Breast Cancer (2018) 25:213–223
https://doi.org/10.1007/s12282-017-0817-5

ORIGINAL ARTICLE

Identification of key gene modules and pathways of human breast cancer by co-expression analysis

Qingnan Zhao[1] · Wenqing Song[2] · Dai yu He[2] · YanSong Li[2]

Received: 10 June 2017 / Accepted: 8 November 2017 / Published online: 23 November 2017

Abstract

Background Breast cancer is the most common and aggressive tumor causing injury to women world wide. Although gene expression analysis had been performed previously, systemic co-expression analysis for this cancer is still lacking to date. We attempted to identify the critical modules of breast cancer.

Methods Co-expression modules were established with the help of WGCNA and the interactions among them were performed by R language. Biological process and pathways analysis of co-expression genes were figured out by GO and KEGG functional enrichment analysis using DAVID dataset.

Results In this study, expression data of 4,000 genes from 136 samples with breast cancer was used for the establishment of co-expression modules. And nine modules were identified. There was much higher scale independence among different modules by interactions analysis. Moreover, there was an obvious difference in adjacency degree among different modules. The most enriched pathways as immune response and ubiquitin-mediated proteolysis were identified as the most critical modules of breast cancer by GO and KEGG enrichment analysis.

Conclusion Our result demonstrated that immune response and ubiquitin-mediated proteolysis could serve as prognostic and predictive markers for the occurrence of breast cancer, providing evidence for further analysis in the prognosis and treatment of breast cancer.

Keywords Breast cancer · Co-expression modules · Metabolic pathways

Introduction

Breast cancer is the most common and aggressive tumor causing great injury to women physically and mentally [1]. This disease largely affects women in their 40s to 60s. Women before or after the period of menopause were more prone to be affected. It is the second most cancer now, just after lung cancer, the principal cause of death from cancer among women both in developing and developed countries [2]. However, the mechanisms of critical pathways and their interactions involved in the occurrence and development of breast cancer, remain largely unknown. Up to now, early diagnosis is still the key to improving the curative effect in the clinical treatment of breast cancer [3, 4]. Therefore, in this study, we aimed to explore the molecule mechanism in the development of breast cancer and thus provide evidence for further research.

Weighted Gene Co-expression Network Analysis (WGCNA) is a method frequently used in the co-expression module correlation analysis by microarray samples [5]. Besides, it is a comprehensive collection of R functions, which is commonly used in various aspects of weighted correlation network analysis. It's widely used in various biological processes, such as cancer, genetics, and brain imaging data analysis [6], which is quite helpful for the identification of candidate biomarkers or therapeutic targets. Not only can it help in the process of comparing differentially expressed genes, but also help in figuring out the interactions among genes in different co-expression modules [7]. It is reported that WGCNA analysis had been performed on publicly available microarray data covering a genome-wide scale of genes. WGCNA was proven to be a promising and reliable tool for clinical diagnosis of breast cancer. In this study, a total of

✉ YanSong Li
yansongli3@hotmail.com

[1] Department of Breast Surgeon, China-Japan Union Hospital of JILIN University, Chang Chun 130033, Jilin Province, China

[2] Daqing Longnan Hospital, Daqing 163453, China

nine co-expression modules were constructed by WGCNA. In this study, the WGCNA analysis identified nine modules of genes with high topological overlap in total.

Kyoto Encyclopedia of Genes and Genomes (KEGG) [8], a bioinformatics resource for better understanding of high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, was widely used in the mechanism research. The result of KEGG analysis in this study showed that the enriched pathways of hsa04120 (ubiquitin-mediated proteolysis) in co-expression module nine were quite meaningful in the occurrence of breast cancer. We hope our study will help in better understanding the discovery of biomarker in the clinical diagnosis of breast cancer.

Materials and methods

Expression value analysis of microarray data of breast cancer samples

Probe values were downloaded from GEO dataset at the https://www.ncbi.nlm.nih.gov/geo/ of NCBI with the key word "breast cancer". Annotation information of microarray data was used to match probes with corresponding gene information. Probes matching with more than one gene were eliminated and the average expression values were calculated out for genes matching with more than one probe. The number of genes was calculated with different expression threshold value of genes so as to determine the appropriate threshold value. WGCNA algorithm was used to evaluate the expression value of genes. What is more, flashClust tool package in R language [9] was used to conduct the cluster analysis of samples at the appropriate threshold value.

Analysis of co-expression modules of breast cancer

Power values were screened out by WGCNA [5] algorithm in the construction of co-expression modules. Scale independence and average connectivity analysis of modules with different power value were performed by gradient test (power value ranging from 1 to 20). Appropriate power value was determined when the scale independence value was equal to 0.8. WGCNA algorithm was then used to construct the co-expression modules and extract the gene information in each module. The smallest number was set as 50 for the reliability of the result.

Interaction analysis of co-expression modules of breast cancer

WGCNA algorithm was used to analyze the interaction relationship among different co-expression modules. Heatmap tool package in R language was used to describe the strength of the relationship (strong or weak degree).

Functional annotation analysis of co-expression genes of breast cancer

Co-expression modules were ranging from the most to the least by the number of genes. Then, functional enrichment analysis was performed on the genes in these modules. Corresponding gene information was mapped to the DAVID dataset (https://david.ncifcrf.gov/summary.jsp) [10]. Gene ontology (GO) [11] and KEGG [8, 12] enrichment analysis were performed. Therefore, the enriched biological processes and metabolic pathways were obtained. The analysis was conducted with the condition of $P < 0.05$. If there were more than five records, then the top five were selected for the further analysis.

Results

Expression values analysis of microarray data of breast cancer

A total of 136 typical breast cancer samples were obtained from NCBI with the accession number of GSE12903 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12903) [13]. The sequencing platform was GPL96 ([HG-U133A] Affymetrix Human Genome U133A Array) and the number of cancer samples was from GSM305129 to GSM30526. This dataset was larger and much newer. These 136 tumors were from Breast cancer patients who had received adjuvant tamoxifen therapy only. Frozen tumor specimens source and clinical information for breast cancer patients are listed in Table 1 [13]. The microarray data was transformed to genes expression information using the original data. On one hand, probes matching with more than one gene were eliminated and the average value of expression value of genes matching with more than one probe was calculated out as the final expression value of the gene. Besides, genes with the negative values were eliminated. As a result, a total of 12,389 expression values of genes were obtained. Then, 4000 genes with the highest average expression value were selected for the cluster analysis by flashClust tool package of WGCNA algorithm (Fig. 1). As can be seen in Fig. 1, 136 breast cancer samples were divided into two clusters, GSM305262 and GSM305263, on the whole. Two samples were included in Cluster I, and 134 samples were included in Cluster II, which can be divided into two sub-clusters, including 124 samples (Sub-Cluster I) and 10 samples (Sub-Cluster II), respectively.

Table 1 Tumor characteristics for breast cancer patients in the present study

Characteristics	Information	Sample number (tamoxifen-treated, $n = 136$)
Source		
	Institute of Oncology, Ljubljana, Slovenia	36 (26%); 1997–1999
	National Cancer Institute, Bari, Italy	28 (21%); 1990–1998
	Technische Universitaet Muenchen, Germany	9 (7%); 1992–1999
	one US institution, Cleveland Clinic Foundation	63 (46%); 1987–2000
Age (years)		
	Mean (SD)	64 (9)
	≤ 40	4 (3%)
	41–55	23 (17%)
	56–70	80 (59%)
	> 70	29 (21%)
	Unknown	0
T stage		
	T1	63 (46%)
	T2	65 (48%)
	T3/4	7 (5%)
	Unknown	1 (1%)
Tumor grade		
	Poor	30 (22%)
	Moderate	43 (32%)
	Good	43 (32%)
	Unknown	55 (40%)
Metastasis within 5 years		
	Yes	12 (9%)
	No	124 (91%)

Construction of co-expression module of breast cancer

Co-expression modules were constructed by the expression values of 4000 genes in 136 breast cancer samples using the WGCNA algorithm. Power value was one of the most critical parameters in the construction process, which mainly affected the scale independence and average connectivity of co-expression modules. Firstly, we screened the appropriate power value. When power value was equal to 8, the scale independence can be up to 0.8 (Fig. 2a) and was with higher average connectivity meanwhile (Fig. 2b). Therefore, power value equal to 8 was determined for further analysis. 4000 genes with highest expression value in 136 breast cancer samples were used for the construction of co-expression modules (Fig. 2c). As a result, a total of nine co-expression modules were constructed by the screened power value (8) and each module was manifested in different colors. These modules were numbered from the most to the least by the number of genes. There were 996 genes in module 1 (gray), 607 genes in module 2 (turquoise), 563 genes in module 3 (blue), 553 genes in module 4 (brown), 403 genes in module 5 (yellow), 371 genes in module 6, (green), 305 genes in module 7 (red), 120 genes in module 8 (black) and 82 genes in module (pink). The average number of genes in these nine modules was 444. The information of module each gene belongs to was listed in supplement Table 2.

Interaction relationship among co-expression modules of genes

Interaction relationship among the nine co-expression modules of genes was further analyzed (Fig. 3). As can been from the result, there was not any obvious difference of the interaction relationship, on the whole, indicating the relative independence expression of genes in each module and the much higher scale independence among different modules. What is more, the connectivity degree of eigengenes was analyzed for the better understanding of interaction relationship among the constructed co-expression modules. First, cluster analysis was performed on these critical genes (Fig. 4a) and we found that these nine modules were enriched in two clusters, one included six samples (module 1, 3, 5, 7, 8, 9) while the other included three samples

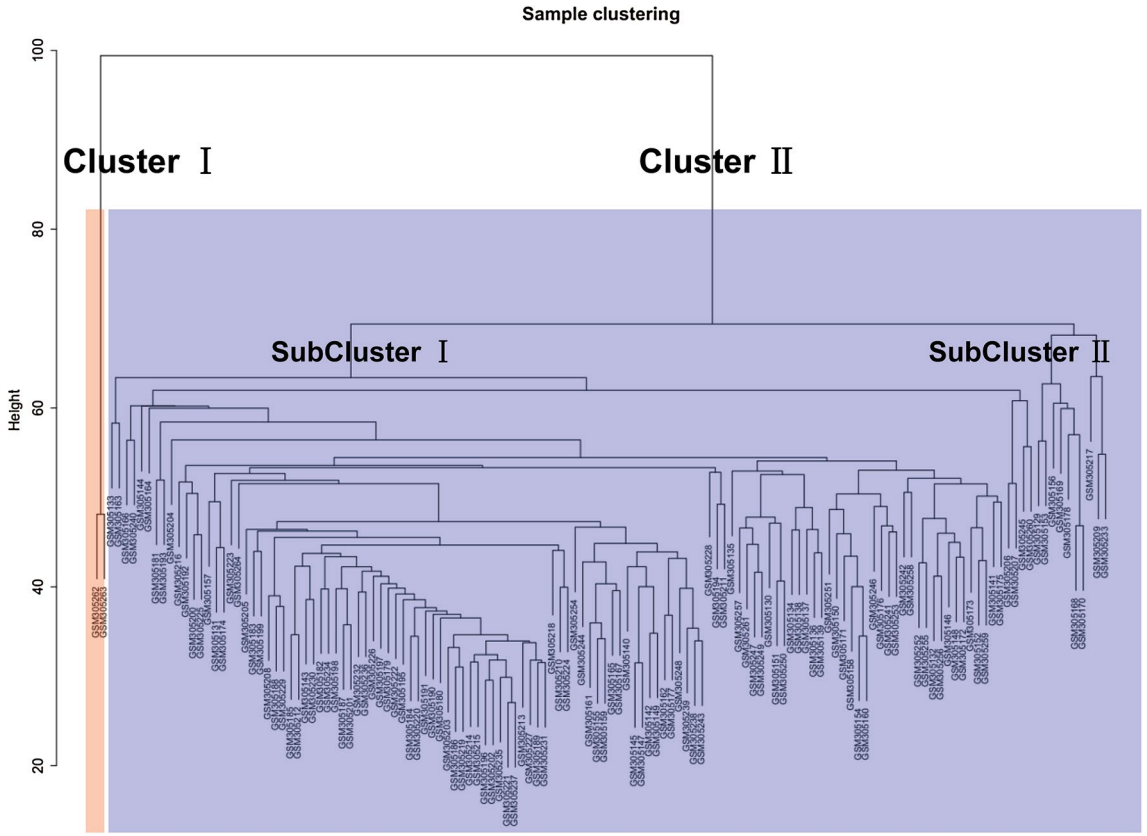

Fig. 1 Cluster analysis of breast cancer samples. The top 4000 genes with the highest average expression values were used for the analysis by WGCNA and flashClust. All samples were divided into two clusters on the whole, cluster I (pale red) and cluster II (pale blue), including two samples and 134 samples, respectively. Two sub-clusters were identified in cluster II. There were 124 samples in sub-cluster I and ten samples in sub-cluster II (color figure online)

(module 2, 4 and 6). Furthermore, there was an obvious difference in the effect of connectivity degree of different modules. Three pairs of module combination had much higher adjacency degree besides the highest self-comparison and adjacency degree. The three pairs had much stronger effects, and they are module 2 and module 6, module 3 and module 5, module 7 and module 8.

Functional enrichment analysis of critical modules

GO and KEGG enrichment analysis was performed on the genes in the constructed nine modules. We found that there was much difference in the enriched functions among different modules by the result of biology process analysis. The enriched GO terms in module 1 were mainly about the cell division and adherence and DNA repairing, including GO:0098609 (cell–cell adhesion), GO:0051301 (cell division) and GO:0006260 (DNA replication). The GO terms in module 2 were mainly enriched in the splicing and regulation of mRNA, mainly including GO:0000398 (mRNA splicing, via spliceosome) and GO:0043488 (regulation of mRNA stability). Genes in module 3 were similar to that in module 2, mainly enriched in the splicing process of mRNA, mainly including GO:0000398 (mRNA splicing, via spliceosome) and GO:0008380 (RNA splicing). Genes in module 4 were significantly enriched in rRNA processing and translation inhibition, mainly including GO:0006364 (rRNA processing) and GO:0006413 (translational initiation). Genes in module 5 were mainly enriched in the process of the mitochondrion, which was associated with energy supplying, mainly including GO:0006120 (mitochondrial electron transport, NADH to ubiquinone). Module 6 and module 7 were similar to module 1, mainly enriched in GO:0098609 (cell–cell adhesion). Module 8 was mainly enriched in immune/defend reactions, including GO:0006955 (immune response), GO:0006954 (inflammatory response) and GO:0051607 (defense response to virus). In module 9, genes were mainly enriched in the process of protein ubiquitination and instability, mainly including GO:0031648 (protein destabilization) and GO:0016925 (protein sumoylation). The result of KEGG enrichment analysis of genes in the nine constructed modules was shown in Fig. 5. The result showed that there were significant enriched metabolic pathways in each module and the enriched degree of metabolic pathways

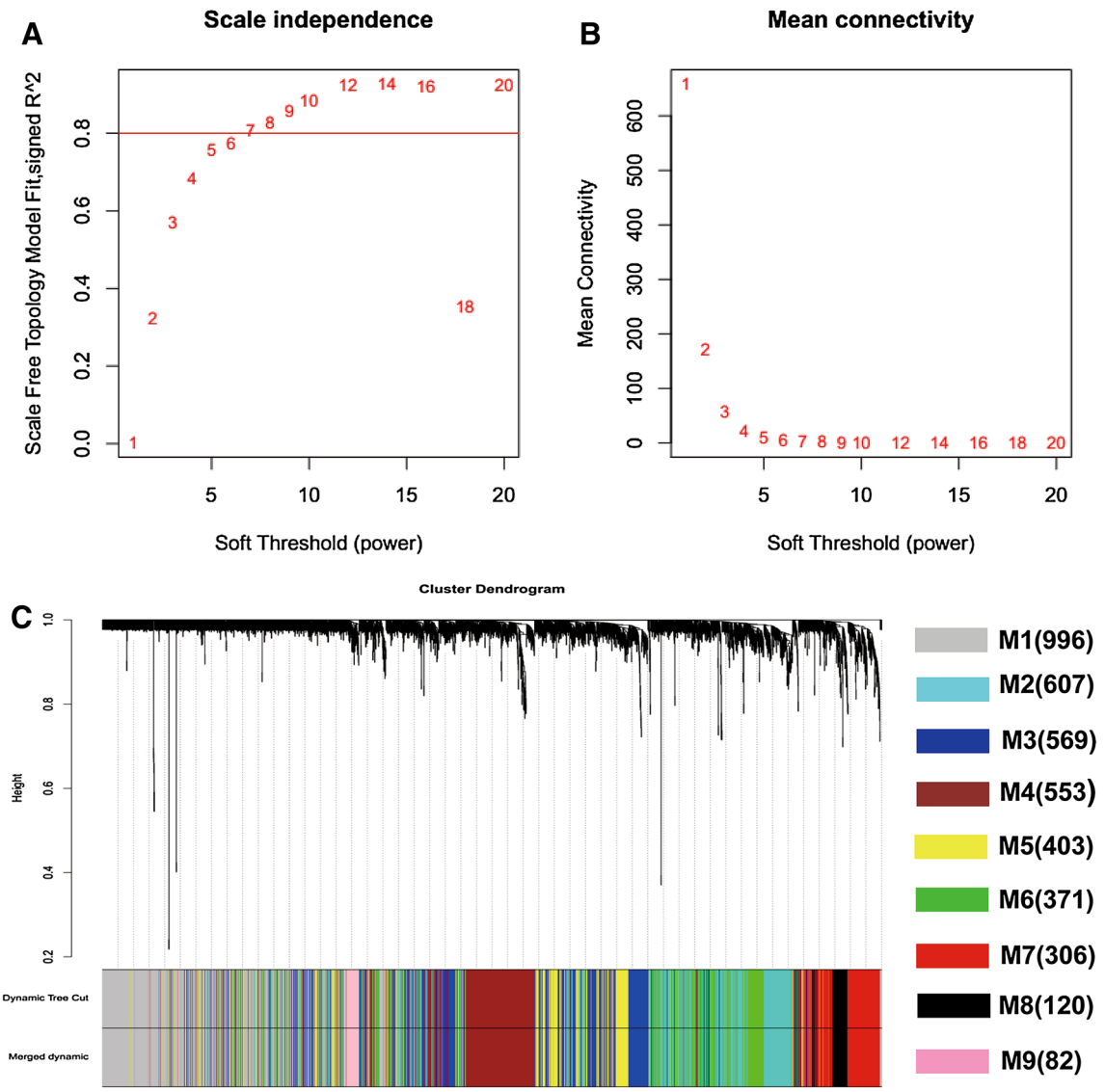

Fig. 2 Construction of co-expression modules of breast cancer-related genes. **a** The effect of different power values on the scale independence of co-expression modules of breast cancer genes. **b** The effect of different power values on the average connectivity degree of co-expression modules of breast cancer genes. **c** The construction of co-expression modules by WGCNA software. Each branch in the figure represented one gene and every color below represented one co-expression module. The icon M on the right stands for the module and number in the brackets represented the number of genes in this module (color figure online)

was quite different. Metabolic pathways in module 8 had the highest enriched degree while module 1 was the lowest. The result of KEGG analysis was illustrated in Table 3. Genes in module 1 were mainly enriched in hsa01100 (metabolic pathways) and hsa04110 (cell cycle). Genes in module 2 were mainly enriched in pathways as hsa03040 (spliceosome) and hsa00190 (oxidative phosphorylation). Genes in module 3 were mainly enriched in pathways as splicing and antibiotic synthesis, mainly including hsa03040 (spliceosome) and hsa01130 (biosynthesis of antibiotics). Genes in module 4 were mainly enriched in hsa03010 (ribosome) and hsa03040 (spliceosome) pathways. Genes in module 5 were mainly enriched in pathways of hsa00190 (oxidative phosphorylation) while genes in module 6 was mainly enriched in pathways as hsa04141 (protein processing in endoplasmic reticulum) and hsa01130 (biosynthesis of antibiotics). Genes in module 7 were mainly enriched in hsa04512 (ECM-receptor interaction) and hsa04510 (focal adhesion) pathways. Genes in module 8 were mainly enriched in hsa04612 (antigen processing and presentation) and hsa04145 (phagosome) pathways, which are in accordance with the result of GO analysis about biological process of immune response. Genes in module 9 were mainly enriched in the biological process of immune response pathways.

Table 2 GO enrichment analysis of genes in the co-expression module

	Term	Count	%	*P* value
M1	GO:0098609 ~ cell–cell adhesion	38	3.815261	3.63E−07
	GO:0051301 ~ cell division	36	3.614458	5.44E−04
	GO:0090200 ~ positive regulation of release of cytochrome c from mitochondria	8	0.803213	6.68E−04
	GO:0006260 ~ DNA replication	20	2.008032	9.60E−04
	GO:0031145 ~ anaphase-promoting complex-dependent catabolic process	13	1.305221	0.001297
M2	GO:0002479 ~ antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	21	3.459638	9.04E−15
	GO:0000398 ~ mRNA splicing, via spliceosome	36	5.930807	2.24E−14
	GO:0038061 ~ NIK/NF-kappaB signaling	17	2.800659	4.04E−10
	GO:0006521 ~ regulation of cellular amino acid metabolic process	15	2.47117	8.53E−10
	GO:0043488 ~ regulation of mRNA stability	20	3.294893	1.39E−09
M3	GO:0016032 ~ viral process	26	4.634581	5.06E−06
	GO:0000398 ~ mRNA splicing, via spliceosome	21	3.743316	1.51E−05
	GO:0008380 ~ RNA splicing	17	3.030303	4.72E−05
	GO:0000209 ~ protein polyubiquitination	18	3.208556	4.73E−05
	GO:0045454 ~ cell redox homeostasis	11	1.960784	1.08E−04
M4	GO:0006614 ~ SRP-dependent cotranslational protein targeting to membrane	46	8.318264	5.15E−43
	GO:0000184 ~ nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	47	8.499096	1.11E−38
	GO:0006413 ~ translational initiation	47	8.499096	1.97E−35
	GO:0006364 ~ rRNA processing	56	10.12658	2.66E−35
	GO:0019083 ~ viral transcription	42	7.594937	1.94E−33
M5	GO:0006120 ~ mitochondrial electron transport, NADH to ubiquinone	10	2.48139	1.17E−06
	GO:1902600 ~ hydrogen ion transmembrane transport	10	2.48139	7.81E−06
	GO:0043161 ~ proteasome-mediated ubiquitin-dependent protein catabolic process	16	3.970223	5.30E−05
	GO:0006123 ~ mitochondrial electron transport, cytochrome c to oxygen	6	1.488834	6.36E−05
	GO:0032981 ~ mitochondrial respiratory chain complex I assembly	9	2.233251	7.61E−05
M6	GO:0098609 ~ cell–cell adhesion	20	5.390836	3.53E−06
	GO:0036498 ~ IRE1-mediated unfolded protein response	8	2.156334	1.98E−04
	GO:0006094 ~ gluconeogenesis	7	1.886792	2.63E−04
	GO:0043488 ~ regulation of mRNA stability	10	2.695418	2.71E−04
	GO:0043161 ~ proteasome-mediated ubiquitin-dependent protein catabolic process	14	3.773585	3.01E−04
M7	GO:0030198 ~ extracellular matrix organization	38	12.45902	2.91E−28
	GO:0007155 ~ cell adhesion	44	14.42623	3.34E−20
	GO:0030574 ~ collagen catabolic process	18	5.901639	3.04E−16
	GO:0002576 ~ platelet degranulation	15	4.918033	2.04E−09
	GO:0016525 ~ negative regulation of angiogenesis	12	3.934426	6.32E−09
M8	GO:0060333 ~ interferon-gamma-mediated signaling pathway	14	11.66667	9.20E−16
	GO:0051607 ~ defense response to virus	13	10.83333	9.22E−10
	GO:0006955 ~ immune response	18	15	2.33E−09
	GO:0006954 ~ inflammatory response	17	14.16667	3.83E−09
	GO:0060337 ~ type I interferon signaling pathway	9	7.5	9.19E−09
M9	GO:0006368 ~ transcription elongation from RNA polymerase II promoter	5	6.097561	5.62E−04
	GO:0015031 ~ protein transport	7	8.536585	0.007875
	GO:0031648 ~ protein destabilization	3	3.658537	0.010377
	GO:0016925 ~ protein sumoylation	4	4.878049	0.014931
	GO:0043066 ~ negative regulation of apoptotic process	7	8.536585	0.015053

The icon M in the first row represented module

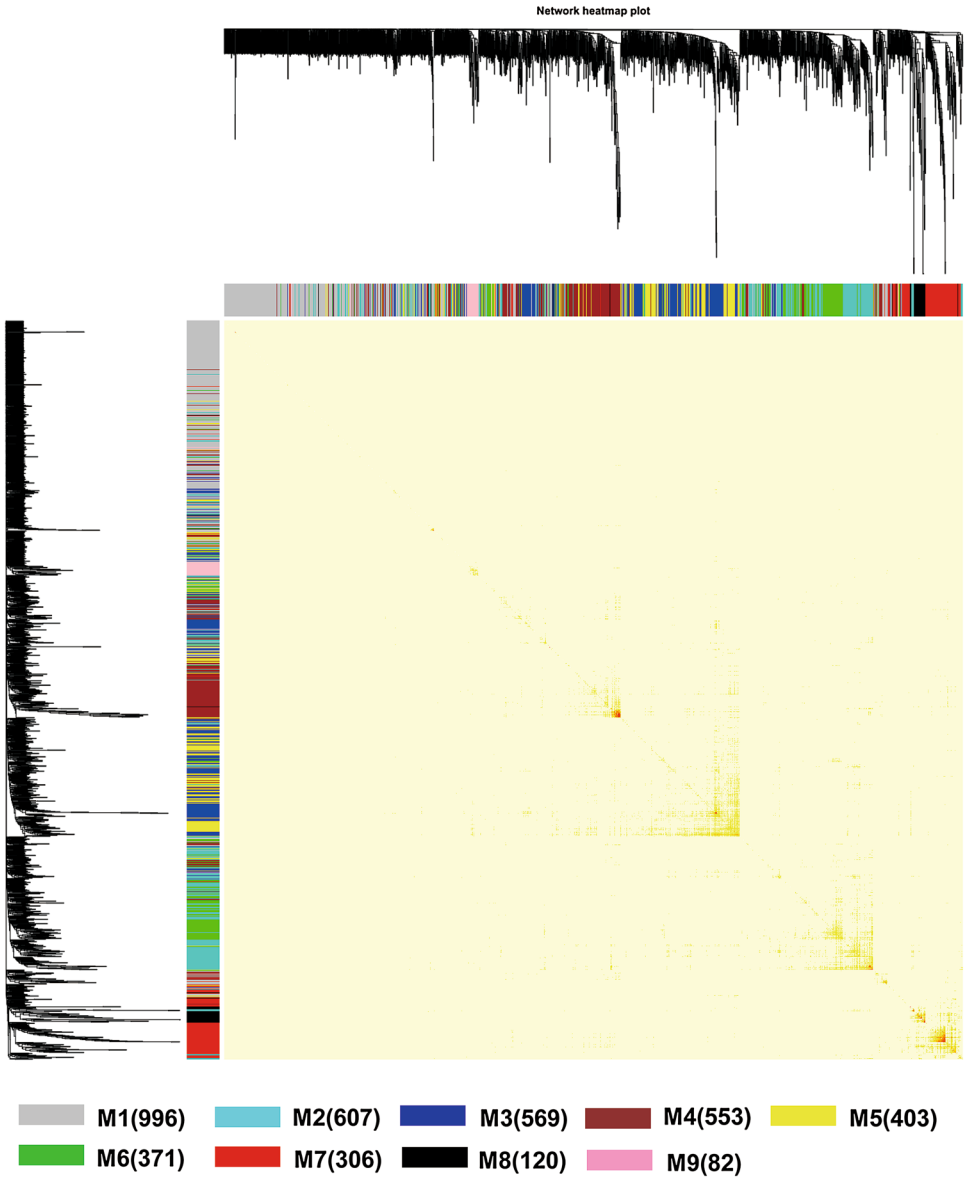

Fig. 3 Interaction relationship analysis of co-expression genes. Different colors of horizontal axis and vertical axis represented different modules. The brightness of yellow in the middle represented the connectivity degree of different modules. There was not much difference in interactions among different modules, indicating the higher scale independence degree among these modules. The icon below represented the module and the number in the brackets represented the number of genes in corresponding modules (color figure online)

Discussion

Breast cancer is the second most common tumors affecting people, especially women around the period of menopause worldwide. It is also one of the most principal causes of death of patients suffering from cancer [14]. Nowadays, there hasn't been any effective treatment for patients with breast cancer and the most effective measure to this disease was prevention [3]. What is worse, patients at the same stage of disease can have quite different treatment responses and overall outcome, which makes the situation more complicated and thus the research on prognostic or predictive markers of breast cancer became more urgent. In this study, we aimed to explore the critical biomarker for a better understanding of the molecular mechanism, which can then be applied in the diagnosis or treatment of breast cancer. In this study, co-expression patterns in breast cancer and matched normal tissues were examined by WGCNA, a powerful method used to extract co-expressed groups of genes from large expression data sets. As a result, a total of nine co-expression modules were screened out by WGCNA in the training dataset GSE12903 from NCBI dataset. Besides, the critical co-expression modules and genes they included were identified by GO and KEGG functional enrichment analysis. Early studies on breast cancer most relied on gene expression profiles, which had some disadvantages. Although genome-wide gene expression breast cancer datasets were available and offered opportunities for translational advances and personalized medicines, the challenges still existed in data analysis. For example, the result of differential expressed gene analysis cannot be in accordance with another which

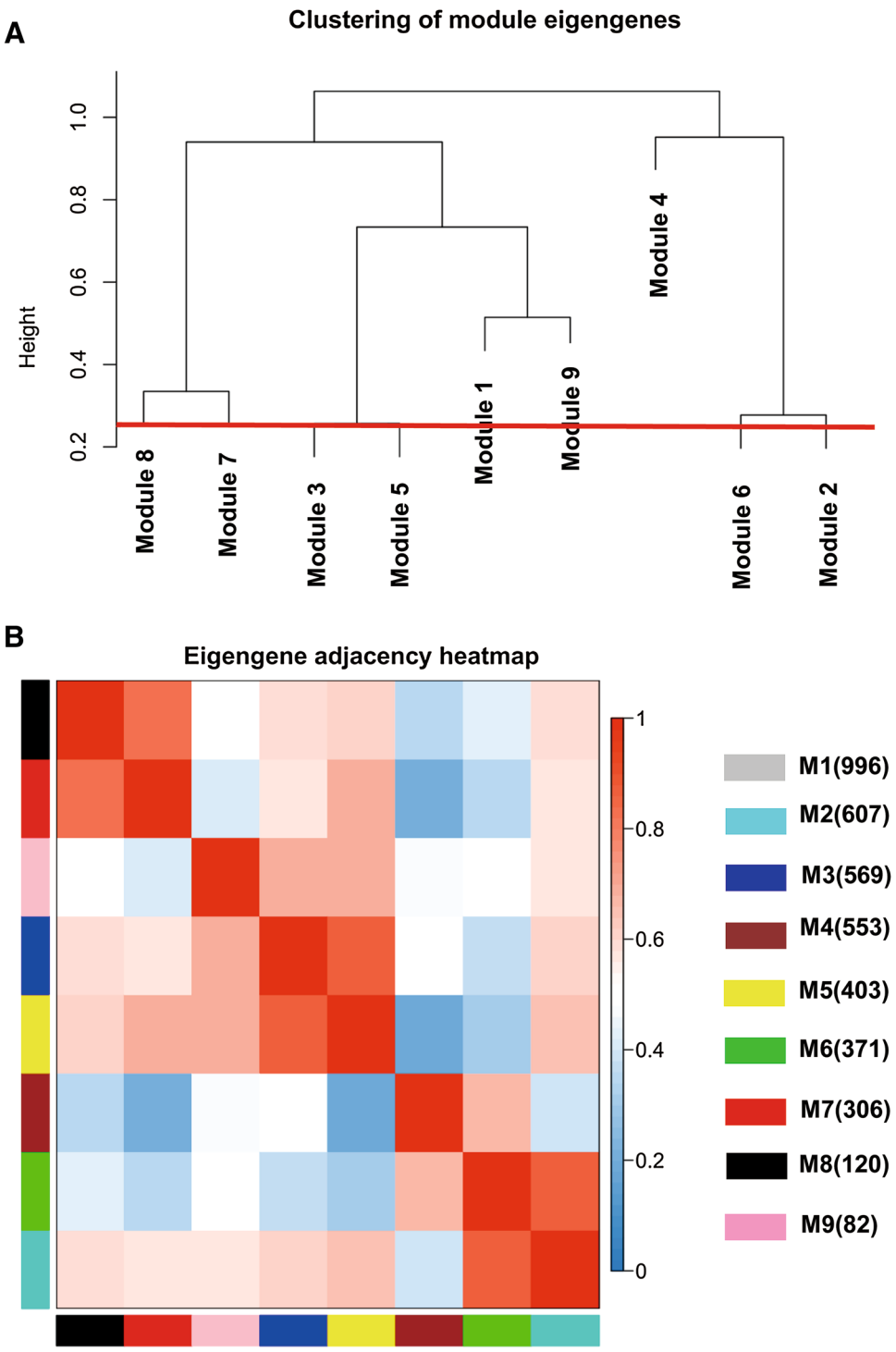

Fig. 4 The connectivity analysis of critical genes in different module. **a** Cluster analysis of critical genes in modules. Two clusters were found out, which included six samples (module 1, 3, 5, 7, 8 and 9) and three samples (module 2, 4 and 6), respectively. **b** The connectivity heatmap of critical genes in modules. The change of color from blue (0) to red (0) in the heatmap represented the connectivity degree of critical genes in different modules from weak to strong. The icon on the right represented the module and the number in the brackets represented the number of genes in this module (color figure online)

was obtained at different platforms, thus making the result unreliable.

However, WGCNA approach can well avoid this disadvantage by performing well across all types of data and focusing on a batch of gene modules rather than individual genes. Besides, it does not rely on a prior assumption about genes or covariates. Therefore, WGCNA can avoid biologically wrong assumptions about independence of gene expression levels since it can also transform gene expression profiles into functional co-expressed gene modules. Up to now, WGCNA method has been applied in many types of cancers, such as lung cancer, brain cancer, and breast cancer. In this study, we found the genes in two co-expression modules, module 8 and module 9, played an essential role

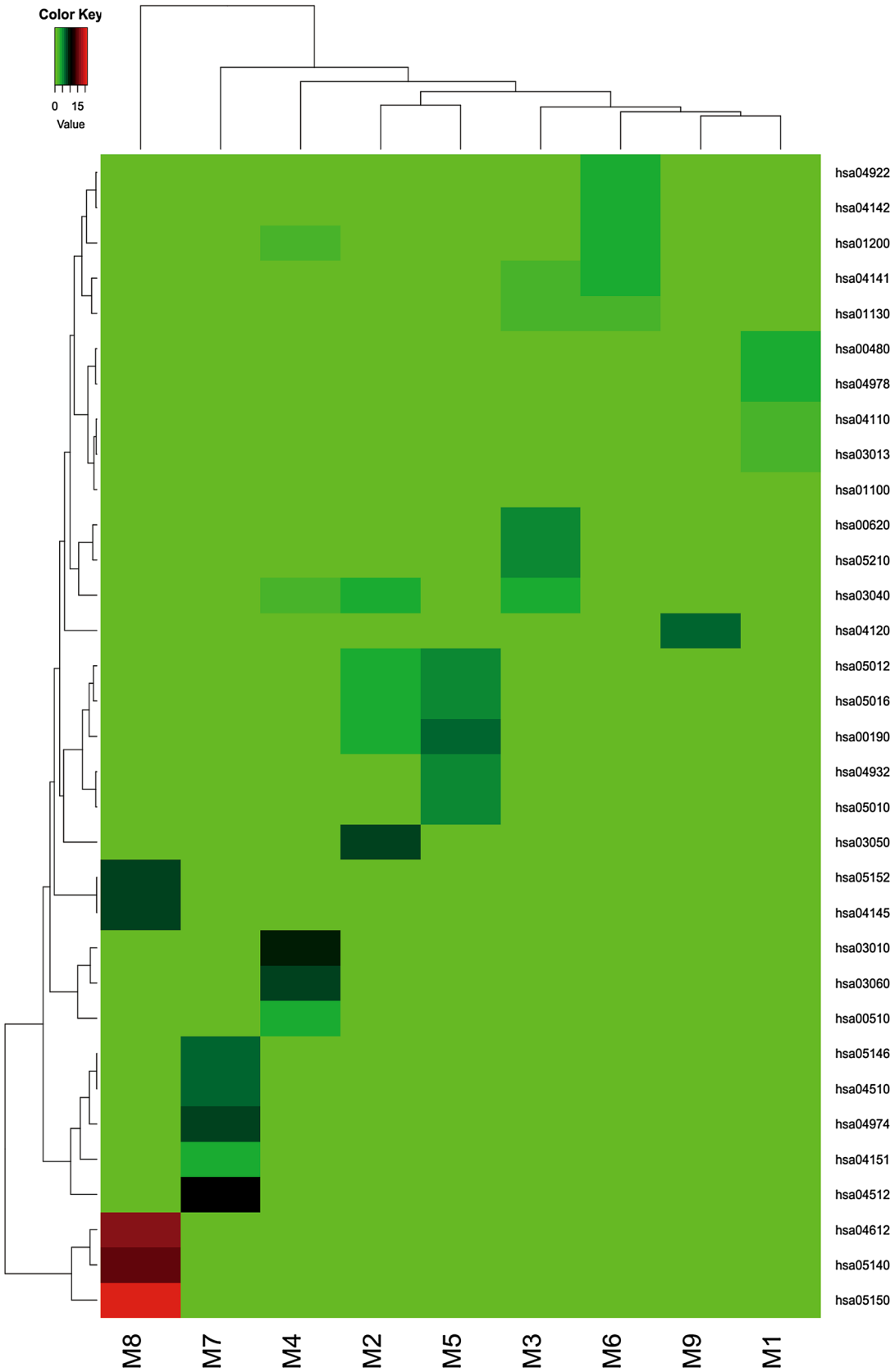

Fig. 5 KEGG enrichment heatmap of genes in the co-expression module. Words on the right represented the number of metabolic pathways of KEGG and the words below represented the constructed modules in this study. The M icon represented the module (color figure online)

Table 3 KEGG enrichment analysis of genes in the co-expression modules

	Term	Count	%	*P* value
M1	hsa00480: glutathione metabolism	11	1.104418	7.15E−04
	hsa04978: mineral absorption	9	0.903614	0.005259
	hsa01100: metabolic pathways	93	9.337349	0.008818
	hsa03013: RNA transport	19	1.907631	0.013726
	hsa04110: cell cycle	15	1.506024	0.015324
M2	hsa03050: proteasome	14	2.306425	3.18E−08
	hsa05016: Huntington's disease	27	4.448105	3.13E−07
	hsa03040: spliceosome	20	3.294893	5.81E−06
	hsa00190: oxidative phosphorylation	19	3.130148	2.19E−05
	hsa05012: Parkinson's disease	19	3.130148	5.36E−05
M3	hsa03040: spliceosome	17	3.030303	1.95E−05
	hsa01130: biosynthesis of antibiotics	21	3.743316	6.70E−05
	hsa05210: colorectal cancer	10	1.782531	3.28E−04
	hsa00620: pyruvate metabolism	8	1.426025	4.76E−04
	hsa04141: protein processing in endoplasmic reticulum	16	2.85205	0.001043
M4	hsa03010: ribosome	44	7.9566	6.36E−30
	hsa03060: protein export	6	1.084991	0.001152
	hsa03040: spliceosome	13	2.350814	0.002943
	hsa01200: carbon metabolism	11	1.98915	0.007356
	hsa00510: *N*-glycan biosynthesis	7	1.265823	0.007863
M5	hsa00190: oxidative phosphorylation	26	6.451613	5.28E−14
	hsa05012: Parkinson's disease	23	5.707196	1.05E−10
	hsa05016: Huntington's disease	26	6.451613	2.46E−10
	hsa04932: non-alcoholic fatty liver disease (NAFLD)	22	5.459057	2.21E−09
	hsa05010: Alzheimer's disease	23	5.707196	2.86E−09
M6	hsa04141: protein processing in endoplasmic reticulum	16	4.312668	1.25E−05
	hsa01130: biosynthesis of antibiotics	13	3.504043	0.005116
	hsa01200: carbon metabolism	9	2.425876	0.005858
	hsa04142: lysosome	9	2.425876	0.00876
	hsa04922: glucagon signaling pathway	8	2.156334	0.009818
M7	hsa04512: ECM-receptor interaction	18	5.901639	3.39E−13
	hsa04510: focal adhesion	25	8.196721	5.24E−13
	hsa04151: PI3K-Akt signaling pathway	26	8.52459	5.35E−09
	hsa04974: protein digestion and absorption	14	4.590164	8.84E−09
	hsa05146: amoebiasis	13	4.262295	6.83E−07
M8	hsa05150: *Staphylococcus aureus* infection	11	9.166667	4.58E−11
	hsa04612: antigen processing and presentation	12	10	7.52E−11
	hsa05152: tuberculosis	14	11.66667	6.77E−09
	hsa05140: leishmaniasis	10	8.333333	1.53E−08
	hsa04145: phagosome	12	10	1.40E−07
M9	hsa04120: ubiquitin-mediated proteolysis	4	4.878049	0.022857

The icon M in the first row represented module

in immune response and ubiquitin-mediated proteolysis process, and these two modules were recognized as the most important modules in the occurrence of breast cancer. GO analysis showed that genes in module 8 were mainly involved in pathways in response to the immune system, inflammatory, and defense. Similarly, we found that genes in module 9 played important roles in response to protein syntheses, such as ubiquitin-mediated proteolysis, protein destabilization, and protein sumoylation processes. Furthermore, KEGG analysis revealed that module 8 was mainly enriched in hsa01130 (Biosynthesis of antibiotics) and hsa00190 (Oxidative phosphorylation) pathways. Most co-expression modules were in close association with immune reaction and ubiquitin-mediated proteolysis process, and

these two pathways were regarded as potential biomarkers in the mechanism study of breast cancer. The enrich pathway of hsa04120 (ubiquitin-mediated proteolysis) was recognized as the most critical prognostic marker in the occurrence of breast cancer. Combined with the result of other two enriched pathways, that is, hsa01130 (biosynthesis of antibiotics) and hsa00190 (oxidative phosphorylation), enriched by more than one co-expression module, which were also in close association with the process of ubiquitin-mediated proteolysis, we have reason to believe these enriched pathways can function as biomarkers in the diagnosis of breast cancer. It is reported that cell proliferation correlate with relapse rate in pre- and postmenopausal women with breast cancer [15], and women around this period experienced changes in hormone levels in vivo. The ubiquitin-mediated proteolysis was in close association with the protein syntheses required for the cell proliferation and hormone synthesis. For example, estrogen and progesterone, two main hormones in menopause period, were largely affected in women with breast cancer [16, 17], combined with their main component of protein, the profound meaning of critical biomarker of ubiquitin-mediated proteolysis pathway was more certain to believe, which required further investigations.

In summary, our study used systems biology-based WGCNA approach to construct co-expression modules, which played a critical role in breast cancer. Ubiquitin-mediated proteolysis pathway, significantly enriched in module 8 and module 9, could function as the prognostic and predictive marker in the clinical management of breast cancer.

Compliance with ethical standards

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

References

1. Berg JW, Robbins G. Factors influencing short and long term survival of breast cancer patients. Surg Gynecol Obstet. 1966;122:1311.
2. Adair F, Berg J, Joubert L, Robbins GF. Long-term followup of breast cancer patients: the 30-year report. Cancer. 1974;33:1145–50.
3. Saez RA, McGuire WL, Clark GM. Prognostic factors in breast cancer. Semin Surg Oncol. 1989;5:102–10.
4. Bloom H, Richardson W. Histological grading and prognosis in breast cancer: a study of 1409 cases of which 359 have been followed for 15 years. Br J Cancer. 1957;11:359.
5. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 2008;9:559.
6. Ivliev AE, AC't Hoen P, Sergeeva MG. Coexpression network analysis identifies transcriptional modules related to proastrocytic differentiation and sprouty signaling in glioma. Cancer Res. 2010;70:10060–70.
7. Clarke C, Madden SF, Doolan P, Aherne ST, Joyce H, O'Driscoll L, et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. Carcinogenesis. 2013;34:2300–8.
8. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
9. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5:299–314.
10. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4:R60.
11. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
12. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277–80.
13. Zhang Y, Sieuwerts AM, McGreevy M, Casey G, Cufer T, Paradiso A, et al. The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. Breast Cancer Res Treat. 2009;116:303–9.
14. Fisher B, Bauer M, Wickerham DL, Redmond CK, Fisher ER, Cruz AB, et al. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update. Cancer. 1983;52:1551–7.
15. Isola J, Visakorpi T, Holli K, Kallioniemi O-P. Association of overexpression of tumor suppressor protein p53 with rapid cell proliferation and poor prognosis in node-negative breast cancer patients. J Natl Cancer Inst. 1992;84:1109–14.
16. Foekens JA, Portengen H, Van Putten WL, Peters HA, Krijnen HL, Alexieva-Figusch J, et al. Prognostic value of estrogen and progesterone receptors measured by enzyme immunoassays in human breast tumor cytosols. Cancer Res. 1989;49:5823–8.
17. Berger U, Wilson P, Thethi S, McClelland RA, Greene GL, Coombes RC. Comparison of an immunocytochemical assay for progesterone receptor with a biochemical method of measurement and immunocytochemical examination of the relationship between progesterone and estrogen receptors. Cancer Res. 1989;49:5176–9.