Co-expression Network Analysis by WGCNA and Identify Potential Prognostic Markers Associated with Lung Metastasis in Breast Cancer

xixun zhang (13411964082@163.com)

The first Affiliated Hospital of Shantou University Medical College, Shantou city, China

Primary research

Keywords: Breast cancer, Lung metastasis, WGCNA, Prognostic markers

Posted Date: February 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-182567/v1

Abstract

Backgroud: Breast cancer (BC) is an aggressive cancer with a high percentage recurrence and metastasis. As one of the most common distant metastasis organ in breast cancer, lung metastasis has a worse prognosis than that of liver and bone. Therefore, it's important to explore some potential prognostic markers associated with the lung metastasis in breast cancer for preventive treatment.

Methods: In our study, transcriptomic data and clinical information of breast cancer patients were downloaded from The Cancer Genome Atlas (TCGA) database. Co-expression modules was built by Weighted gene co-expression network analysis (WGCNA) to find out the royalbule modules which is significantly associated with lung metastasis in breast cancer. Then, co-expression genes were analyzed for functional enrichment. Furthermore, the prognostic value of these genes was assessed by GEPIA Database and Kaplan-Meier Plotter.

Results: Results showed that the hub genes, LMNB and CDC20, were up-regulated in breast cancer and indicated worse survival. Therefore, we speculate that these two genes play crucial roles in the process of lung metastasis in breast cancer, and can be used as potential prognostic markers in lung metastasis of breast cancer.

Conclusion: Collectively, our study identified two potential key genes in the lung metastasis of breast cancer, which might be applied as the prognostic markers of the precise treatment in breast cancer with lung metastasis.

Introduction

Breast cancer (BC) is the most common malignancy in the female population, accounting for about 30% of all female cancers [1]. Though the diagnosis and treatment methods have developed rapidly in recent decades,the mortality of BC still remains high due to the frequent distal metastasis [2–4]. The metastases at distant sites are the main cause of breast cancer death, therefore, the presence and location of breast cancer metastasis has been the critical diagnose to the clinical course and prognosis of patients [5, 6]. Nevertheless, the prognosis of distant metastasis is significantly affected by the site of initial spread [7]. Among the priority sites of breast cancer transmission, metastasis to the lung has a worse prognosis than those to the liver or bone [8]. This is mainly due to little or no symptoms of early lung metastasis, and large metastases and severe symptoms have been caused when found the breast cancer lung metastases [9]. However, to date, there is no effective treatment for different metastatic sites, especially for the lung metastasis.

Different primary tumors have a proclivity to metastasize to distinct organs. The "seed" and "soil" theory has put forward that the distant metastasis of tumor was not accidental, but a certain biocompatibility existed between tumor cells and target organs, and tumor metastasis was affected by the effect of driver genes and target organ microenvironment [10]. However, identifying the specific driver genes is still challenging.

WGCNA is a systematic biological method which can analyze multiple gene expression patterns in multiple samples [11]. Through constructing gene co-expression modules according to their expression patterns, WGCNA can analyze the relationship between modules and specific phenotypes, for example the clinical symptoms of patients. Comparing with traditional molecular biology, the unique advantage of WGCNA is that it can convert gene expression data into co-expression modules, and provide insight into the signaling networks responsible for phenotypic traits of interest. It has been successfully used to study various biological processes, such as gastric cancer [12], colon adenocarcinoma [13], liver hepatocellular carcinoma [14] and glioma [15], proving its effective to identify the potential biomarkers and therapeutic targets. It not only helps to compare the expression of different genes, but also helps to calculate and analyze the interactions between genes in different co-expression modules. WGCNA deliver a more comprehensive exploration of the whole biological system in diseases and will be quite helpful to identify the candidate biomarkers or therapeutic targets.

Here we used WGCNA to explore the relationship between gene expression and metastatic site of breast cancer. We find out the meaningful module correlated with lung metastasis of breast cancer. GO enrichment and KEGG pathway analysis were performed to figure out the main functions of genes in the main module. Cox regression analysis was performed to pick out the hub genes of the main module. GEPIA and Kaplan-Meier Plotter were used to identify the expression value and prognostic value of these genes. We believe that our work will lay an important foundation for the future treatment of lung metastasis of breast cancer and improve the overall survival rate of breast cancer.

Material And Methods

Patient samples data collection and processing

Public gene-expression data and clinical annotation were downloaded from the cBioportal online database (http://www.cbioportal.org/).The data we selected is the Metastatic Breast Cancer Project. The samples we selected were met the following criteria: 1) Include the sample with the character like bone metastasis, brain metastasis, live metastasis, lung metastasis and ovary metastasis. 2) Exclude the sample with missing data. And then we choose the top 30% most variable genes for our study. In the end, a total of 80 patient samples and 10083genes were included in our study.

Weighted gene co-expression network construction

In the present study, the soft-threshold β was set as 7. Subsequently, the adjacency matrix was transformed into a topological overlap matrix (TOM). Next, we performed hierarchical clustering to identify modules, each module included at least 20 genes (min Module Size = 20). Finally, we calculated the eigen gene, hierarchically clustered the modules, and merged similar modules.

Clinically significant modules identification

The co-expression module is defined as a class of genes with high topological overlap similarity, and genes in the same module generally have a higher degree of co-expression. In this study, two methods were used to identify the important modules associated with clinical traits. First, the module eigengene (ME) represents the principal component of the module to describe the expression pattern of the module in each sample. Second, module membership (MM) refers to the correlation coefficient between genes and module eigengenes to describe the reliability of a gene belonging to a module. Finally, the correlation was calculated between the modules and the clinical data to identify significantly clinical modules.

Gene Ontology Enrichment and KEGG Pathway Analysis

WebGestalt (http://www.webgestalt.org/) is a functional enrichment analysis web tool for users to comprehend biological function information of genes and proteins, which supports three well-established and complementary methods for enrichment analysis. We used WebGestalt online tools to perform the GO enrichment and KEGG pathway analysis of the genes in royalbule module. “adjusted $P < 0.05$” was used as the threshold value to identify the significant terms.

Multivariate Cox regression

Multivariate Cox regression was performed using SPSS software. 39 genes were selected for screening the optimal prognostic signatures for breast cancer with lung metastasis. All data was evaluated by the Pearson’s Chi-Square method with SSPS software.

GEPIA Database

GEPIA (http://gepia.cancer-pku.cn/index.html) is an online database which facilitates the standardized analysis of RNA-seq data from 9,736 tumor samples and 8,587 normal control samples in the TCGA and GTEx data sets. In our study, we used this database to analyze the transcription levels of hub genes in breast cancer sample and normal sample. The P value was cut off at 0.01.

Kaplan-Meier Plotter Analysis

In our study, we used Kaplan-Meier plotter (http://kmplot.com/analysis/), which is an online database to explore the impact of genes on patient survival in different types of cancer, to verify the prognostic value of hub genes in breast cancer patients.

bc-GenExMiner v4.0

Breast Cancer Gene-Expression Miner v4.0 (http://bcgenex.centregauducheau.fr/BC-GEM/GEM-Accueil.php?js=1) is an online dataset containing published annotated genomic data, which includes 36 annotated genomic datasets and 5861 patients with breast cancer. Based on these, it can be used as a statistical mining tool to estimate the Pearson's correlation module.

Results

Construction of co-expression modules of breast cancer.

The data was processed and analyzed following the flowchart (Fig. 1). Expression values of top 30% most variable genes (10083 genes) of breast cancer were used to construct the co-expression module by WGCNA package tool. The FlashClust tools package was used to perform the cluster analysis on these samples and the result was shown in (Fig. 2A). The power value, which mainly affected the independence and the average connectivity degree of co-expression modules, was screened out and equal to seven (Fig. 2B) and the independence degree was up to 0.8. Therefore, the power value was used to build gene co-expression module and the results showed that there were 39 modules in breast cancer. These co-expression modules were displayed in different colours (Fig. 3). These modules were ranged from large to small by the number of genes they included. The number of genes in the 39 modules was shown in (Table 1). Interactions of the 39 co-expression modules were analyzed and shown in (Fig. 4).

Table 1
The number of genes in the 39 module was counted as follows:

Module	Count
blue	914
brown	801
brown4	43
cyan	3334
darkgrey	116
darkolivegreen	84
darkorange	167
darkorange2	46
darkred	308
darkslateblue	37
floralwhite	48
green	599
grey	54
grey60	261
honeydew1	20
ivory	913
lavenderblush3	23
lightcyan	203
lightgreen	185
lightpink4	24
lightsteelblue1	61
maroon	25
mediumpurple3	62
navajowhite2	25
paleturquoise	86
palevioletred3	25

Module	Count
pink	280
plum2	31
purple	241
red	286
royalblue	140
saddlebrown	92
sienna3	69
skyblue	102
skyblue3	64
steelblue	91
thistle1	26
thistle2	28
violet	86

Gene Co-expression Modules Correspond to Clinic Traits.

The clinical information was provided by TCGA database. And we selected the metastasis organs as research targets and removed the patient sample information that was meaningless or lacking in our study. According to the correlation between module eigengene and clinic traits, the interaction of co-expression modules and particular traits were identified (Fig. 5). We found that the royalblue module (R = 0.34, p = 0.04) were significantly associated with breast cancer with lung metastasis. The eigengene dendrogram and heatmap were used to identify groups of correlated eigengenes (Fig. 6).

Functional enrichment analysis of genes in royalblue modules.

Go enrichment analysis and KEGG analysis were performed on the genes in the royal blue module. The biological process of most genes was metabolic process, biological regulation and cellular component organization. The cellular component of most genes was nucleus, and the molecular function of most genes was protein binding. According to KEGG analysis, genes in royal blue module were mainly enriched in systemic lupus erythematosus, reproduction, alcoholism, RHO GTPase effectors, viral carcinogenesis, signaling by Rho GTPase, cell cycle, developmental biology, generic transcription pathway.

Module visualize and hub genes.

The genes in royalblue module were calculated the intramodular connectivity. The intramodular connectivity was calculated for each gene by summing the connection strengths with other module genes and dividing this number by the maximum intramodular connectivity. The number of genes with connectivity greater than 0.1 was 39. These genes were selected as hub genes and then analyzed using the Cytoscape software (Fig. 9). Next, we integrated the expression profiles of 39 genes in the module with the occurrence of pulmonary metastasis in the corresponding 80 patients and conducted cox regression analysis. Finally, KRTAP4-1, LMNB1 and CDC20 were independent risk factors for breast cancer with lung metastasis. The relative risk between these three genes and lung metastasis of breast cancer were 1.146, 1.269 and 0.885, respectively (Table 2). The omnibus test of model coefficients showed that the overall test of the model was statistically significant (Table 3). And then we found LMNB1 and CDC20 were up-regulated in breast cancer tissue through the GEPIA database. and we also found that the over-expression of this two genes predict worse survival of breast cancer using the Kaplan-Meier plotter website (Fig. 10). We further verified the correlation between LMNB1 and CDC20 using the bc-GenExMiner v4.0, the result showed that the correlation coefficientwas up to 0.9.

Table 2
The multivariate cox regression analysis between markers and lung metastasis status.

	B	SE	Wald	df	p-value	Exp(B)	95%CI
KRTAP41	0.136	0.047	8.309	1	0.004	1.146	1.045–1.257
CDC20	0.238	0.074	10.430	1	0.001	1.269	1.098–1.466
LMNB1	-0.122	0.045	7.519	1	0.006	0.885	0.811–0.966

Table 3
Omnibus test of model coefficients.

-2 Logarithmic probable value	Overall (score)			Changes from a previous block		
	Chi-square	df	p-value	Chi-square	df	p-value
44.955	39.244	3	.000	28.429	3	.000
Method = Step by step forward(Logistic regression)						

Discussion

Lung metastasis is a pernicious outcome of breast cancer. About 17% of patients with BC have a propensity developing to lung metastases [16]. Based on the selective evolution of organs, metastatic localization does not occur randomly, but is a preferred location controlled by numerous micro-environmental, cellular, and molecular factors [9]. Understanding the mechanism of lung metastasis of

breast cancer is helpful for further prevention and even targeted therapy. There are many theories to explain the process of lung metastasis in breast cancer: for example, the interaction between CSCs forming breast cancer cells and pulmonary vessels [17–19],the histologic and intrinsic genomic profiles of breast cancer [20]barrier drive of host organs [21]. Individual genes are associated with organ-specific metastasis. Multiple individual genes constitute a complex, dynamic and interactive network and govern the process of breast cancer cell metastasis to specific organs [4, 22]. To date, no studies have correlated gene expression with clinical traits and used it to predict lung metastasis of breast cancer. The advantage of our study is that we found out a predictive model that is beneficial to clinical decision making. Our result indicated that CDC20 and LMNB1 were co-expressed in breast cancer and played important roles in the process of lung metastasis in breast cancer, suggesting their important value for further study.

LMNB1 is a nuclear membrane protein that can build a framework for the nuclear envelope [23]. LMNB1 is also essential for cell senescence [24]. For example, down-regulation of LMNB1 induces cellular senescence through activating either the p53 or pRB tumor suppressor pathway [25]. Increasing evidence shows that up-regulation or down-regulation of LMNB1 affects the clinical behavior of cancer. Furthermore, previous study also showed LMNB1 is associated with lung metastasis in breast cancer in mice [26]. Therefore, we conclude that LMNB1 plays an important role in the process of lung metastasis in breast cancer, whereas further confirming evidence of this result in human breast cancer is still need.

CDC20 is a spindle assembly checkpoint molecule, which is critical in cell cycle progression [27]. Aberrant expression of CDC20 is associated with malignant progression and poor prognosis in various types of cancer [28–30]. In addition, there is a significant correlation between a high expression of CDC20 and advanced tumor stage in carcinoma [31]. To the best of our knowledge, our study is the first time to verify that CDC20 is related with the progress of lung metastasis in breast cancer, indicating that CDC20 can be a prognostic marker or a potential therapy target for breast cancer with lung metastasis.

In conclusion, our study is the first study to screen the characteristic hub genes and construct a prognostic model based on hub genes in breast cancer with lung metastasis using WGCNA. The potential prognostic predictive genes, LMNB1 and CDC20, have been proved to be able to investigate the occurrence of lung metastasis in breast cancer. Therefore, these two genes might be used as the potential biomarkers to identify the high-risk patients and assess the prognosis to facilitate the precise treatment in breast cancer with lung metastasis.

Declarations

Acknowledgments

This study was supported by the Youth Program of National Natural Science Foundation of China (81802956).

Declaration of Interest Statement

The authors declare that they have no conflicts of interest with the contents of this article.

References

1. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2020.* CA Cancer J Clin, 2020. **70**(1): p. 7-30.
2. Gupta, G.P. and J. Massague, *Cancer metastasis: building a framework.* Cell, 2006. **127**(4): p. 679-95.
3. Rabbani, S.A. and A.P. Mazar, *Evaluating distant metastases in breast cancer: from biology to outcomes.* Cancer Metastasis Rev, 2007. **26**(3-4): p. 663-74.
4. Valastyan, S. and R.A. Weinberg, *Tumor metastasis: molecular insights and evolving paradigms.* Cell, 2011. **147**(2): p. 275-92.
5. Weigelt, B., J.L. Peterse, and L.J. van 't Veer, *Breast cancer metastasis: markers and models.* Nat Rev Cancer, 2005. **5**(8): p. 591-602.
6. Solomayer, E.F., et al., *Metastatic breast cancer: clinical course, prognosis and therapy related to the first site of metastasis.* Breast Cancer Res Treat, 2000. **59**(3): p. 271-8.
7. Leone, B.A., et al., *Stage IV breast cancer: clinical course and survival of patients with osseous versus extraosseous metastases at initial diagnosis. The GOCS (Grupo Oncologico Cooperativo del Sur) experience.* Am J Clin Oncol, 1988. **11**(6): p. 618-22.
8. Lee, Y.T., *Breast carcinoma: pattern of metastasis at autopsy.* J Surg Oncol, 1983. **23**(3): p. 175-80.
9. Yousefı, M., et al., *Organ-specific metastasis of breast cancer: molecular and cellular mechanisms underlying lung metastasis.* Cell Oncol (Dordr), 2018. **41**(2): p. 123-140.
10. Fidler, I.J., *The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited.* Nat Rev Cancer, 2003. **3**(6): p. 453-8.
11. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
12. Zhou, Q., et al., *Identification of subtype-specific genes signature by WGCNA for prognostic prediction in diffuse type gastric cancer.* Aging (Albany NY), 2020. **12**(17): p. 17418-17435.
13. Wang, H., et al., *Identification of gene modules and hub genes in colon adenocarcinoma associated with pathological stage based on WGCNA analysis.* Cancer Genet, 2020. **242**: p. 1-7.
14. Bai, K.H., et al., *Identification of cancer stem cell characteristics in liver hepatocellular carcinoma by WGCNA analysis of transcriptome stemness index.* Cancer Med, 2020. **9**(12): p. 4290-4298.
15. Hsu, J.B., et al., *Identification of potential biomarkers related to glioma survival by gene expression profile analysis.* BMC Med Genomics, 2019. **11**(Suppl 7): p. 34.
16. Lu, X. and Y. Kang, *Organotropism of breast cancer metastasis.* J Mammary Gland Biol Neoplasia, 2007. **12**(2-3): p. 153-62.
17. Wicha, M.S., S. Liu, and G. Dontu, *Cancer stem cells: an old idea--a paradigm shift.* Cancer Res, 2006. **66**(4): p. 1883-90; discussion 1895-6.

18. Croker, A.K., et al., *High aldehyde dehydrogenase and expression of cancer stem cell markers selects for breast cancer cells with enhanced malignant and metastatic ability.* J Cell Mol Med, 2009. **13**(8B): p. 2236-52.
19. Sheridan, C., et al., *CD44+/CD24- breast cancer cells exhibit enhanced invasive properties: an early step necessary for metastasis.* Breast Cancer Res, 2006. **8**(5): p. R59.
20. Schito, L. and G.L. Semenza, *Hypoxia-Inducible Factors: Master Regulators of Cancer Progression.* Trends Cancer, 2016. **2**(12): p. 758-770.
21. Chen, W., et al., *Organotropism: new insights into molecular mechanisms of breast cancer metastasis.* NPJ Precis Oncol, 2018. **2**(1): p. 4.
22. Nevins, J.R. and A. Potti, *Mining gene expression profiles: expression signatures as cancer phenotypes.* Nat Rev Genet, 2007. **8**(8): p. 601-9.
23. Hutchison, C.J., *B-type lamins in health and disease.* Semin Cell Dev Biol, 2014. **29**: p. 158-63.
24. Tsai, M.Y., et al., *A mitotic lamin B matrix induced by RanGTP required for spindle assembly.* Science, 2006. **311**(5769): p. 1887-93.
25. Freund, A., et al., *Lamin B1 loss is a senescence-associated biomarker.* Mol Biol Cell, 2012. **23**(11): p. 2066-75.
26. Kurpinska, A., et al., *Proteomic characterization of early lung response to breast cancer metastasis in mice.* Exp Mol Pathol, 2019. **107**: p. 129-140.
27. Weinstein, J., *Cell cycle-regulated expression, phosphorylation, and degradation of p55Cdc. A mammalian homolog of CDC20/Fizzy/slp1.* J Biol Chem, 1997. **272**(45): p. 28501-11.
28. Ding, Y., et al., *CDC20 with malignant progression and poor prognosis of astrocytoma revealed by analysis on gene expression.* J Neurooncol, 2017. **133**(1): p. 87-95.
29. Ding, Z.Y., et al., *Expression characteristics of CDC20 in gastric cancer and its correlation with poor prognosis.* Int J Clin Exp Pathol, 2014. **7**(2): p. 722-7.
30. Yan, H., et al., *Aberrant expression of cell cycle and material metabolism related genes contributes to hepatocellular carcinoma occurrence.* Pathol Res Pract, 2017. **213**(4): p. 316-321.
31. Paul, D., et al., *Cdc20 directs proteasome-mediated degradation of the tumor suppressor SMAR1 in higher grades of cancer through the anaphase promoting complex.* Cell Death Dis, 2017. **8**(6): p. e2882.

Figures

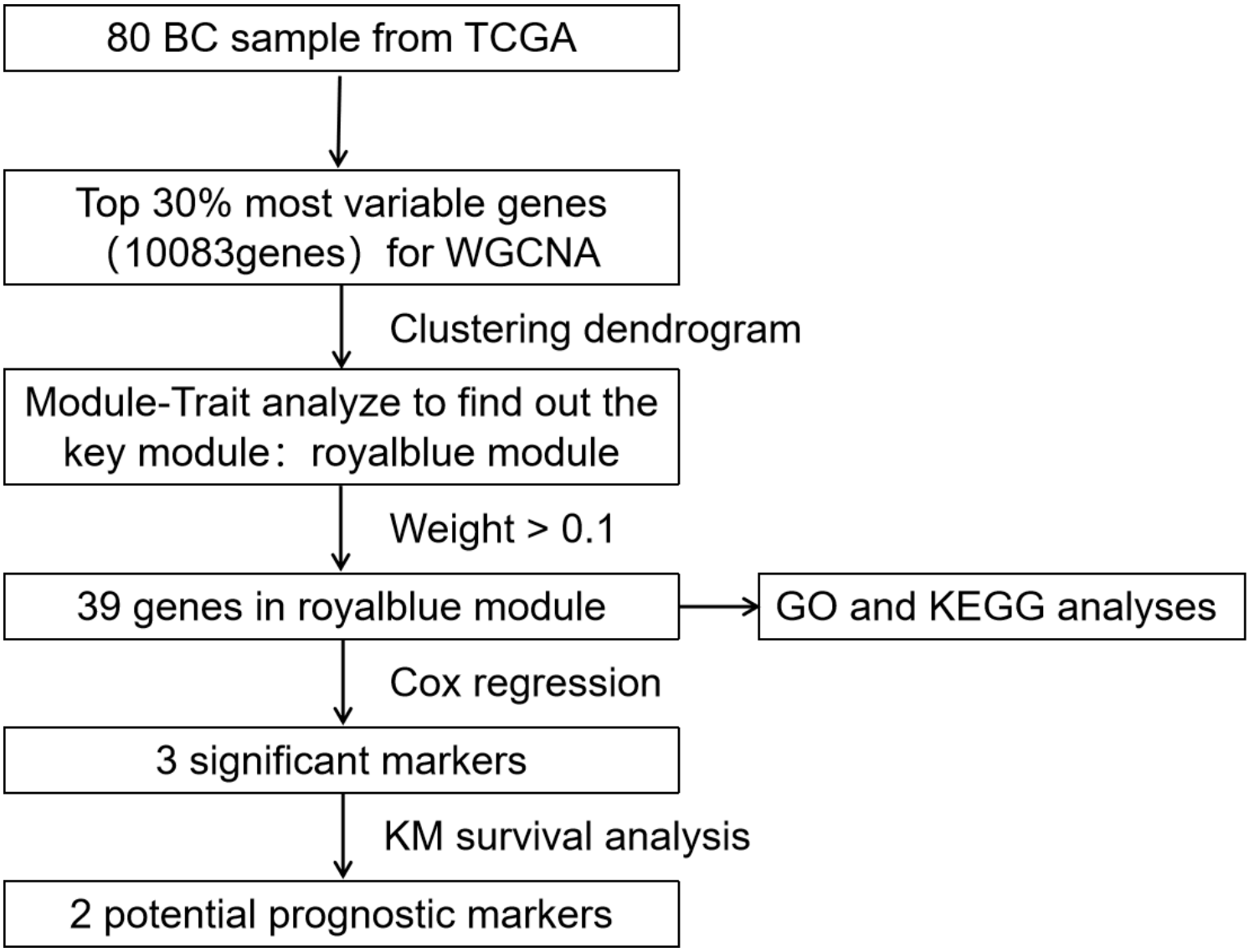

Figure 1

The flowchart of identifying procedure for the multi-gene signatures with breast cancer lung metastasis.

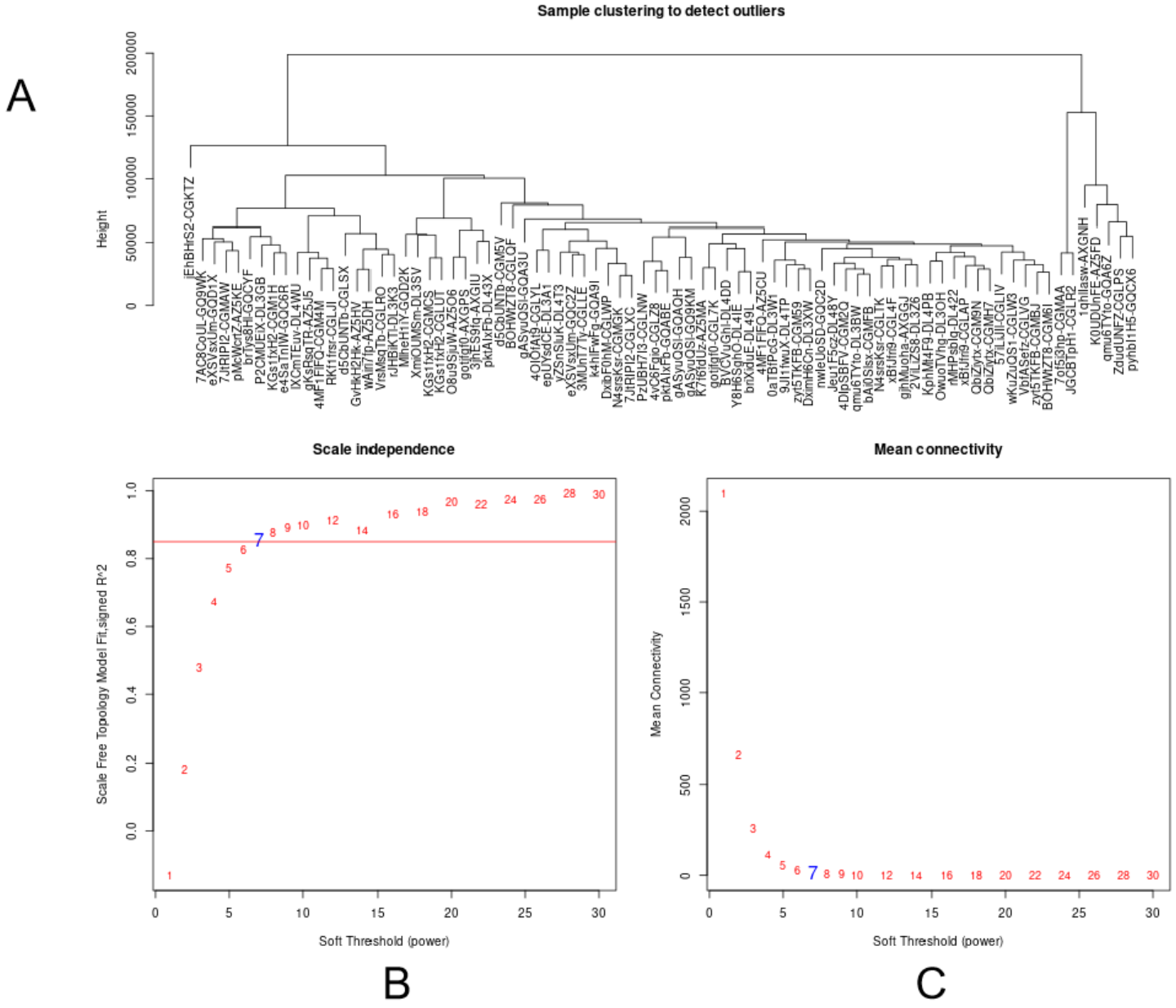

Figure 2

Determination of soft-threshold power in the WGCNA. (A) Clustering tree based on the module eigengenes of modules to detect outliers.(B) Scale-free index for various soft-threshold powers. (C) Mean connectivity for various soft-threshold powers.

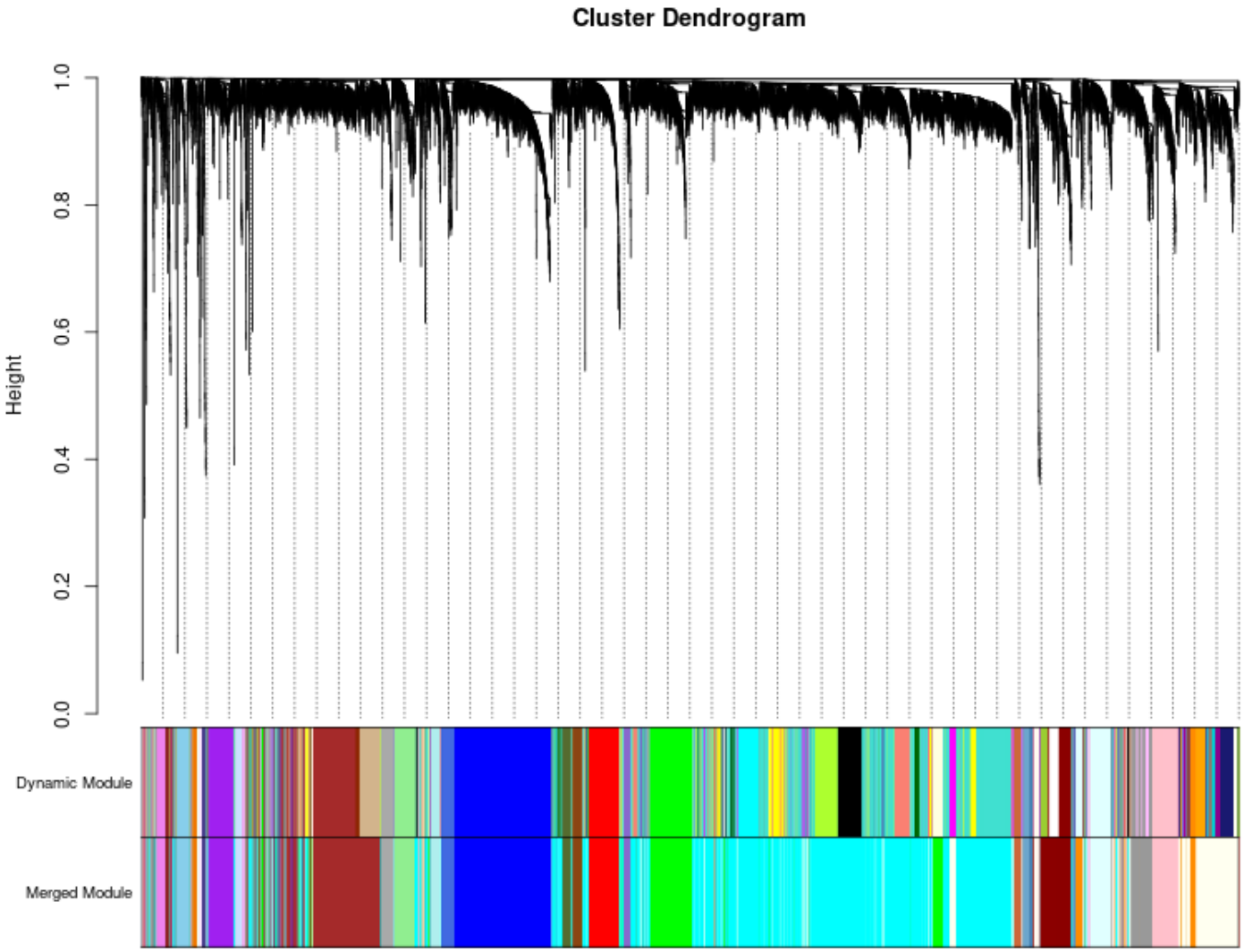

Figure 3

Clustering dendrograms of genes, with dissimilarity based on topological overlap, together with assigned module colors.As a result, 39 co-expression modules were constructed and was shown in different color. These modules were ranged from large to small by the number of genes they included. The number of genes in the 39 modules was listed in Table1.

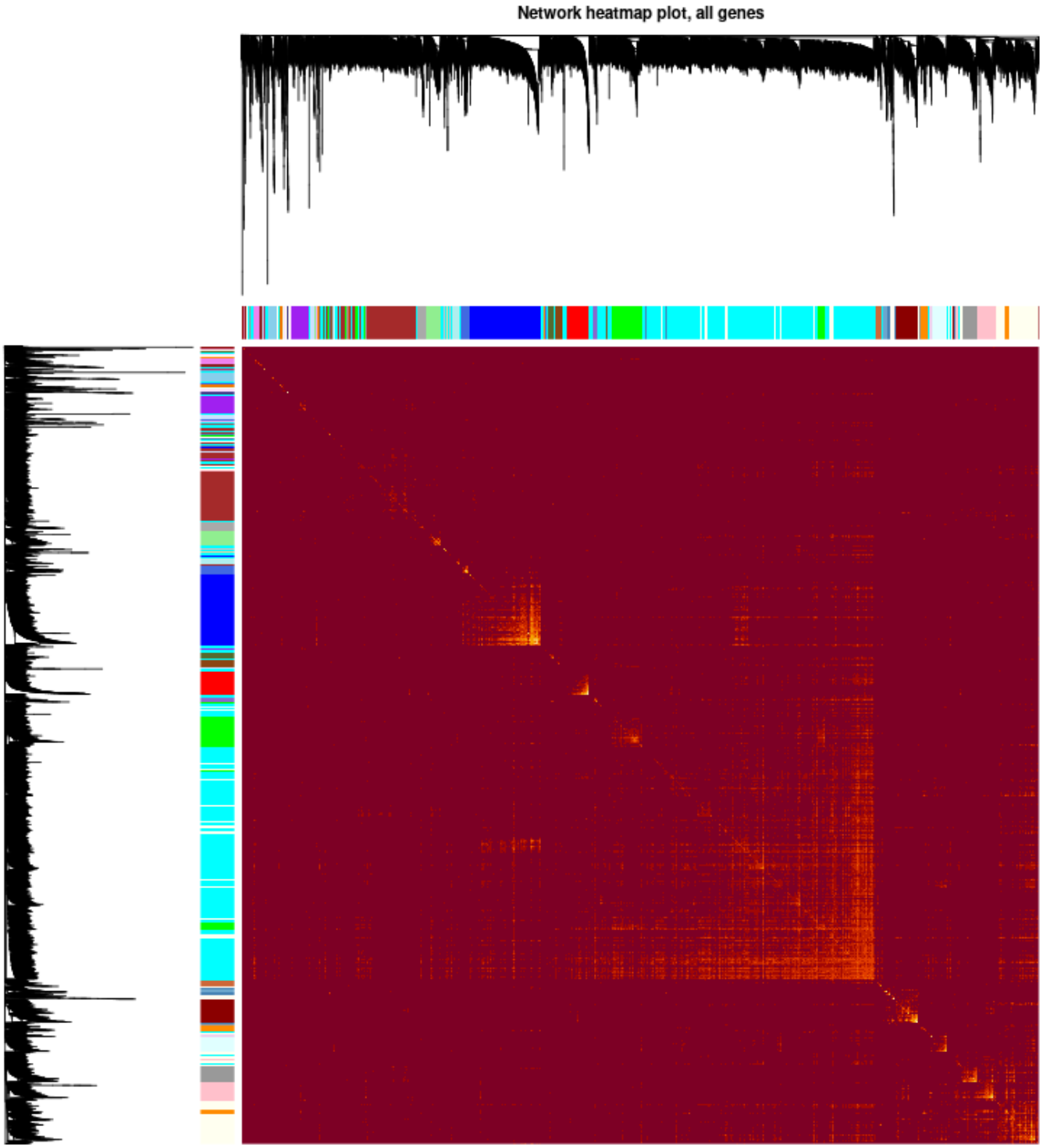

Figure 4

Construction of the gene correlation network. Each color represents a certain gene module. Visualizing the gene network using a heatmap plot. The heatmap depicts the Topological Overlap Matrix(TOM) among all genes in the analysis. Light color represents low overlap and progressively darker red color represents higher overlap.Blocks of darker colors along the diagonal are the modules. The gene dendrogram and module assignment are also shown along the left side and the top.

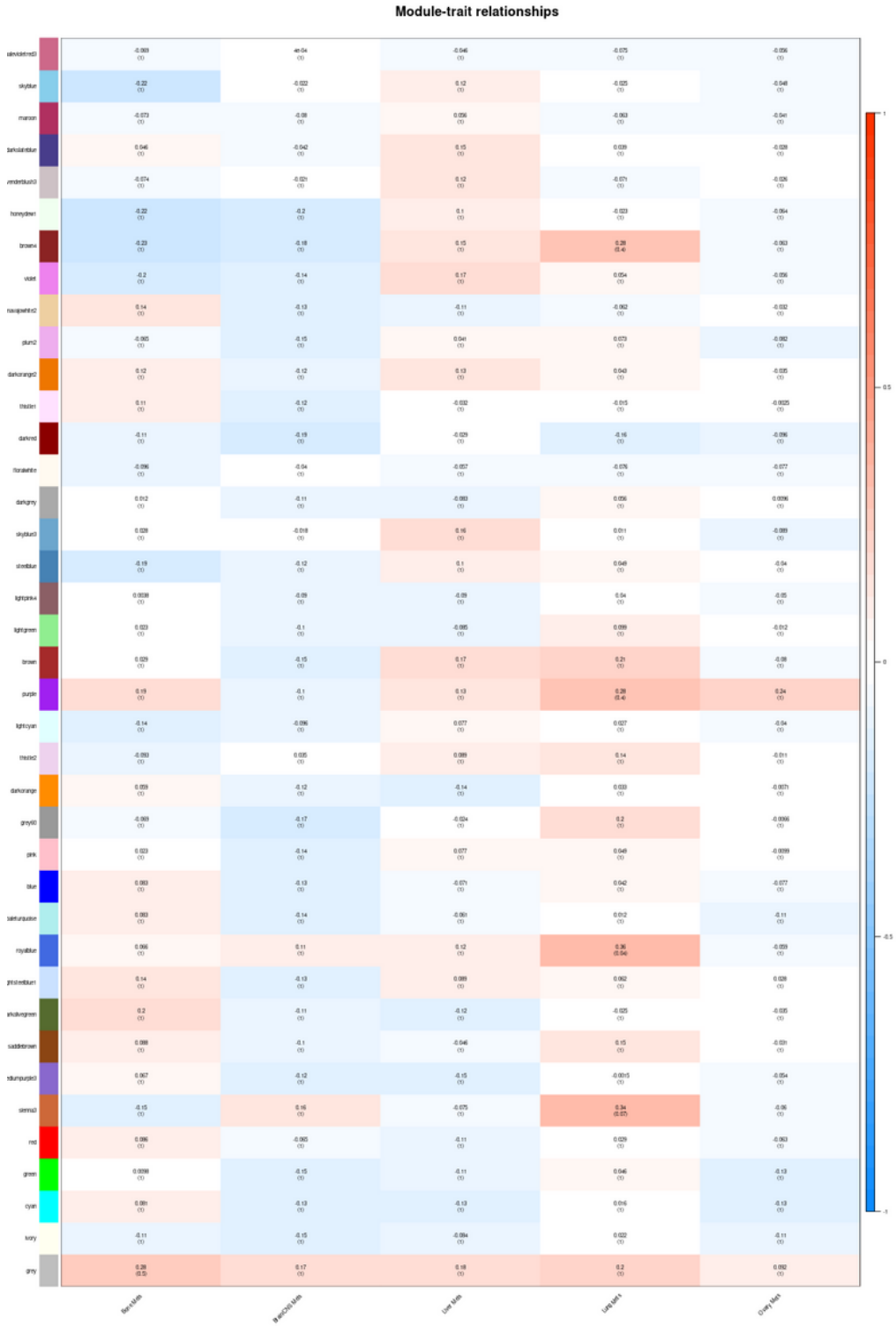

Figure 5

Module-trait associations. Each row corresponds to a module eigengene, column to a trait. Each cell contains the corresponding correlation and p-value. The table is color-coded by correlation according to the color legend.

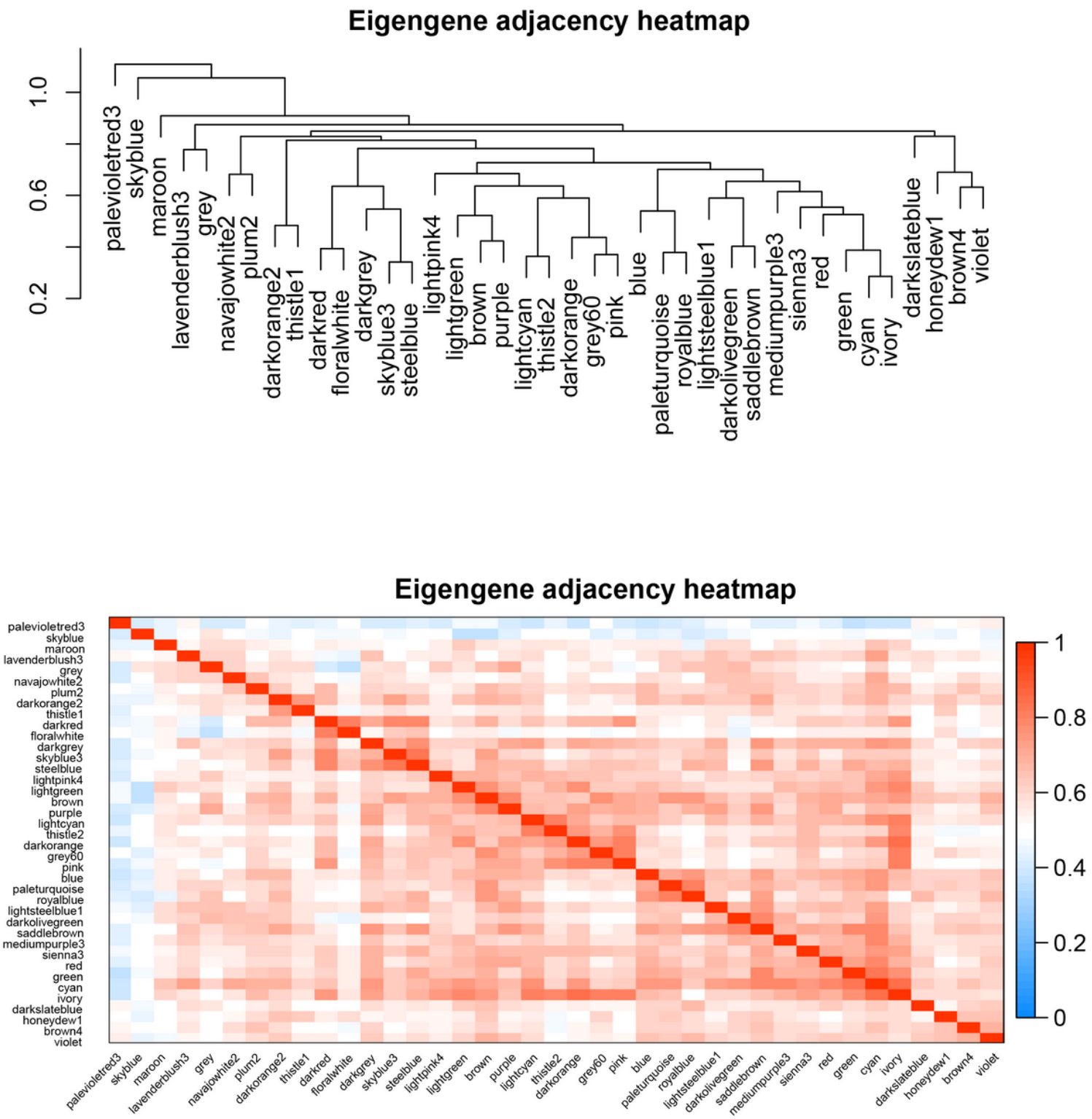

Figure 6

Clustering tree based on modules' eigengenes.The eigengene dendrogram and heatmap identify groups of correlated eigengenes termed meta-modules. As a result, (a): the dendrogram indicates that red, purple and magenta modules are highly related to patient's life status. (c): the dendrogram indicates that red and purple modules are highly related to recurrence time. (e): green module is highly related to recurrence. (b),(d) and (f): The heatmap in panel shows the eigengene adjacency.

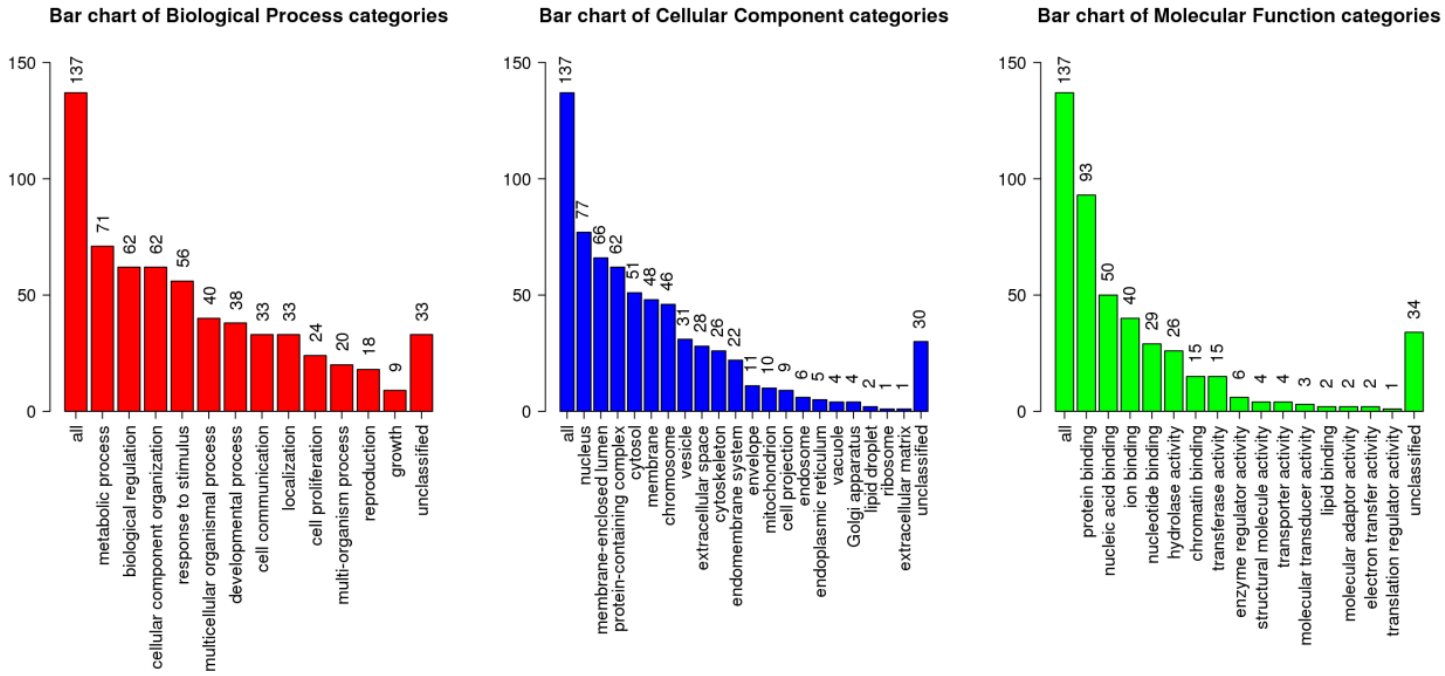

Figure 7

Gene ontology (GO) analyses. Bar plot of represent the biological process、cellular component and molecular function of hub genes in the royalbule module.

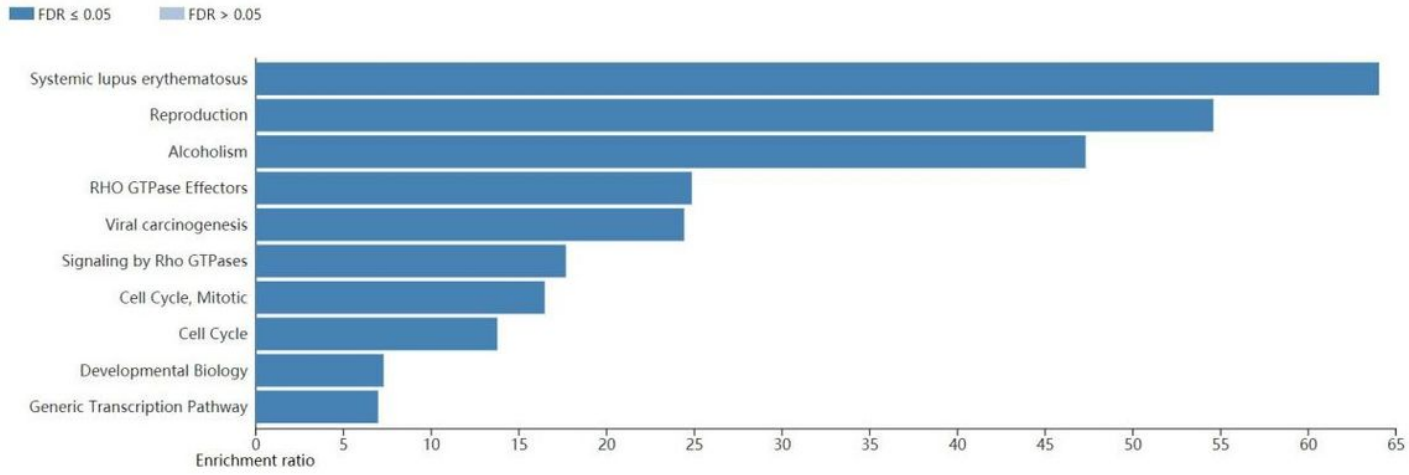

Figure 8

KEGG analyses. Bar plot of represent the pathways that the hub genes of the royalbule module involved in.

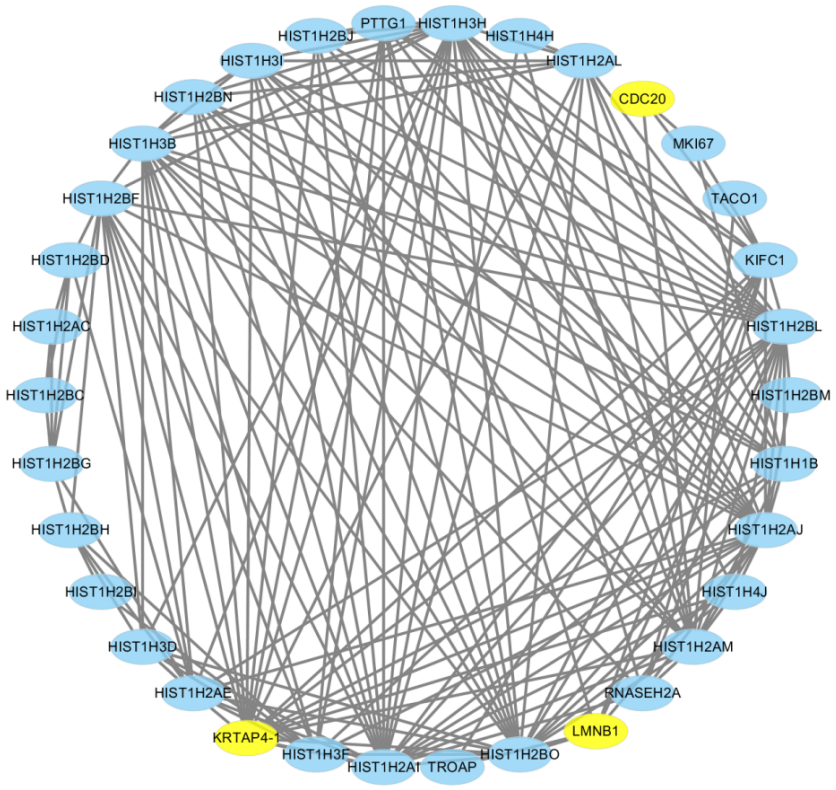

Figure 9

Construction of PPI network of genes in royalbule module.The PPI network of hub genes were analyzed by Cytoscape software.

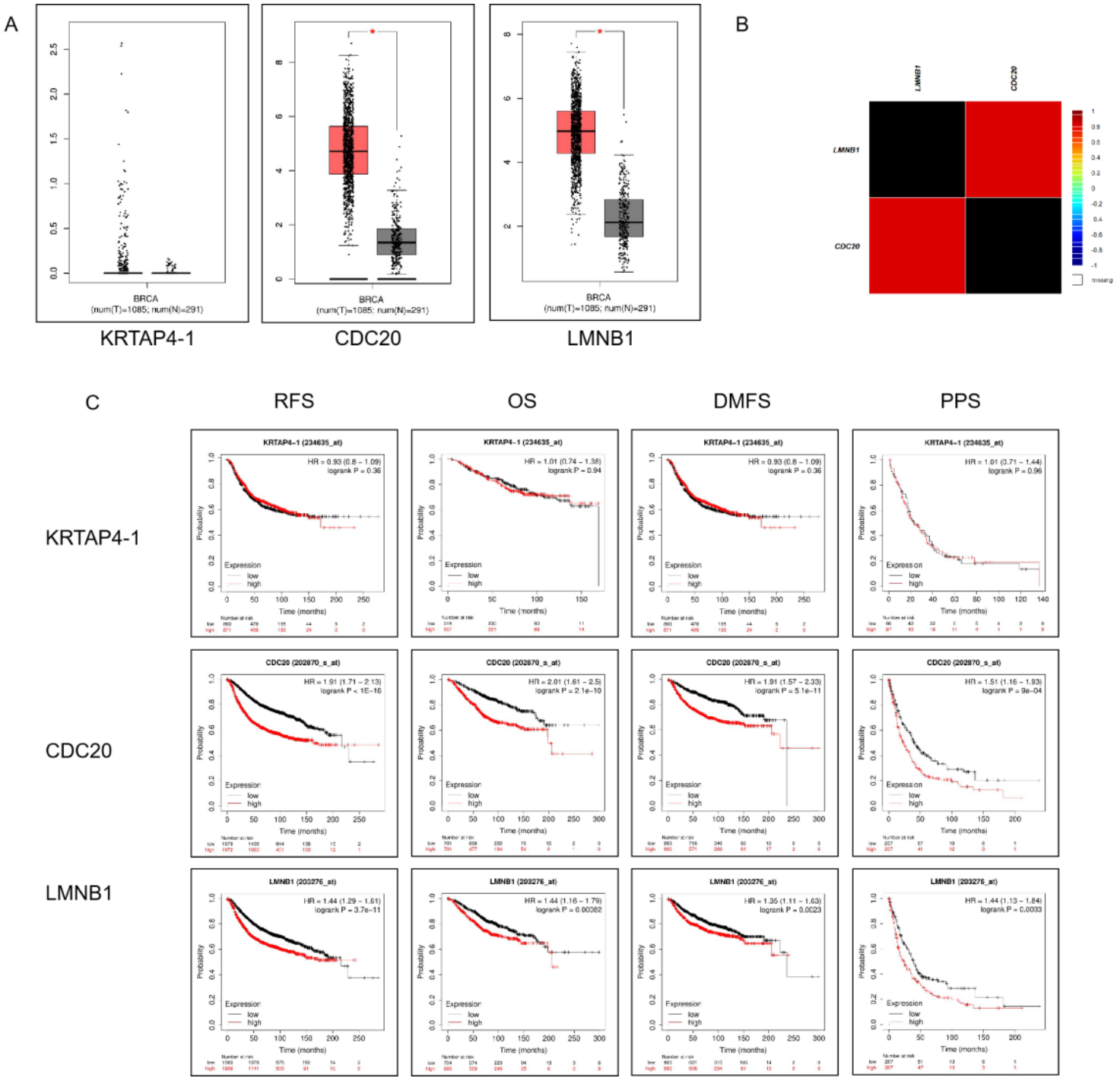

Figure 10

A.The GEPIA database verified that LMNB1 and CDC20 mRNA expression was significantly up-regulated in breast cancer tissues (BRCA) (n = 1085) compared with normal breast tissues (n = 291), *P < .05. B.The relationship between LMNB1 and CDC20 in breast cancer analyzed using bc-GenExMiner v4.0. C.The prognostic value of KRTAP4-1, LMNB1 and CDC20 in breast cancer patients using Kaplan-Meier Plotter.