Study on the Pathogenesis of Prostate Cancer Based on Network Analysis

Hongyu Chen, Tingchun Shi*

School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China
**Corresponding Author*

Abstract. *Prostate cancer(PRAD) is one of the most common malignant tumor diseases in men around world. In this study, we chose PRAD as research object, establish a set of bioinformatics algorithm on the basis of the network analysis, and excavate the main biological processes as well as key targets for the occurrence and development of PRAD from a new perspective, so as to provide a new solution for the research on the treatment of PRAD. Specifically, this study chose the transcriptome data of PRAD contained in the Cancer Genome Atlas(TCGA) database as the research object, and apply the FC-t algorithm and Pearson correlation coefficients to analyze and build the gene co-expression network. Moreover, we performed Mcode algorithm and the GO/Reactome enrichment analysis so as to excavate the main biological processes in the development of PRAD. Consequently, we find five genes (C2orf72, CAMKK2, FGFRL1, HPN, UCN) which play key roles in PRAD, and six types of biological processes related to the development of the PRAD. These provide new ideas for subsequent research on PRAD.*

Keywords: *bioinformatics algorithms, gene coexpression networks, prostate cancer, network analysis*

1. Introduction

Research on the treatment of PRAD has always been a hot-spot in cancer research. Raj Satkunasivam et al performed cox proportional hazards models to find that radical prostatectomy and intensity modulated radiation therapy without conformal radiation therapy is closely related to the survival benefit of patients with metastatic PRAD[1]. Archana et al found that LSD1 blocked important demethylase independent functions and inhibited the survival of castration-resistant PRAD cells through a small molecule LSD1 inhibitor SP-2509[2]. Chakravarthi et al performed next-generation sequencing, qRTPCR, and western blot analysis to find that PAICS genes play an key role in the proliferation and invasion of PRAD cells[3]. It was infered that the current research on PRAD mainly focuses on two aspects, one was to performed regression models to mine key factors closely related to prognosis, and the other is to mine the key genes closely related to survival prognosis through transcriptomics or next-generation sequencing technology.

The research on tumor data mainly uses the WGCNA algorithm to analyze the main mechanisms and key pathogenic targets of tumor development nowadays. Xue et al performed the WGCNA algorithm to analyze GEO datasets(GSE53819, GSE12452 and GSE64634) and found that c9orf24, PCDP1 and LRRC46 were closely related to nasopharyngeal carcinoma[4]. Xiao et al performed the WGCNA algorithm to analyse renal cell carcinoma transcriptome data on GEO database and TCGA database and found that EHHADH, ACADM and AGXT2 are closely related to T stage and prognostic survival status[5]. Huang et al performed WGCNA algorithm to explore the main mechanism of peripheral nerve infiltration, and found that the gene module positively related to peripheral nerve infiltration were mainly highly enriched during the cell cycle, and determined that cancer cell proliferation was common to the neurocancer microenvironment reaction[6].

The occurrence and development of tumors is an intricate biological process. In this study, a bioinformatics algorithm based on network analysis were performed to dig out the main biological processes and key targets of the occurrence and development of PRAD. Firstly, we chose the transcriptome data of PRAD in the Cancer Genome Atlas(TCGA) database[7] as the research subject to remove low-expressing genes from the datasets. Then, the FC-t algorithm was performed to identify significantly differentially expressed genes(DEGs) between cancer tissues and paracancer tissues. Furthermore, we computed the Pearson correlation coefficients between DEGs, and built a gene co-expression network(GCN). Later, We performed the Mcode plug-in in cytoscape software to divide the gene co-expression network into gene modules. Next, we performed GO / Reactome enrichment analysis on gene modules to mine the main biological processes in the occurrence and development of PRAD. Finally, we chose 20 key genes with the smallest p values as the research object and the oncomine database was performed to mine the main functions of the key genes.

2. Methodology

2.1 Data preprocessing and sample evaluation

The original dataset of this study was chosen the PRAD as the research object that were derived from TCGA database (https://portal.gdc.cancer.gov/projects/) which including cancer tissues transcriptome data and paracancer tissues transcriptome data. Cancer tissues transcriptome data chose 551 samples including the FPKM value of 60,483 genes. Paracancer tissues transcriptome data chose 52 samples including the FPKM value of 60,483 genes(Supplementary Material, Table S1-S2). Then, zero-expressed genes in the transcriptome data of cancer and paracancer tissues were remove

2.2 Bioinformatics algorithm

To explore the pathogenesis of PRAD, we used a bioinformatics algorithm based on network analysis(Fig. 1) to analyze the gene expression profile of PRAD.

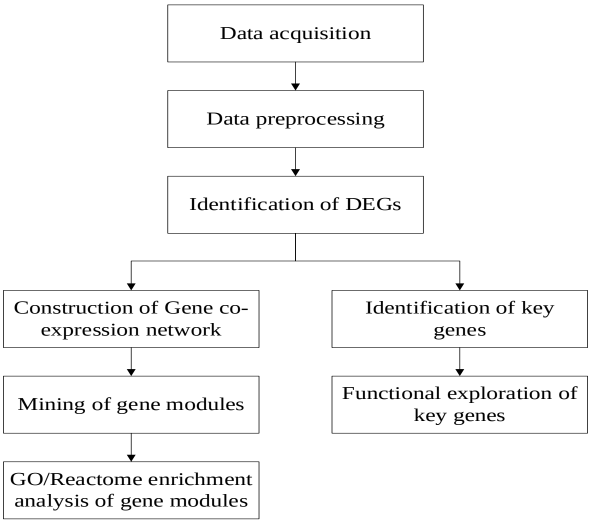
Data acquisition
Data preprocessing
Identification of DEGs
Construction of Gene co-expression network
Identification of key genes
Mining of gene modules
Functional exploration of key genes
GO/Reactome enrichment analysis of gene modules

Figure. 1 Flow chart of target tracking process

2.3 Analysis of differentially expressed genes

FC-t algorithm was used to identify the differentially expressed genes (DEGs) in PRAD. The cut-off value was set at Fold change >= 1.5 or Fold change <= 0.67 and P value <= 0.05.

2.4 Construction of gene co-expression network

In this study, Pearson correlation analysis was used to construct a interaction matrix among DEGs. The cut-off value was set at |Pearson correlation coefficient|>= 0.7 and p value < 0.05 to screen the interactions between the pairs of genes. Two genes that met this qualification were considered to be co-expressed, and a GCN was constructed based on these interactions.

2.5 Mining of gene modules

To obtain gene modules composed of genes with similar functions, the Mcode plug-in in Cytoscape software were used to divided the GCN into modules.

2.6 GO / Reactome enrichment analysis of gene modules

To explore the biological significance of each gene module, the genes contained in each module were enriched with the biological processes (BPs) provided by the GO database (http://geneontology.org/) and the signaling pathways provided by the Reactome database (https://reactome.org/). The 10 BP Terms and 10 signaling pathways with the smallest P value in each module were selected for further research.

2.7 Identification of key genes and exploration of their functions

The 20 genes with the lowest P value on the T-test were defined as key genes in PRAD. Further, to explore the functions of these key genes, the Oncomine database (https://www.oncomine.org/) was used to query the expression of key genes in common cancers.

3. Results and discussion

3.1 Sample quality control analysis

We removed zero-expressed genes in the transcriptome data of cancer and paracancer tissues and kept 20,046 genes (Supplementary Material, Table S3-S4).

3.2 Identification of DEGs using FC-t algorithm

We performed FC-t algorithm to identify the DEGs and got 4865 DEGs which met appropriate thresholds (Fold change >= 1.5 or Fold change <= 0.67 and P value <= 0.05) filtering (Supplementary Material, Table S5). Notably, all of DEGs were up-regulated.

3.3 Construction of GCN using Pearson correlation analysis

We performed Pearson correlation analysis on the the FPKM values of DEGs (5153) in cancer tissue transcriptome data. Then, the interactions failed to meet the chosen appropriate threshold (|Pearson correlation coefficient| < 0.7, P value >= 0.05) were removed, and 3521 genes were kept (Fig. 2).

These genes formed a large net and several small nets. The large net contained 3,155 genes, and each small net contained less than 20 genes. We removed all of small nets and remained the large net (GCN) for further research.

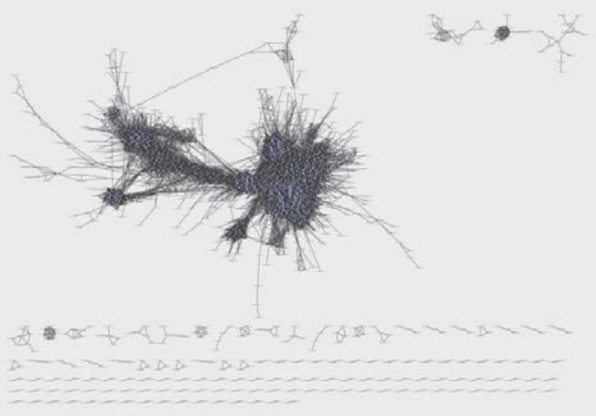

Figure. 2 Pearson correlation analysis of DEGs

3.4 Mining of gene modules

We used Mcode to divide the GCN into 78 modules (Supplementary Material, Table S6). We removed the modules containing less than 50 genes, leaving 9 gene modules. The number of genes contained in each gene module is shown in Table 1.

Table 1 The number of genes contained in each gene module

Module	Gene Number
m1	240
m2	289
m3	89
m4	221
m5	70
m6	136
m7	68
m8	100
m9	55

3.5 GO/Reactome enrichment analysis of gene modules

We'd like to check out the biological functions of the 9 modules above. GO enrichment analysis was performed on each one of them. the BPs obtained from module m1 are mainly related to the metabolism of extracellular matrix. The BPs obtained by module m2 are mainly related to the development of muscle tissue and the proliferation and differentiation of muscle cells. The BPs obtained from module m3 are mainly related to the process of cell mitosis. The BPs obtained from module m4 are mainly related to the cardiac tissue contraction. The BPs obtained from module m5 are mainly related to the energy metabolism and the expression of genetic material. The BPs obtained from module m6 are mainly related to the formation of tissues and organs. The BPs obtained from the m8 module are mainly related to the development of the urogenital system. It is worth noting that modules m7 and m9 have no enrichment results.

Then, Reactome enrichment analysis was performed on each one of them. The signaling pathways obtained by module m1 are mainly related to epidermal growth factor-mediated signaling. The signaling pathways obtained by module m2 are mainly related to muscle contraction and laminin interaction. The signaling pathways obtained by module m3 are mainly related to the process of mitosis. The signaling pathways obtained by module m4 are mainly related to TGFBR-mediated signal transmission and VEGF-mediated signal transmission. The signaling pathways obtained by module m5 are mainly related to the transcription and translation of genetic material. The signaling pathways obtained by module m6 are mainly related to the differentiation of osteoblasts and the Hippo signaling pathway. The signaling pathways obtained by module m9 is mainly related to the metabolism of lipids. It is worth noting that modules m7 and m8 have no enrichment results.

3.6 Identification of key genes and exploration of their functions

We defined the 20 genes with the lowest P value on the T-test as key genes in PRAD (Table 2).

Table 2 key genes in PRAD

Ensembl_gene_id	External_gene_name	FC	P-value
ENSG00000163794	UCN	4.076182115	7.99E-55
ENSG00000223400	AP006748.1	9.770659301	1.28E-53
ENSG00000272894	AC004982.2	3.642143313	3.51E-48
ENSG00000105707	HPN	4.79485988	9.77E-48
ENSG00000231806	PCAT7	5.222285384	3.23E-47
ENSG00000272732	AC004982.1	3.659136047	4.22E-42
ENSG00000127418	FGFRL1	2.812889308	8.84E-42
ENSG00000233493	TMEM238	3.022472471	1.13E-41
ENSG00000165689	ENTR1	1.718280576	2.96E-38
ENSG00000204128	C2orf72	4.33489103	3.74E-38
ENSG00000110931	CAMKK2	2.829978991	7.18E-38
ENSG00000131188	PRR7	3.297285205	1.08E-37
ENSG00000206630	SNORD60	5.737894926	2.26E-37
ENSG00000182154	MRPL41	2.177708539	2.59E-37
ENSG00000162241	SLC25A45	2.134868537	2.70E-37
ENSG00000142544	CTU1	2.138342989	5.00E-37
ENSG00000168993	CPLX1	2.683731642	6.47E-37
ENSG00000280927	CTBP1-AS	5.351784937	8.07E-37
ENSG00000007264	MATK	4.911109586	1.05E-36
ENSG00000232442	MHENCR	2.307129146	3.13E-36

Further, these key genes were imported into the Oncomine database to query the expression of key genes in common cancers (Fig. 3). According to Fig. 3, C2orf72, CAMKK2, FGFRL1, HPN and UCN showed significant differential expression in PRAD. In addition, the key genes showed significant differential expression in breast cancer, liver cancer, lung cancer, pancreatic cancer and other common cancers.

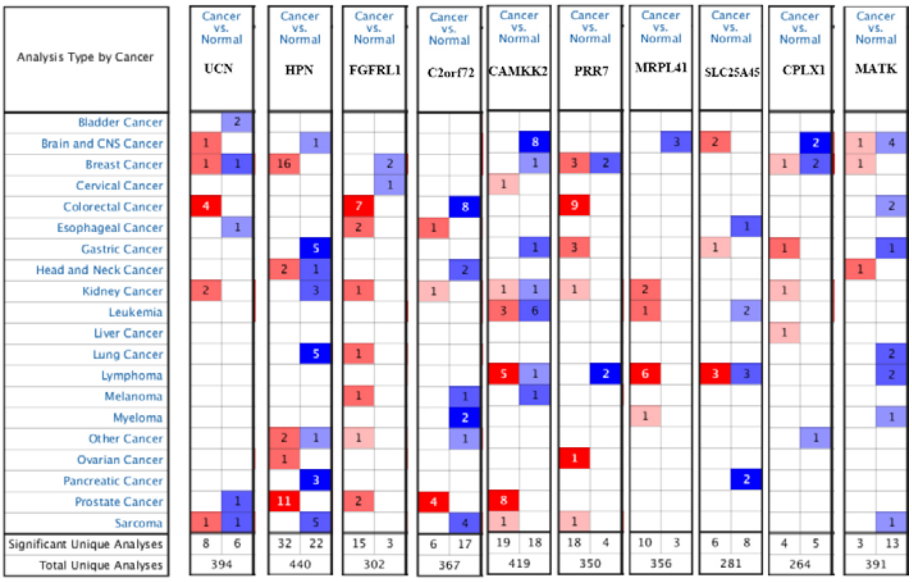

Figure. 3 The expression of key genes in common cancers

4. Conclusion

In this study, we performed the Mcode algorithm to analyse the GCN to mine important biological processes in the development of PRAD. It provided new ideas for the treatment of PRAD.

We preformed the FC-t algorithm to analyse transcriptome data sets after preprocessing, and obtained 5153 DEGs. Furthermore, we used the Mcode algorithm to divide the GCN into modules, and found that these gene modules are mainly involved in regulating the development of muscle tissue and the proliferation and differentiation of muscle cells, regulating the process of cell mitosis, regulating the metabolism of energy and the expression of genetic materials, and regulating tissues and organs, and regulation of the development of the urogenital system. Therefore, it was infered that the occurrence and development of PRAD were closely related to the above six types of biological processes. Finally, we used the oncomine database to predict key genes, and found that C2orf72, CAMKK2, FGFRL1, HPN, and UCN were significantly expressed in PRAD, and also significantly expressed in other tumors.

CAMMK2 is a member of the serine and threonine protein kinase family and plays a key role in cell signaling transduction. FGFRL1 was the member of the fibroblast growth factor receptor family. It played a vital role in embryonic development, growth and development, neuroregulation, and metabolic regulation. It has significant high-level expression in a variety of tumors. HPN is an acidic mucopolysaccharide, mainly produced by mast cells and basophils, and has anticoagulant and blood lipid regulating effects. UCN plays a role in a variety of downstream signaling pathways, plays an important role in a variety of cell metabolism, and there is multiple evidence that UCN has a promoting role in cell protection.

In summary, this study performed a bioinformatics algorithm based on network analysis to find that C2orf72, CAMKK2, FGFRL1, HPN, and UCN played key roles in PRAD, and six types of biological processes related to development of the PRAD. These provide new ideas for subsequent research on PRAD.

Acknowledgment

This paper is supported by National Natural Science Foundation of China, No.61873078.

References

[1] Satkunasivam, R., Kim, A. E., Desai, M., Nguyen, M. M., Quinn, D. I., Ballas, L., … Gill, I. S. (2015). Radical Prostatectomy or External Beam Radiation Therapy vs No Local Therapy for Survival Benefit in Metastatic Prostate Cancer: A SEER-Medicare Analysis. The Journal of urology, 194(2), 378-385.

[2] Sehrawat, A., Gao, L., Wang, Y., Bankhead, A., 3rd, McWeeney, S. K., King, C.J., … Alumkal, J. J. (2018). LSD1 activates a lethal prostate cancer gene network independently of its demethylase function. Proceedings of the National Academy of Sciences of the United States of America, 115(18), E4179-E4188.

[3] Chakravarthi, B. V., Goswami, M. T., Pathi, S. S., Dodson, M., Chandrashekar,D. S., Agarwal, S., … Varambally, S. (2017). Expression and Role of PAICS, a De Novo Purine Biosynthetic Gene in Prostate Cancer. The Prostate, 77(1), 10-21.

[4] Xue K, Cao J, Wang Y, Zhao X, Yu D, Jin C, Xu C. (2019). Identification of Potential Therapeutic Gene Markers in Nasopharyngeal Carcinoma Based on Bioinformatics Analysis. Clin Transl Sci. 2019 Dec 21.

[5] Xiao H, Chen P, Zeng G, Xu D, Wang X, Zhang X. (2019). Three novel hub genes and their clinical significance in clear cell renal cell carcinoma. J Cancer. Nov 1;10(27):6779-6791.

[6] Huang T, Wang Y, Wang Z, Cui Y, Sun X, Wang Y. (2019). Weighted Gene Co-Expression Network Analysis Identified Cancer Cell Proliferation as a

Common Phenomenon During Perineural Invasion. Onco Targets Ther. Nov 28;12:10361-10374.

[7] Hutter C, Zenklusen JC. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell. Apr 5;173(2):283-285.