Identification of Key Genes and Pathways in Colorectal Cancer by Integrated Bioinformatics Analysis

Chaochao Wang
The Affiliated Hospital of Southwest Medical University https://orcid.org/0000-0002-0637-6696
Li Zhang (zhanglizhangli762@gmail.com)
Affiliated hospital of Southwest Medical University
Yingchun Hu
The Affiliated Hospital of Southwest Medical University

Research

Keywords: Colorectal Cancer, Bioinformatics, Differentially expressed genes

Posted Date: November 20th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-111316/v1

Abstract

Purpose

In order to understand the mechanism of colorectal cancer occurrence and development, we screened related core genes and provided new targets for clinical diagnosis and treatment of colorectal cancer.

Methods

We downloaded CRC-associated gene expression profile of GSE110233 from Gene Expression Omnibus (GEO) dataset. There were 26 samples in this dataset, all the differentially expressed genes (DEGs) with $p<0.05$ and fold change ≥ 1 or ≤ -1 were identified. Gene ontology (GO) and "Kyoto Encyclopedia of Genes and Genomes" (KEGG) were used to search for these DEG enrichment methods. In addition, the protein-protein interaction (PPI) network was also used to construct visual interactions between proteins. At last, we used GEPIA to conduct the survival analysis 4 down-regulation and 8 up-regulation genes for clarify the potential effects on CRC.

Results

A total of 866 differentially expressed genes were obtained, including 360 up-regulated genes and 506 down-regulated genes. These genes were involve in Cell proliferation; Extracellular exosome; Protein binding; Chemokine activity. Genes were mainly involved in the KEGG pathway termed Cell cycle; PI3K-Akt signaling pathway; Mineral absorption; MicroRNAs in cancer; Cytokine-cytokine receptor interaction. We finally found 12 hubgenes by PPI connective degree whom named PRKACB, FGFR2, FGFR3, CKB, TIMP1, CCNA2, CCNB1, CDC20, CDC6, CCND1, CDK4 and CDK1.

Conclusion

Bioinformatics is helpful for comprehensive and in-depth study of the occurrence and development mechanism of diseases, to screen possible core targets, and to provide a reliable basis for clinical diagnosis and treatment of colorectal cancer.

1. Introduction

Colorectal cancer (CRC) is one of the most common malignant tumors in the world. According to the statistics of the World Health Organization, it ranks third in the global morbidity and mortality of tumors. Nearly 1.4 million new cases of CRC and there are 700,000 CRC-related deaths reported annually worldwide[1-2]. It was reported that the 5-year survival rate only remains at approximately 60%. However, the CRC mortality still remains high[3]. Although surgical treatment of colorectal cancer has been widely used in clinic, there is still a lack of effective treatment for advanced colorectal cancer with distant metastasis[4-5]. In recent years, there are a large number of research data on colorectal cancer formation and metastasis, and more and more high-throughput gene chips have been widely used to screen

differentially expressed genes (DEG) between normal samples and human tumor samples , which also allows us to further explore the molecular mechanism of tumors[6].

This article intends to comprehensively analyze the disease state of colorectal cancer based on GEO gene expression data, thereby constructing a genetic network and screening out potential key molecular targets for disease. These molecular targets may provide us with new insights into the pathogenesis of colorectal cancer, and pave the way to diagnose colorectal cancer and provide new ideas for clinical treatment.

2. Materials And Methods

2.1 Microarray Data:

The GEO database belongs to the National Center for Biotechnology Information (NCBI) of the National Institutes of Health[7]. It is currently the world's largest and most complete public gene expression data resource. The gene expression profile GSE110223 was downloaded from the GEO database. The chip platform GPL96 and the species is human. The chip data includes gene expression array data of 13 colorectal cancer patients and 13 healthy people.

2.2 Screening of differentially expressed genes:

GEO 2R was used to identify DEGs between the CRC samples and matched with normal samples. The data which $P < 0.05$ and the absolute log2 fold change (log2FC) ≥ 1 or ≤ -1 was considered statistically significant.

2.3 Gene ontology and KEGG pathway analysis

Gene ontology (GO) and KEGG pathway analysis for DEGs contribute to a better understanding of the mechanisms of disease. The Database for Annotation,

Visualization and Integrated Discovery (DAVID, https://david.ncifcrf.gov/), an online web based on the bioinformatics, is routinely applied for annotating genes and protein function [8]. Finally we put the gene symbol into DAVID to acquire the GO and KEGG analysis, $P < 0.05$ as choice criterion.

2.4 *Protein-protein interaction (PPI) network analysis:*

The protein-protein interaction (PPI) network is based on two proteins reported in the literature. Batch analysis of the direct relationship between them, the more the connection, the closer to the central area, the more helpful for screening out possible key genes. The common differential genes were analyzed by STRING [9] (https://string-db.org/) to select genes mainly related to immunity, proliferation and apoptosis.

2.5 *Comparison of the up-regulated and down-regulated Genes:*

GEPIA is a newly developed interactive web server aiming at analyzing the RNA sequencing expression data of 9736 tumors and 8587 normal samples from the

GTEx and TCGA projects in a standard processing pipeline [10]. In this study, we employed the boxplot to visualize the mRNA expression of up-regulated and down-regulated Genes in CRC and normal tissues.

2.6 S*urvival analysis:*

Similarly, we used the GEPIA database to get the overall survival information of these DEGs. The logrank P value and hazard ratio (HR) with 95% confidence intervals were showed on the plot. $P < 0.05$ was statistically significant.

2.7 *Analysis of regulatory relationships between genes:*

In order to further understand the mutual regulation relationship between differential genes, common differential genes have formed a mutual regulatory network between genes through GCBI (https://www.gcbi.com.cn/) analysis. To further screen out potential key genes related to colorectal cancer.

3. Results

3.1 *Differential gene analysis:*

GSE110223 was selected and underwent diferentially expressed genes (DEGs) analysis using GEO2R. The grouping data of specimens in this study was favorable, and the two groups were comparable(Figure1). By analyzing the differential genes of the two gene chip datasets, a total of 866 differential genes were selected in the GSE110223 dataset, 360 genes were up-regulated and 506 genes were down-regulated in the tumor group(Figure 2).

3.2 *GO Term Enrichment Analysis and KEGG Pathway Analysis of DEGs:*

All the signifcantly changed genes were submitted to the DAVID website, we have selected the top ten meaningful ones to do the following analysis. For the the biological process(BP) are Cell proliferation; Male gonad development; One-carbon metabolic process; Positive regulation of cell proliferation; DNA unwinding involved in DNA replication. The cellular component (CC) are Extracellular exosome; Extracellular space; Cytosol; Membrane; Cell surface. For the molecular function(MF) are Protein binding; ATP binding; Chemokine activity; NAD binding; CXCR chemokine receptor binding and the KEGG analysis was performed and showed that genes are mainly involved in Cell cycle; PI3K-Akt signaling pathway; Mineral absorption; MicroRNAs in cancer; Cytokine-cytokine receptor interaction; DNA replication; Pancreatic secretion and Pathways in cancer, as shown in (Figure 3).

3.3 *Protein-Protein Interaction Network*

All Genes were submitted to STRING according to the MF, BP, CC and KEGG pathway we finally found out 15 hubgenes by PPI connective degree(Figure 4). And then we used PCA to filter the samples of the 15 key genes through linear transformation to exclude the difference samples (Figure 5). Finally, we screened out 12 key genes PRKACB, FGFR2, FGFR3, CKB, TIMP1, CCNA2, CCNB1, CDC20、CDC6, CCND1, CDK4 and CDK1, and used heat map to verify the expression in normal tissues and cancer tissues(Figure 6).

3.4 *Regulatory network:*

After 12 genes were analyzed by GCBI with multi-gene radar, we found that the regulatory relationship between these core genes is complex, and these genes play an important mutual regulation role in the survival network. They expected to be a new targets for research（Figure 7）.

3.5 *Validation of DEGs:*

To ensure the credibility of the microarray of GSE110223 and proceed further credible analysis, we validated up-regulated genes and down-regulated genes by GEPIA. The results showed that the mRNA expression of FGFR2, FGFR3 and CKB were significantly lower in Cancer groups compared to Normal groups, while the mRNA expression of TIMP1, CCNA2, CCNB1, CDC20, CDC6, CCND1, CDK4 and CDK1 in Cancer groups were significantly increased than Normal groups (P＜0.05).(Figure 8)

3.6 *Survival curve analysis*

Additionally, we analyzed the potential association between the expression levels of 8 up-regulated genes as well as 4 down-regulated genes and the OS of patients with CRC(Figure 9). It showed that CCNA2 (Logrank p=0.0095,HR=0.53), CCNB1 (Logrank p=0.041,HR=0.6), TIMP1 (Logrank p=0.034, HR=1.7) displayed significantly correlation with the OS of patients with CRC. The high level of CCNA2, CCNB1 and TIMP1 may contribute to a poorer prognosis of CRC.

4. Discussion

Colorectal cancer (CRC) is a malignant tumor with high morbidity and mortality. Most CRC patients are already in the advanced stage when they are diagnosed, because there are no symptoms in the early stage[11-12]. Over the past few years, an increasing number of researches have focused on the molecular pathogenesis of CRC, in order to better define the biological path of CRC development and heterogeneity, and provide more reliable methods for early clinical diagnosis and treatment[13]. However, there is currently no particularly effective and specific CRC diagnosis method and treatment plan, which is mainly due to complex pathogenesis and symptoms that are difficult to diagnose in the early stage[14].

In our study, an integrated bioinformatics analysis was performed to identify the DEGs in CRC. Based on the gene expression profiles of GSE110223 a total of 866 significantly DEGs were identified, including 306 up-regulated and 506 down-regulated genes. The results of GO and the KEGG pathway enrichment

analysis suggested that these genes were significantly enriched in different cancer related functions and pathways.

According to the PPI network by the STRING database and PCA samples analysis, finally we screened out 12 key genes PRKACB, FGFR2, FGFR3, CKB, TIMP1, CCNA2, CCNB1, CDC20□CDC6, CCND1, CDK4 and CDK1, and made use of heat map to verify the expression in normal tissues and cancer tissues. The robust DEGs associated with the carcinogenesis of CRC were screened through the GEO database, and the integrated bioinformatics analysis was conducted in our study. We validated up-regulated genes and down-regulated genes by GEPIA. The results showed that the mRNA expression of FGFR2, FGFR3 and CKB were significantly lower in Cancer groups compared to Normal groups while the mRNA expression of TIMP1, CCNA2, CCNB1, CDC20, CDC6, CCND1, CDK4 and CDK1 in Cancer groups were statistically higher than the Normal groups.

To further research, we have reviewed the relevant literature and have verified the key genes we screened. FGFR signaling drives many downstream pathways, including the mitogen-activated protein kinase (MAPK) and AKT pathways[15], which are crucial for cell proliferation, survival and migration. Therefore, FGFR2 is regarded as a therapeutic target across tumor types[16]. Moreover, FGFR3 silencing inhibited proliferation, migration and invasion of breast cancer cells[17]. In addition, TIMPs are dimers consisting of a smaller C-terminal domain and an N-terminal domain binding to the MMPs substrate and thereby essential for MMPs degradation[18]. TIMP1 is consistently upregulated in the pathological process of CRC and can be a potential biomarker for the worse prognosis of CRC[19]. However, TIMP1 can not only inhibit cancer by repressing MMP expression and activation, but also promotes cancer via angiogenesis, cell growth promotion and tumour inflammation[20]. These studies all suggest the complicated role TIMP1 plays in cancer development. We also found out that the other up-regulated genes including CCNA2, CCNB1, CDC20, CDC6, CCND1, CDK4 and CDK1 are implicated with chemokine signaling pathway and also capable to regulate tumour angiogenesis [21], and survival [22-23].

5. Conclusion

In conclusion, through bioinformatics analysis, we have identified key genes and their important signaling pathways involved in the occurrence and development of CRC. These findings can give us a better understanding of the molecular mechanism of CRC progress. According to our results, we found CCNA2, CCNB1 and TIMP1 were significantly correlated with the overall survival of CRC patients. And they play a vital role in the development of colorectal cancer, which may provide new ideas for future diagnosis and treatment.

6. Abbreviations

GEO:Gene Expression Omnibus GEO; KEGG: Kyoto Encyclopedia of Genes and Genomes; PPI:Protein-protein interaction; GO: Gene Ontology; BP: Biological process; MF: Molecular function; CC: Cellular component;

Declarations

Acknowledgement

The authors gratefully acknowledge financial support from China Scholarship Council.

Authors'contributions

Chaochao Wang:Collected and analyzed the data.Yingchun Hu:Supervised the research. Li Zhang:Guided the instructions and provided revisions. All authors read and approved the final version of the manuscript.

Funding

None.

Availability of data and materials

The datasets used and analyzed in the current study are available from the corresponding author in response to reasonable requests.

Ethics approval and consent

Not applicable

Consent for publication

Not applicable

Competing interests

All authors declare no conflicts and interests in this paper.

Author details

[1]Affiliated hospital of Southwest Medical University, Department of Emergency Medicine,Lu Zhou 646000,SiChuan, China.

[2]Affiliated hospital of Southwest Medical University, Department of Health Management,Lu Zhou 646000,SiChuan, China.

References

1. Testa, U., Pelosi, E., Castelli, G. Colorectal cancer: genetic abnormalities, tumor progression, tumor heterogeneity, clonal evolution and tumor-initiating cells. Med. Sci. (Basel) 2018,6.

2. Torre, L.A., Bray, F., Siegel, R.L. Global cancer statistics, 2012. CA Cancer J Clin.2015 Mar;65(2):87-108.
3. Moghimi-Dehkordi B1, Safaee A. An overview of colorectal cancer survival rates and prognosis in Asia. World Journal of Gastrointestinal Oncology, 2012 Apr 15;4(4):71-5.
4. Zheng He, Lianhua Yu, Shiyi Luo. miR-296 inhibits the metastasis and epithelial-mesenchymal transition of colorectal cancer by targeting S100A4, BMC Cancer, 2017 Feb 16:140–147.
5. Sun GG, Wang YD, Cui DW. Epithelial membrane protein 1 negatively regulates cell growth and metastasis in colorectal carcinoma.World J Gastroenterol,2014 ,20: 4001-4010.
6. Z. Tang, C. Li, B. Kang. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses, Nucleic Acids Res, (2017), 98–102.
7. L. Ling, L. Ning, C. He. Proteomic analysis of differentially expressed proteins in kidneys of brain dead rabbits, Mol. Med. Rep, (2017), 215–223.
8. G. Dennis, B. T. Sherman, D. A. Hosack. DAVID: Database for Annotation, visualization, and Integrated Discovery, Genome Biol, (2003), 3.
9. Szklarczyk D, Morris JH, Cook H. The STRING database in 2017: quality-controlled protein-protein association networks□made broadly accessible□Nucleic Acids □es□2017□45(D1) : D362-D368.
10. Z. Tang, C. Li, B. Kang. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses, Nucleic Acids Res, (2017), 98–102.
11. Menon M, Cunningham C, Kerr D. Addressing unwarranted variations in colorectal cancer outcomes: a conceptual approach. Nat Rev Clin Oncol. 2016;13(11):706–712.
12. Markowitz SD, Bertagnolli MM. Molecular origins of cancer: molecular basis of colorectal cancer. N Engl J Med. 2009;361(25):2449– 2460.
13. Pekow J, Meckel K, Dougherty U. miR-193a-3p is a key tumor suppressor in ulcerative colitis-associated colon cancer and promotes carcinogenesis through upregulation of IL17RD. Clin Cancer Res. 2017;23(17):5281–5291.
14. P. Deenadayalu and D. K. Rex. Colorectal cancer screening: a guide to the guidelines, Rev. Gasteroenterol. 2007 Fall;7(4):204-13.
15. M. Lau, E. Teng, K.K. Huang. Acquired resistance to FGFR inhibitor in diffffuse-type gastric Cancer through an AKT Independent PKC-Mediated phosphorylation of GSK3beta. Mol. Cancer Ther.(2018) 232–242.
16. Joon Young Hur , Joseph Chao, Kyung Kim. High-level FGFR2 amplifification is associated with poor prognosis and Lower response to chemotherapy in gastric cancers. Pathol Res Pract.2020 Apr;216(4):152878.
17. Xinghua Long, Yu Shi, Peng Ye. MicroRNA-99a Suppresses Breast Cancer Progression by Targeting FGFR3. Front Oncol. 2019; 9: 1473.

18. Grunnet M, Mau-Sorensen M, Brunner N. Tissue inhibitor of metalloproteinase 1 (TIMP-1) as a biomarker in gastric cancer: a review. Scand J Gastroenterol. 2013;48(8):899–905.
19. Ru Huang, Kaijing Wang, Lei Gao. TIMP1 Is A Potential Key Gene Associated With The Pathogenesis And Prognosis Of Ulcerative Colitis-Associated Colorectal Cancer. OncoTargets and Therapy 2019:12 8895–8904.
20. Jackson HW, Defamie V, Waterhouse P. Khokha R. TIMPs: versatile extracellular regulators in cancer. Nat Rev Cancer. 2017;17(1):38-53.
21. Chuanyong Wu, Xiao-ting Zhu, Lei Xia. High Expression of Long Noncoding RNA PCNA-AS1 Promotes Non-Small-Cell Lung Cancer Cell Proliferation and Oncogenic Activity via Upregulating CCND1. Journal of Cancer 2020;11(7):1959-1967.
22. Jianxin Li, Yinchun Wang, Xin Wang. CDK1 and CDC20 overexpression in patients with colorectal cancer are associated with poor prognosis:evidence from integrated bioinformatics analysis. World Journal of Surgical Oncology (2020) 18:50.
23. Tian Gao, Yong Han, Ling Yu. CCNA2 Is a Prognostic Biomarker for ER+ Breast Cancer and Tamoxifen Resistance. PLoS One. 2014 Mar 12;9(3):e91771.

Figures

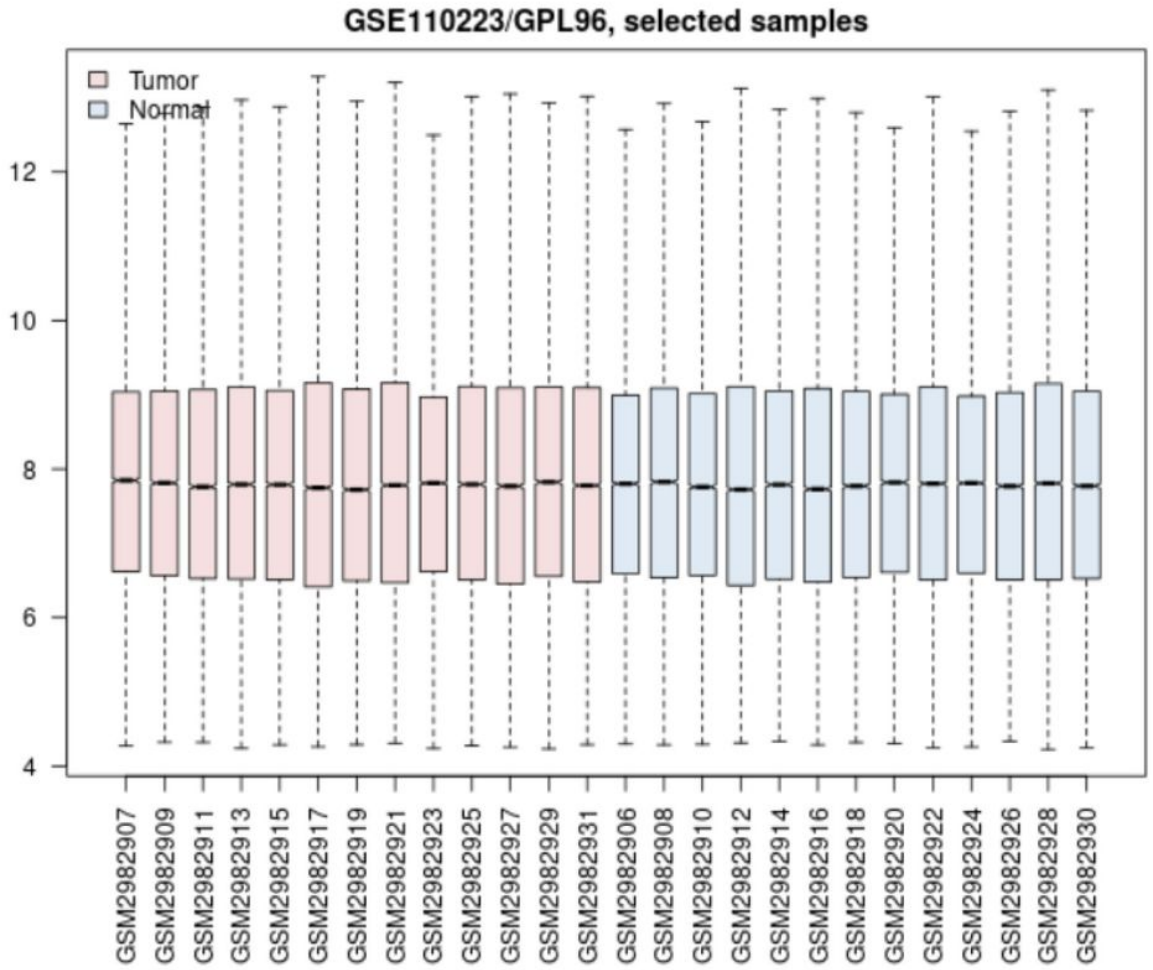

Figure 1

The grouping data of specimens shows the two groups are comparable.

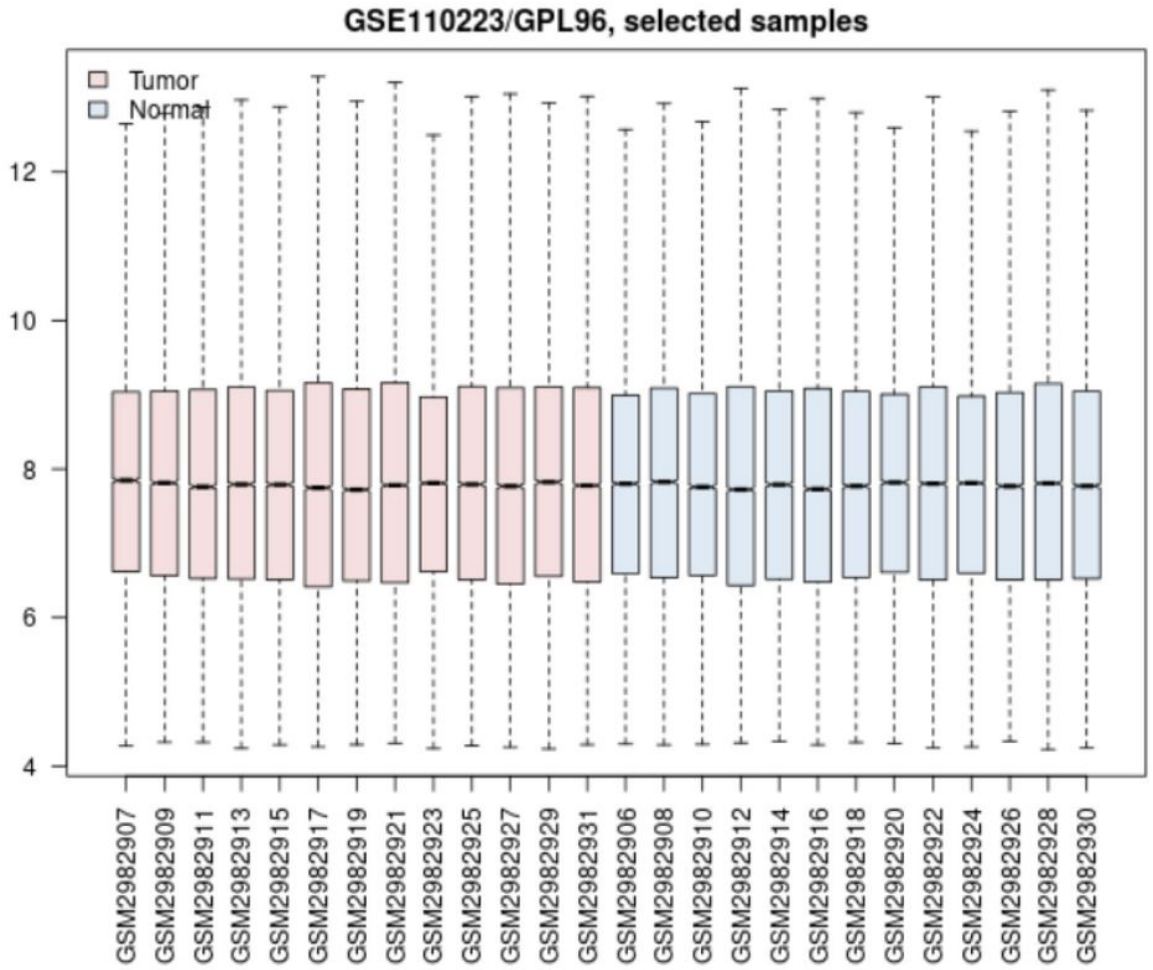

Figure 1

The grouping data of specimens shows the two groups are comparable.

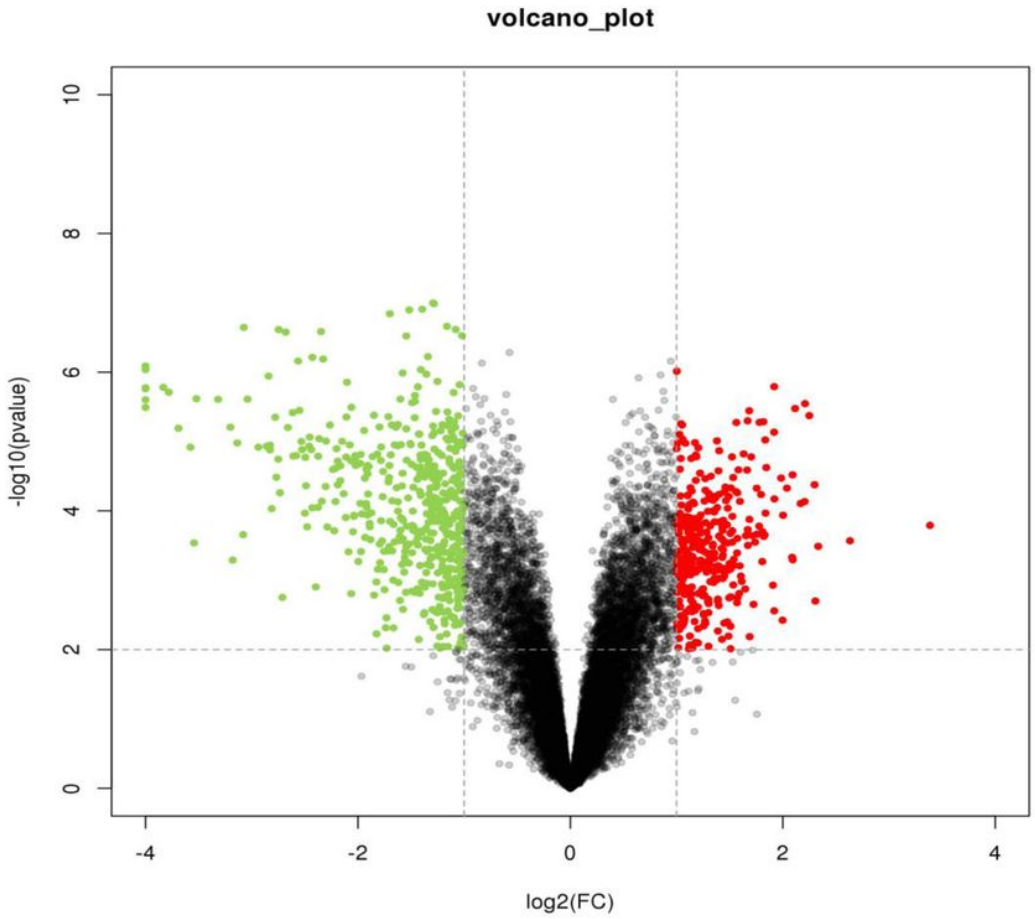

Figure 2

Volcano plot of genes detected in CRC. Green means down-regulated DEGs; Red means up-regulated DEGs; Black means no difference.

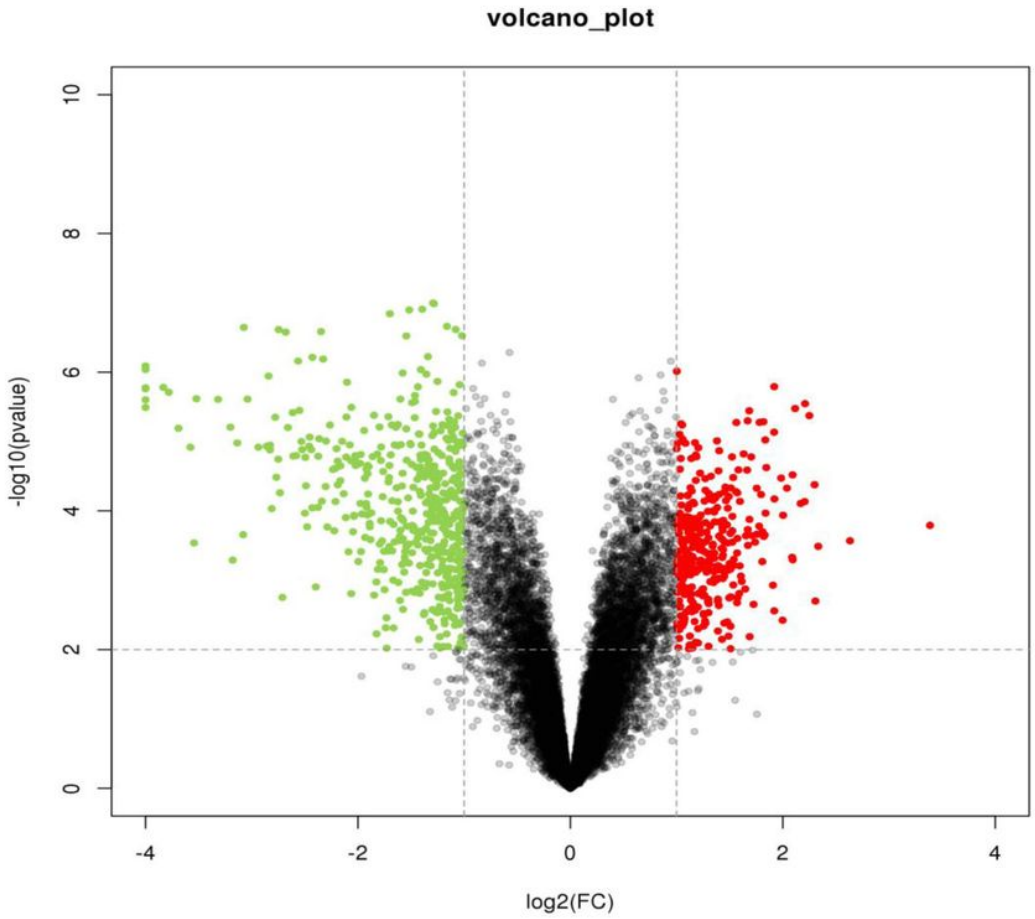

Figure 2

Volcano plot of genes detected in CRC. Green means down-regulated DEGs; Red means up-regulated DEGs; Black means no difference.

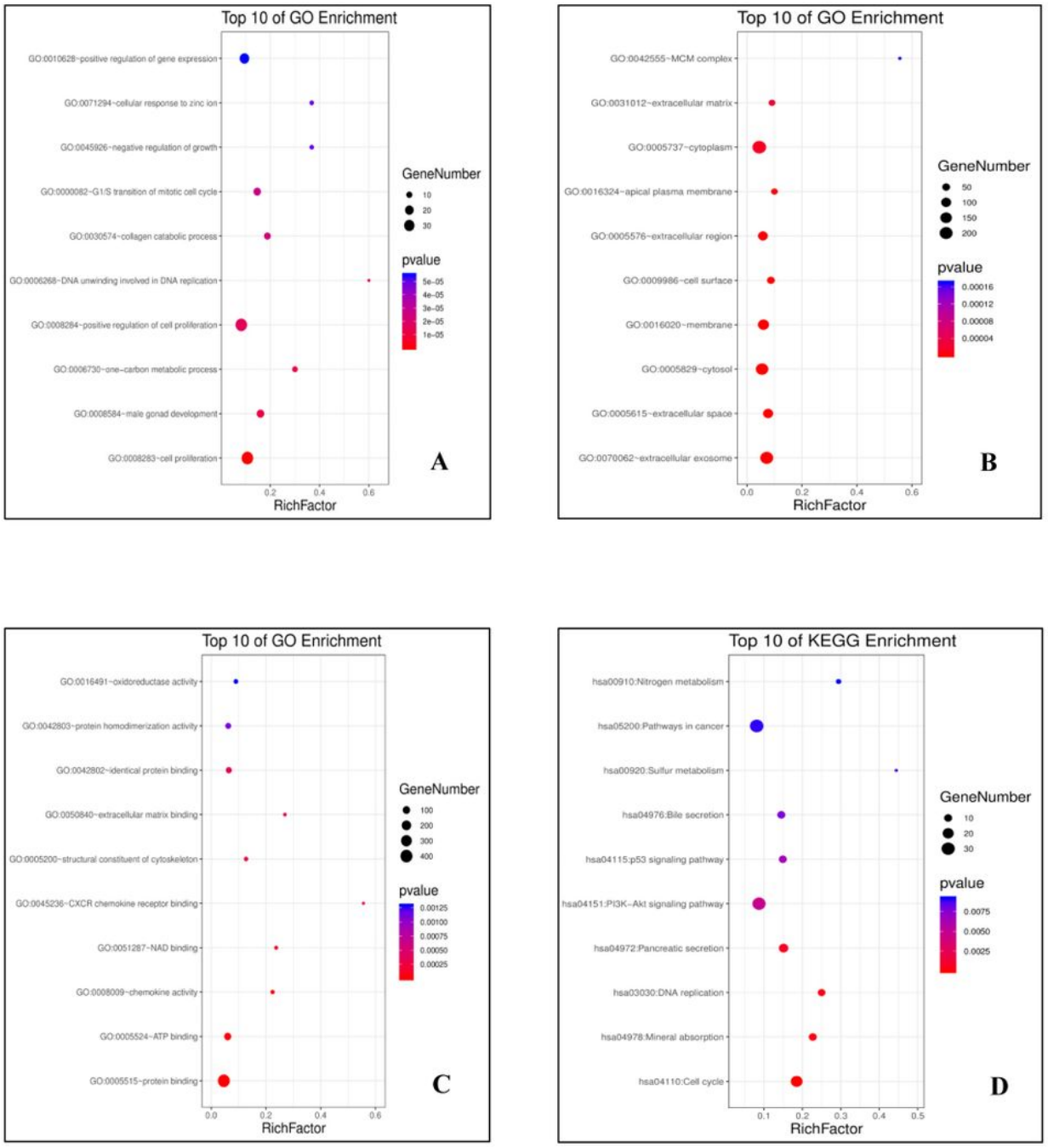

Figure 3

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of CRC. (A) The enriched GO terms in the biological process (BP); (B) The enriched GO terms in the cellular component (CC); (C) The enriched GO terms in the molecular function (MF); (D) The enriched KEGG pathway in CRC.

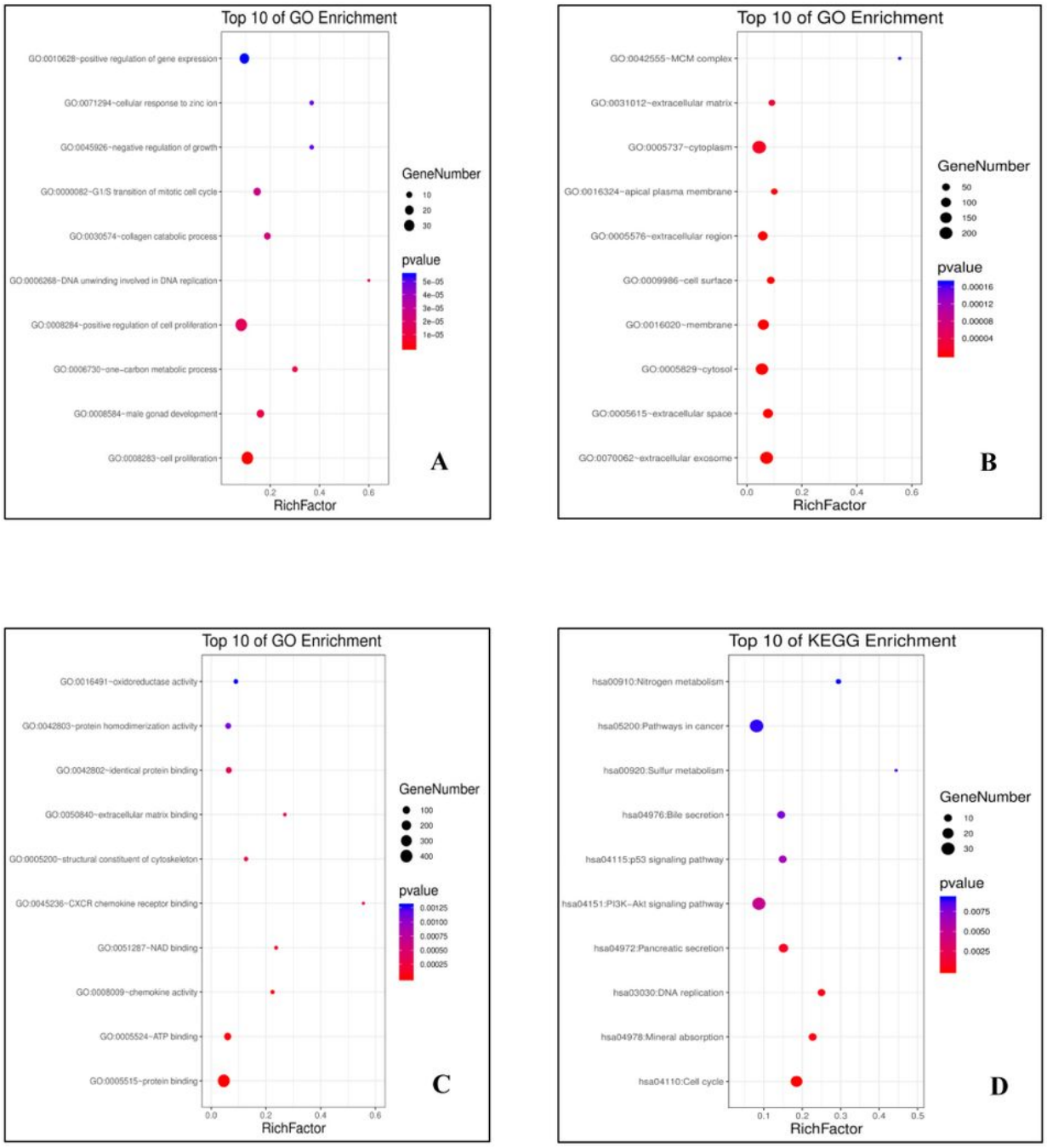

Figure 3

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of CRC. (A) The enriched GO terms in the biological process (BP); (B) The enriched GO terms in the cellular component (CC); (C) The enriched GO terms in the molecular function (MF); (D) The enriched KEGG pathway in CRC.

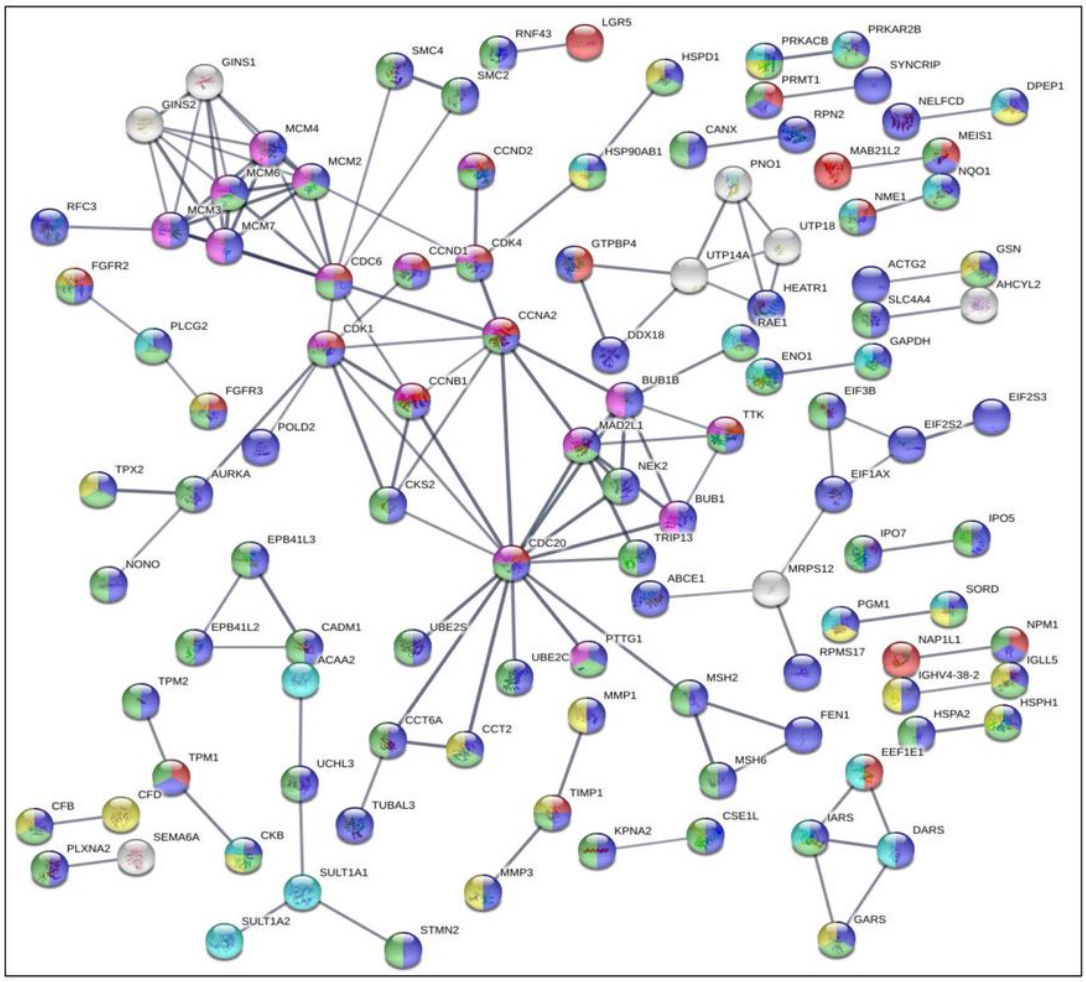

Figure 4

Construction of protein-protein interaction (PPI) network. PPI enrichment p-value:< 1.0e-16

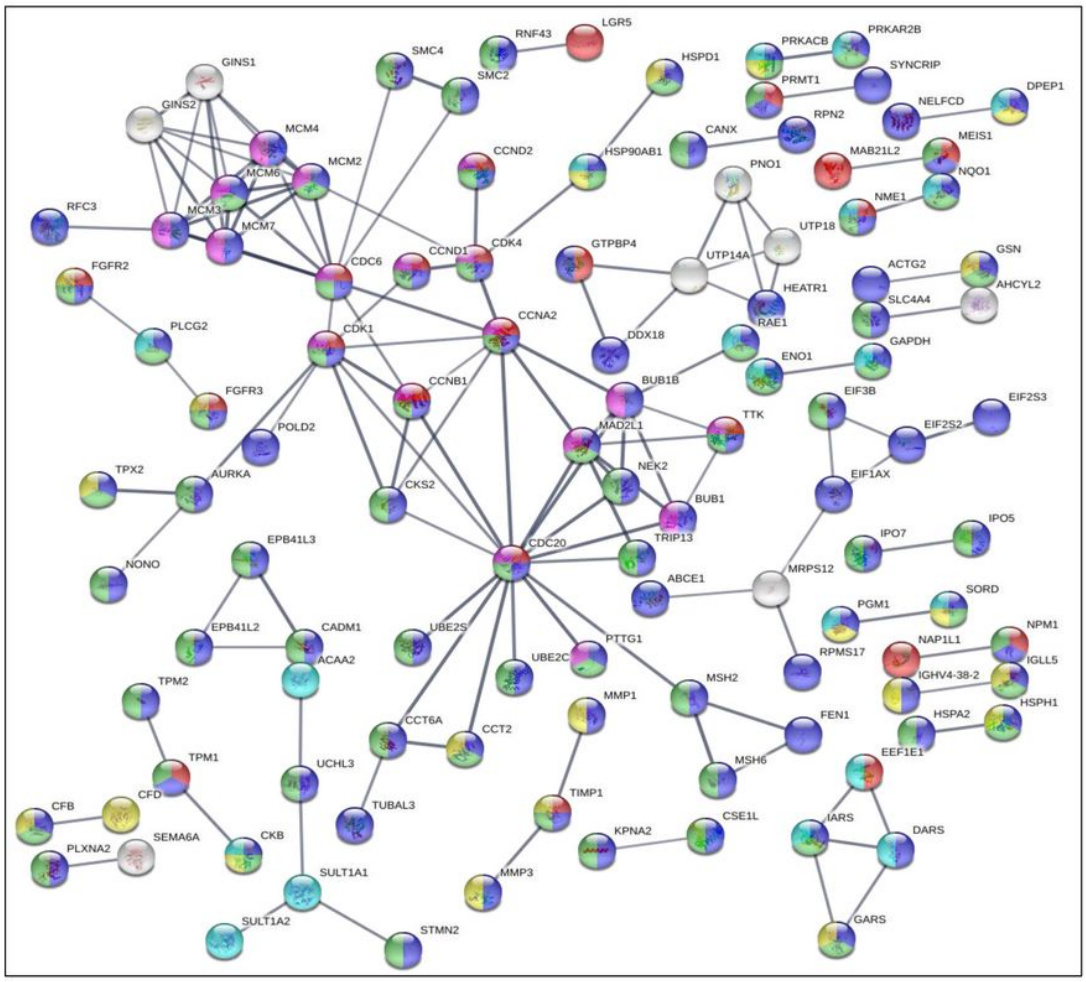

Figure 4

Construction of protein-protein interaction (PPI) network. PPI enrichment p-value:< 1.0e-16

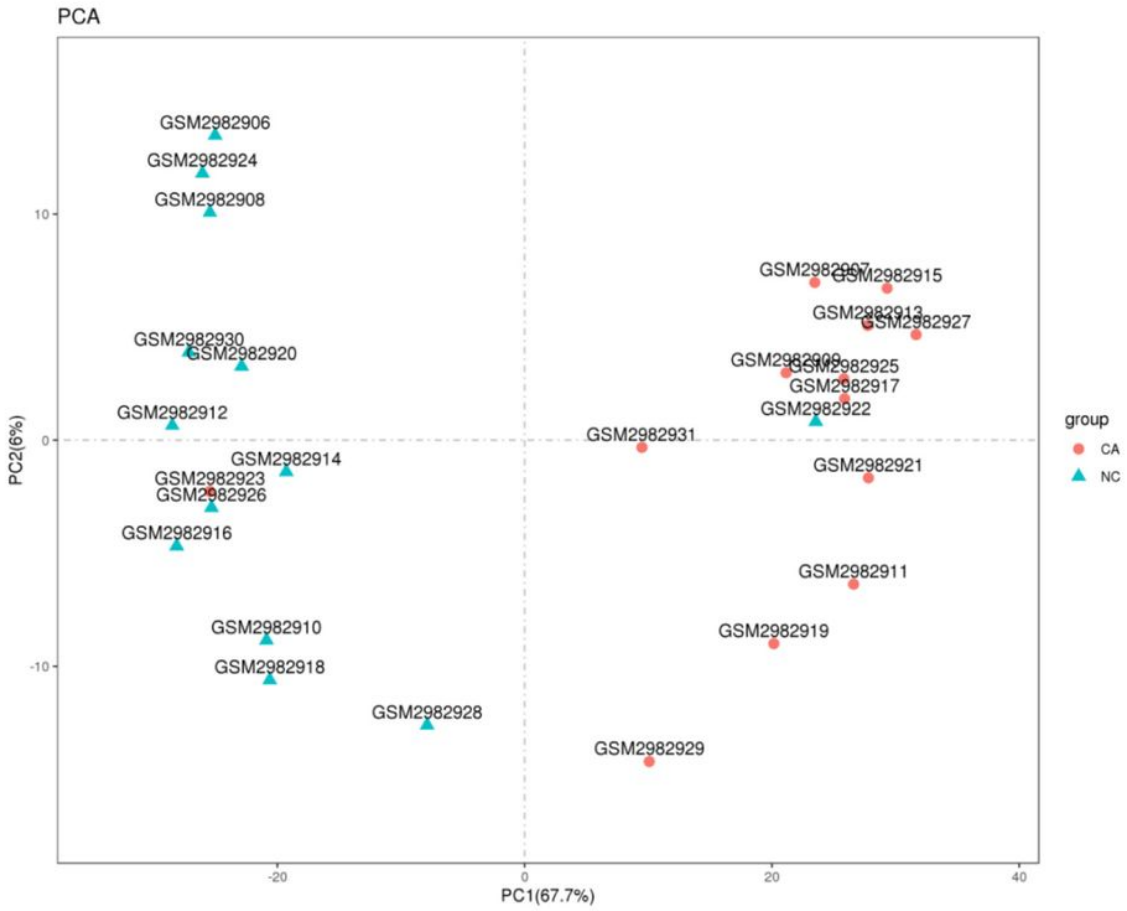

Figure 5

Samples GSM2982923 and GSM2982922 have an abnormal distribution

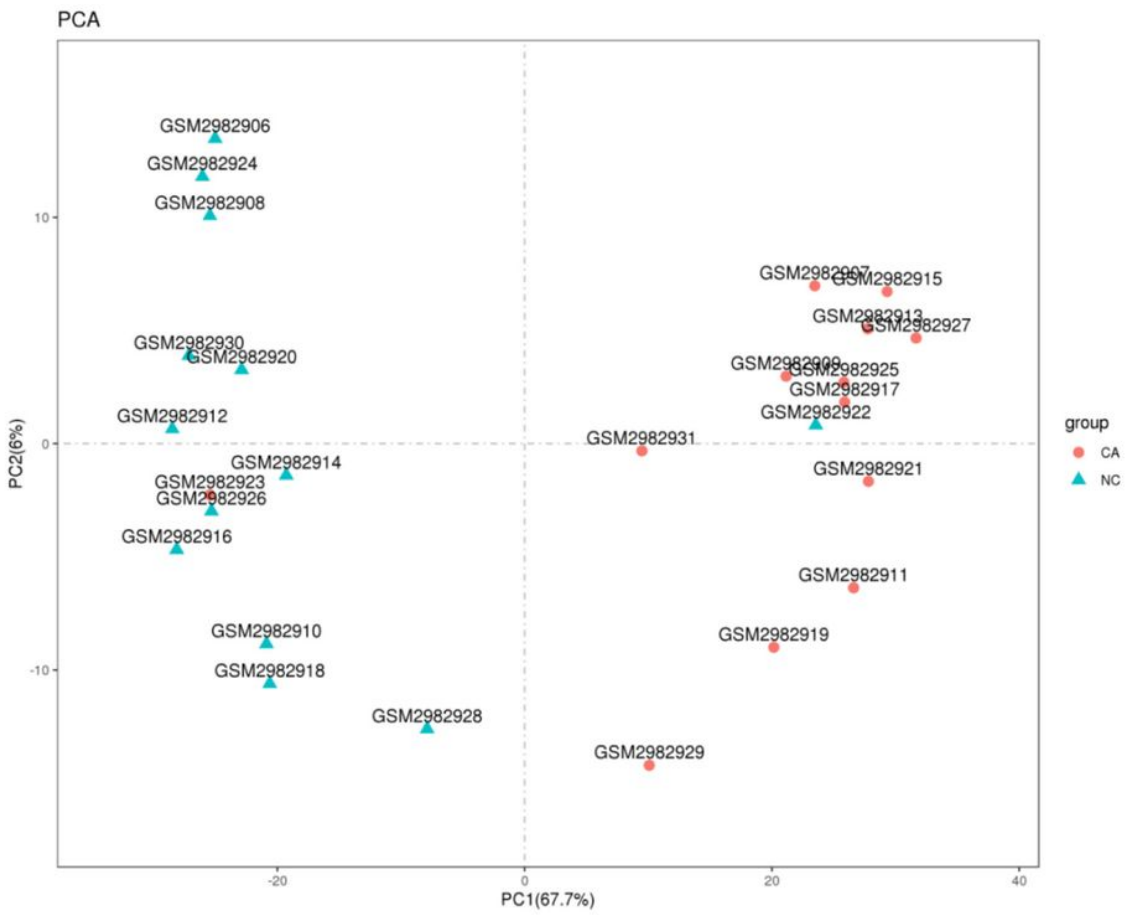

Figure 5

Samples GSM2982923 and GSM2982922 have an abnormal distribution

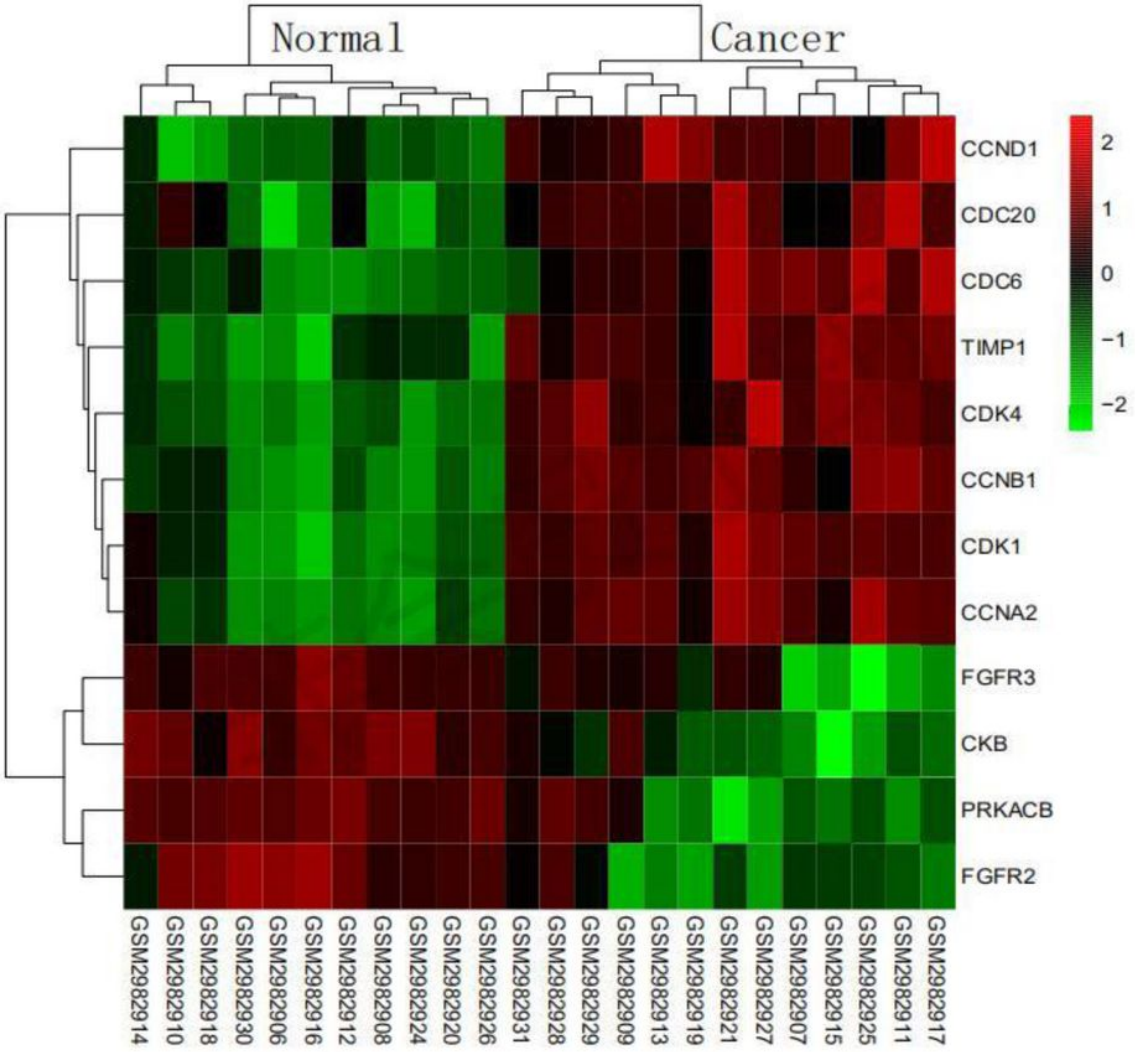

Figure 6

Heatmap of the 12 key genes. Red ones represented up-regulation and Green ones represented down-regulation.

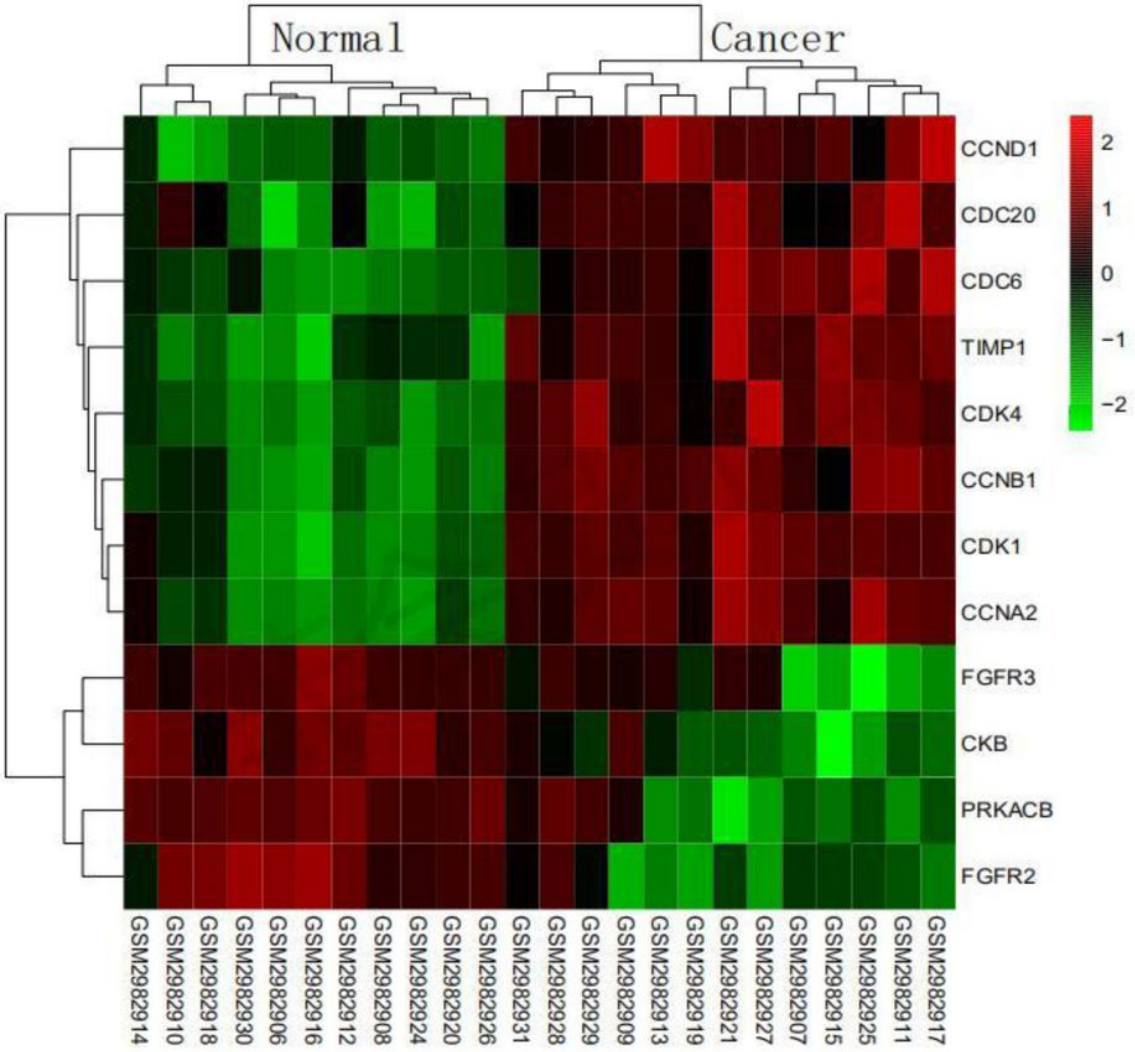

Figure 6

Heatmap of the 12 key genes. Red ones represented up-regulation and Green ones represented down-regulation.

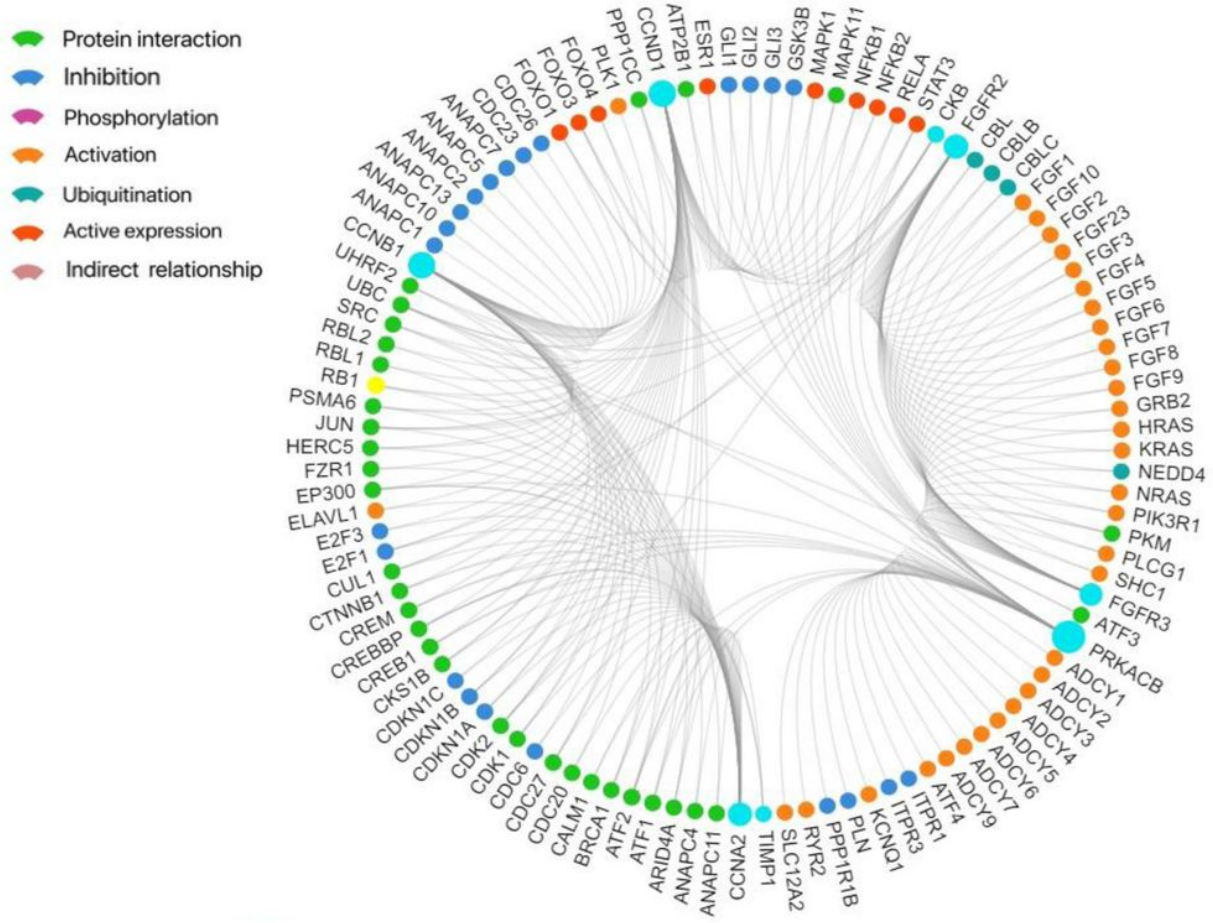

Figure 7

Key genes polygene regulatory network

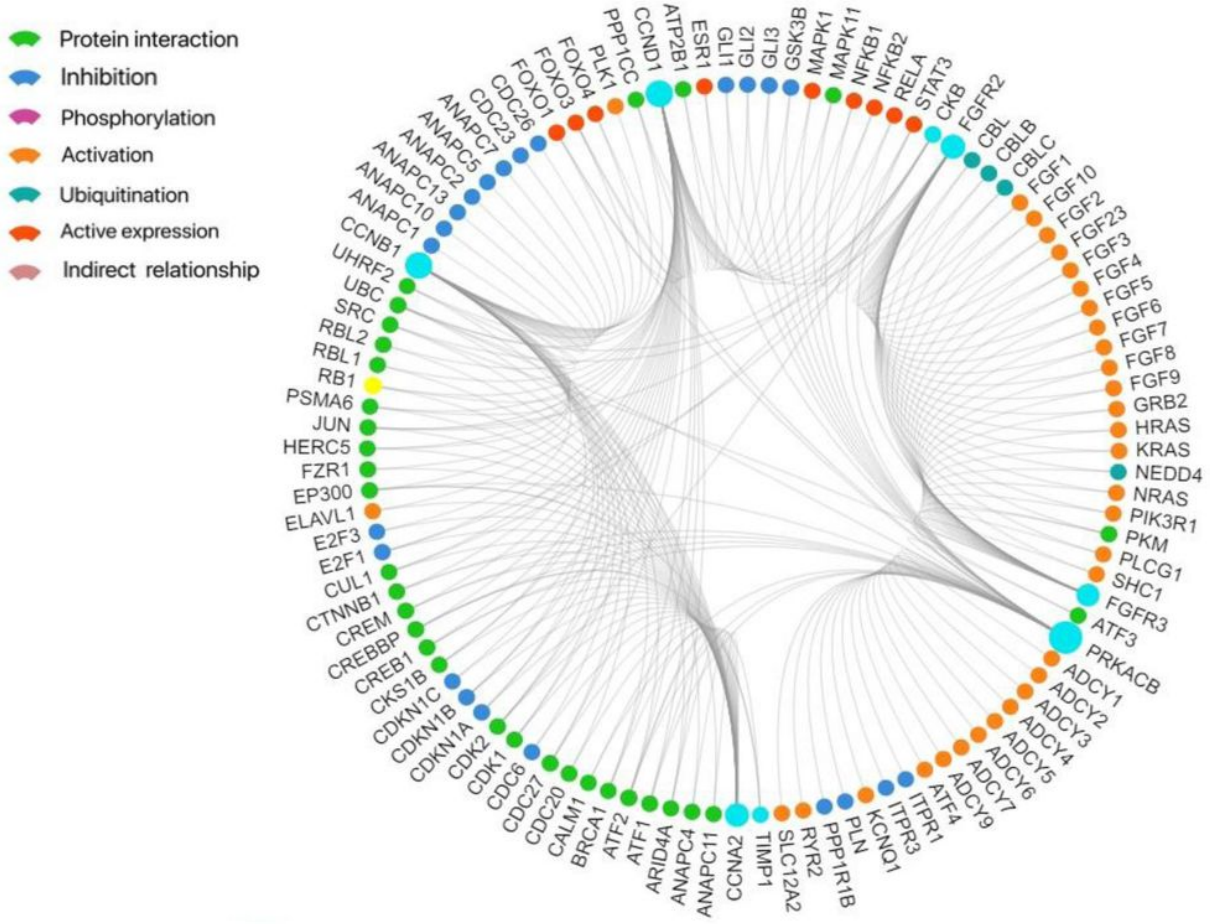

Figure 7

Key genes polygene regulatory network

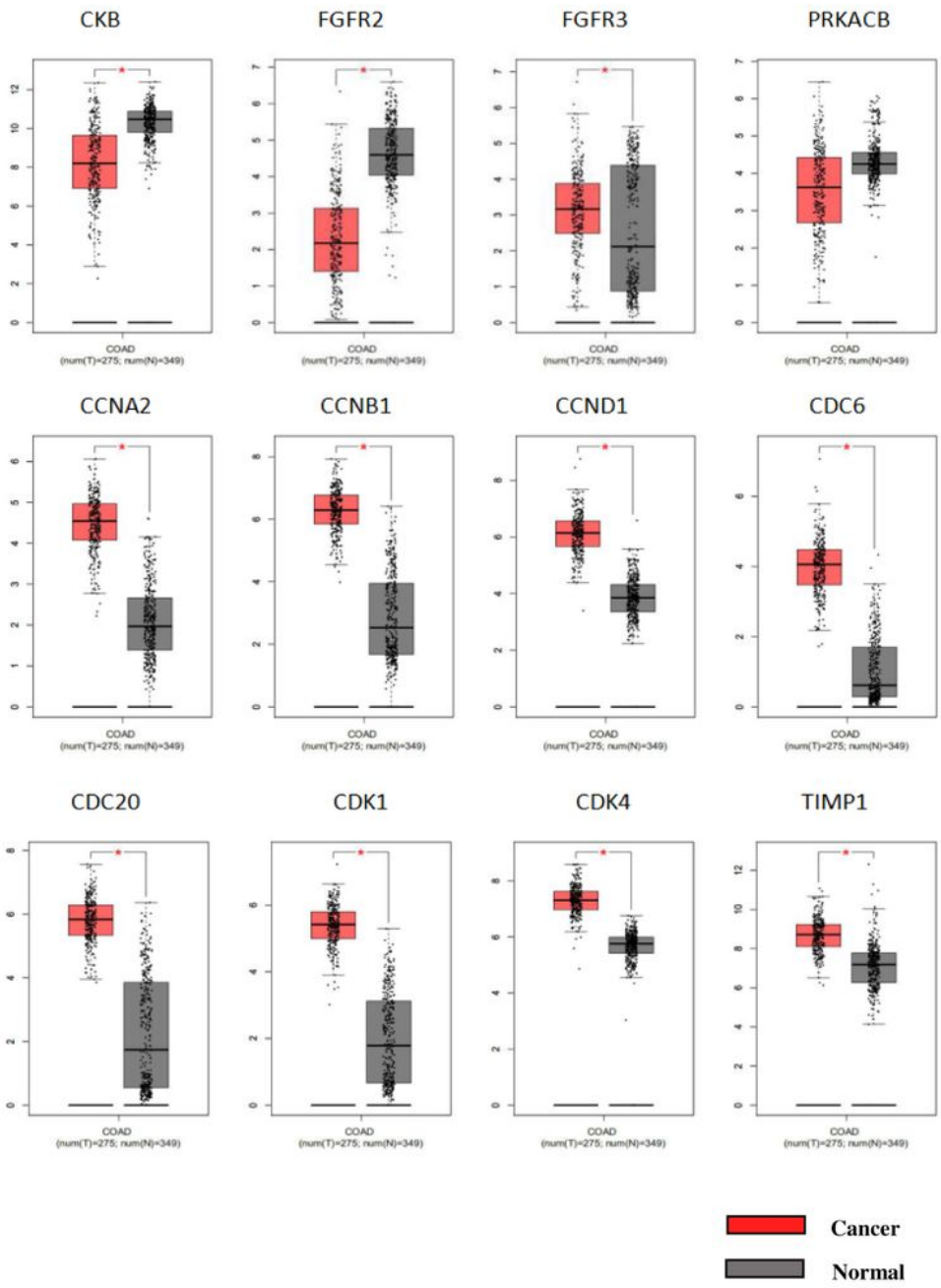

Figure 8

The mRNA expression of 8 up-regulated and 4 down-regulated genes based on TCGA database.* P < 0 05 was regarded statistically different.

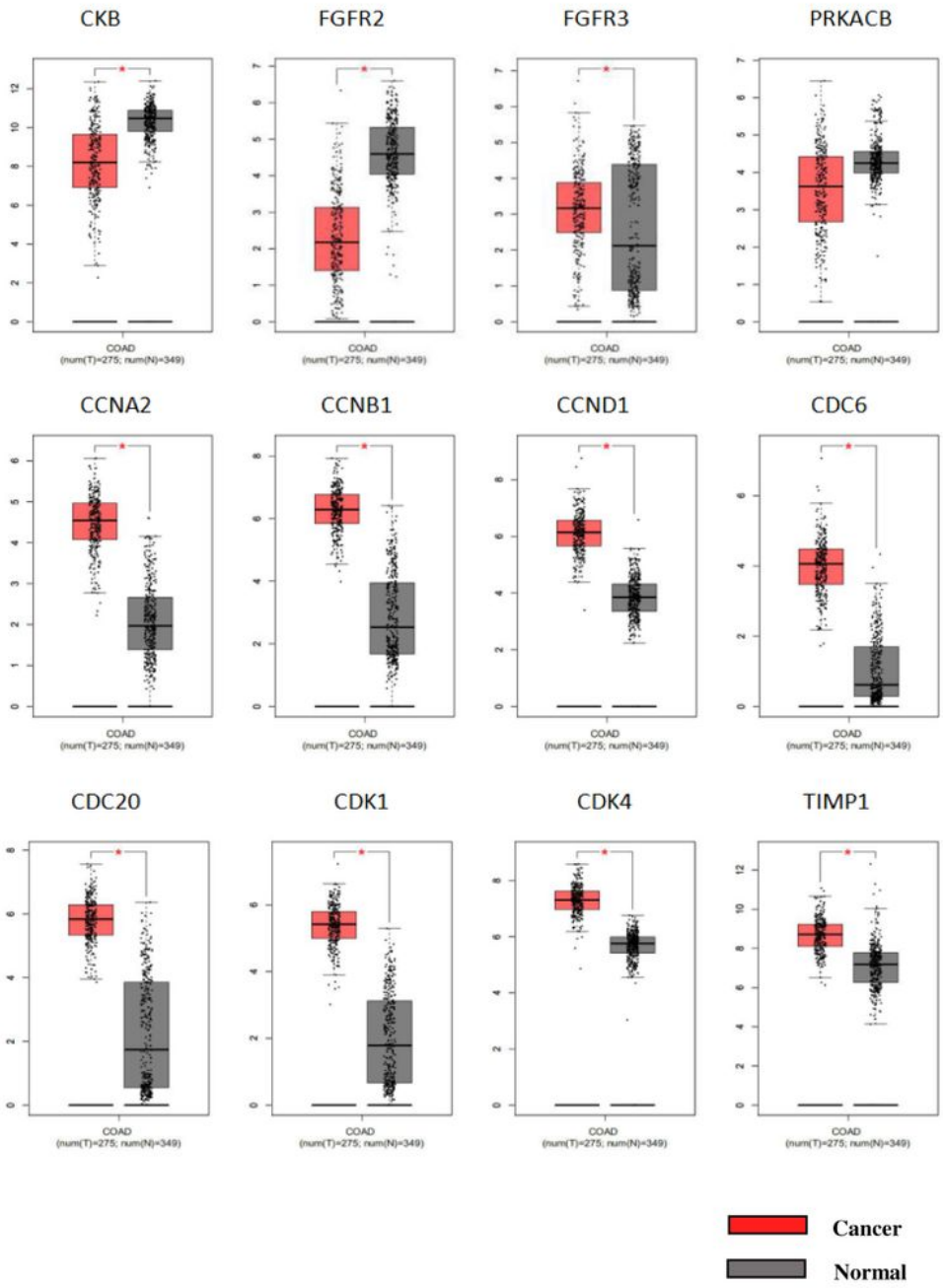

Figure 8

The mRNA expression of 8 up-regulated and 4 down-regulated genes based on TCGA database.* P < 0 05 was regarded statistically different.

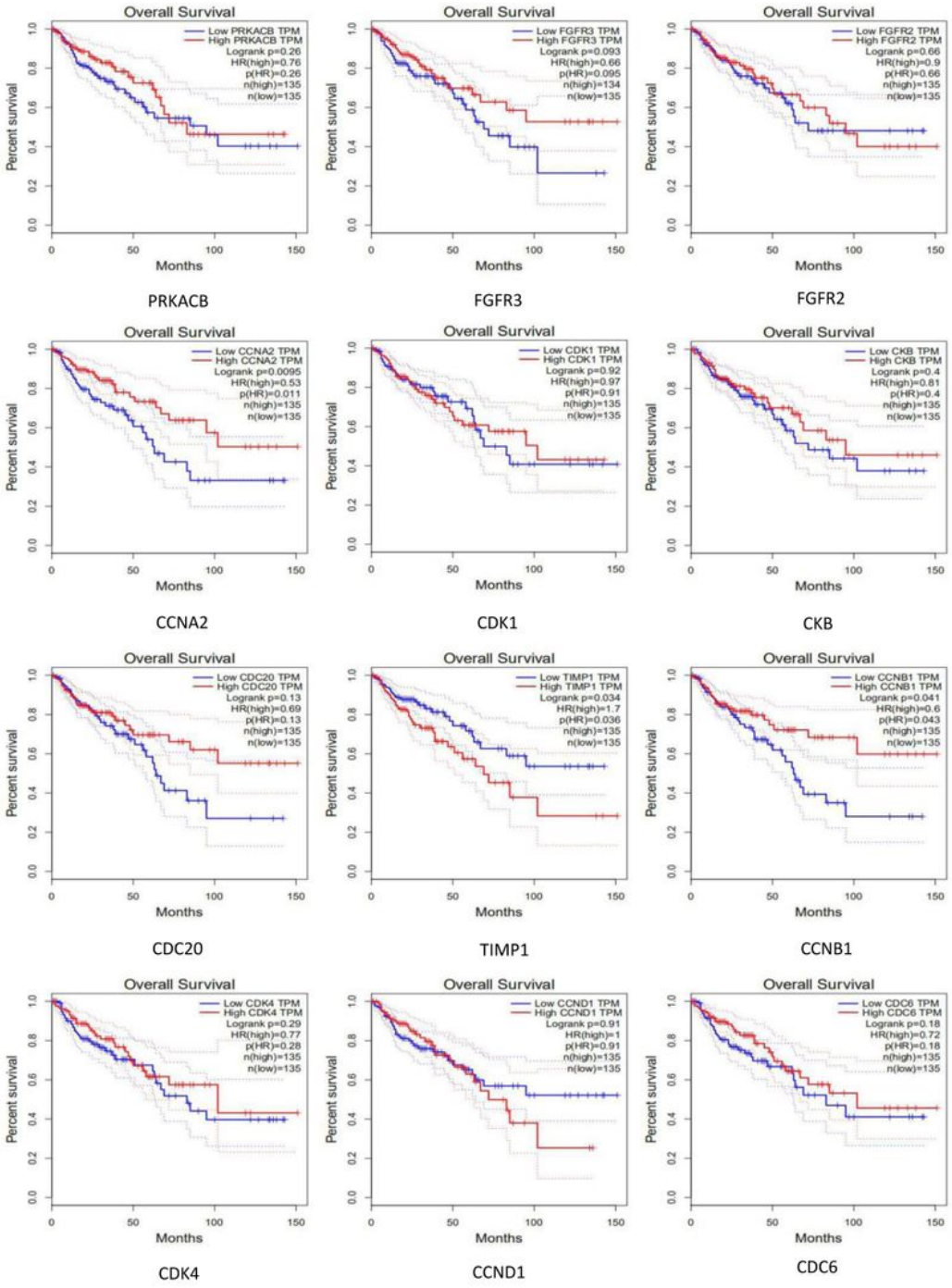

Figure 9

Prognostic value of 8 up-regulated and 4 down-regulated genes. P < 0.05 was regarded statistically different.

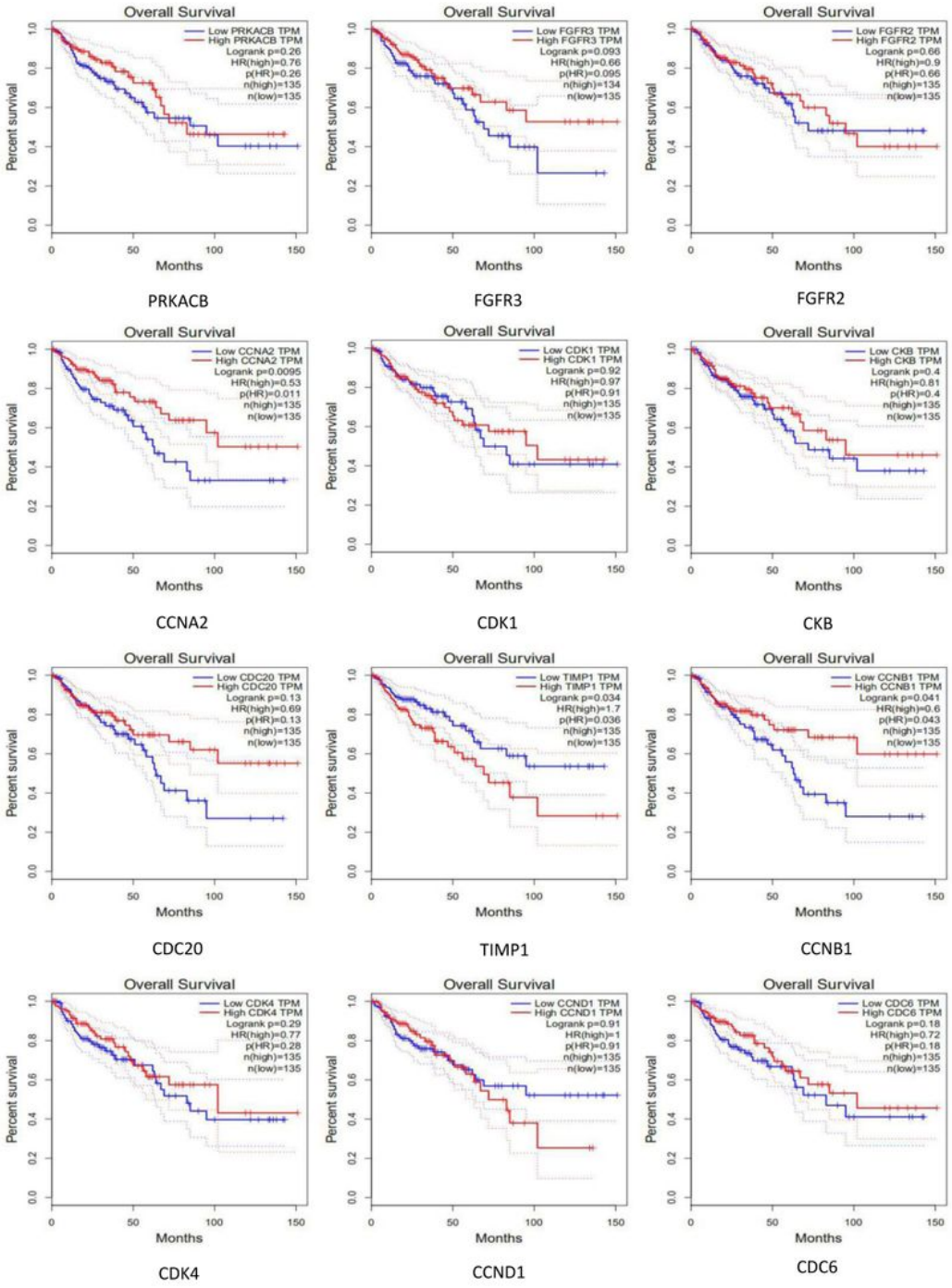

Figure 9

Prognostic value of 8 up-regulated and 4 down-regulated genes. P < 0.05 was regarded statistically different.