Identification of genes associated with lymph node metastasis in papillary thyroid carcinoma by weighted gene co-expression network analysis

Zhen Liu (zhenliudr@163.com)
Shengjing Hospital of China Medical University https://orcid.org/0000-0003-1599-1487
Mingyue Guo
Shengjing Hospital of China Medical University
Lidong Wang
Shengjing Hospital of China Medical University
Chenxi Liu
Shengjing Hospital of China Medical University
Yonglian Huang
Shengjing Hospital of China Medical University

Primary research

Keywords: Thyroid papillary carcinoma, weighted gene co-expression network analysis, Hub gene, differential genes, lymph node metastasis

Posted Date: February 18th, 2020

DOI: https://doi.org/10.21203/rs.2.23820/v1

Abstract

Background: Thyroid cancer (TC) is the most common endocrine malignant tumor, in which thyroid papillary carcinoma has the highest incidence of (PTC). However, some patients will relapse or even die because of tumor metastasis. Therefore, for the treatment of PTC, it is extremely important to find gene targets and their potential biomarkers. We obtained the RNA sequence matrix of PTC and the clinical information of the corresponding samples in the TCGA database.

Methods: The weighted gene co-expression network analysis was utilized to construct the co-expression network, and the gene modules with high correlation among genes and the relationship between genes and clinical traits were established.

Results: We utilized the screened 5271 genes for WGCNA analysis and constructed 19 gene modules. Genes in the magenta module with the highest correlation with lymph node metastasis were selected for enrichment analysis to construct a protein interaction network. After that, we selected 12 hub genes of this module by using the intersection of the differential genes in the data set of cytoscape and GEO database. 12 genes were visualized in GEPIA. After that, the gene expression was verified on the oncomine page, and 8 genes with high correlation with TC lymph metastasis (ARNTL, CDH3, CLDN1, COL8A1, COL8A2, IL1RAP, LPAR5) were screened.

Conclusion: In this paper, the differential genes are screened by combining the datasets of TCGA, GEO and Oncomine in TC for the first time, which makes the screened differential genes have more biological significance. These genes may offer new insights for the treatment of PTC lymph node metastasis.

Background

Thyroid cancer (TC) is a very frequent endocrine system malignant tumor. In recent years, its incidence is increasing year by year, and is generally younger[1]. The most common histological and pathological classification of TC is papillary thyroid carcinoma (PTC), accounting for 80% of all TC[2]. However, although the growth of PTC is slow and most of the patients is a real good prognosis, there will be some patients with poor surgical results and insensitive to drug treatment, resulting in recurrence or metastasis or even death[3]. Researchers use genetic analysis to study the molecular mechanism of TC to explore the process of TC transfer. In the course of continuous exploration, many new genes related to the TC have been found[4–6]. However, at present, the understanding of the molecular mechanism of PTC is not certain. Despite the fact that many genes have clearly played an important role in the occurrence and development of PTC, cancer metastasis is a complex process, which is the result of the interaction of multiple genes. For example, mutation activation of Ras gene is the most classical gene mutation in TC. Because Ras gene is closely related to the classical signal pathway-PI3K/Akt signal pathway, it controls many cellular processes, such as cell growth, proliferation and survival, cell metabolism and autophagy[7]. The mutations of BRAF, RAS and RET-PTC fusion genes in the activated state can cause the activation of MAPK signal pathway, and the activation of p38MAPK/ERK signal transduction

pathway will lead to cell proliferation and increase the apoptosis, autophagy and invasion of TC cells[8]. The study of the interaction between genes is more helpful for us to continue to explore the molecular mechanism of PTC.

Weighted co-expression network analysis (WGCNA) is a systems biology method used to describe gene association patterns among different samples. It can be used to establish highly synergistic gene sets, and to identify candidate biomarker genes or therapeutic targets according to the interconnectedness of gene sets and the association between gene sets and phenotypes. WGCNA R software package is a comprehensive collection of R functions, which are used to perform all aspects of weighted correlation network analysis. The software package includes network construction, module detection, gene selection, topology calculation, data simulation, visualization and interface with external software. It is just an algorithm for constructing scale-free gene co-expression network, which can not only classify different gene modules, but also find out the relationship between clinical features and gene modules. In this network, the “node” represents the gene expression vector in the organization, and the “edge” is weighted by the correlation between the nodes (pearson coefficient)[9]. WGCNA can classify highly related genes into a module, and soft thresholds can be used in determining whether to draw edges between two nodes. In the process of calculation, the identified modules can be summarized by module feature genes, and the feature gene networks can be linked with the shape of external samples, and finally the correlation of module members can be calculated. This method has been widely used for the screening of cancer genes[10]. WGCNA has been applied to analysis in breast cancer[11], pancreatic cancer[12] and many other cancers. In PTC, some people use WGCNA to analyze the relationship between PTC gene matrix and lymph node metastasis and histopathological grade of patients[13, 14]. In this paper, we mainly analyze the genes that are closely linked with the lymph node metastasis of PTC.

In this study, we used WGCNA to explore the public transcriptome data and corresponding clinical information of patients with PTC. The key gene modules related to PTC lymph node metastasis were identified and 8 final differential genes were found by combining multiple databases. These genes may play an important role in lymph node metastasis of PTC and may be used as clinical biomarkers of lymph node metastasis and recurrence of PTC.

Materials And Methods

Download of data

The data and clinical information of PTC patients were downloaded from the cancer genome map (TCGA) database (https://cancergenome.nih.gov/) by R package TCGAbiolinks, including the RNA sequences of 568 samples of 507 patients and their related data. The human PTC mRNA expression profile data set was downloaded from the gene expression comprehensive database (http://www.ncbi.nlm.nih.gov/geo/). The GSE3678 dataset contains 7 PTC samples and 7 normal samples.

Construct co-expression network and obtain gene module

First of all, the selected genes were used R WGCNA package to construct the scale-free network of 5271 genes utilizing weighted expression correlation, and these genes were analyzed by Pearlson correlation matrix. We use the power function a power function amn = | cmn | β to build a weighted adjacency matrix. After selecting the appropriate β, the network connectivity of genes can be calculated by transforming the adjacency matrix into topological overlap matrix (TOM). The connectivity of the network of this gene, the sum of its adjacency with all genes, is used to set up the network[15]. Soft thresholds are used to make sure of scale-free networks. In the co-expression network, genes with extreme absolute correlation are clustered in the same module. In order to classify the genes with similar expression profiles into module genes, the average linkage hierarchical clustering was performed on the gene tree map. In addition, the functional module is chosen by using the topology overlap between the module and the adjacent module. Select the most suitable power value of the soft threshold, so that the correlation coefficient is the best, the fitting effect is the best, and the co-expression network can be realized. Then cluster analysis was performed out, and the genes with similar patterns were divided into gene modules[16].

Correlation between construction phenotype and modular characteristic genes

Highly linked module genes are represented and summarized by their first principal component, which is called module feature gene (MEs), module number-trait association, that is, the correlation between clinical phenotype and module feature genes. Modular characteristic gene is an effective and biologically meaningful tool to study the relationship between modules of gene co-expression network[17]. Therefore, we calculated the correlation between phenotype and MEs, and identified the gene modules with high correlation with clinical traits.

Gene functional enrichment Analysis (GO) and Gene set enrichment Analysis (GSEA) of key modules

In order to further understand the gene function in the key module, we use clusterProfiler package which has been used by many researchers in their paper to do GO analysis of the genes in the key module to annotate the gene function and KEGG analysis to see the related pathways[18]. And use this package to analyze the gene set enrichment of the key modules.

Construction and Identification of key genes in protein interaction Network

The genes related to lymph node metastasis were obtained by WGCNA, and the pathways of these genes were analyzed and the protein interaction network was established. We upload the genes in the key co-expression gene module to STRING database (http://www.STRING-db.org/) to construct PPI network analysis, and use cytoscape software for visualization[19]. In the network, the node represents the gene, and the edge represents the interaction between the nodes. According to k-core, a very crucial node in the network, namely Hub nodes, is obtained. The Hub gene represents the genes with key functions and has great interconnection with the genes in the module. Therefore, we utilize Cytoscape software to filter out Hub gene. We used limma packages to screen out differential genes in the dataset GSE3678.

Visualization of Hub Gene

Gene expression profile Interactive Analysis (GEPIA) (Http:/gepia.cancer-pku.cn) is a highly visual analysis website based on TCGA transcription database developed by Peking University. This website has rich functions, including single gene analysis, tumor type analysis, multi-gene analysis, differential expression, survival analysis, correlation analysis and so on. The differential expression of these 10 genes in GEPIA (Http://gepia.cancer-pku.cn) was analyzed in thyroid papillary carcinoma and adjacent thyroid carcinoma[20]. The screening criteria for differential gene expression analysis are |Log2FC|>1, p-value<0.01, and the figure is box map. Marked on the box diagram with significant differences.

Screening of differential genes (DEGs)

Oncomine (https://oncomine.org/) database is a cancer microarray database and Web-based data mining platform, which can be used to analyze the differential expression of most major cancer types and corresponding normal tissues and various cancer subtypes based on clinical and pathological analysis. On this site, all data can be queried and visualized, or genes can be chosen for analysis and visualization[21]. Finally, differential genes were screened according to the standard of p-value>0.05, FDR>1.

Result

Data download and preprocessing

The data and clinical information of PTC patients were downloaded from Cancer Genome Map (TCGA) Database (https://cancergenome.nih.gov/) by R package TCGAbiolinks on October 31, 2019, including the RNA sequences and related data of 568 samples from 507 patients. Use Deseq2 to convert the

original data of the gene expression matrix to the standardized data. The expression matrix containing duplicated genes was removed. The genes with low expression were filtered out, and the genes with the first 75% median absolute deviation of (MAD) were screened. The outliers were excluded and the remaining 5271 genes were analyzed. The clinical character file used for association analysis must be a numerical file, so the sample data in the clinical information are converted into a numerical matrix in 0–1 format. 1 indicates that it belongs to this group or has this group attribute, and 0 indicates that it not part of this group or does not have this group attribute. In the matrix of lymph node metastasis, 1 indicates lymph node metastasis and 0 indicates no lymph node metastasis. By removing the duplicate samples, samples of the gene matrix are completely consistent with the clinical samples, and finally 501 samples and 5271 genes are obtained.

Construct weighted co-expression network and identify key modules

After data preprocessing and quality evaluation, the expression matrix of 5271 genes in TCGA data set was used only for WGCNA analysis. When the soft threshold power = 5, the threshold of the correlation coefficient reaches 0.86, and the topological network is closer to the scale-free (figure 1A). We use power = 5 as the power value. As showing the figure, the cluster graph consists of 19 co-expression modules. We have made a hierarchical clustering tree to show the heat map of each module (figure 1B) and the correlation between each module, and the number of each module is sorted into a table (Table 1). The gray module is the gene that is not considered into the module. Except for the gray module, other co-expressed gene aggregation modules are associated with clinical traits in varying degrees. For instance, magenta module is closely related to lymph node metastasis.

Figure 1

Insert Table 1 here.

Correlation between construction phenotype and modular characteristic genes

Delete worthless information from the clinical data of 501 PTC patients in the TCGA database and use this data set to determine the correlation between MEs and PTC lymph node metastasis (figure 2A). The results showed that the co-expression module was correlated with some clinical phenotypes. For example, module magenta (Cor = 0.12, p = 0.009) is most closely related to PTC lymph node metastasis. For each gene module, gene significance (GS) represents the correlation level between expression pattern and phenotype, and module membership (MM) represents the correlation level between the expression pattern and MEs. The correlation between this module MM and GS is 0.66 ~ 2.2e ~ 23 (figure 2B), so we choose the one in this module for follow-up analysis.

Figure 2

Gene functional enrichment Analysis and Gene set enrichment Analysis of key Modules

Genes in the magenta module were analyzed by GO using clusterProfiler package to annotate the gene function. GSEA enrichment analysis was carried out (figure 3F). The most important GO term in GO analysis is organelle division, mitosis, microtubule cytoskeleton, mitosis and chromosome segregation figure 3 A□B□C□D□. In KEGG pathway analysis, the highest enrichment pathways are cell cycle, proteasome, DNA replication, IL–17 signal pathway, ribosome occurrence in eukaryotes and so on(figure 3E).

Figure 3

Construction of protein interaction network and screening of Hub gene

Use the software Cytoscape to visualize the PPI network. The plug-in CytoHubba in Cytoscape software calculates Hub gene. At the same time, 579 differential genes in the dataset GSE3678 were screened by limma package, and the volcano map of these differential genes was made (figure 4A). The standard was $p<0.05$. Genes calculated by the plug-in CytoHubba in Cytoscape software are intersected with the differential genes in GSE3578, and 12 hub genes, are obtained. Heat maps of the 12 genes are made (figure 4B, C, D). These genes are ARNTL, CDH3, CLDN1, COL8A1, COL8A2, IL1RAP, LPAR5, PARP4, PERP, PPL, TIAM1 and TNFRSF21.

Figure 4

Visualization of Hub gene

GEPIA (Http://gepia.cancer-pku.cn) was utilized to analyze the differential expression and survival of 12 selected genes. The screening criteria for differential gene expression analysis are $|Log2FC|>1$, p-value<0.01, and the figure is box map. The box-plot with significant differential expression is marked * (figure 5).

Figure 5

Screening of DEGs

We input the above screened genes into the search box of oncomine to analyze the difference between cancer and adjacent cancer, as well as visualization. The difference analysis selected from the analysis type, and the TC in head and neck cancer was selected as the cancer type. The genes with p-value<0.05， what is more, |FDR|>1 were screened out and verified by subsequent experiments. Finally, it was proved that there were significant differences in the expression of 8 genes between PTC and normal thyroid samples（p<0.05,FDR>1）(figure 6), and their P values and Fold change were shown in the table (Table 2). At last, three genes with the most significant differences, CDH3, CLDN1 and TIAM1, were chosen for follow-up experiment.

Figure 6

Insert Table 2 here.

Conclusion

In this study, through the combination of multiple databases, we finally screened out 8 DEGs. In the TC analysis, it is the first time to use the data in the TCGA database to use WGCNA analysis and combined with the GEO database and oncomine data to verify the differential genes. The weighted analysis method originally improves the accuracy of bioinformatics screening, and the combination of multiple databases will further improve the biological significance of the selected genes.

Many literature have confirmed that TIAM2 may promote lymph node metastasis of solid tumors[22]. Chuen Hsuch et al tracked the patients with PTC and verified the elevated expression of TIAM1 in PTC by immunohistochemistry. Liu Lin and other studies have shown that TIAM1 can promote the epithelial-mesenchymal transformation of TC by activating Wnt/ β-catenin signal pathway, thus promoting the metastasis of TC[23].

All in all, in further research, other potentially differential genes will also be tested in the future. These genes may offer new insights for the treatment of PTC lymph node metastasis.

Declarations

Acknowledgements

The authors would like to acknowledge the supported by the National Natural Science Foundation of China.The authors also thank the contributors of The Cancer Genome Atlas (https://cancergenome.nih.gov/) and the gene expression comprehensive database (http://www.ncbi.nlm.nih.gov/geo/), for sharing the COAD dataset on open access. This study was sponsored in part by National Natural Science Foundation of China.

Authors’s contribution

Mingyue Guo and Zhen Liu designed the research; all the authors were engaged into the performance of the research data analysis and wrote the literal editing. All authors read and approved the final manuscript.

Funding information

National Natural Science Foundation of China, Grant/Award Number:81672644

Availability of data and materials

Authors can provide all of datasets analyzed during the study on reasonable request.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of General Surgery Shengjing Hospital of China MedicalUniversity, Shenyang, Liaoning, P. R. China 110004,[2]Department of General Surgery Shengjing Hospital of China MedicalUniversity, Shenyang, Liaoning, P. R. China, 110004,[3]Department of General Surgery Shengjing Hospital of China MedicalUniversity, Shenyang, Liaoning, P. R. China, 110004,[4]Department of General Surgery Shengjing Hospital of China MedicalUniversity, Shenyang, Liaoning, P. R. China, 110004

Reference

1.Seib CD, Sosa JA: *Evolving Understanding of the Epidemiology of Thyroid Cancer. Endocrinol Metab Clin North Am* 2019, *48(*1):23–35.

2.Xu C, Wang Y, Yang H, Hou J, Sun L, Zhang X, Cao X, Hou Y, Wang L, Cai Q *et al: Association Between Cancer Incidence and Mortality in Web-Based Data in China: Infodemiology Study. J Med Internet Res* 2019, *21(*1):e10677.

3.Kampo S, Ahmmed B, Zhou T, Owusu L, Anabah TW, Doudou NR, Kuugbee ED, Cui Y, Lu Z, Yan Q *et al: Scorpion Venom Analgesic Peptide, BmK AGAP Inhibits Stemness, and Epithelial-Mesenchymal Transition by Down-Regulating PTX3 in Breast Cancer. Front Oncol* 2019, *9:*21.

4.Liang W, Sun F: *Identification of key genes of papillary thyroid cancer using integrated bioinformatics analysis. J Endocrinol Invest* 2018, *41(*10):1237–1245.

5.Liu Q, Pan LZ, Hu M, Ma JY: *Molecular Network-Based Identification of Circular RNA-Associated ceRNA Network in Papillary Thyroid Cancer. Pathol Oncol Res* 2019.

6.Zhao H, Li H: *Network-based meta-analysis in the identification of biomarkers for papillary thyroid cancer. Gene* 2018, *661:*160–168.

7.Zhang L, Xiao X, Arnold PR, Li XC: *Transcriptional and epigenetic regulation of immune tolerance: roles of the NF-kappaB family members. Cell Mol Immunol* 2019, *16(*4):315–323.

8.Fu H, Cheng L, Jin Y, Cheng L, Liu M, Chen L: *MAPK Inhibitors Enhance HDAC Inhibitor-Induced Redifferentiation in Papillary Thyroid Cancer Cells Harboring BRAF (V600E): An In Vitro Study. Mol Ther Oncolytics* 2019, *12:*235–245.

9.Langfelder P, Horvath S: *WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics* 2008, *9:*559.

10.Li S, Li B, Zheng Y, Li M, Shi L, Pu X: *Exploring functions of long noncoding RNAs across multiple cancers through co-expression network. Sci Rep* 2017, *7(*1):754.

11.Guo X, Xiao H, Guo S, Dong L, Chen J: *Identification of breast cancer mechanism based on weighted gene coexpression network analysis. Cancer Gene Ther* 2017, *24(*8):333–341.

12.Zhou Z, Cheng Y, Jiang Y, Liu S, Zhang M, Liu J, Zhao Q: *Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. Int J Biol Sci* 2018, *14(*2):124–136.

13.Zhai T, Muhanhali D, Jia X, Wu Z, Cai Z, Ling Y: *Identification of gene co-expression modules and hub genes associated with lymph node metastasis of papillary thyroid cancer. Endocrine* 2019.

14.Tang X, Huang X, Wang D, Yan R, Lu F, Cheng C, Li Y, Xu J: *Identifying gene modules of thyroid cancer associated with pathological stage by weighted gene co-expression network analysis. Gene* 2019, *704:*142–148.

15.Jiang C, Wu S, Jiang L, Gao Z, Li X, Duan Y, Li N, Sun T: *Network-based approach to identify biomarkers predicting response and prognosis for HER2-negative breast cancer treatment with taxane-anthracycline neoadjuvant chemotherapy. PeerJ* 2019, *7:*e7515.

16.Guo L, Zhang K, Bing Z: *Application of a coexpression network for the analysis of aggressive and nonaggressive breast cancer cell lines to predict the clinical outcome of patients. Mol Med Rep* 2017, *16(*6):7967–7978.

17.Langfelder P, Horvath S: *Eigengene networks for studying the relationships between co-expression modules. BMC Syst Biol* 2007, *1:*54.

18.Jiang Y, He J, Guo Y, Tao H, Pu F, Li Y: *Identification of genes related to low-grade glioma progression and prognosis based on integrated transcriptome analysis. J Cell Biochem* 2019.

19.Wettenhall JM, Simpson KM, Satterley K, Smyth GK: *affylmGUI: a graphical user interface for linear modeling of single channel microarray data. Bioinformatics* 2006, *22(7*):897–899.

20.Tang Z, Kang B, Li C, Chen T, Zhang Z: *GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res* 2019, *47(*W1):W556-W560.

21.Daniel R. Rhodes JY, K. Shankerz, Nandan Deshpandez, Radhika Varambally, Debashis Ghosh, Terrence Barrette, Akhilesh Pandeyb and Arul M. Chinnaiyan: *ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform. Neoplasia* 2004, *6(*1):1–6.

22.Ding J, Yang F, Wu W: *Tiam1 high expression is associated with poor prognosis in solid cancers: A meta-analysis. Medicine (Baltimore)* 2019, *98(*45):e17529.

23.Liu L, Wu B, Cai H, Li D, Ma Y, Zhu X, Lv Z, Fan Y, Zhang X: *Tiam1 promotes thyroid carcinoma metastasis by modulating EMT via Wnt/beta-catenin signaling. Exp Cell Res* 2018, *362(*2):532–540.

Tables

Table 1. The number of genes in 19 modules

module	Gene number
Black	214
Blue	499
Brown	318
Cyan	92
Green	328
Greenyellow	103
Grey	634
Grey60	48
Lightcyan	40
Lightgreen	38
Magenta	159
Midnightblue	58
Pink	139
Purple	57
Red	207
Salmon	87
Tan	85
Turquoise	468
Yellow	406

Table 2. The p-value and of DEGs in Oncomine

gene	P-value	Fold change
ARNTL	0.008	2.132
CDH3	6.79E-25	3.461
CLDN1	6.58E-08	12.859
COL8A1	5.33E-05	1.392
COL8A2	2.40E-05	2.777
IL1RAP	1.40E-06	2.016
LPAR5	9.72E-06	2.312
TIAM1	8.45E-05	4.072

Figures

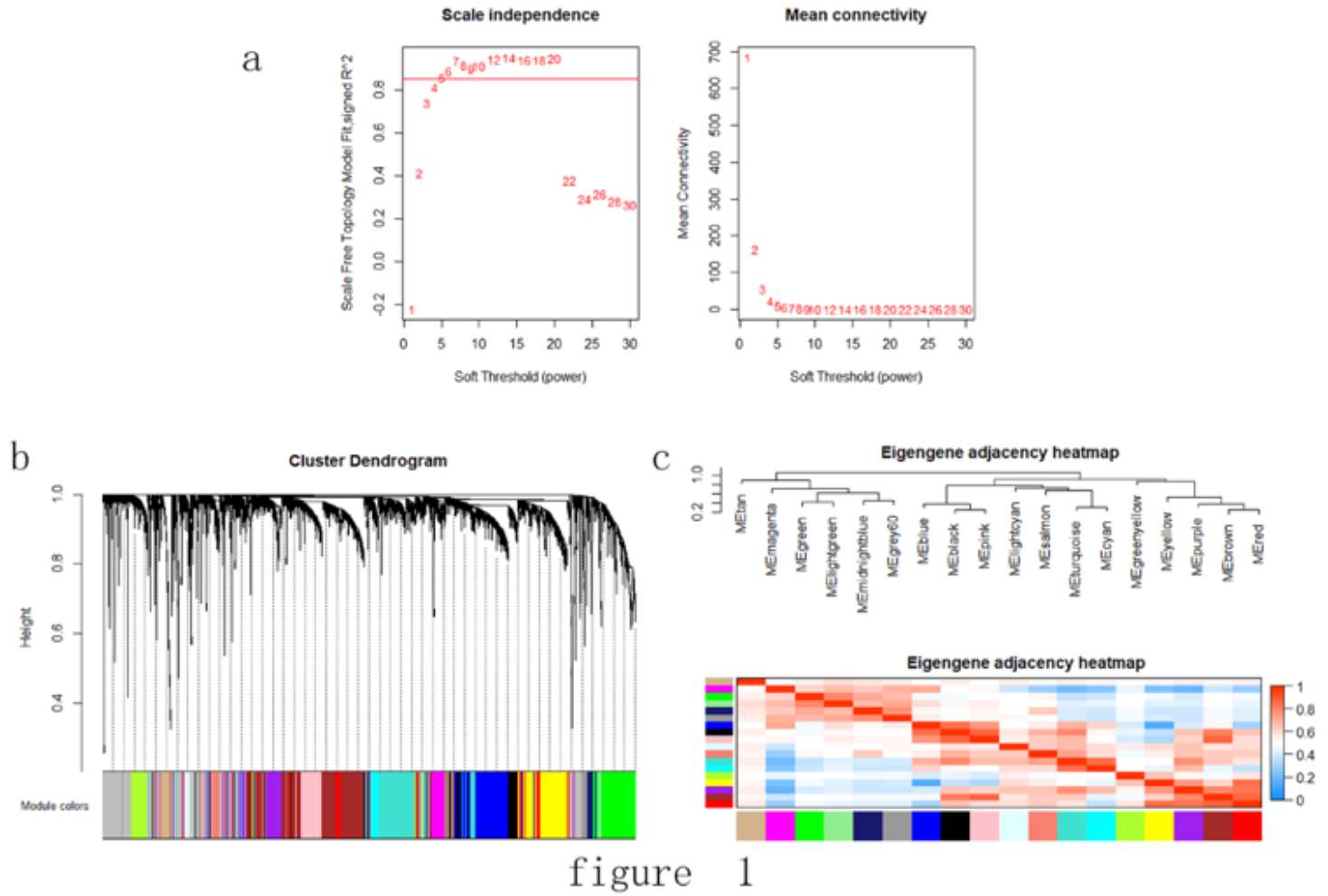

figure 1

Figure 1

After data preprocessing and quality evaluation, the expression matrix of 5271 genes in TCGA data set was used only for WGCNA analysis. When the soft threshold power=5, the threshold of the correlation coefficient reaches 0.86, and the topological network is closer to the scale-free (figure 1A). We use power=5 as the power value. As showing the figure, the cluster graph consists of 19 co-expression modules. We have made a hierarchical clustering tree to show the heat map of each module (figure 1B) and the correlation between each module, and the number of each module is sorted into a table (Table 1).

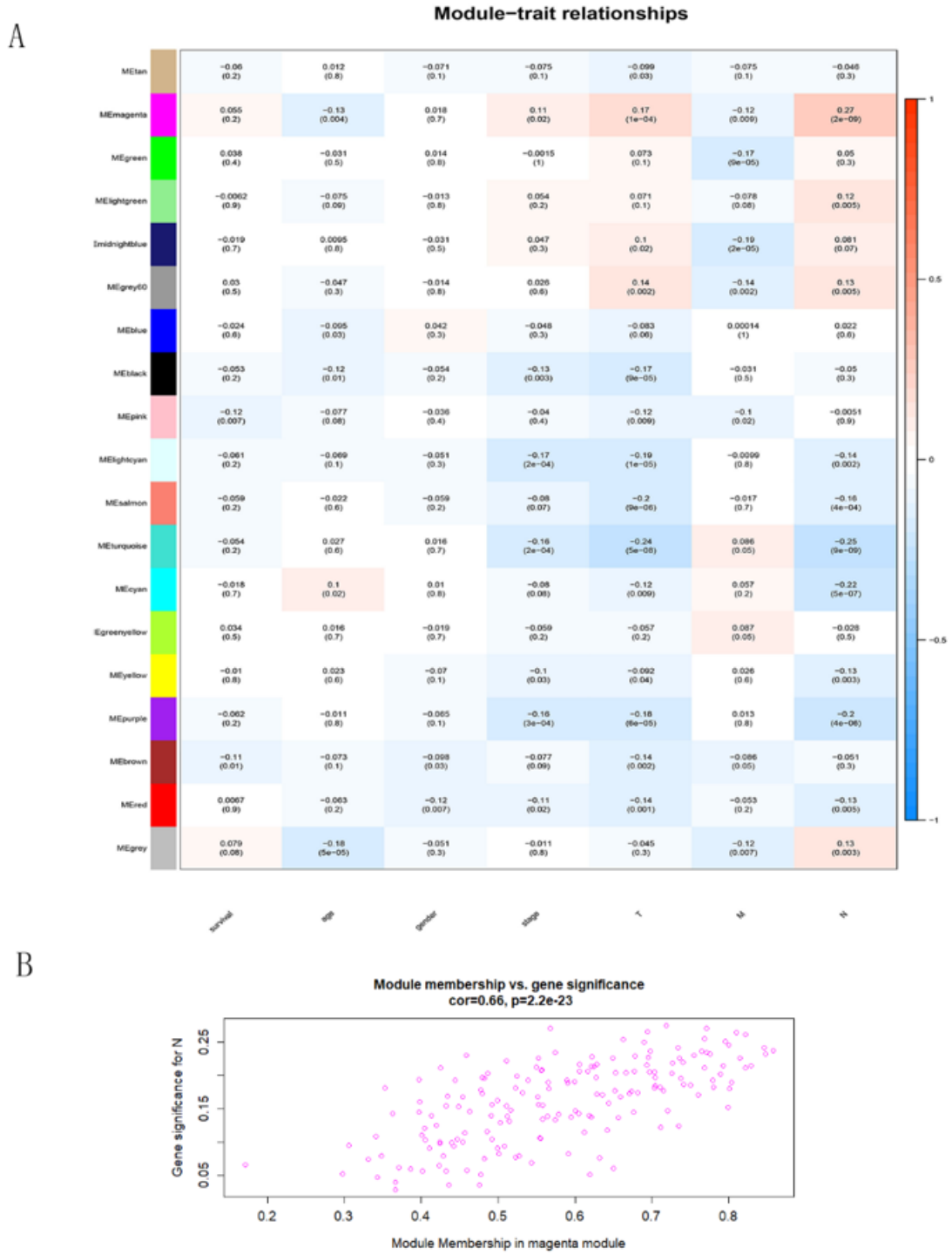

Figure 2

Delete worthless information from the clinical data of 501 PTC patients in the TCGA database and use this data set to determine the correlation between MEs and PTC lymph node metastasis (figure 2A). The results showed that the co-expression module was correlated with some clinical phenotypes. For example, module magenta (Cor=0.12, p=0.009) is most closely related to PTC lymph node metastasis. For each gene module, gene significance (GS) represents the correlation level between expression pattern and phenotype, and module membership (MM) represents the correlation level between the expression pattern and MEs. The correlation between this module MM and GS is 0.66 ~ 2.2e ~ 23 (figure 2B), so we choose the one in this module for follow-up analysis.

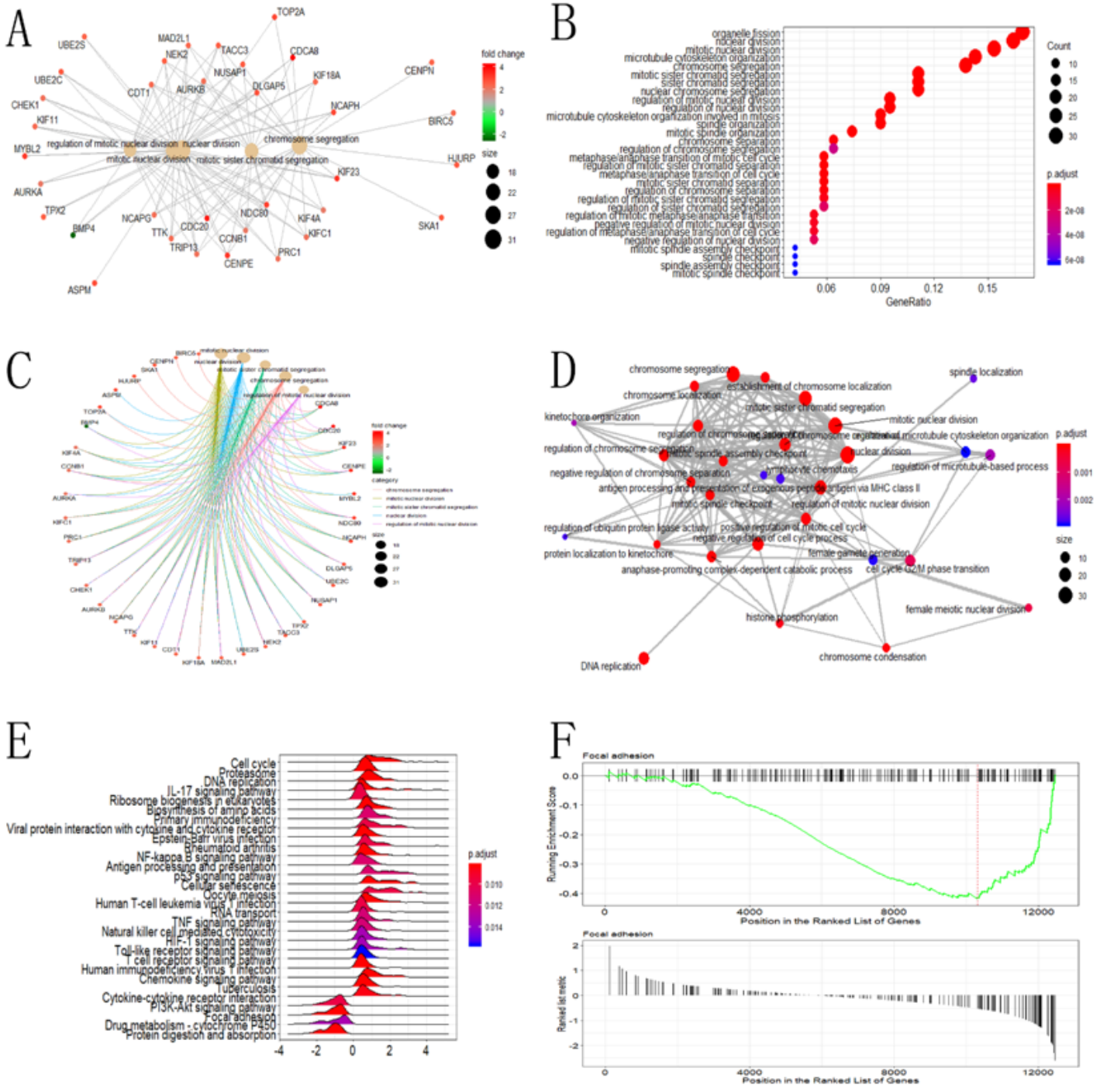

Figure 3

Genes in the magenta module were analyzed by GO using clusterProfiler package to annotate the gene function. GSEA enrichment analysis was carried out (figure 3F). The most important GO term in GO analysis is organelle division, mitosis, microtubule cytoskeleton, mitosis and chromosome segregation （figure 3 A、B、C、D）. In KEGG pathway analysis, the highest enrichment pathways are cell cycle, proteasome, DNA replication, IL-17 signal pathway, ribosome occurrence in eukaryotes and so on(figure 3E).

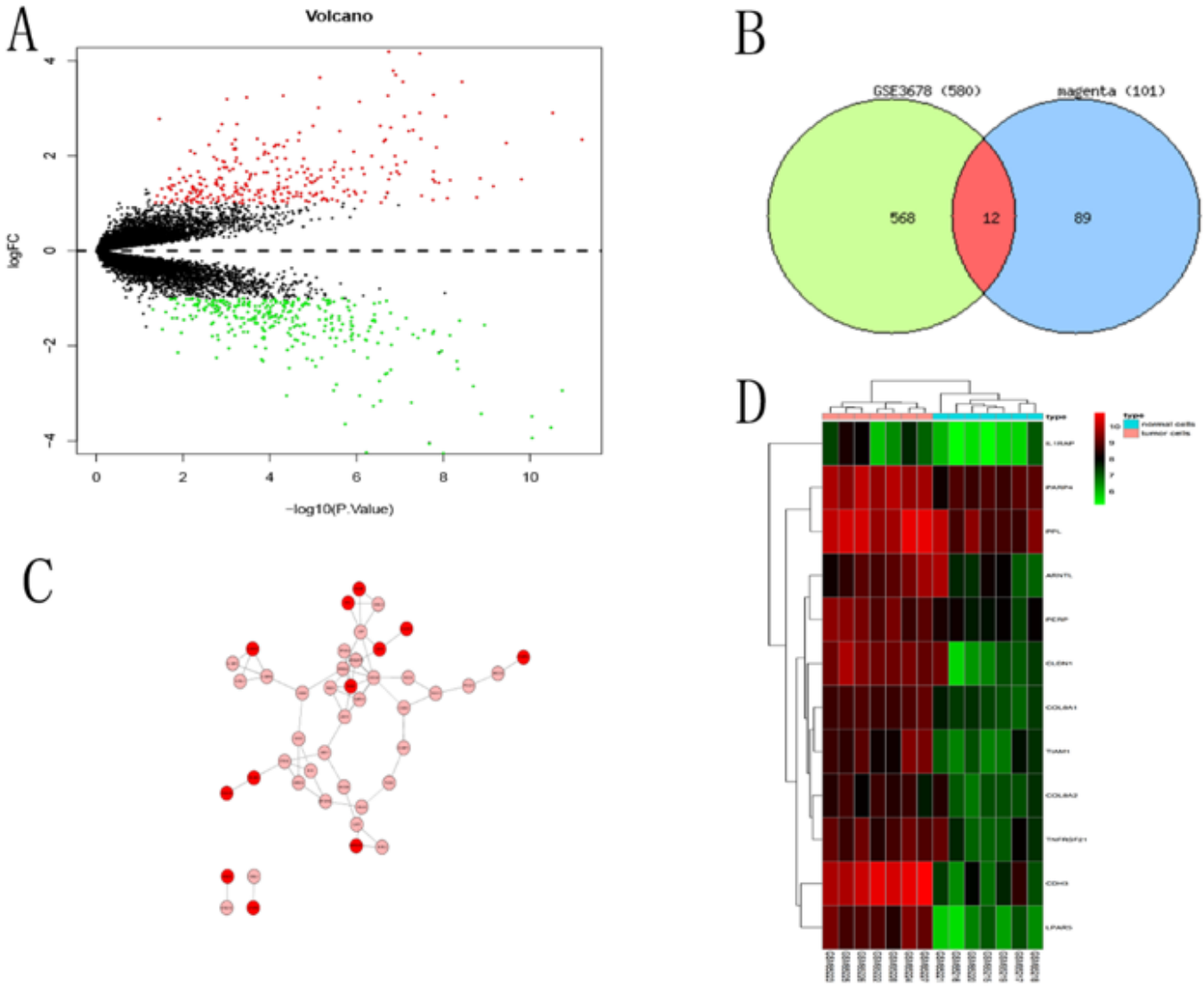

Figure 4

579 differential genes in the dataset GSE3678 were screened by limma package, and the volcano map of these differential genes was made (figure 4A). The standard was p<0.05. Genes calculated by the plug-in CytoHubba in Cytoscape software are intersected with the differential genes in GSE3578, and 12 hub genes, are obtained. Heat maps of the 12 genes are made (figure 4B, C, D).

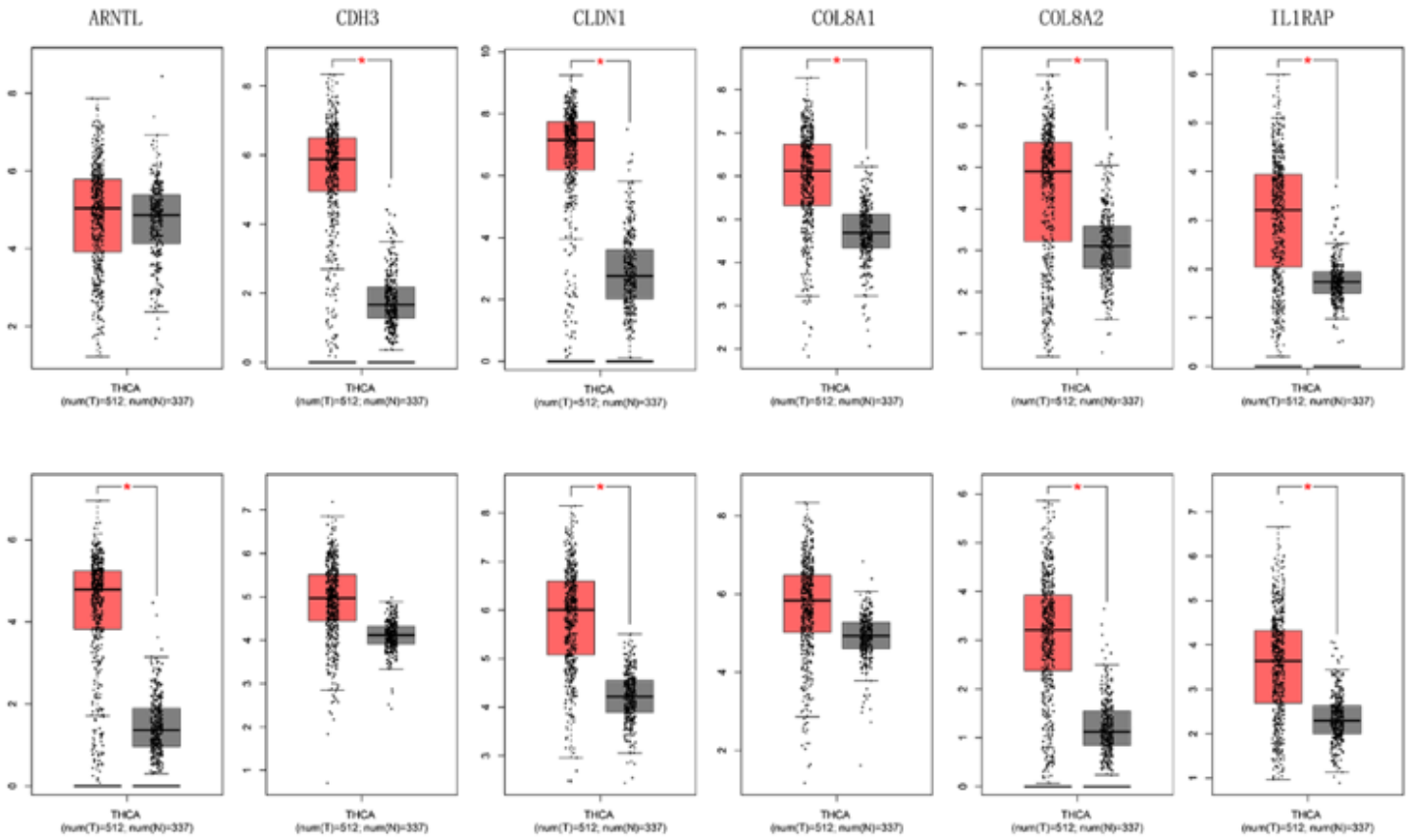

Figure 5

GEPIA (Http://gepia.cancer-pku.cn) was utilized to analyze the differential expression and survival of 12 selected genes. The box-plot with significant differential expression is marked *.

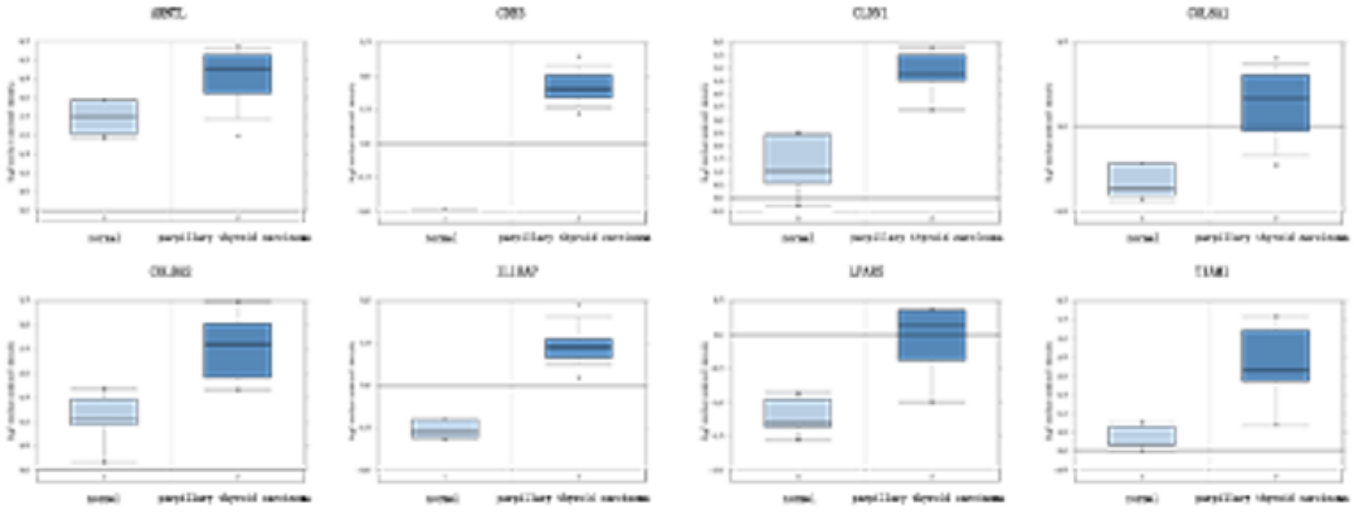

Figure 6

There were significant differences in the expression of 8 genes between PTC and normal thyroid samples （p<0.05,FDR>1）(figure 6), and their P values and Fold change were shown in the table (Table 2).