Technology in Cancer Research & Treatment
Volume 19: 1-9
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1533033820957011
journals.sagepub.com/home/tct
SAGE

Transcriptomics and Prognosis Analysis to Identify Critical Biomarkers in Invasive Breast Carcinoma

Jun Wu, MB[1], Xiao-Jun Liu, MB[2], Jia-Nan Hu, MB[3], Xu-Hui Liao, MB[1], and Fei-Fei Lin, MB[4]

Abstract

Objective: Invasive breast cancer (BRCA) is one of the prevalent types of invasive tumors with high mortality worldwide. Due to the lack of effective treatment to control the recurrence of distant metastases, the prognosis of BRCA is still very unsatisfactory. We aimed to find some biomarkers by bioinformatics analysis for survival prediction. **Methods:** Differentially expressed genes (DEGs) were screened out based on tumor group and normal group. Then, the weighted gene correlation network analysis (WGCNA) was employed to identify the clinically associated gene sets. Meanwhile, the enrichment analyses were performed for the functional annotation of the critical genes. The Kaplan Meier analysis calculated the essential genes' prognostic value. **Results:** After threshold screening, 1655 DEGs were obtained for subsequent analysis. 51 out of 1655 DEGs were significantly associated with BRCA patients' estrogen receptor status via WGCNA. Three genes (FABP7, CXCL3, and LOC284578) out of the 51 genes were associated with overall survival, and 3 genes were relapse-free survival associated. Finally, we obtained 5 essential prognostic associated genes (FABP7, CXCL3, LOC284578, CAPN6, and NRG2), which could be used as prognostic factors for BRCA. **Conclusion:** Our findings obtained a gene module associated with BRCA clinical trait and several key genes that acted as essential components in the prognostic of cancer, which may improve its treatment.

Keywords

invasive breast cancer, differential analysis, WGCNA, survival analysis, prognostic biomarkers

Received: June 2, 2019; Revised: July 8, 2020; Accepted: August 18, 2020.

Introduction

Breast cancer is one of the malignant tumors with high mortality worldwide.[1] It is a heterogeneous disease with different molecular subtypes, cell content, clinical manifestation, and treatment response.[2,3] According to reports, after breast protection therapy (BCT) for invasive breast cancer (BRCA), the patient's treatment and pathological factors are associated with an increased risk of recurrence of ipsilateral breast tumors.[4,5] It is estimated that there will be 276,480 new cases and 42,170 deaths from BRCA in the United States in 2020.[6] At present, BRCA treatment mainly includes surgery, radiotherapy, chemotherapy, and endocrine therapy.[7,8] However, due to the lack of effective treatment to control the recurrence of distant metastases, especially in advanced patients, the prognosis of BRCA is still very unsatisfactory.[9] In recent years, biologically targeted therapy has been proven to be useful for various cancers, significantly improving the prognosis.[10-13] Therefore, the identification of new biomarkers is essential for the specific treatment of patients.

WGCNA is based entirely on scale-free networks and is used to determine the relationship between genes, thereby identifying modules (clusters) of highly related genes.[14] WGCNA is an ideal method for identifying gene modules and determining essential genes for phenotypic traits. For example, Zou et al. used WGCNA to detected the loss of MAGI2 promotes chronic kidney disease, which regulates cytoskeletal rearrangement in podocytes.[15] Liang W et al. also used WGCNA to identified several key genes that acted as essential components

[1] Pathology Department, The People's Hospital of Lishui, Zhejiang, China
[2] External Liaison Office, The Central Hospital of Lishui City, Zhejiang, China
[3] The Oncology Department, The People's Hospital of Lishui, Zhejiang, China
[4] Department of Clinical laboratory, The People's Hospital of Lishui, Zhejiang, China

Corresponding Author:
Fei-Fei Lin, Department of Clinical Laboratory, The People's Hospital of Lishui, Lishui, Zhejiang 323000, China.
Email: lslff1@163.com

of diabetes-associated cardiovascular disease.[16] Therefore, in this study, our goal is to use the WGCNA algorithm to identify highly relevant gene modules related to breast cancer development and then detect the hub gene (network center gene) to discover new proven effective breast cancer diagnosis and treatment biomarkers.

Methods

Data Sources

The high-throughput RNA-seq data of 1241 patients with BRCA and 113 normal samples were downloaded from the TCGA database. The gene expression profiles were quantified by fragments per kilobase of transcript per million mapped reads (FPKM) normalized estimation and log2-based transformation. We selected 838 BRCA patients with overall survival and relapse-free survival information for further analysis. The following corresponding clinical characteristics were also extracted from the TGGA.

Differential Analysis

The differential expression analyses were conducted between 838 BRCA patients and 113 normal samples using the "limma" R package.[17] In this study, genes with an absolute log2 fold change > 0.585, and adjusted p-value < 0.05 are differentially expressed.

Co-Expression Module Detection

We used the WGCNA R package to construct the co-expression network. At first, a network construction function was used to construct the co-expression network of all the differentially expressed genes (DEGs). The "pickSoftThreshold" R function was used to calculate the soft threshold power, and the "softConnectivity" function was used to calculate the network's scale-free value. Second, hierarchical clustering and the dynamic tree cut function were used to detect modules. Next, all modules were related to clinical information by "cor" function based on correlation analysis. Gene significance (GS) and module membership (MM) were calculated to relate modules to clinical traits. Finally, the key genes from the preserved module were explored. The correlation (cor.) Gene GS > 0.2 and cor. Gene MM > 0.7 was the inclusive criterion for screening key genes.[18]

Function Enrichment Analysis

For the DEGs, KEGG pathway analysis[19] and GO enrichment analysis[20] was performed using Metascape (http://metascape.org) website tool to explore the potential function.[21] Only the tasks with an adjust-p-value <0.05 were selected.

Survival Analysis

Kaplan-Meier analysis with log-rank test was conducted by "survival" R package to screen prognostic associated vital genes. Survival curves were drawn to illustrate the associations of expression levels of these genes with BRCA. The OSbrca Tool was used to testify the survival ability of potential prognostic biomarkers.[22]

Statistical Analysis

All analyses were conducted using R software (version 3.6.3) with related packages. T-test was conducted to compare the differences. And adjusted p-value < 0.05 was regarded as significant.

Results

Identification of Differentially Expressed Genes in BRCA

We downloaded transcriptome and clinical data of 838BRCA cases from the TCGA database. Based on our threshold value (absolute log2 fold change > 0.585 and adjusted p-value <0.05), there were 1655 DEGs between BRCA patients and normal groups (Figure 1A). Then, the function analysis showed that these DEGs involved in the IL-17 signaling pathway, PPAR signaling pathway and cell adhesion molecules (CAMs), and so on, and they also associated with receptor regulator activity and growth factor activity, etc (Figure 1B&C). It showed that these DEGs might play a cure role in invasive breast carcinoma.

Construction and Analysis of Gene Co-Expression Modules

After differential analyses, a co-expression network of 1655 DEGs were conducted (Figure 2). We first calculated the soft threshold power, which could help us construct a more suitable system based on the co-expression similarity. The function "pickSoftThreshold" in the WGCNA package was used to perform the analysis of network topology. We constructed the gene network and identified modules using the 1-step network construction function. To cluster splitting, the minimum module size was set at 30. The soft threshold power was set at 5 in the subsequent analysis because the scale independence reached 0.85 (Figure 2A and B). The soft power also can assure a scale-free topology model (scale-free R2 = -0.95, Figure 2C). Finally, 5 gene co-expression modules (turquoise, blue, yellow, brown, and green) were constructed (Figure 2D). We analyzed the connectivity of eigengenes. Eigengenes can provide information about the relationship between the gene co-expression modules. The results showed that 5 modules could be clustered into 2 clusters (Figure 2E). Four combinations (modules turquoise, modules blue, modules yellow, and modules brown and green) had a high degree of interaction connectivity.

We mapped the relationships between the identified modules (Figure 3). The heatmap depicts the topological overlap matrix (TOM) among all genes included in the analysis. The darker red color represents a low overlap, and the progressively light color represents an increasing overlap. The results of this

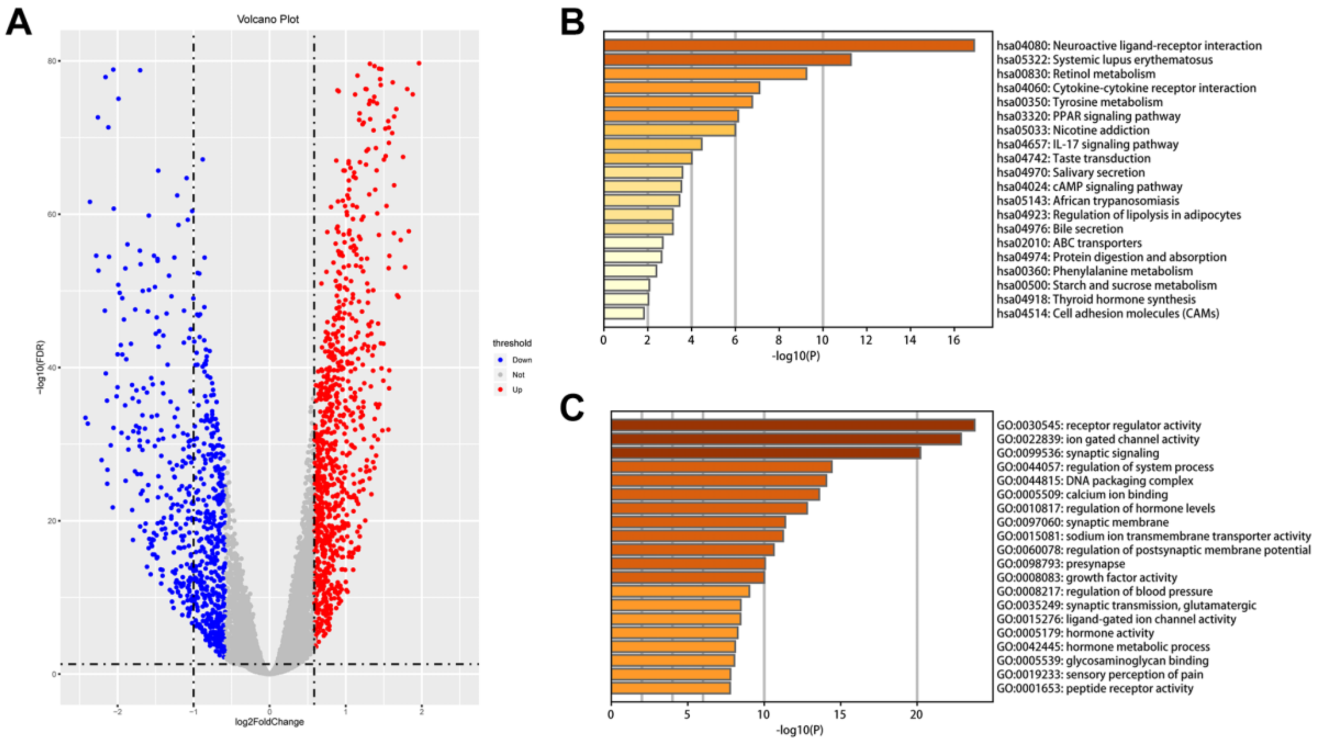

Figure 1. Differential Analysis of TCGA BRCA samples. **A,** Volcano plot representing differentially expressed genes between BRCA patients andnormal samples. The significantly upregulated genes are shown in red while downregulated genes are shown in blue ($P < 0.05$). **B&C,** Enrichment KEGG, and GO analysis of differentially expressed genes.

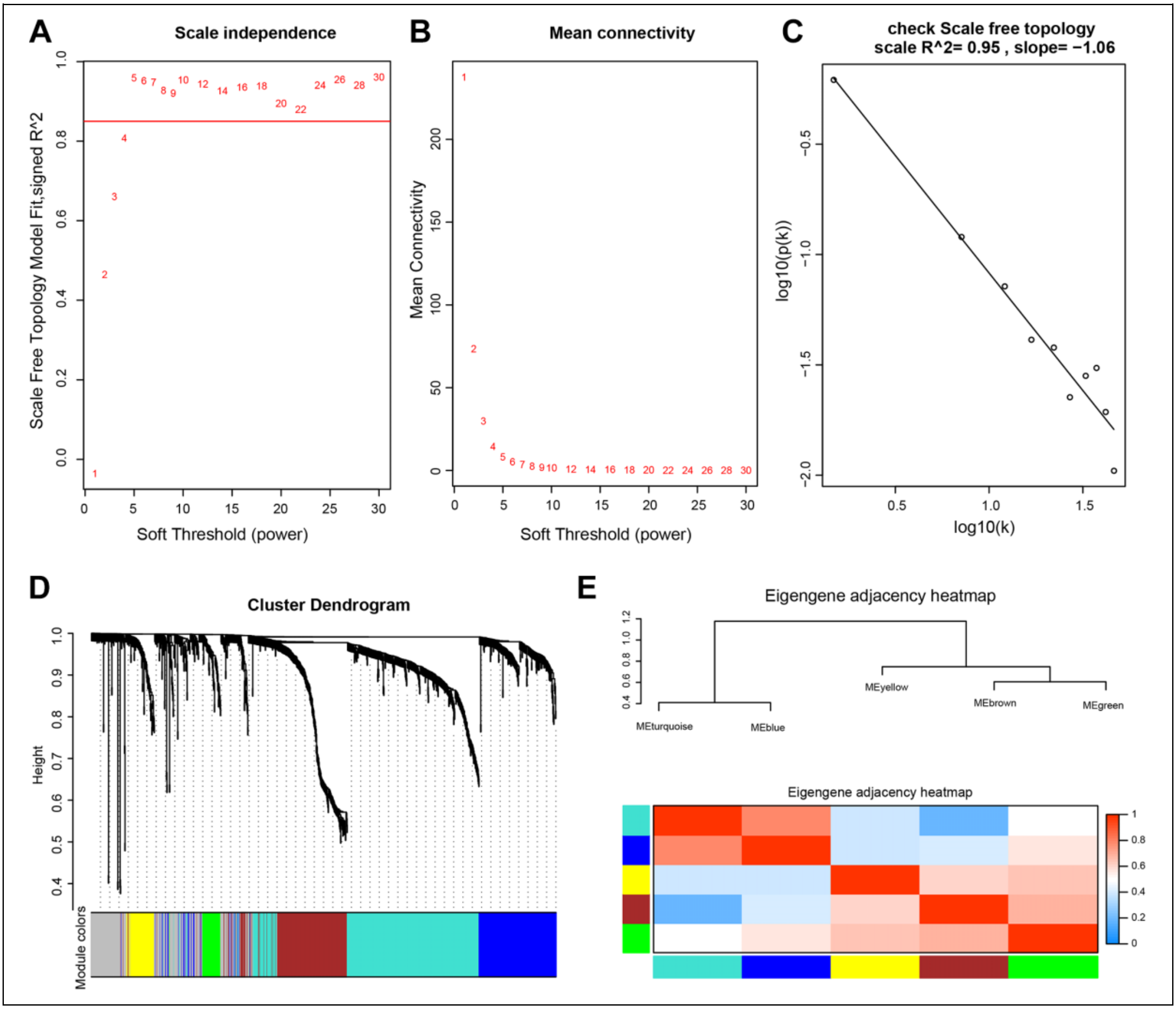

Figure 2. Analysis of the weighted gene correlation network analysis (WGCNA) of DEGs.A&B Analysis of network topology for various soft-threshold powers. The x-axis reflects the soft-thresholding power. The y-axis indicates the scale-free topology model fit index. **C,** Scale-free topology model under the soft threshold powers. **D,** Clustering dendrogram of genes, with dissimilarity based on the topological overlap, together with assigned module colors. **E,** Eigengene network representing the relationships among the modules.

analysis indicated that the gene expression was relatively independent between modules.

Identification of Clinically Related Modules

We correlated modules with clinical characteristics (Figure 4A) and searched for the most significant associations. This analysis showed that the blue module was most significantly correlated with estrogen receptor status (Figure 4B). With the cor. gene GS > 0.2 and cor. Gene MM > 0.7 threshold limits, 51 out of 238 hub genes were identified (Figure 4C).

The enrichment analysis was executed to describe the function of the critical genes (Supplemental Figure 1). The results indicated that the essential genes were significantly enriched in Hallmark KRAS signaling and IL-17 signaling pathway. These genes also enriched in several tumor-related terms, such as epithelial cell proliferation, germ cell development, and protein kinase activity, etc. It suggests the potential regulatory mechanism of these critical genes in BRCA.

Survival Analysis of Crucial Genes

To determine the prognostic performance of these genes, all of the 51 essential genes were tested by Kaplan-Meier analysis. It was found that 3 out of the 51 essential genes (FABP7, CXCL3, and LOC284578) were significantly associated with overall survival

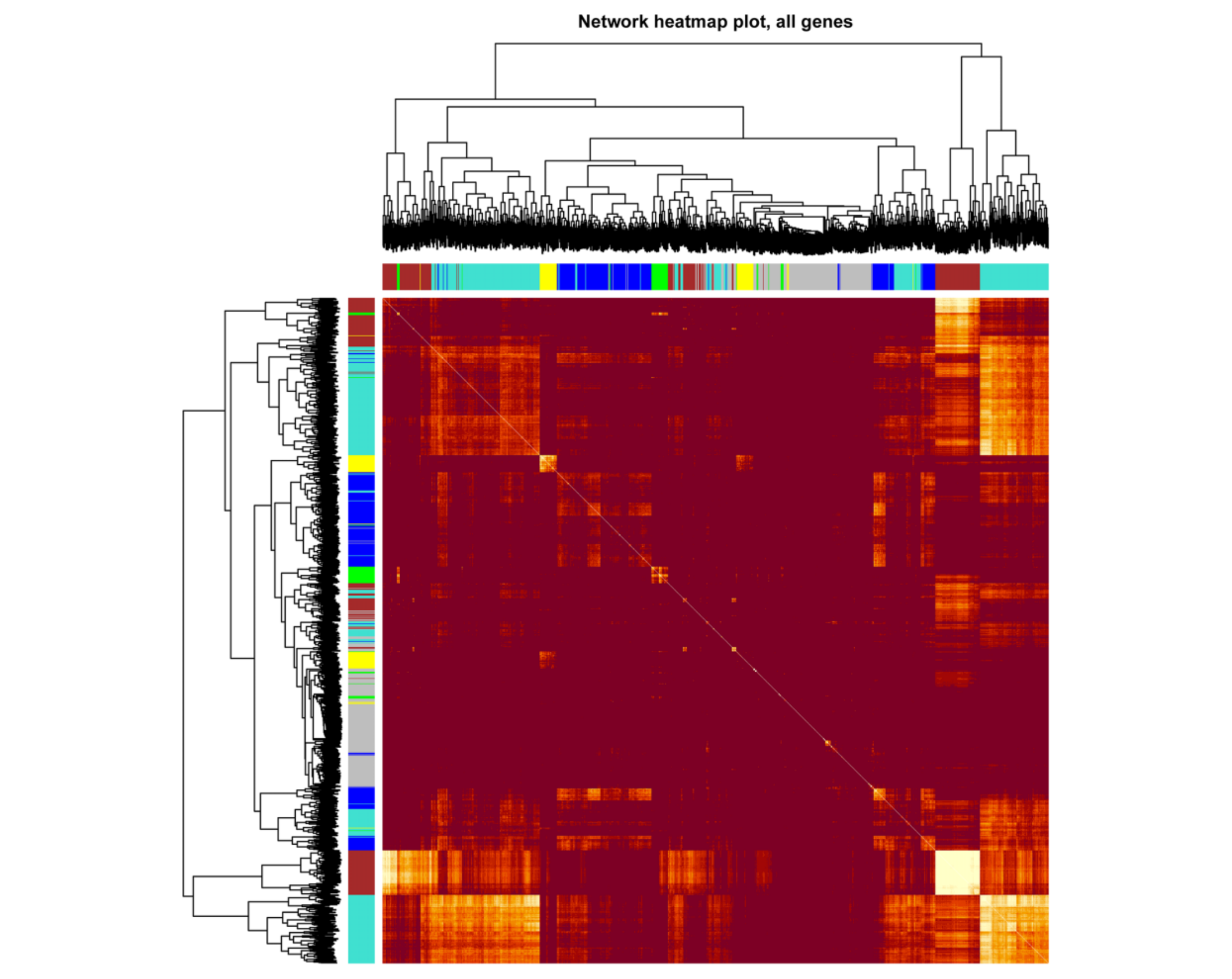

Figure 3. Visualization of the WGCNA network. Heatmap plot representing the gene network. The heatmap depicts the topological overlap matrix among all genes in the analysis.

(OS) (Figure 5A). And 3 essential genes (FABP7, CAPN6, and NRG2) were significantly related to Relapse-free survival (RFS), too (Figure 5B). In particular, key gene FABP7 has dramatically associated with both OS and RFS. Finally, we obtained 5 essential prognostic associated genes (FABP7, CXCL3, LOC284578, CAPN6, and NRG2), which could be used as prognostic factors for BRCA. To further determine the prognostic efficacy of the 5 crucial genes, we used the OSbrca Tool to testify their survival ability. In overall survival group, we found FABP7 and CXCL3 were also associated with overall survival (OS) in GSE18229 and GSE39004. LOC284578 were not detected in these data sets (Supplemental Figure 2). In addition, FABP7, CAPN6, and NRG2 were related to Relapse-free survival (RFS) in the independent test data sets (GSE18229, GSE10893, GSE21653, and GSE2607) (Supplemental Figure 3).

We also used GEPIA database[23] to testify the expression level of the 5 genes in BRCA tumor patients compare with normal samples (Supplemental Figure 4). We can see that all of the 5 genes were down-regulated in tumor, and the low expression of them in BRCA was related to the poor prognosis. Then, we used the HPA (The Human Protein Atlas) database to explore the expression of these genes in clinical patients.[24] Only FABP7 and NRG2 were found expressed in the BRCA tumor tissues by immunohistochemistry (IHC) analysis (Supplemental Figure 5). This suggests that the further study of us is using IHC to verify the expression of these genes in clinical patients. All these results showed that the 5 genes played a crucial role in BRCA prognosis.

Discussion

Invasive breast cancer is the most common malignancy in females and seriously threatens physical and mental health.[25-27] In recent years, bioinformatics analysis has been widely used in cancer research.[28-30] This kind of research using a public

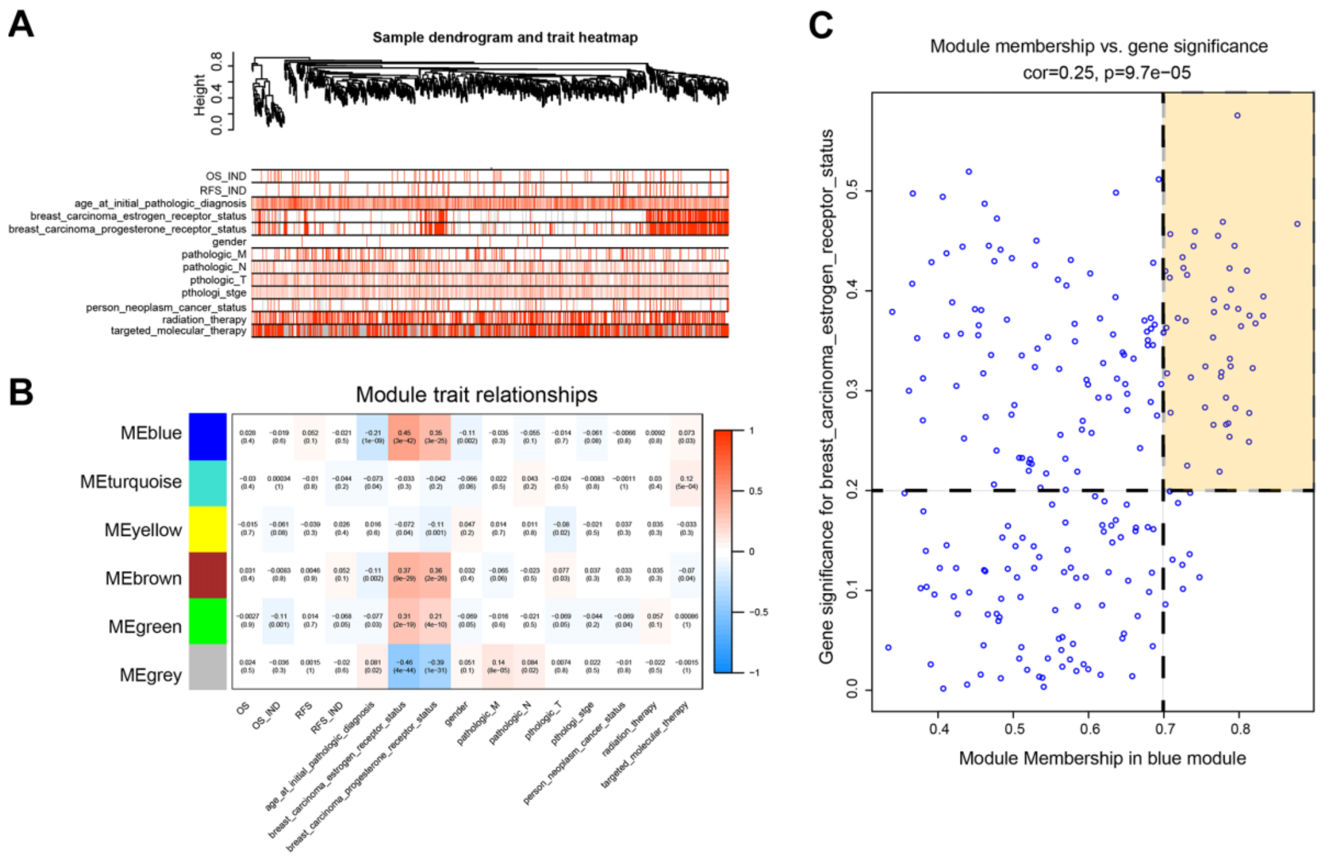

Figure 4. Module-trait associations. A, The trait heatmap of 838 BRCA patients. B, Module-trait relationships. Each row corresponds to a module, and each column corresponds to a trait. Each cell contains the corresponding correlation and P-value. C, Dot plot representing the genes' gene significance and module membership in estrogen receptor status associated blue module.

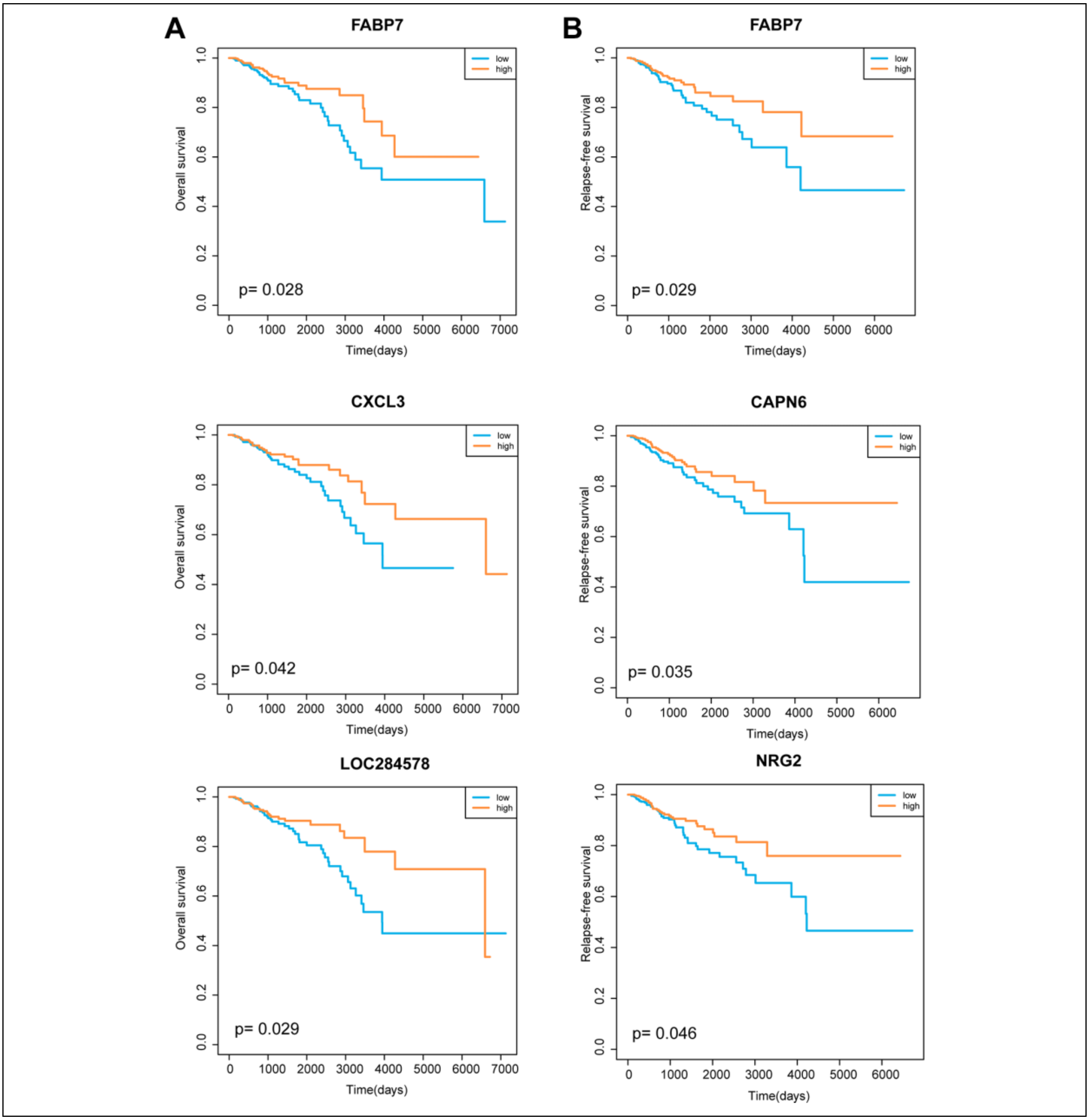

Figure 5. Survival analysis of estrogen receptor status associated with crucial genes.

database containing many samples provides an excellent way to identify tumor biomarkers.[31-33]

In this study, we built the co-expression modules via WGCNA using the TCGA BRCA data. We identified genes that were significantly differentially expressed in BRCA patients compared with normal samples. Then the function enrichment was investigated to show these genes' functions. After WGCNA analysis, we obtained a gene module that associated with estrogen receptor status, and several hub genes in the module were survival associated. FABP7 has been reported to be a key metabolic regulator in HER2+ breast cancer brain metastasis[34] and could mediate triple-negative breast cancer cell death via PPAR-α signaling.[35] CXCL3 was found to be a potential target for breast cancer metastasis[36] and inhibition of CXCL3 could reduce STAT3 activation.[37] LOC284578 has not been reported to be related to the prognosis of breast cancer in previous studies, which may be a new prognostic biomarker of BRCA. CAPN6 could regulate RAC1 activity and cell motility through interaction with GEF-H1.[38] It also can be regulated by PI3K-Akt pathway and may be a therapeutic target of cancer.[39] TGFα, HB-EGF, and NRG2 have been reported to be related to the biological aggressiveness of the tumors.[40] The genetic variants in NRG2 also can influence breast cancer risk.[41] In summary, the 5 genes identified in our study may use as new biomarkers for improving the prognosis of BRCA patients.

Our research does have some limitations. First of all, our research results are based on public online database information, so we need to prove it through experiments. In the future, we will build a cell model to express the above 5 genes to verify our conclusions differentially. Second, although we conducted gene enrichment analysis, our research did not clarify the mechanism of these critical genes participating in BRCA. Therefore, this is one of the focuses of our future examinations.

Authors' Note

Jun Wu and Xiao-Jun Liu contributed equally to this work. The study was based on an analysis of public data and did not address any ethical issues.

Acknowledgments

The authors wish to thank the 4 anonymous reviewers whose constructive comments were helpful to strengthen the presentation of this study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Fei-Fei Lin https://orcid.org/0000-0002-2388-5011

Supplemental Material

Supplemental material for this article is available online.

References

1. DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. *CA Cancer J Clin*. 2011;61(6):409-418.
2. Lu YS, Kuo SH, Huang CS. Recent advances in the management of primary breast cancers. *J Formos Med Assoc*. 2004;103(8): 579-598.
3. Desmedt C, Zoppoli G, Sotiriou C, Salgado R. Transcriptomic and genomic features of invasive lobular breast cancer. *Semin Cancer Biol*. 2017;44:98-105.
4. Houssami N, Macaskill P, Marinovich ML, et al. Meta-analysis of the impact of surgical margins on local recurrence in women with early-stage invasive breast cancer treated with breast-conserving therapy. *Eur J Cancer*. 2010;46(18):3219-3232.
5. Schnitt SJ, Moran MS, Giuliano AE. Lumpectomy margins for invasive breast cancer and ductal carcinoma in situ: current guideline recommendations, their implications, and impact. *J Clin Oncol*. 2020;38(20):2240-2245. JCO1903213.
6. Siegel RL, Miller KD, Jemal A. Cancer statistics. *CA Cancer J Clin*. 2020;70(1):7-30.
7. Goetz MP, Gradishar WJ, Anderson BO, et al. NCCN guidelines insights: breast cancer, version 3.2018. *J Natl Compr Canc Netw: JNCCN*. 2019;17(2):118-126.
8. Possanzini M, Greco C. Stereotactic radiotherapy in metastatic breast cancer. *Breast*. 2018;41:57-66.
9. Stacker SA, Williams SP, Karnezis T, Shayan R, Fox SB, Achen MG. Lymphangiogenesis and lymphatic vessel remodelling in cancer. *Nat Rev Cancer*. 2014;14(3):159-172.
10. Miao R, Chen HH, Dang Q, et al. Beyond the limitation of targeted therapy: improve the application of targeted drugs combining genomic data with machine learning. *Pharmacol Res*. 2020; 159:104932.
11. Chuang YH, Lee CH, Lin CY, et al. An integrated genomic strategy to identify chrnb4 as a diagnostic/prognostic biomarker for targeted therapy in head and neck cancer. *Cancers*. 2020;12(5): 1324.
12. Tse BWC, Volpert M, Ratther E, et al. Neuropilin-1 is upregulated in the adaptive response of prostate tumors to androgen-targeted therapies and is prognostic of metastatic progression and patient mortality. *Oncogene*. 2017;36(24):3417-3427.
13. Prasad CP, Manchanda M, Mohapatra P, Andersson T. WNT5A as a therapeutic target in breast cancer. *Cancer Metastasis Rev* 2018;37(4):767-778.
14. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559.
15. Zuo Z, Shen JX, Pan Y, et al. Weighted gene correlation network analysis (WGCNA) detected loss of MAGI2 promotes chronic kidney disease (CKD) by podocyte damage. *Cell Physiol Biochem*. 2018;51(1):244-261.
16. Liang W, Sun F, Zhao Y, Shan L, Lou H. Identification of susceptibility modules and genes for cardiovascular disease in diabetic patients using WGCNA analysis. *J Diabetes Res*. 2020; 2020:4178639.
17. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7): e47.
18. Niemira M, Collin F, Szalkowska A, et al. Molecular signature of subtypes of non-small-cell lung cancer by large-scale transcriptional profiling: identification of key modules and genes by weighted gene co-expression network analysis (WGCNA). *Cancers*. 2019;12(1):37.
19. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 2017;45(D1): D353-D361.
20. Dennis G Jr, Sherman BT, Hosack DA, et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*. 2003;4(5): P3.
21. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10(1):1523.
22. Yan Z, Wang Q, Sun X, et al. A web server for breast cancer prognostic biomarker investigation with massive data from tens of cohorts. *Front Oncol*. 2019;9:1349.

23. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*. 2017;45(W1): W98-W102.
24. Uhlen M, Zhang C, Lee S, et al. A pathology atlas of the human cancer transcriptome. *Science*. 2017;357(6352):eaan2507.
25. Lai BW, Tsang JY, Poon IK, et al. The clinical significance of neuroendocrine features in invasive breast carcinomas. *Oncologist*. 2020.
26. Dania V, Liu Y, Ademuyiwa F, Weber JD, Colditz GA. Associations of race and ethnicity with risk of developing invasive breast cancer after lobular carcinoma in situ. *Breast Cancer Res*. 2019; 21(1):120.
27. Spronk I, Schellevis FG, Burgers JS, de Bock GH, Korevaar JC. Incidence of isolated local breast cancer recurrence and contralateral breast cancer: a systematic review. *Breast*. 2018;39:70-79.
28. Millstein J, Budden T, Goode EL, et al. Prognostic gene expression signature for high-grade serous ovarian cancer. *Ann Oncol*. 2020;S0923-7534(20):39841.
29. Chen Y, Liao LD, Wu ZY, et al. Identification of key genes by integrating DNA methylation and next-generation transcriptome sequencing for esophageal squamous cell carcinoma. *Aging*. 2020;12(2):1332-1365.
30. Tolios A, De Las Rivas J, Hovig E, Trouillas P, Scorilas A, Mohr T. Computational approaches in cancer multidrug resistance research: identification of potential biomarkers, drug targets and drug-target interactions. *Drug Resist Updat*. 2020;48:100662.
31. Hu X, Cong Y, Luo HH, et al. Cancer stem cells therapeutic target database: the first comprehensive database for therapeutic targets of cancer stem cells. *Stem Cells Transl Med*. 2017;6(2):331-334.
32. Tang Q, Zhang Q, Lv Y, Miao YR, Guo AY. SEGreg: a database for human specifically expressed genes and their regulations in cancer and normal tissue. *Brief Bioinform*. 2019;20(4): 1322-1328.
33. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*. 2004;6(1):1-6.
34. Cordero A, Kanojia D, Miska J, et al. FABP7 is a key metabolic regulator in HER2+ breast cancer brain metastasis. *Oncogene*. 2019;38(37):6445-6460.
35. Kwong SC, Jamil AHA, Rhodes A, Taib NA, Chung I. Metabolic role of fatty acid binding protein 7 in mediating triple-negative breast cancer cell death via PPAR-alpha signaling. *J Lipid Res*. 2019;60(11):1807-1817.
36. See AL, Chong PK, Lu SY, Lim YP. CXCL3 is a potential target for breast cancer metastasis. *Curr Cancer Drug Targets*. 2014; 14(3):294-309.
37. Marotta LL, Almendro V, Marusyk A, et al. The JAK2/STAT3 signaling pathway is required for growth of CD44(+)CD24(-) stem cell-like breast cancer cells in human tumors. *J Clin Invest*. 2011;121(7):2723-2735.
38. Tonami K, Kurihara Y, Arima S, et al. Calpain-6, a microtubule-stabilizing protein, regulates Rac1 activity and cell motility through interaction with GEF-H1. *J Cell Sci*. 2011;124(pt 8): 1214-1223.
39. Liu Y, Mei C, Sun L, et al. The PI3K-Akt pathway regulates calpain 6 expression, proliferation, and apoptosis. *Cell Signal*. 2011;23(5):827-836.
40. Revillion F, Lhotellier V, Hornez L, Bonneterre J, Peyrat JP. ErbB/HER ligands in human breast cancer, and relationships with their receptors, the bio-pathological features and prognosis. *Ann Oncol*. 2008;19(1):73-80.
41. Slattery ML, John EM, Stern MC, et al. Associations with growth factor genes (FGF1, FGF2, PDGFB, FGFR2, NRG2, EGF, ERBB2) with breast cancer risk and survival: the breast cancer health disparities study. *Breast Cancer Res Treat*. 2013;140(3): 587-601.