromosome 9q21.32. *J Natl Cancer Inst* 97(18):1377–1382
- Oshita H, Nishino R, Takano A et al (2013) RASEF is a novel diagnostic biomarker and a therapeutic target for lung cancer. *Mol Cancer Res* 11(8):937–951
- Wu H, Liu G, Li C et al (2003) bri3, a novel gene, participates in tumor necrosis factor-alpha-induced cell death. *Biochem Biophys Res Commun* 311(2):518–524
- Supiot S, Gouraud W, Campion L et al (2013) Early dynamic transcriptomic changes during preoperative radiotherapy in patients with rectal cancer: a feasibility study. *World J Gastroenterol* 19(21):3249–3254
- Alvarez C, Aravena A, Tapia T et al (2016) Different Array CGH profiles within hereditary breast cancer tumors associated to BRCA1 expression and overall survival. *BMC Cancer* 16:219
- Shi M, Ma F, Liu J et al (2017) A functional BRCA1 coding sequence genetic variant contributes to prognosis of triple-negative breast cancer, especially after radiotherapy. *Breast Cancer Res Treat* 166(1):109–116
- Johnson RM, Vu NT, Griffin BP et al (2015) The alternative splicing of cytoplasmic polyadenylation element binding protein 2 drives anoikis resistance and the metastasis of triple negative breast cancer. *J Biol Chem* 290(42):25717–25727
- Grudzien-Nogalska E, Reed BC, Rhoads RE (2014) CPEB1 promotes differentiation and suppresses EMT in mammary epithelial cells. *J Cell Sci* 127(Pt 10):2326–2338
- Elenbaas B, Spirio L, Koerner F et al (2011) Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev* 15(1):50–65
- Toy W, Shen Y, Won H et al (2013) ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nat Genet* 45(12):1439–1445
- Ramos EA, Camargo AA, Braun K et al (2010) Simultaneous CXCL12 and ESR1 CpG island hypermethylation correlates with poor prognosis in sporadic breast cancer. *BMC Cancer* 10:23
- Smid M, Wang Y, Klijn JG et al (2006) Genes associated with breast cancer metastatic to bone. *J Clin Oncol* 24(15):2261–2267
- Miyoshi N, Yamamoto H, Mimori K et al (2014) ANXA9 gene expression in colorectal cancer: a novel marker for prognosis. *Oncol Lett* 8(5):2313–2317
- Malapeira J, Esselens C, Bech-Serra JJ et al (2011) ADAM17 (TACE) regulates TGFbeta signaling through the cleavage of vasorin. *Oncogene* 30(16):1912–1922
- Buck MB, Knabbe C (2006) TGF-beta signaling in breast cancer. *Ann N Y Acad Sci* 1089:119–126

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.