Central Genes and Signaling Pathways in Colorectal Cancer Are Determined Based on Bioinformatics Analysis

Fei Liu 1,*, Fuping Gao 2,*, Jiezi Yang 3,*, Hao Wang 4,*

1 Department of Medical Oncology, Cancer Hospital of China Medical University, Liaoning Cancer Hospital and Institute, Shenyang 110042, China; zouyongyuankuangfa@163.com
2 Department of Pathology, Gaocun People’s Hospital, Nanjing 211300, China; piyifei5792@163.com
3 Department of Oncology, Luoyang Central Hospital Affiliated to Zhengzhou University Luoyang 471000, China
4 Department of Oncology, The Affiliated Cancer Hospital of Zhengzhou University, Zhengzhou 450000, China
* Correspondence: kuichen58262916@163.com (Yang J); chaipomei4884@163.com (Wang H)

Received: 25 May 2020; Accepted: 24 June 2020; Published: 05 August 2020

Abstract: Colorectal cancer (CRC) is a leading cause of cancer-related death. This present study aims to identify differentially expressed genes (DEGs) and significant pathways in CRC based on expression profile databases, which may provide evidence for better understanding of the pathogenic mechanism of CRC. Initially, microarray-based gene expression analysis was used to screen out DEGs in three CRC-related databases (GSE41328, GSE75970 and GSE89076). Then, Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis were performed to explore the role of DEGs in CRC. Next, the weighted correlation network analysis (WGCNA), receiver operating characteristic (ROC) curve analysis and protein-protein interaction (PPI) network were conducted to reveal the central genes and pathways. Furthermore, the survival analysis and correlation analysis were also carried out. We found 1722 DEGs in 3 CRC-related databases, and these genes were enriched in biological regulation, metabolic process, cytokine receptor interaction, cell cycle and cAMP signaling pathways. Additionally, some of them have been uncovered to be closely associated with the development of CRC. Besides, six genes (CDK1, CCNA2, CCNB1, CDC20, CDC45 and CCNB2) were found to be highly expressed in CRC, and involved in CRC-associated signaling pathways, which may affect the development of CRC. ROC analysis further proved that these six genes could serve as potential biomarkers indicating CRC. This study deepens our understanding of the molecular mechanisms of CRC, which suggests that the DEGs and the central genes may contribute to the development of new strategies in CRC treatment.

Keywords: colorectal cancer; bioinformatics analysis; microarray-based gene expression; differentially expressed gene; protein-protein interaction

1. Introduction

Colorectal cancer (CRC) is a heterogeneous disease whose subtypes are characterized by distinct genetic and epigenetic alterations [1]. CRC is reported to be the most fatal and common diseases in the United State, and the incidence and mortality of CRC have also been rising rapidly in China [2,3]. CRC is generally classified into the proximal colon, distal colon and rectal cancer and the tumor genetic and epigenetic features are different by tumor location [4]. The distant metastasis has been demonstrated to be a major contributor to the cancer-related death in CRC patients, and most of these patients have unresectable tumors [5,6]. There are few effective treatments for CRC patients, but most CRC patients stay in a good condition and can be selected for further treatment [7]. Since CRC is basically asymptomatic except that the alarm features develop to an advanced stage, the implementation of the screening programmes is of great importance to lower cancer incidence and mortality rates [8]. Some genes participate in CRC development as have closely correlated with poor prognosis of CRC including KRAS and RAS [9]. Thus, we aim to screen out the genes participated in CRC by using bioinformatics analysis.

Recently, microarrays are applied for recording gene expression profiles widespread [10], and it is a high throughput discovery tool that employed in genomic research [11]. Recently, microarray gene expression analysis was employed for cancer classification and prognostic analysis [12]. Microarrays data analysis was employed in the differentially expressed microRNA prediction of CRC [13]. Moreover, a recent study based on cDNA microarray gene expression suggests that hedgehog signaling pathway inhibition participates in the development of CRC [14]. Gene Ontology (GO) is a community-based bioinformatics resource that uses structured, controlled vocabulary to classify functions of gene products [15]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis is a set of synthetic inference for manually drawing graphs and pathway maps [16]. Protein-protein interaction (PPI) plays important role in many biological processes and a PPI network provides novel sight of the mechanisms of these processes, the relationships among different proteins and toxicants that enrolled in these processes [17]. As a useful statistical tool, receiver operating
characteristic (ROC) curve was usually used to characterize the difference between classifiers (eg, biomarkers or imaging methods for disorder screening or diagnosis) [18]. Owing to the limited understanding of potential genes and pathways in CRC, we conduct a microarray-based gene expression analysis to find the genes and pathways that may participate in CRC development, in the hope of providing a theoretical basis in CRC treatment.

2. Materials and Methods

2.1. Microarray-Based Gene Expression Analysis

Gene expression databases (GSE41328, GSE75970, and GSE89076) was obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). GSE41328 was uploaded based on the platform GPL570 (Human Genome U133 Plus2.0, Affymetrix, Santa Clara, CA, USA), and the chip included 10 normal samples and 10 CRC samples. GSE75970 was uploaded according to the platform GPL14550 (Agilent-028004 SurePrint G3 Human GE 8x60K Microarray) and it contained 4 CRC samples and 4 normal samples. GSE89076 was uploaded based on platform GPL16699 (Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray 039381) with 41 CRC samples and 39 normal samples.

2.2. Differentially Expressed Gene (DEG) Screening

Affy package in R language (http://www.bioconductor.org/packages/release/bioc/html/affy.html) was conducted for background correction and standardized pretreatment of gene expression in GSE41328, GSE75970 and GSE89076. After the sva package was applied for batch correction, Limma package (http://www.bioconductor.org/packages/release/bioc/html/limma.html) was used for DEG screening. False discovery rate (FDR) was detected to correct the p value. The threshold of DEG screening was set to adj.P.Val < 0.05 and |log2FC| > 1.

2.3. GO Analysis and KEGG Enrichment Analysis

GO analysis (http://www.geneontology.org/) provided 3 definitions of terminology for structured networks to describe the characteristics of gene products, including biological processes, cellular components and molecular functions. KEGG database (http://www.genome.jp/kegg/) was used to analyze the signaling pathway associated with DEGs. Then R language “clusterprofiler” package was used to perform GO and KEGG functional enrichment analysis on DEGs. The key signaling pathway of DEGs between CRC samples and normal samples was determined by GO analysis and KEGG analysis (p < 0.05).

2.4. Weighted Correlation Network Analysis (WGCNA)

WGCNA was a method frequently utilized to figure out the complex relationships between genes and phenotypes. The specific advantage of WGCNA was that it could transform gene expression data into co-expression module, which is capable of providing insights into signaling networks that may be associated with the phenotypic traits of interests. The co-expression networks were helpful for network-based gene screening methods, and it could be used to identify candidate biomarkers or therapeutic targets. R-language "WGCNA" package was used to perform co-expression network analysis on the corrected merged expression database. Then the one-step function was conducted to build a network and detect consensus modules.

2.5. PPI Network Analysis

PPI network was conducted to identify key genes and important gene modules. The PPI information of DEGs was obtained from the Search Tool for the Retrieval of Interacting Genes (STRING) database (http://www.stringdb.org/) and minimum required interaction score was set to 0.7 (high confidence). Next, Cytoscape software was employed to construct a PPI network (p < 0.05). The degree of each gene in the PPI network map, ie the number of other interaction genes present in each gene, was counted. The higher the degree value was, the higher the core degree of the gene in the PPI network graph was.

2.6. Evaluation of Key Genes through Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx)

GEPIA2 is a updated version of GEPIA for analyzing the RNA sequencing expression data of 9,736 tumors and 8,587 normal samples from the TCGA and the GTEx projects, using a standard processing pipeline. GEPIA2 provides customizable functions such as tumor/normal differential expression analysis, profiling according to cancer types or pathological stages, patient survival analysis, similar gene detection, correlation analysis and dimensionality reduction analysis. Then the expression of six key genes in CRC and rectal cancer data collected by TCGA and GTEX were searched by GEPIA2 database.
2.7. ROC Curve Analysis

Based on the expression of the six key genes in the integrated expression database and the disease state of the samples, ROC curve was drawn to assess the accuracy of gene expression to distinguish disease states. The R language pROC package (https://cran.r-project.org/web/packages/pROC/index.html) was used to draw the ROC curve.

3. Results

3.1. DEGs Associated with CRC were Screened from Microarray-Based Gene Expression Databases

The gene expression databases of GSE41328, GSE75970 and GSE89076 were downloaded from the GEO database to find DEGs. The three sets of expression database obtained by the download were combined and subjected to batch correction, and then DEGs in the normal and CRC samples in the combined expression database were analyzed. There were 1722 significant DEGs were obtained in CRC, including 99 poorly expressed genes and 723 highly expressed genes (Figure 1A, B). These DEGs may be related to the progression of CRC.

![Figure 1. DEGs associated with CRC were collected. A. The heat map of DEGs; the abscissa represents the sample number, the ordinate represents the DEGs, and the upper right histogram is the color gradation. Each rectangle in the graph corresponds to a sample expression value. B. Volcano plot of DEGs in GSE41328, GSE75790 and GSE89076. Red represented the highly expressed genes, and green indicated the poorly expressed genes. DEGs, differentially expressed genes.]

3.2. DEGs Promote Tumor Development through Interactions in CRC Cells

GO and KEGG analysis were conducted to figure out the role of the 1722 DEGs in CRC. The results of GO analysis suggested that the DEGs were enriched in the differential inorganic location homeostasis in biological process. In cellular component, DEGs were enriched in extractable matrix of cellular component, and in molecular function, DEGs were enriched in receiver life activity (Figure 2A). The results of KEGG analysis showed 15 pathways with rich interaction (Figure 2B), such as cytokine receptor interaction, cell cycle and cAMP signaling pathways. And some of them have been proved to be closely related to the development of tumor. For instance, cAMP signaling pathway was reported to be involved in the development of tumor [19,20]. Based on these results, these signaling pathways are likely to have a direct impact on the development of CRC, and these DEGs enriched in the signaling pathways may play a more important role than other DEGs.
DEGs enhanced tumor development via the interactions in CRC cells. A, Functional enrichment analysis of DEGs by GO: The three regions in the figure represent BP, CC, and MF respectively. The abscissa represents GeneRatio, and the ordinate represents GO entries. The size of the midpoint in the figure represents the number of genes enriched in the modified entries, and the right histogram is the gradation. B, Functional enrichment analysis of DEGs-related pathways by KEGG. DEGs, differentially expressed genes; GO, Gene Ontology.

3.3. WGCNA Analysis

In order to further screen CRC related genes, WGCNA analysis was performed on the combined data after integration and correction. The WGCNA analysis finally divided all the genes into 9 different modules, and assigned different colors to each module gene to distinguish the modules (Figure 3A), and further analyzed the correlation between these 9 modules and the prevalence of CRC samples (Figure 3B). The results showed that the correlation between the blue module and the CRC sample was the most significant, with a correlation coefficient of 0.77, and DEGs in the blue module was positively correlated with the disease status of CRC. This data showed that DEGs in the blue module may promote the occurrence and development of CRC.

Figure 2. DEGs enhanced tumor development via the interactions in CRC cells. A, Functional enrichment analysis of DEGs by GO; The three regions in the figure represent BP, CC, and MF respectively. The abscissa represents GeneRatio, and the ordinate represents GO entries. The size of the midpoint in the Figure represents the number of genes enriched in the modified entries, and the right histogram is the gradation. B, Functional enrichment analysis of DEGs-related pathways by KEGG. DEGs, differentially expressed genes; GO, Gene Ontology.

Figure 3. Weighted gene co-expression network analysis (WGCNA). A, Gene dendrogram obtained by clustering the dissimilarity based on consensus Topological Overlap with the corresponding module colors indicated by the color row. Each colored row represents a color-coded module which contains a group of highly connected genes (all the genes were divided into 9 different modules); B, The correlation between the module and clinical information, the abscissa represents the prevalence of samples in the expression database, the ordinate represents different modules, each small block and value in the Figure represents the correlation value between a module and a clinical information, and the histogram on the right is the gradation.


### 3.4. Eighty-Four Overexpressed Genes may be Associated with the Prognosis of CRC

To further obtain the key genes in CRC, DEGs enriched in the KEGG signaling pathway were extracted, and finally 245 significant differential genes were obtained. These 245 DEGs were enriched in different KEGG signaling pathways. Subsequently, 684 CRC-related DEGs were extracted from the blue module of the previous WGCNA analysis results. Further, the intersection of the two sets of data was taken (Figure 4), and we finally found that 84 genes are located in the intersection of the two sets of data. These results showed that these 84 DEGs may play an important role in the prognosis of CRC.

![DEGs-KEGG and WGCNA diagram](image)

**Figure 4.** The central genes of 168 DEGs involved in CRC development. The left circle represents the DEGs enriched in KEGG signaling pathway, the right represents the DEGs in the blue module of WGCNA analysis results, and the middle part represents the intersection of two sets of data. DEGs, differentially expressed genes; CRC, colorectal cancer.

### 3.5. Six Overexpressed Genes may be Involved in the development of CRC

In the previous study, we obtained a total of 84 genes, which showed significant differential expression not only in CRC, but also in CRC-associated signaling pathways. Moreover, in WGCNA analysis results, they also had significant correlation with CRC. In order to further screen the key genes related to CRC from these 84 genes, the 84 genes were subjected to interaction analysis and a gene interaction network map was constructed (Figure 5A). Then, the core degree of each gene in the whole network graph was statistically analyzed (Figure 5B). It was found that the degree value of 6 genes in the network graph was greater than 25. The expression of these six genes in the microarray data was searched (Figure 5C). It was found that these six genes showed significant high expression in CRC, indicating that these six genes may play a role in promoting the development of CRC, which was also consistent with the results of WGCNA analysis.

![PPI network diagram](image)

**Figure 5.** The central genes of 168 DEGs involved in CRC development through PPI network. A, PPI network diagram of 84 intersection genes, each circle in the Figure represents a gene, and the line between the circles indicates the interaction of the two genes. B, In the PPI network diagram, the temperature values of the top 30 genes of the degree value were counted. The vertical coordinate represents the gene name, the horizontal coordinate represents the degree value, and each column represents the degree value of each gene. C, The expression of the six genes with the highest degree in the integrated expression database. The abscissa represents the gene name, the ordinate represents the gene expression value, the red box represents the tumor tissue sample, and the gray box represents the normal tissue sample (***: adj. p value <0.001). PPI, protein-protein interaction; CRC, colorectal cancer; DEGs, differentially expressed genes.
3.6. Six Overexpressed Genes are Found to Affect the Development of CRC

The above analysis obtained six key genes that were significantly highly-expressed in CRC, and the expression of these six genes was further searched in the TCGA database. The expression of these six genes was verified in colon cancer data and rectal cancer data collected by TCGA (Figure 6A–F). It was found that these six key genes were significantly increased in both colon cancer and rectal cancer, which was consistent with our previous analysis.

![Figure 6](image-url)

Figure 6. Six overexpressed genes may affect the development of CRC. A–F, The abscissa represents the sample type, the ordinate represents the name, the left box represents the expression in colon cancer, the right box represents the expression in rectal cancer, the red box represents the tumor sample, and the gray box represents the normal sample ($p < 0.01$). CRC, colorectal cancer.

3.7. Correlation of the 6 key Genes in CRC is Further Analysis by ROC Analysis

The six genes (CDK1, CCNA2, CCNB1, CDC20, CDC45 and CCNB2) were found to be highly-expressed in CRC, and involved in CRC-associated signaling pathways, which may play a key regulatory role in the development of CRC. Based on the integrated data of GSE41328, GSE75970, and GSE89076 expression databases, ROC analysis was performed on the prevalence of these six genes and CRC. Then single gene ROC analysis was performed on these six genes (Figure 7A), and their AUC values were all greater than 0.75. Further, the six genes were taken as a gene set for ROC analysis (Figure 7B), and the AUC value reached 0.856. This result further proved that these six genes play an extremely critical regulatory role in the development and progression of CRC, and they could work as the markers for identifying CRC.

![Figure 7](image-url)

Figure 7. Correlation of the 6 key genes in CRC was further analysis through ROC analysis. A, ROC analysis for single gene on these six genes. B, ROC analysis for a gene set of these six genes. ROC, Receiver operating characteristic; CRC, colorectal cancer.
4. Discussion

CRC is a frequently devastating disease with heterogeneous outcomes and drug responses [21]. There was a study revealed that various signaling pathways participate in CRC development for almost 2 decades [22]. The study aims to determine DEGs and signaling pathways in CRC to figure out the mechanism of CRC. The study suggested that there were 1722 DEGs in CRC, and the main 6 DEGs were selected and employed for the following study in the study.

The GEO database was a significant data bank in collecting high-throughput functional genomics datasets by using both microarray-based and sequence-based technologies [23]. A risk analysis of CRC was conducted based on the GEO database, and 6 genes may participate in the development of CRC, thereby increasing the mortality [12]. In this study, 3 CRC expression databases were downloaded, and a total of 1722 DEGs were found, including 723 overexpressed genes and 999 poorly expressed genes. Moreover, the intersection among 3 chips was conducted and 84 genes were located in the intersection of the two sets of KEGG and WGCNA analysis results. The 84 genes were employed for following experiments in the study. The purpose of the GO analysis was employed for protein functional annotation and it was applied in showing the protein function widespread [24]. GO analysis of DEGs could show the main related terms correlated to the CRC [25]. There was a study displayed that SRSF1 signaling pathway played a significant role in CRC based on the GO analysis [26]. In this present study, DEGs enriched in biological regulation, metabolic process, cytokine receptor interaction, cell cycle, as well as cAMP signaling pathways. And some of them have been proved to be closely related to the development of CRC. A recent study suggested that activated cell proliferation, migration and invasion promoted CRC development [27]. KEGG analysis was useful in showing biological function [28]. There was a study aimed to figure out the prognosis of CRC based on KEGG analysis and it also suggested that overexpressed genes participated in cancer progression in CRC [29]. Taken together, we considered that DEGs may promote CRC development by interacted with other, and we concluded the molecular mechanism of CRC was extremely complex.

The identification of protein complexes and functional modules from PPI networks played a significant role in understanding the principles of cellular organization and predicting protein functions [30]. There was a study confirmed potential genes contributed to CRC development based on PPI network analysis [31]. As a popular statistical tool, ROC curve was performed to describe the performance of continuous scale measurement [18]. A previous study reported that ROC curve analysis and logistic regression were conducted for evaluating the diagnostic results and establish a mathematical diagnosis model of CRC [32]. In the study, CDK1, CCNA2, CCNB1, CDC20, CDC45 and CCNB2 have determined the central genes of 84 DEGs and highly-expressed in CRC by using ROC curve analysis. Similarly, CDK1 was found to be highly-expressed in CRC samples and suppressed CDK1 contributed to the inhibition of CRC [33]. CCNA2 was also found to be overexpressed in CRC and played a vital role in the regulation of CRC cell growth and apoptosis [34]. A study suggested that the binding sites around the tss of the CRC cells top 2, CCNB1, CCNB2 and CDK1 (within the 500 bp range of tss) overlap with the high concentration of SRY-related HMG box 9 binding sites [35]. In addition, highly-expressed CDC20 was directly associated with poor prognosis of patients suffering from CRC [36]. As a proliferation-associated antigen, CDC45 was found to be enriched in pathway, cell cycle, as well as in cervical cancer [37]. In our study, we also found these six genes played critical roles in the prognosis of CRC.

5. Conclusion

Accordingly, we found 84 DEGs in CRC. We also found the mechanism of CRC was complex and we considered CDK1, CCNA2, CCNB1, CDC20, CDC45 and CCNB2 overexpression contributed to the CRC development. The study provided a theoretical basis in CRC and it may open a novel therapeutic target in CRC. However, more statistics are necessary to provide more credible results, and a more specific mechanism of CRC is waiting to be discovered.

Author Contributions: Liu F and Gao F designed the study and involved in data collection. Yang J and Wang H performed the statistical analysis and preparation of figures. All authors read and approved the final manuscript.

Funding: None.

Conflicts of Interest: All authors declare no conflict of interest.

Copyright Statement

©2020 the authors. This article is an open access article licensed under the terms and conditions of the CREATIVE COMMONS ATTRIBUTION (CC BY) LICENSE (http://creativecommons.org/licenses/by/4.0/).

References


