Identification of Hub Genes Associated with Development of Colon Carcinoma by Integrated Bioinformatics Analysis

Lebin Yuan
The Second Affiliated Hospital of Nanchang University
Fei Cheng
The Second Affiliated Hospital of Nanchang University
Zao Wu
The Second Affiliated Hospital of Nanchang University
Xiaodong li
The Second Affiliated Hospital of Nanchang University
Weiyang Xia
The Second Affiliated Hospital of Nanchang University
Zhigang Li
The Second Affiliated Hospital of Nanchang University
Shengping Mao
The Second Affiliated Hospital of Nanchang University
Zeyu Huang
The Second Affiliated Hospital of Nanchang University
Wei Shen (shenweiniu@163.com)
The Second Affiliated Hospital of Nanchang University

Research Article

Keywords: colon cancer, differential gene expression analysis, weighted gene co-expression network analysis, differentially co-expressed genes, biomarkers

Posted Date: March 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1415220/v1

Abstract

Colon cancer (COAD) stem cells are resistant to cancer treatment, so they are easy to lead to tumor progression after routine treatment failure. However, little is known about the molecular mechanism of cancer cells. The purpose of our study is to identify differential gene modules and representative candidate biomarkers for the clinical prognosis of patients, help to predict the prognosis and reveal the mechanism of cancer progression. In our study, based on the tumor Genome Atlas (TCGA) COAD database and gene expression profiles of GSE44076 from the Gene Expression Omnibus (GEO), the differentially co-expressed genes in COAD and normal tissues were explored. Combined with weighted gene co-expression network analysis (WGCNA) and differential gene expression analysis, 451 differential co-expression genes were screened out. As shown by functional annotation analysis using R clusterProfiler software package, these genes are mainly enriched in lipid catabolic process (biological process), cell apical part (cellular component) and phosphate hydrolase activity (molecular function). In addition, in the protein-protein interaction (PPI) network, 20 hub genes (ACAA2, FABP1, ACOX1, EHHADH, CPT2, ACADS, CPT1A, MT1G, MT1E, MT1X, MT1H, PPARGC1A, ACSS2, MT2A, MT1F, CRAT, UGT2B17, B3GNT6, and MUC4) were identified using Cytoscape's CytoHubba plug-in. The expression of 20 hub genes was down regulated in colon cancer compared with normal control group. Based on survival analysis, lower expression of CPT1A and B3GNT6 was associated with better overall survival (OS) and lower expression of ACADS and CPT2 was associated with disease-free survival (DFS) in cancer patients. Finally, we verified the protein levels of CPT1A and B3GNT6 through HPA database, which was consistent with the mRNA level in colon cancer samples. Also, TIMER and CIBERSORT database showed that there was a great correlation between CPT1A and B3GNT6 gene expression and immune infiltrating cells in tumor progression. In conclusion, our study shows that two survival related genes are highly correlated with the development of colon cancer. Therefore, CPT1A and B3GNT6 can play an important role in the progression of colon cancer and serve as potential biomarkers for future diagnosis and treatment.

Introduction

Colon adenocarcinoma (COAD) is a common malignant tumor, which endangers human health[1]. In 2020, there were more than 1.9 million new cases of colon cancer in the world with epidemiological analysis[2]. It is estimated that by 2030, there will be more than 2.2 million new cases and 1.1 million deaths[3]. The tumorigenesis and progression of COAD is considered to be a multi-step, multi-stage and multi-gene cellular genetic disease. Although the progress of treatment including surgical resection, neoadjuvant radiotherapy and chemotherapy, postoperative radiotherapy and chemotherapy, targeted therapy and immunotherapy have significantly improved the prognosis of patients with COAD, its prognosis is far from ideal especially in patients with advanced diseases[4]. Therefore, more research is needed on useful new biomarkers that may have therapeutic or prognostic value for patients with colon cancer.

Nowadays, microarray technology is widely used in life science research purposes. Gene expression profiling is extensively used to study the molecular mechanism of disease and finding disease-specific

biomarkers, which has given scholars more perspectives to study the characteristics and treatment of cancer, especially colon cancer[5]. Weighted Gene Co-expression Network Analysis (WGCNA), which is a systems biology method to describe the correlation patterns among genes via microarray samples, help physicians to understand the gene function and gene association from genome-wide expression[6]. Co-expression modules of highly correlated genes and interested modules associated with clinical traits can be ascertained by WGCNA[7, 8], which providing great insight into predicting the functions of co-expression genes and finding hub genes that play key roles in human diseases[9, 10]. Moreover, another powerful analysis within transcriptomics is differential gene expression analysis, which is an amicable online data processing tool for studying molecular mechanisms underlying genome regulation and discovering quantitative changes in expression levels between experimental groups and control groups[7, 11]. Potential biomarkers of some specific diseases can be found through gene expression differences. Thus, two methods are applied, combining the results of WGCNA and differential gene expression analysis to improve the recognition ability of highly related genes as candidate biomarkers to assist us better research cancer.

In this study, we downloaded the original mRNA expression data of COAD from the TCGA and GEO databases, differential co-expression gene was obtained by sophisticated WGCNA and differential gene expression analysis. In order to better understand COAD development, functional enrichment and protein-protein interaction (PPI) analysis combined with survival analysis were explored by integrated online database. Furthermore, taking advantage of Cytoscape software to construct the module gene network and identify the hub genes. By analyzing differentially co-expressed genes, this study provides a latent basis for clinical diagnosis or treatment to comprehend the etiology and potential molecular events of COAD.

Materials And Methods

The workflow of the gene extraction management pipeline of the analysis center is shown in Fig. 1. We expound on each step in the following sub-sections.

Data collection From TCGA and GEO Database

The gene expression profiles of COAD were downloaded from TCGA (https://portal.gdc.cancer.gov/) and GEO (https://www.ncbi.nlm.nih.gov/gds). In the TCGA database, all data on COAD and corresponding clinical information were freely downloaded by R package TCGA biolinks[12]. There were 437 COAD samples, including 398 tumor and 39 normal tissues, and the RNA-seq data were generated by the Illumina HiSeq platform. A total of the data had been generated by using the Illumina HiSeq 2,000 platform, and were annotated to a reference transcript set of Human hg38 gene standard track. As suggested by the edgeR package tutorial[13], genes of low read counts are usually not of interest for further analysis. So, we kept the genes with a cpm (count per million) ≥ 1 in this study. After filtering using function rpkm in edgeR package, which is calculated by dividing gene counts by gene length, a total of 14,280 genes with RPKM values were subject to our next analysis. On the other hand, the

normalized expression profiles of GSE44076, using R package GEO query[14] to obtain another gene expression profile of COAD from GEO. GSE44076 consisted of 98 tumor samples and 148 paired normal tissues from patients with COAD, which were studied with the GPL13667 platform [HG-U219] Affymetrix Human Genome U219 Array. Accord to the annotation file provided by the manufacturer, converting the probe into gene symbol and duplicating the probe of the same gene by measuring the median expression value of all corresponding probes removed. Subsequent, a total of 19,014 genes were selected for follow-up analysis.

Construction of gene co-expression network and detection of WGCNA module

Co-expression networks boost network-based gene screening methods which can be used to identify candidate biomarkers and therapeutic targets. In our study, the gene expression data profiles of TCGA-COAD and GSE44076 were constructed to gene co-expression networks using the WGCNA package in R[6]. An appropriate soft powers β = 3 and 20 were selected using the function pick Soft Threshold to build a scale-free network. Next, the adjacency matrix was created by the following formula: aij = |Sij| β(aij: adjacency matrix between gene i and gene j, Sij: similarity matrix which is done by Pearson correlation of all gene pairs,β: softpower value), and was transformed into a topological overlap matrix (TOM) as well as the corresponding dissimilarity(1-TOM). Later, the hierarchical clustering tree of 1-tom matrix is constructed, and the similar gene expression is divided into different gene co-expression modules. To further identify functional modules in a co-expression network, the module-trait associations between modules, and clinical feature information were calculated according to the previous study[15, 16]. Selected for subsequent analysis based on Modules with high correlation coefficients which are regarded as candidate modules related to clinical features. A more detailed description of the WGCNA method was reported in previous study[15, 16].

Differential expression analysis and interaction with modules of interest

Differential expression analyses on RNA-Sequencing and microarray data were integrated by the R limma (linear models for microarray data)[17]. TCGA-COAD and GSE44076 dataset were used to obtain the differentially expressed genes (DEGs) by limma. The p-value was adjusted by the Benjamini–Hochberg method to control for the false discovery Rate (FDR). Genes with the cutoff criteria of $|logFC| \geq 1.0$ and adj. $P < 0.05$ were regarded as DEGs. The DEGs of the TCGA-COAD and GSE44076 dataset were visualized as a volcano plot by using the R package ggplot2[18, 19]. Subsequently, the overlapping genes between DEGs and co-expression genes that were extracted from the co-expression network were used to identify potential prognostic genes, which were presented as a Venn diagram using the R package Venn Diagram[20].

Functional Enrichment Analysis via Gene Ontology and Kyoto Encyclopedia of Genes and Genomes

The functional enrichment analysis included the signal pathway enrichment analysis and Gene Ontology (GO) enrichment, which comprised biological process (BP), molecular function (MF) and cellular component (CC), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways analysis by DEGs. R

package ClusterProfiler package is a convenient tool to explore the function between genes of interest, with a cut-off criterion of adjusted $p < 0.05$.

Analysis of the PPI network of the DEGs and hub genes identification

The STRING (Search Tool for the Retrieval of Interacting Genes) online database allowed us to build a protein-protein interaction network (PPI) network consisting of selected genes. Based on a minimum required interaction score of 0.4, We used Cytoscape[21] (version 3.9.0) software to complete the analysis and visualization of the PPI network. In a co-expression network, Maximal Clique Centrality (MCC) algorithm was reported to be the most effective method of finding hub nodes. The MCC of each node is calculated by CytoHubba[22], plug-in in Cytoscape, computed the MCC of each node to perform the hub genes analysis. In this study, the genes with the top 20 MCC values were considered as the hub genes.

Hub genes expression analysis and Survival analysis

Using the survival package in R software, we preformed Kaplan–Meier univariate survival analysis to investigate the relationship between overall survival rate (OS) and hub genes. GEPIA2 (http://gepia2.cancer-pku.cn/#index) is an Internet-based Interactive Platform for detailed analysis of TCGA gene expression data. We used GEPIA for quantification of two genes relative expression across different grades of tumors and stages of cancer between healthy tissue and tumor tissue samples, and several other types of clinicopathological characteristics[23]. The online tool GEPIA2 was used to determine the association between disease-free survival (DFS) and hub gene expression in COAD patients. The survival related hub genes with log rank $P < 0.05$ was statistically significant.

Validation of Protein Expressions and Immune-related of Survival-Related Hub Genes by the HPA Database and TIMER Databa

The Human Protein Atlas (HPA) (https://www.proteinatlas.org/) is a worthy database that provides a large number of transcriptomic and proteomic data of specific human tissues and cells[24]. IHC based protein expression pattern is the most common application of immunostaining to detect the relative position and abundance of proteins[25]. HPA was used to directly view immunohistochemical pictures of protein expression of B3GNT6 between normal and tumor samples in this research. Moreover, We used online TIMER (https://cistrome.shinyapps,io.timer/) to further explore the potential immunomodulatory mechanism of CPT1A and B3GNT6 in the regulation of tumor-infiltrating immune cells. R's CIBERSORT[26] software package was used to detect the proportion of 22 immune cells with high and low significant gene expression in COAD samples.

Results

Construction of weighted gene co-expression network analysis

A gene co-expression network was constructed from TCGA COAD and GSE44076 data sets using WGCNA software package to find functional clusters in COAD patients. In the case of assigning color to each

module, 10 modules of TCGA-COA (Fig. 2A) gain color assignment (remove a gray module that is not assigned to any cluster), and 7 modules in GSE44076 (Fig. 3A) get different colors in this article. Subsequently, we mapped the module feature relationship to assess the association between each module and two clinical features (cancer and normal), The outcomes of the module trait relationship are shown in Fig. 2B,3B. It was found that the brown module (r = 0.86, p = 3e-131) in TCGA-COAD and the turquoise module (r = 0.89, p = 2e-87) in GSE44076 had the positively highest correlation with normal tissues respectively.

Gene identification between DEG datdabases and co-expression modules

Based on the cut-off criteria of $|logFC| \geq 1.0$ and adj. $P < 0.05$, It was found that 1913 DEGs in TCGA dataset and 206 DEGs in GSE44076 dataset were dislocate in tumor tissues,480 and 417 co-expressed genes were found in the brown module of TCGA dataset and the turquoise module of GSE44076 Separately. Finally, a total of 451 intersection genes were extracted to verify the genes of the co-expression modules (Fig. 4).

Functional Enrichment Analyses for the 451 common Genes

The clusterProfiler software package was used for gene enrichment analysis to further understand the potential functions of 451 genes overlapping with DEG lists and two co-expression modules. After screening the GO enrichment analysis, we observed that biological processes (BP) mainly focus on lipid catabolic process (Fig. 5 and Table 1). The results of cell composition (CC) showed that these genes were mainly involved in the apical part of cell. In addition, in molecular function (MF) analysis, phosphoric ester hydrolase activity was considered to be related to 451 genes.

Table 1
The GO function enrichment analysis of DEGs in COAD

GO	Category	Description	Count	%	Pvalue	Qvalue
GO:0016042	BP	Lipid catabolic process	30	0.073	1.90E-11	6.17E-08
GO:0006631	BP	Fatty acid metabolic process	29	0.071	8.83E-06	7.83E-06
GO:0006066	BP	alcohol metabolic process	27	0.066	1.03E-05	9.13E-06
GO:0044242	BP	cellular lipid catabolic process	23	0.056	5.02E-07	5.02E-07
GO:0007586	BP	digestion	16	0.039	4.98E-08	1.62E-05
GO:0045177	CC	apical part of cell	39	0.114	1.01E-13	3.31E-11
GO:0016324	CC	apical plasma membrane	34	0.100	1.66E-12	2.70E-10
GO:0031253	CC	cell projection membrane	25	0.073	2.20E-07	1.19E-05
GO:0098862	CC	cluster of actin-based cell projections	19	0.055	2.84E-09	1.85E-07
GO:0005903	CC	brush border	18	0.052	1.67E-11	1.81E-09
GO:0042578	MC	phosphoric ester hydrolase activity	25	0.073	1.10E-06	1.50E-04
GO:0008509	MC	anion transmembrane transporter activity	21	0.050	1.10E-05	6.01E-04
GO:0008081	MC	phosphoric diester hydrolase activity	15	0.035	1.41E-09	7.69E-07
GO:0016616	MC	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	14	0.033	8.74E-07	1.50E-04
GO:0016616	MC	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	14	0.033	8.74E-07	1.50E-04

Construction of PPI network and identification of hub gene

The PPI network between overlapping genes was established by using STRING database. The hub genes selected in the PPI network shows the MCC algorithm using the CytoHubba plug-in (Fig. 6A,6B and Table 2). According to the statistics of MCC, the top 20 genes with the highest scores are selected as the hub genes include Acetyl-CoA acyltransferase2 (ACAA2), Acyl-CoA dehydrogenase short chain (ACADS), Fatty acid binding protein 1 (FABP1), Carnitine palmitoyltransferase 1A (CPT1A), Acyl-CoA oxidase 1(ACOX1), Metallothionein 1G (MT1G), Metallothionein 1E (MT1E), Enoyl-CoA hydratase and 3-hydroxyacyl CoA dehydrogenase (EHHADH), Carnitine palmitoyltransferase 2 (CPT2), Acyl-CoA dehydrogenase medium chain (ACADM), Metallothionein 1X (MT1X), Metallothionein 1H (MT1H), PPARG coactivator 1 alpha (PPARGC1A), Acyl-CoA synthetase short chain family member 2 (ACSS2), Metallothionein 2A (MT2A), Metallothionein 1F (MT1F), Carnitine O-acetyltransferase (CRAT), UDP glucuronosyltransferase family 2 member B17 (UGT2B17), Beta-1,3-N-acetylglucosaminyltransferase 6 (B3GNT6), and Mucin 4 (MUC4).

Table 2
The KEGG function enrichment analysis of DEGs in COAD

GO	Description	Count	%	Pvalue	Qvalue
hsa04978	Mineral absorption	10	0.043	6.23E-06	0.000764207
hsa03320	PPAR signaling pathway	10	0.043	4.71E-05	0.003854135
hsa04530	Tight junction	10	0.043	2.24E-02	0.189207
hsa00071	Fatty acid degradation	9	0.038	2.55E-06	0.000625852
hsa00830	Retinol metabolism	9	0.038	1.20E-04	0.004907811
hsa00982	Drug metabolism - cytochrome P450	9	0.038	1.88E-04	0.005812421
hsa04976	Bile secretion	9	0.038	9.23E-04	0.022638297
hsa04936	Alcoholic liver disease	9	0.038	1.99E-02	0.178525176
hsa01212	Fatty acid metabolism	8	0.034	1.90E-04	0.005812421
hsa00983	Drug metabolism - other enzymes	8	0.034	1.90E-03	0.038794788

Validation of hub gene expression pattern, prognostic value and protein expression

After screening 20 hub genes through the CytoHubba plug-in, we verified the expression level of hub gene in patients by convenient online database GEPIA2. Compared with normal tissues, 20 hub genes in COAD were significantly down regulated as shown in Fig. 6C,6D and supplement 1,2. Also, Kaplan Meier plotter used R survival software package and GEPIA2 database to analyze OS and DFS of 20 hub genes to investigate the prognostic value of hub genes in HNSCC patients. Among the 20 hub genes, Kaplan Meier analysis showed that the low expression level of CPT1A and B3GNT6 was significantly correlated with better OS in patients with COAD ($P < 0.05$) (Fig. 7A,7B and supplement 3,4). For DFS patients, the low expression level of CPT2 and ACADS was significantly correlated with better DFS in patients with COAD

(Fig. 7C,7D), no significant difference in CPT1A and B3GNT6 expression was observed in COAD patients ($P < 0.05$). Moreover, according to the HPA database, the protein level of CPT1A, B3GNT6 gene in tumor tissues was significantly lower than that in normal tissues (Fig. 8). All the above observations confirmed that the down-regulation of CPT1A and B3GNT6 expression was interrelated with better prognosis and higher overall survival in patients with COAD.

Correlation between two meaningful genes and tumor Immune infiltrating cells

We applied the TIMER database to analyze the correlation between CPT1A and B3GNT6 expression and immune infiltrating cells in COAD (Fig. 9,10). The results showed that the positive expression of the B3GNT6 gene was related to the expression levels of B cells ($r = 0.1708$, $P < 0.05$) and CD4 + T cells ($r = 0.0986$, $P < 0.05$).Also, The results showed that the expression of CPT1A gene was positively correlated with the expression levels of B cells ($r = 0.1856$, $P < 0.05$), CD4 + T cells ($r = 0.2684$, $P < 0.05$), macrophages ($r = 0.1253$, $P < 0.05$), neutrophils ($r = 0.1347$, $P < 0.05$) and dendritic cells ($r = 0.1442$, $P < 0.05$). CIBERSORT analysis showed that the expression level of B3GNT6 was related to tumor immune cell infiltration include NK cell activated ($p = 0.030$), Macrophages M1 ($P = 0.006$) and Mast cells activated ($p = 0.037$). Expression level of CPT1A was related to tumor immune cell infiltration include B cell naïve ($p = 0.004$), Dendritic cells activated ($P = 0.035$) and NK cells activated ($p = 0.011$).

Discussion

COAD is the most common intestinal tumor in the world, among the 36 cancers estimated worldwide in 2018, the number of new cases and related deaths of colon cancer ranked fourth, with an estimated 1100000 new cases[27, 28]. Although the treatment of colon cancer has improved, the prognosis of patients is usually poor due to tumor recurrence and metastasis, even if the tumor tissue is removed before tumor metastasis. Precise molecular targeted therapy gives patients more treatment opportunities. For example, Yu et al discovered that ZIC2 gene knockout inhibited the growth of colon cancer cells, prevented the transformation of cell cycle from G0 / G1 phase to S phase, and inhibited tumor formation[29]. A review explains the prognostic characteristics of lncRNA in the diagnosis of colon cancer associated with Ferroptosis[30]. Therefore, it is of great significance to reveal the molecular mechanism of colon cancer.

In this study, we identified 451 important genes COAD differentially expressed genes (DEG) using integrated bioinformatic analysis, these results suggested that these modular genes play an important role in the occurrence and development of COAD. The functional annotation analysis of cluster profiler software package showed that these genes were mainly enriched in metabolic processes and signaling transport pathways, which is basic process of cell function change and progress. Furthermore, based on the MCC scoring function of CytoHubba plug-in in Cytoscape, the first 20 COAD related genes were screened out (namely ACAA2, FABP1, ACOX1, EHHADH, CPT2, ACADS, CPT1A, MT1G, MT1E, MT1X, MT1H, PPARGC1A, ACSS2, MT2A, MT1F, CRAT, UGT2B17, B3GNT6, and MUC4). It was found that all their expression patterns were down regulated in cells when compared with normal control group and tumor

tissue. We also applied GEPIA database to verify the reliability of the data. Among them, the down-regulated expression of CPT1A and B3GNT6 was significantly correlated with the high overall survival rate of colon cancer. Finally, we analyzed the survival rate and immunohistochemistry of CPT1A and B3GNT6.

Carnitine palmitoyltransferase 1A(CPT1A) is one of the protein families CPT1, a key enzyme controlling fatty acid oxidation (FAO). CPT1A involves the rate limiting step of liver subtype catalyzing the conversion of acyl CoA to acyl carnitine[31]. The regulation of CPT1A is complex and has multiple levels involving genetic, epigenetic, physiological and nutritional regulators[32]. Some studies have pointed out that CPT1A mutation is related to human diseases such as keto hypoglycemia[33, 34]. This makes CPT1A an attractive target for therapeutic intervention. Recent studies have shown that CPT1A is a key component of cancer cell growth, survival and drug resistance[35–37]. The upregulation of CPT1A is very important for the tumor promoting effect of adipocytes in colon cancer[38]. IRGD modified exosomes effectively transfer CPT1A siRNA to colon cancer cells and reverse oxaliplatin resistance by regulating fatty acid oxidation[39].However, the role of CPT1A in cancer is not fully understood, especially in colon cancer. In our study, the expression of CPT1A in tumor tissues was down-regulated which was significantly correlated with COAD when compared with normal tissues. The decreased expression of CPT1A also affects the ability of cells to produce aspartate, a nucleotide precursor for DNA production[40], and affects the effects observed in human umbilical vein endothelial cells with impaired proliferation. Recent study found the effect of CPT1A mediated fat addictive oxidation on cell cycle progression and anchor independent growth of invasive ovarian cancer cells[41]. Wang et al. found that CPT1A mediated fat oxidation increased the metastatic ability and anoikis resistance of colorectal cancer cells. In addition, in the clinical samples, they found that the expression of CPT1A at the metastatic site was increased compared with the primary site[42]. It is suggested that CPT1A may be a prognostic biomarker in patients with colon cancer.

B3GNT6, the protein encoded by this gene is β-1,3-n-acetylglucosamine transferase, which adds an N-acetylglucosamine moiety to N-acetylgalactosamine modified serine or threonine[43]. The encoded enzyme is responsible for creating the core 3 structure of O-glycan, which is an important component of mucin glycoprotein[44]. At present, there are few studies on the molecular mechanism of B3GNT6. Research reports say that diseases associated with B3GNT6 include childhood onset schizophrenia. The increase of B3GNT6 mRNA plays an important role in the formation and maintenance of impaired intestinal mucus stability during zinc deficiency[45]. Expression of B3GNT6 in human prostate cancer cells inhibits tumor growth and metastasis[46]. Wang et al show that core B3GNT6 affect EMT-MTE plasticity of colorectal cancer cells through MUC1/p53/mir-200c dependent signal cascade[47]. B3GNT6 is down regulated in colon cancer and deeply inhibits the metastatic potential of cancer cells[48]. That was consistent with our finding of survival analysis. All these studies have strengthened the link between B3GNT6 expression and cancer progression.

Tumor immunotherapy is a hot spot in tumor therapy, which is widely used in the research and treatment of COAD[49]. The immune infiltration of tumor cells and lymph node metastasis are closely related to the

prognosis of colon cancer. TIMER database analysis showed that the expression level of B3GNT6 in COAD was positively correlated with the expression level of CD4 + T cells and B cells. The expression level of B3GNT6 in COAD was positively correlated with the expression level of B cells, CD4 + T cells, macrophages, neutrophils and dendritic cells. The proportion of 22 tumor immune cells in COAD was determined by CIBERSORT analysis. We found that there were significant differences in the expression rate of hub gene with different expression levels. The result showed that B3GNT6 may affect the immune infiltration of COAD by affecting the expression of NK cell activated, Macrophages M1 and Mast cells activated. CPT1A may affect the immune infiltration of COAD by affecting the expression of B cell naïve, Dendritic cells activated and NK cells activated. B cells and NK cells are important immune cells in the body and have a wide range of anti-tumor effects[50, 51]. Enhancement of NK cells in lung cancer can block the signal pathway of transforming growth factor B and play an anti-tumor role[52]. Germain et al. found that lung cancer patients with high-density B cells had a better prognosis[53]. We deem that the low expression of these genes may trigger antitumor immune response, indicating they may play an significant role in the immune system. However, the molecular mechanism of immunity needs more experimental and prospective studies to explore and testify our hypothesis.

Like all studies, our study has limitations. Although we provide an integrated bioinformatics analysis to identify potential diagnostic genes between cancer and normal tissues, especially CPT1A and B3GNT6, this may not be very accurate for each COAD patient and needs to be verified by more data and prospective controlled trials. In addition, the molecular mechanism of survival related genes affecting the prognosis of COAD patients is lack of experimental verification, which needs to be further studied.

In conclusion, by integrating WGCNA and differential gene expression analysis, our study mined two important survival related genes B3GNT6 and CPT1A, which may predict the prognosis of COAD.

Declarations

DATA AVAILABILITY STATEMENT

All available data were analyzed in this study. These can be found here: TCGA (https://portal.gdc.cancer.gov/), GEO (https://www.ncbi.nlm.nih.gov/gds), STRING (https://cn.string-db.org/),GEPIA2(http://gepia2.cancer-pku.cn/#index), R(https://www.r-project.org/), TIMER(https://cistrome.shinyapps.io/timer/) and HPA (https://www.proteinatlas.org/).

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

FUNDING

This research was supported by the Natural Science Foundation of Jiangxi Provincial [20171BAB205064].

AUTHOR CONTRIBUTIONS

LB, FC wrote the paper. ZW and XD edited the paper. SP, ZY, WY, and analyzed the data. ZG made the images out. All authors contributed to the article and approved the submitted version.

Acknowledgments

We acknowledge the TCGA, GEO, HPA, GEPIA2, R, STRING and other tools for free use.

References

1. Siegel, R.L., et al., *Colorectal cancer statistics, 2020*. CA Cancer J Clin, 2020. **70**(3): p. 145–164.
2. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries*. CA Cancer J Clin, 2021. **71**(3): p. 209–249.
3. Arnold, M., et al., *Global patterns and trends in colorectal cancer incidence and mortality*. Gut, 2017. **66**(4): p. 683–691.
4. Dekker, E., et al., *Colorectal cancer*. Lancet, 2019. **394**(10207): p. 1467–1480.
5. Can, T., *Introduction to bioinformatics*. Methods Mol Biol, 2014. **1107**: p. 51–71.
6. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis*. BMC Bioinformatics, 2008. **9**: p. 559.
7. Long, J., et al., *Transcriptional landscape of cholangiocarcinoma revealed by weighted gene coexpression network analysis*. Brief Bioinform, 2021. **22**(4).
8. Langfelder, P. and S. Horvath, *Fast R Functions for Robust Correlations and Hierarchical Clustering*. J Stat Softw, 2012. **46**(11).
9. Shi, M., et al., *APC(CDC20)-mediated degradation of PHD3 stabilizes HIF-1a and promotes tumorigenesis in hepatocellular carcinoma*. Cancer Lett, 2021. **496**: p. 144–155.
10. Yang, Y., et al., *Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types*. Nat Commun, 2014. **5**: p. 3231.
11. Guillotin, D., et al., *Transcriptome analysis of IPF fibroblastic foci identifies key pathways involved in fibrogenesis*. Thorax, 2021. **76**(1): p. 73–82.
12. Colaprico, A., et al., *TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data*. Nucleic Acids Res, 2016. **44**(8): p. e71.
13. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*. Bioinformatics, 2010. **26**(1): p. 139–40.
14. Davis, S. and P.S. Meltzer, *GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor*. Bioinformatics, 2007. **23**(14): p. 1846–7.

15. Yang, Z., et al., *Identification of AUNIP as a candidate diagnostic and prognostic biomarker for oral squamous cell carcinoma*. EBioMedicine, 2019. **47**: p. 44–57.
16. Cheng, Y., et al., *Identification of candidate diagnostic and prognostic biomarkers for pancreatic carcinoma*. EBioMedicine, 2019. **40**: p. 382–393.
17. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies*. Nucleic Acids Res, 2015. **43**(7): p. e47.
18. Ito, K. and D. Murphy, *Application of ggplot2 to Pharmacometric Graphics*. CPT Pharmacometrics Syst Pharmacol, 2013. **2**(10): p. e79.
19. Postma, M. and J. Goedhart, *PlotsOfData-A web app for visualizing data together with their summaries*. PLoS Biol, 2019. **17**(3): p. e3000202.
20. Gao, C.H., G. Yu, and P. Cai, *ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram*. Front Genet, 2021. **12**: p. 706907.
21. Otasek, D., et al., *Cytoscape Automation: empowering workflow-based network analysis*. Genome Biol, 2019. **20**(1): p. 185.
22. Chin, C.H., et al., *cytoHubba: identifying hub objects and sub-networks from complex interactome*. BMC Syst Biol, 2014. **8 Suppl 4**(Suppl 4): p. S11.
23. Chandrashekar, D.S., et al., *UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses*. Neoplasia, 2017. **19**(8): p. 649–658.
24. Colwill, K. and S. Gräslund, *A roadmap to generate renewable protein binders to the human proteome*. Nat Methods, 2011. **8**(7): p. 551–8.
25. Cyriac, G. and L. Gandhi, *Emerging biomarkers for immune checkpoint inhibition in lung cancer*. Semin Cancer Biol, 2018. **52**(Pt 2): p. 269–277.
26. Newman, A.M., et al., *Robust enumeration of cell subsets from tissue expression profiles*. Nat Methods, 2015. **12**(5): p. 453–7.
27. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2019*. CA Cancer J Clin, 2019. **69**(1): p. 7–34.
28. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA Cancer J Clin, 2018. **68**(6): p. 394–424.
29. Xu, Z., et al., *Multilevel regulation of Wnt signaling by Zic2 in colon cancer due to mutation of β-catenin*. Cell Death Dis, 2021. **12**(6): p. 584.
30. Tang, R., et al., *Ferroptosis, necroptosis, and pyroptosis in anticancer immunity*. J Hematol Oncol, 2020. **13**(1): p. 110.
31. Kunau, W.H., V. Dommes, and H. Schulz, *beta-oxidation of fatty acids in mitochondria, peroxisomes, and bacteria: a century of continued progress.* Prog Lipid Res, 1995. **34**(4): p. 267–342.
32. Schlaepfer, I.R. and M. Joshi, *CPT1A-mediated Fat Oxidation, Mechanisms, and Therapeutic Potential*. Endocrinology, 2020. **161**(2).
33. Gobin, S., et al., *Functional and structural basis of carnitine palmitoyltransferase 1A deficiency*. J Biol Chem, 2003. **278**(50): p. 50428–34.

34. Dai, J., et al., *Identification by mutagenesis of conserved arginine and tryptophan residues in rat liver carnitine palmitoyltransferase I important for catalytic activity*. J Biol Chem, 2000. **275**(29): p. 22020–4.
35. Aiderus, A., M.A. Black, and A.K. Dunbier, *Fatty acid oxidation is associated with proliferation and prognosis in breast and other cancers*. BMC Cancer, 2018. **18**(1): p. 805.
36. Galicia-Vázquez, G. and R. Aloyz, *Ibrutinib Resistance Is Reduced by an Inhibitor of Fatty Acid Oxidation in Primary CLL Lymphocytes*. Front Oncol, 2018. **8**: p. 411.
37. Kuo, C.Y. and D.K. Ann, *When fats commit crimes: fatty acid metabolism, cancer stemness and therapeutic resistance*. Cancer Commun (Lond), 2018. **38**(1): p. 47.
38. Xiong, X., et al., *Upregulation of CPT1A is essential for the tumor-promoting effect of adipocytes in colon cancer*. Cell Death Dis, 2020. **11**(9): p. 736.
39. Lin, D., et al., *iRGD-modified exosomes effectively deliver CPT1A siRNA to colon cancer cells, reversing oxaliplatin resistance by regulating fatty acid oxidation*. Mol Oncol, 2021.
40. Schoors, S., et al., *Fatty acid carbon is essential for dNTP synthesis in endothelial cells*. Nature, 2015. **520**(7546): p. 192–197.
41. Shao, H., et al., *Carnitine palmitoyltransferase 1A functions to repress FoxO transcription factors to allow cell cycle progression in ovarian cancer*. Oncotarget, 2016. **7**(4): p. 3832–46.
42. Wang, Y.N., et al., *CPT1A-mediated fatty acid oxidation promotes colorectal cancer cell metastasis by inhibiting anoikis*. Oncogene, 2018. **37**(46): p. 6025–6040.
43. Iwai, T., et al., *Molecular cloning and characterization of a novel UDP-GlcNAc:GalNAc-peptide beta1,3-N-acetylglucosaminyltransferase (beta 3Gn-T6), an enzyme synthesizing the core 3 structure of O-glycans*. J Biol Chem, 2002. **277**(15): p. 12802–9.
44. Ota, T., et al., *Complete sequencing and characterization of 21,243 full-length human cDNAs*. Nat Genet, 2004. **36**(1): p. 40–5.
45. Maares, M., et al., *Zinc Deficiency Disturbs Mucin Expression, O-Glycosylation and Secretion by Intestinal Goblet Cells*. Int J Mol Sci, 2020. **21**(17).
46. Lee, S.H., et al., *Core3 O-glycan synthase suppresses tumor formation and metastasis of prostate carcinoma PC3 and LNCaP cells through down-regulation of alpha2beta1 integrin complex*. J Biol Chem, 2009. **284**(25): p. 17157–17169.
47. Ye, J., et al., *Core 3 mucin-type O-glycan restoration in colorectal cancer cells promotes MUC1/p53*/miR-200c-*dependent epithelial identity.* Oncogene, 2017. **36**(46): p. 6391–6407.
48. Iwai, T., et al., *Core 3 synthase is down-regulated in colon carcinoma and profoundly suppresses the metastatic potential of carcinoma cells*. Proc Natl Acad Sci U S A, 2005. **102**(12): p. 4572–7.
49. Liu, X.S., et al., *NPM1 Is a Prognostic Biomarker Involved in Immune Infiltration of Lung Adenocarcinoma and Associated With m6A Modification and Glycolysis*. Front Immunol, 2021. **12**: p. 724741.

50. Bruno, T.C., *New predictors for immunotherapy responses sharpen our view of the tumour microenvironment*. Nature, 2020. **577**(7791): p. 474–476.
51. Shevtsov, M. and G. Multhoff, *Immunological and Translational Aspects of NK Cell-Based Antitumor Immunotherapies*. Front Immunol, 2016. **7**: p. 492.
52. Germain, C., et al., *Presence of B cells in tertiary lymphoid structures is associated with a protective immunity in patients with lung cancer*. Am J Respir Crit Care Med, 2014. **189**(7): p. 832–44.
53. Yang, B., et al., *Blocking transforming growth factor-β signaling pathway augments antitumor effect of adoptive NK-92 cell therapy*. Int Immunopharmacol, 2013. **17**(2): p. 198–204.

Figures

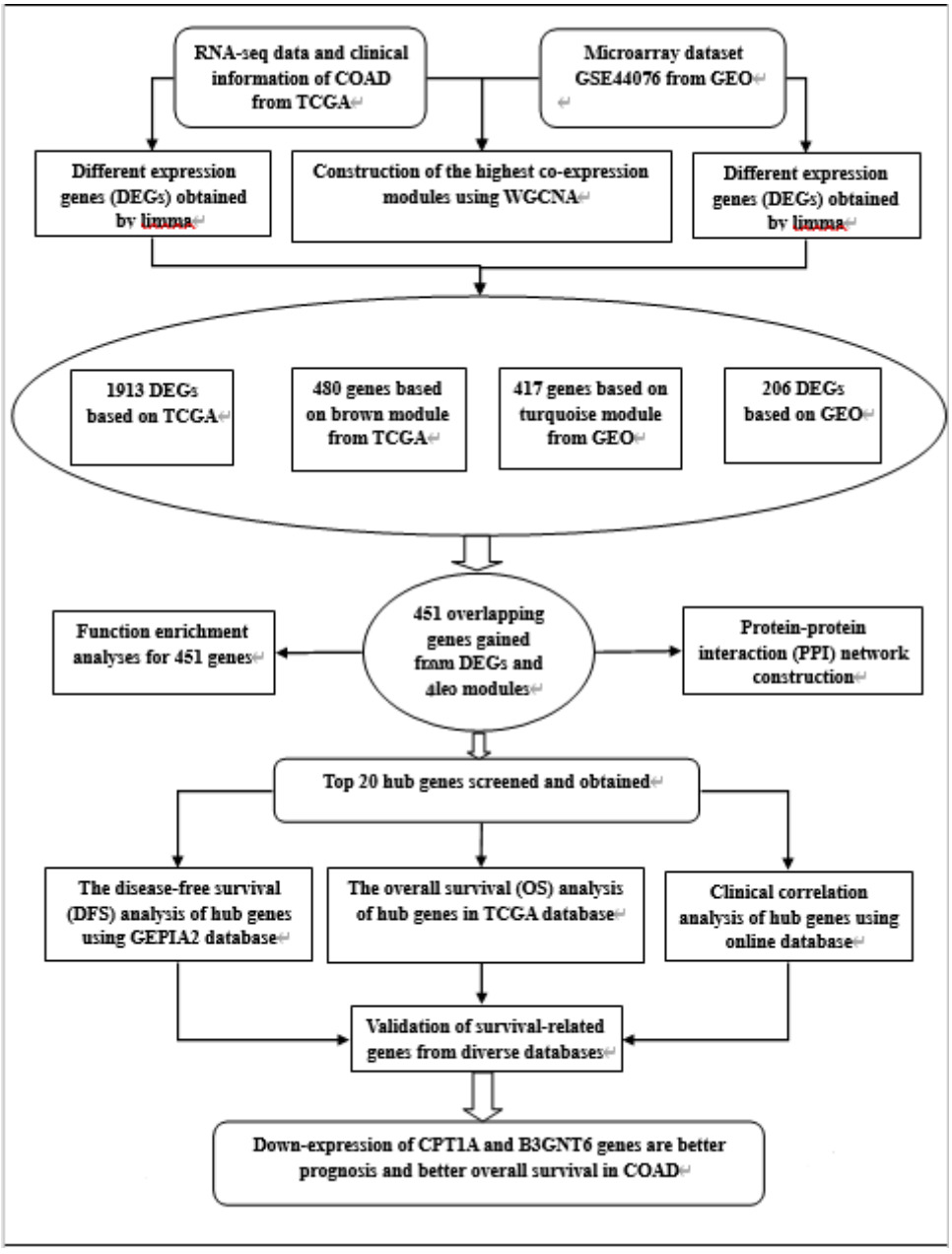

Figure 1

The research design and workflow of this paper

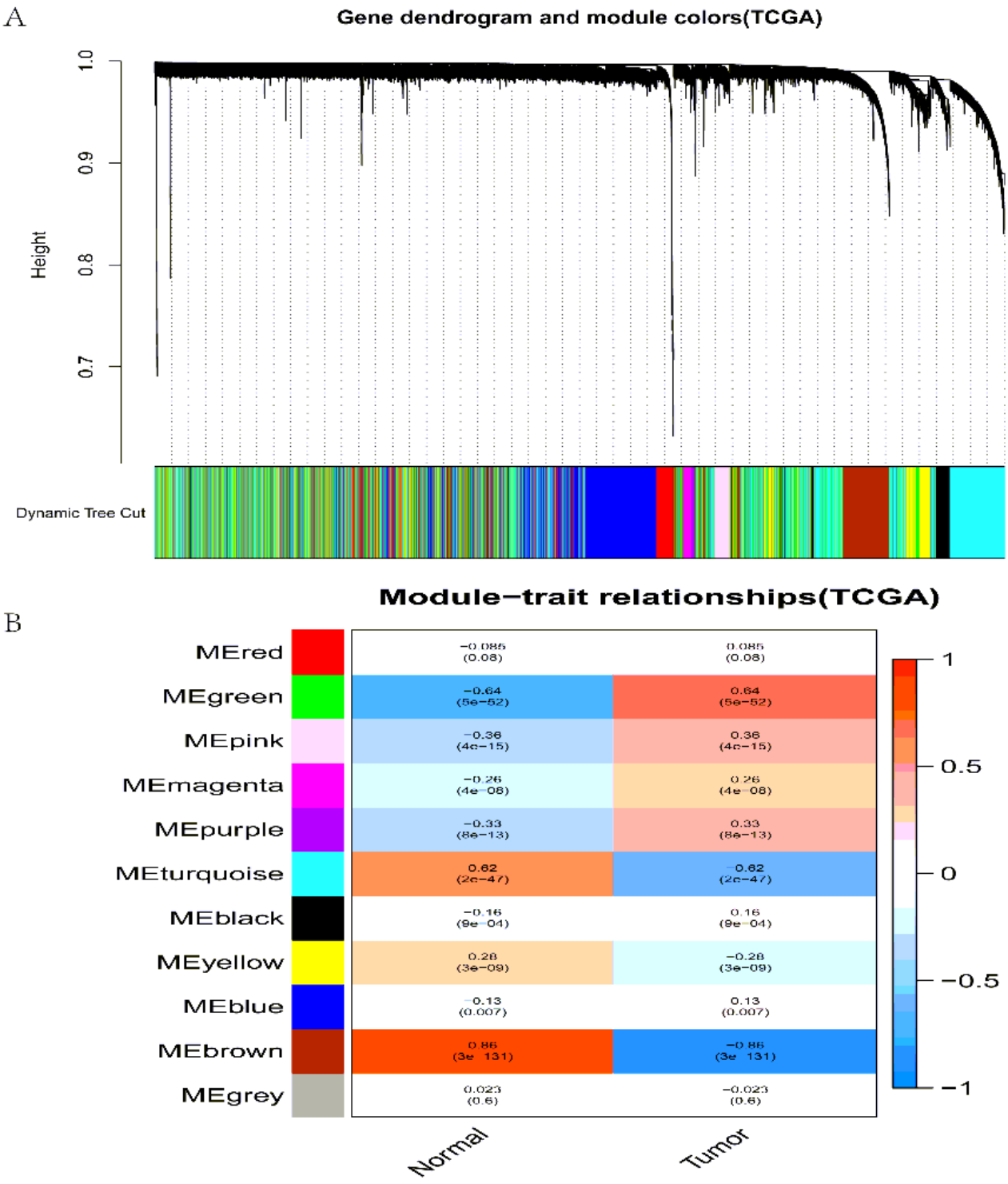

Figure 2

Identify the modules related to clinical information in the TCGA-COAD dataset. (A) The clustering tree module of co expression network is sorted by gene hierarchical clustering method based on 1-TOM matrix. Each module is assigned a different color. (B) Module-trait relationship, each row corresponds to a

color module, and each column corresponds to a clinical feature (cancer and normal). Each cell contains the corresponding correlation and P value.

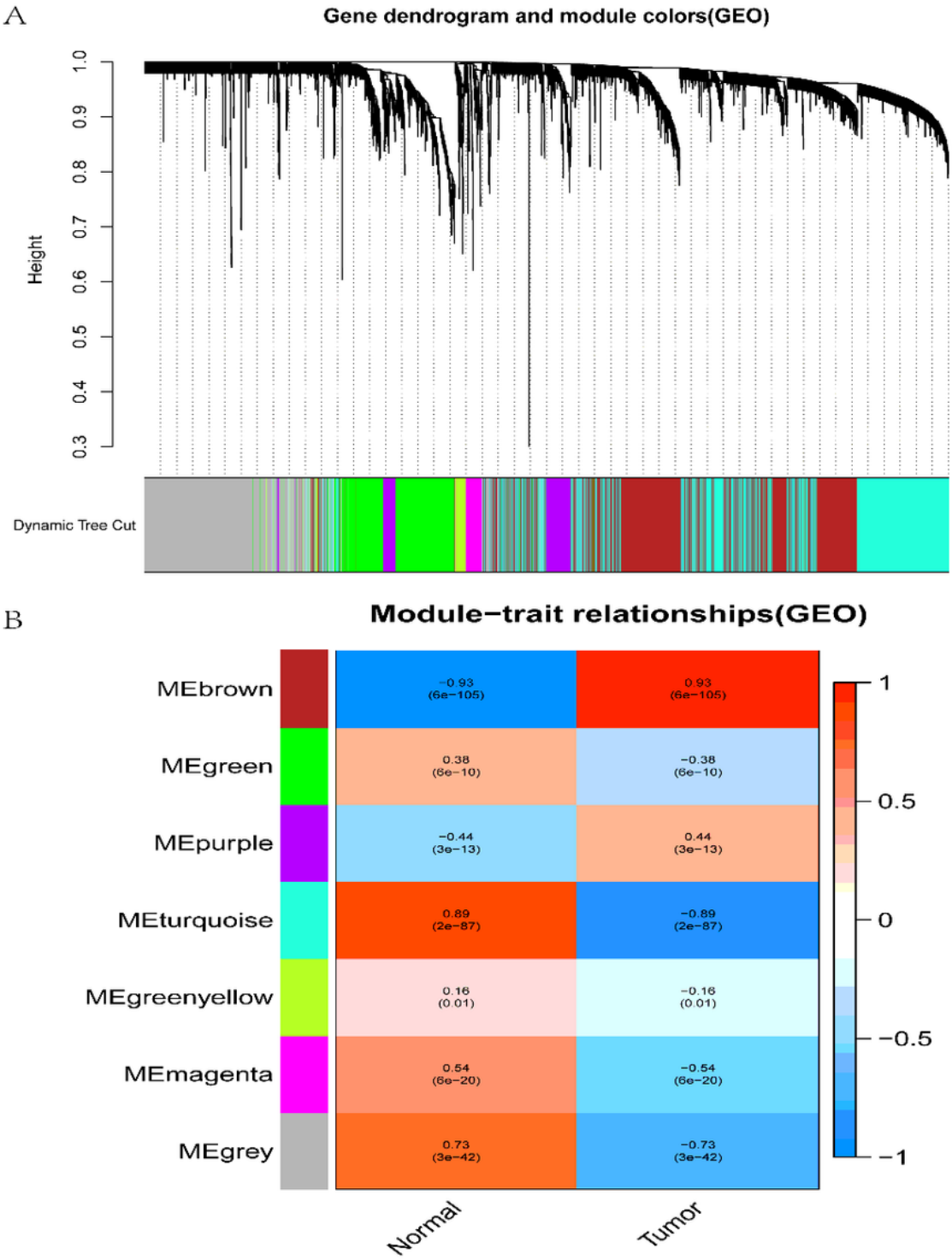

Figure 3

Identify modules related to clinical information in the GSE44076 dataset. (A) The clustering tree of co-expression network modules was sorted by gene hierarchical clustering based on 1-TOM matrix. Each

module was assigned a different color. (B) Module trait relationship. Each row corresponds to a color module and each column corresponds to a clinical feature (cancer and normal). Each cell contains the corresponding correlation and P value.

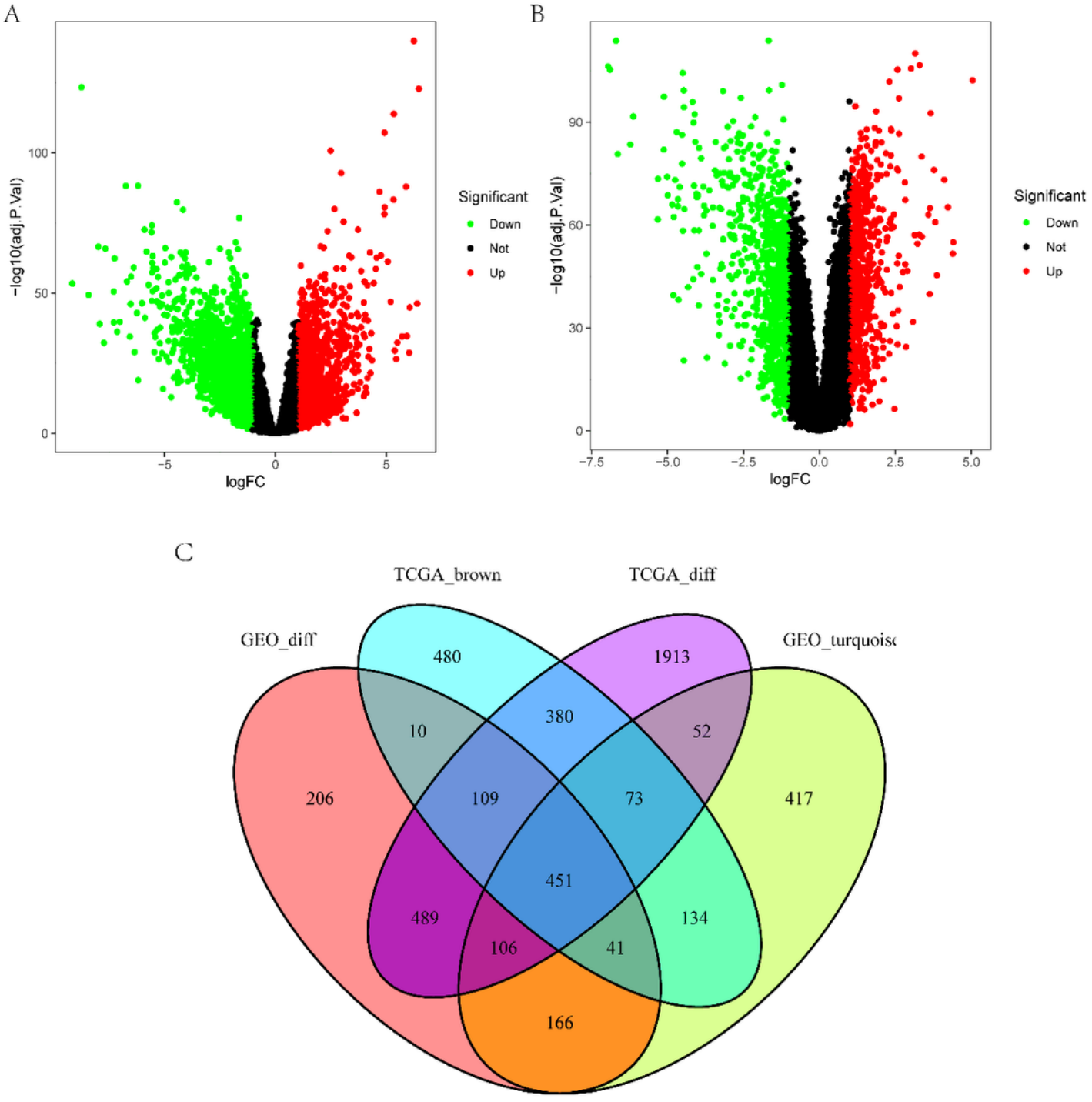

Figure 4

Differentially expressed genes (DEGs) were obtained. (A) Volcano plot of DEGs in the TCGA dataset by Using | logFC | ≥ 1.0 and P < 0.05 as cut-off criteria. (B) Volcano plot of DEGs in the GSE44076 dataset. (C) The genes in DEG tables and were co-expressed by Venn diagram. A total of 451 overlapping genes were located at the intersection of DEG list and two co-expression modules.

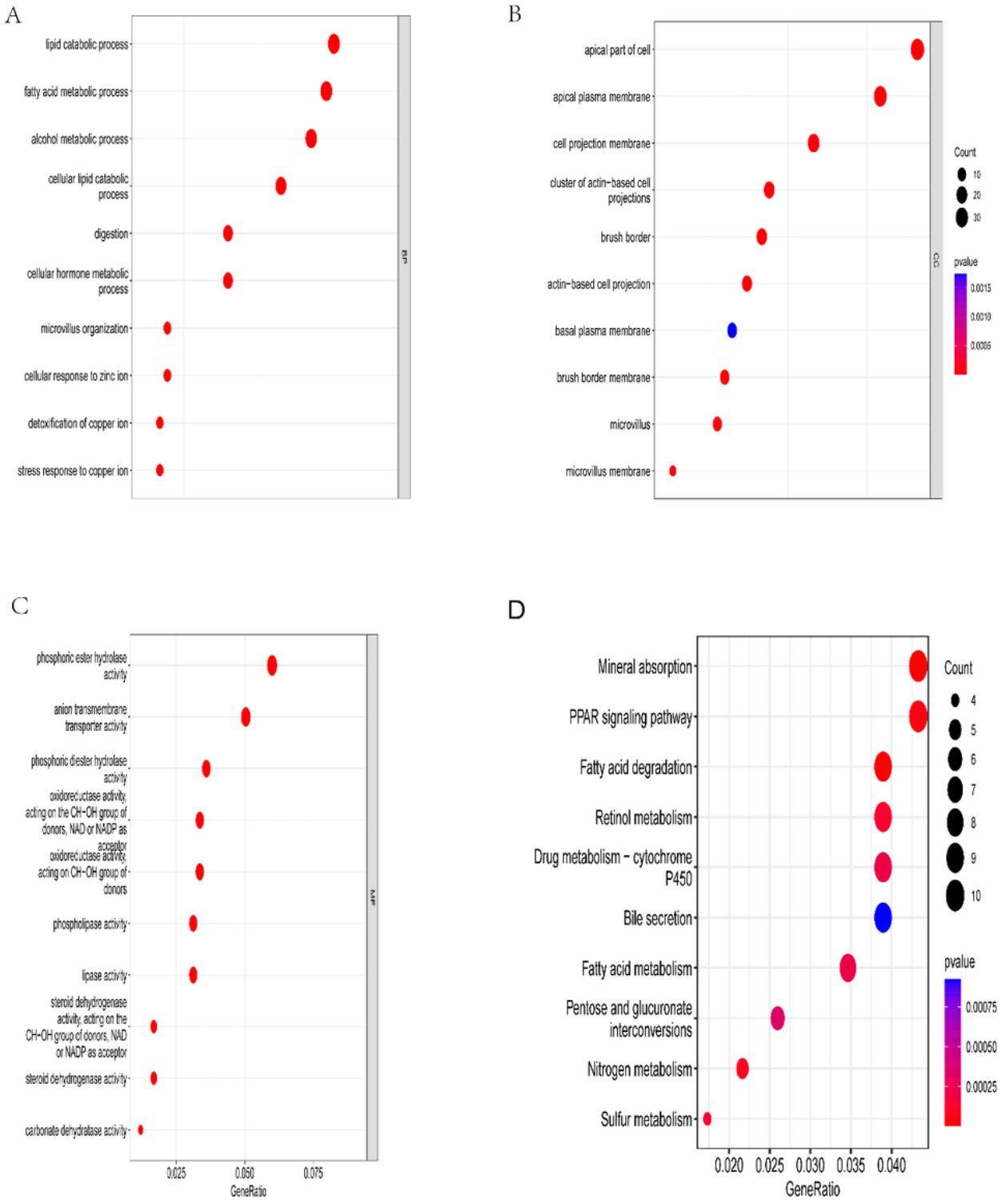

Figure 5

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of genes in DEGs. The color represents the adjusted p value (BH), and the size of the spot represents the number of genes. (A) Top ten in BP. (B) Top ten in CC. (C) Top ten in MF. (D) Top ten in KEGG.

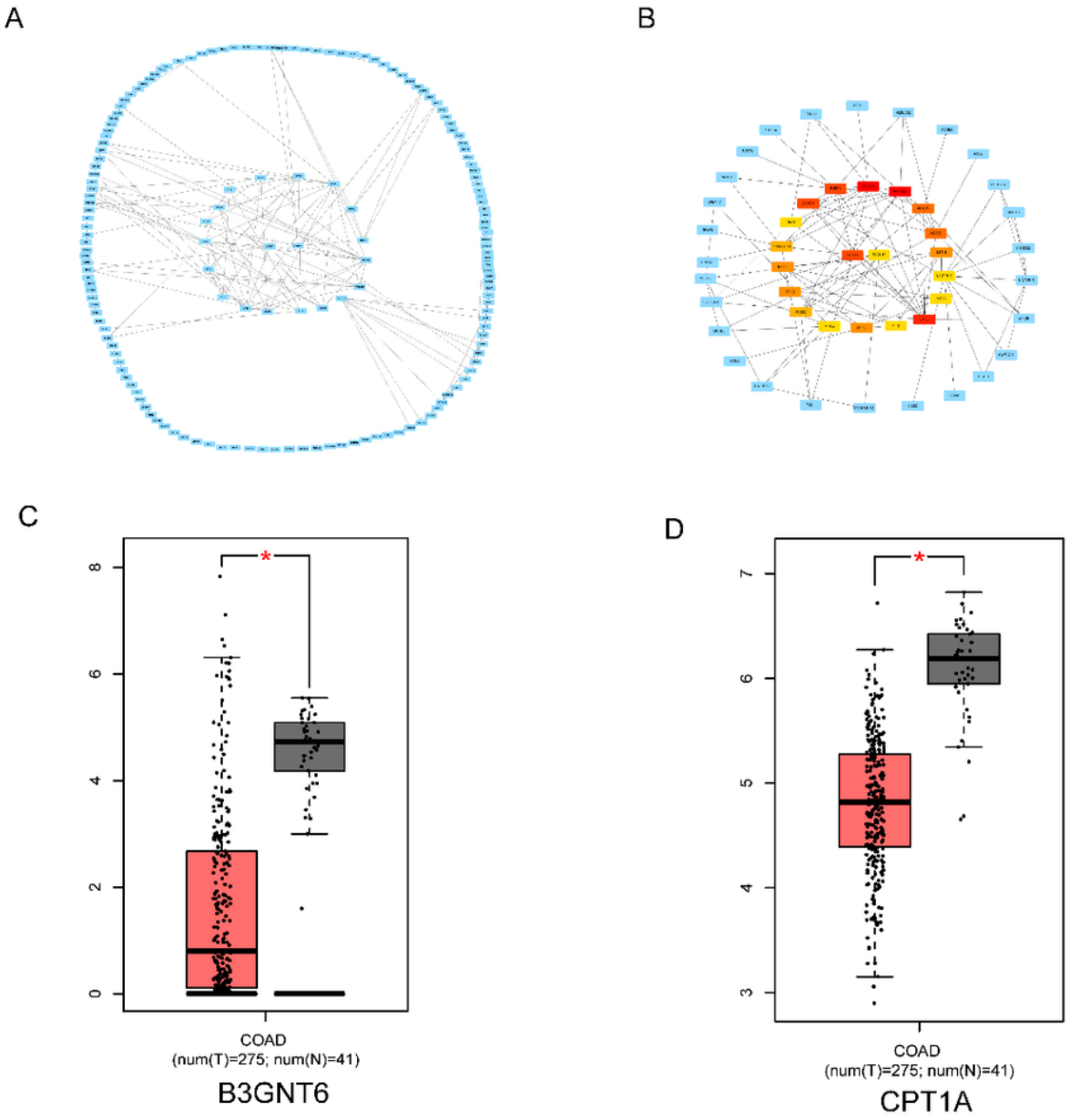

Figure 6

Visualization of protein-protein interaction (PPI) networks and candidate hub genes. (A) The gene PPI network co-expression module between DEGs. Blue nodes represent genes. Edges represent interactive associations between nodes. (B) The maximum cluster center (MCC) algorithm is used to identify hub gene network from PPI. Edges represent protein-protein associations. Red nodes represent genes with high MCC soreness, while yellow nodes represent genes with low MCC values. (C) B3GNT6 expression was verified by GEPIA2. (D) CPT1A expression was verified by GEPIA2.

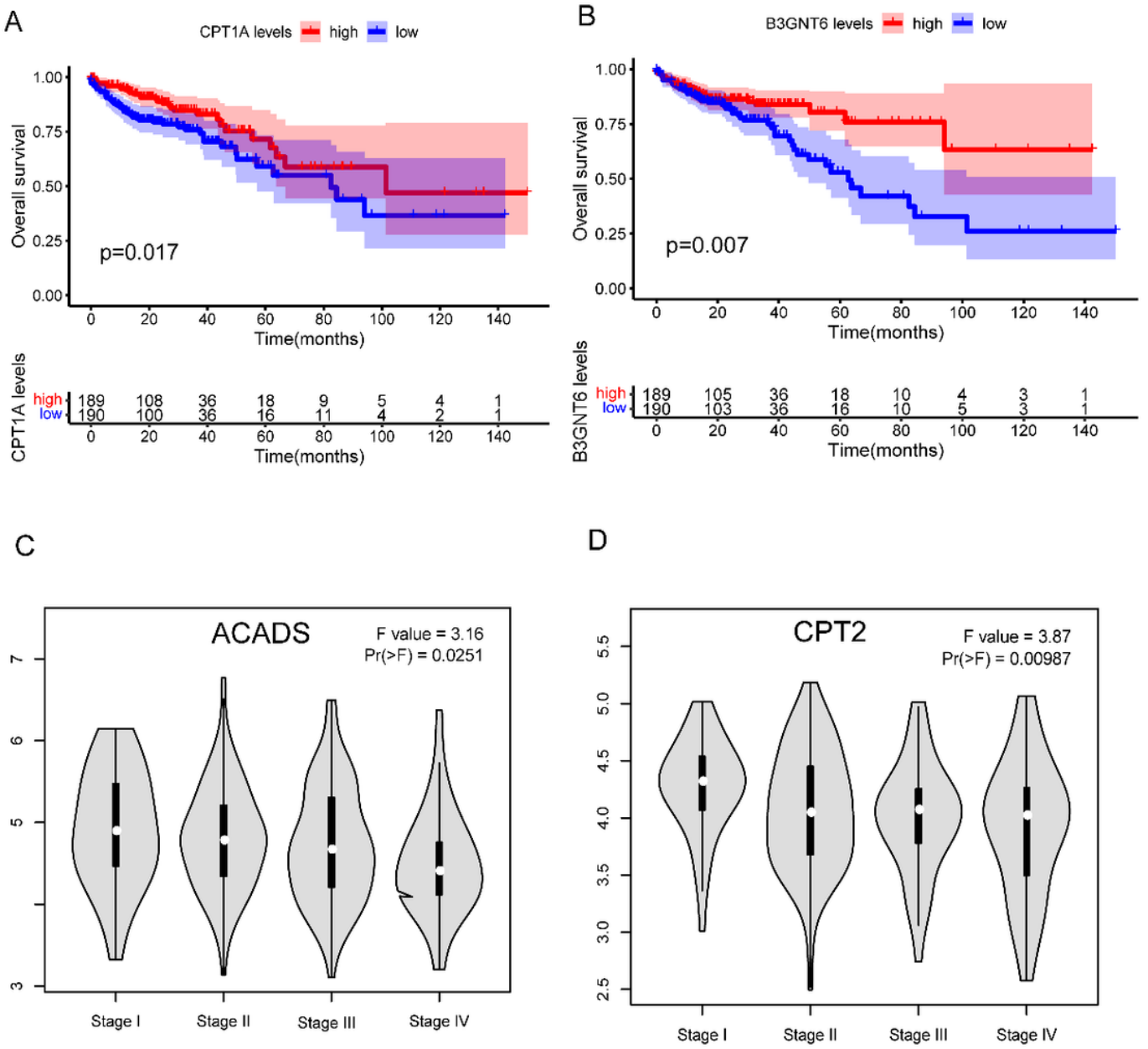

Figure 7

(A) Overall survival (OS) analysis of meaningful CPT1A gene in COAD patients from Kaplan–Meier univariate survival analysis. (B) Overall survival (OS) analysis of meaningful B3GNT6 gene in COAD patients from Kaplan–Meier univariate survival analysis. (C) Disease free survival (DFS) analysis of significant ACADS gene in COAD patients in GEPIA2 database. (D) Disease free survival (DFS) analysis of meaning CPT2 gene in COAD patients in GEPIA2 database.

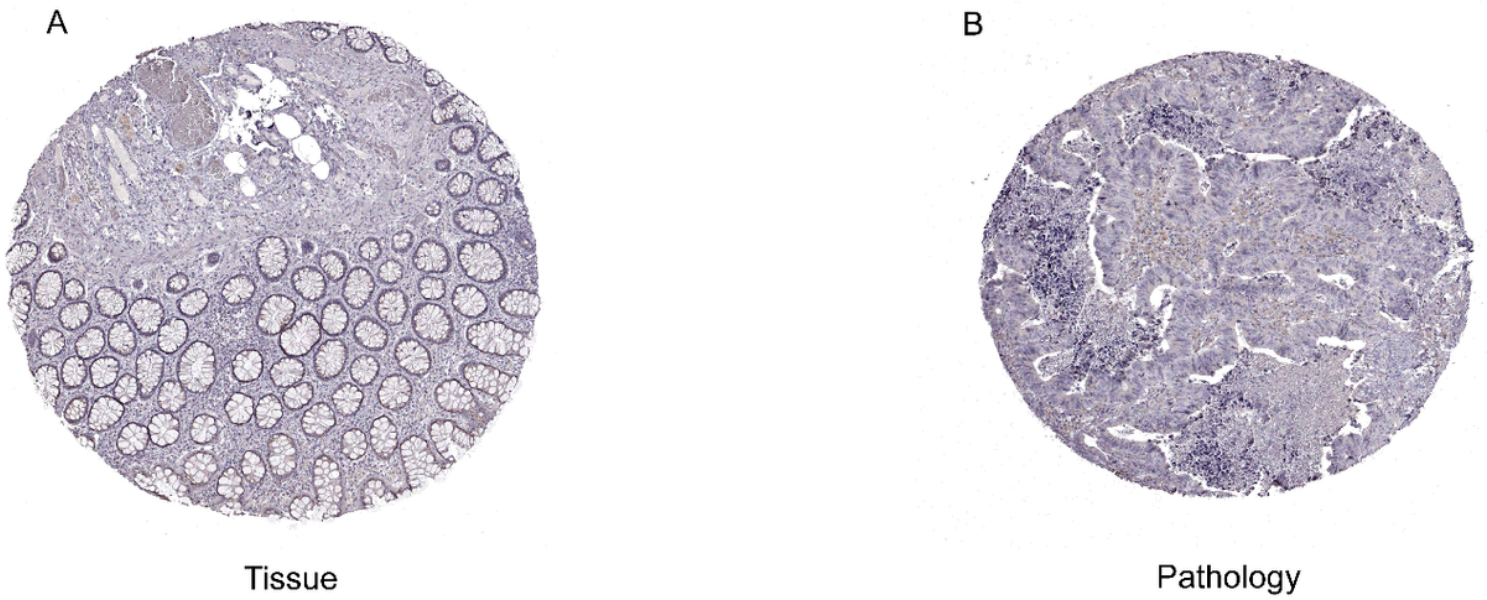

Figure 8

Immunohistochemistry of B3GNT6 gene in COAD and normal tissues in human protein atlas (HPA) database. (A) Protein level of B3GNT6 in normal colonic mucosa. (B) Protein level of B3GNT6 in COAD.

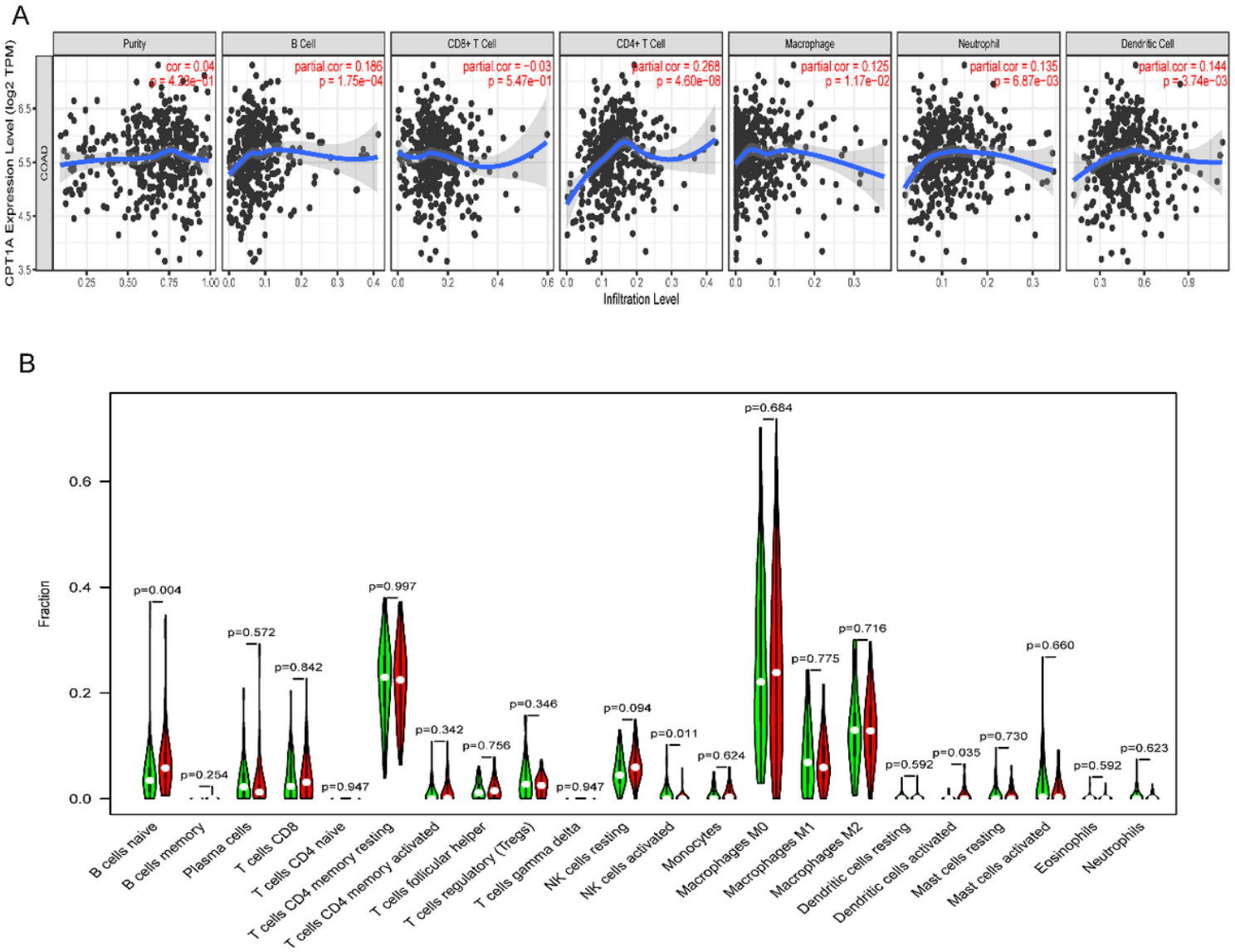

Figure 9

Correlation between CPT1A and tumor immune infiltrating cells. (A) Correlation between CPT1A expression and immune infiltrating cells in COAD. (B) Changes of 22 immune cell subtypes in high and low expression groups of CPT1A in COAD.

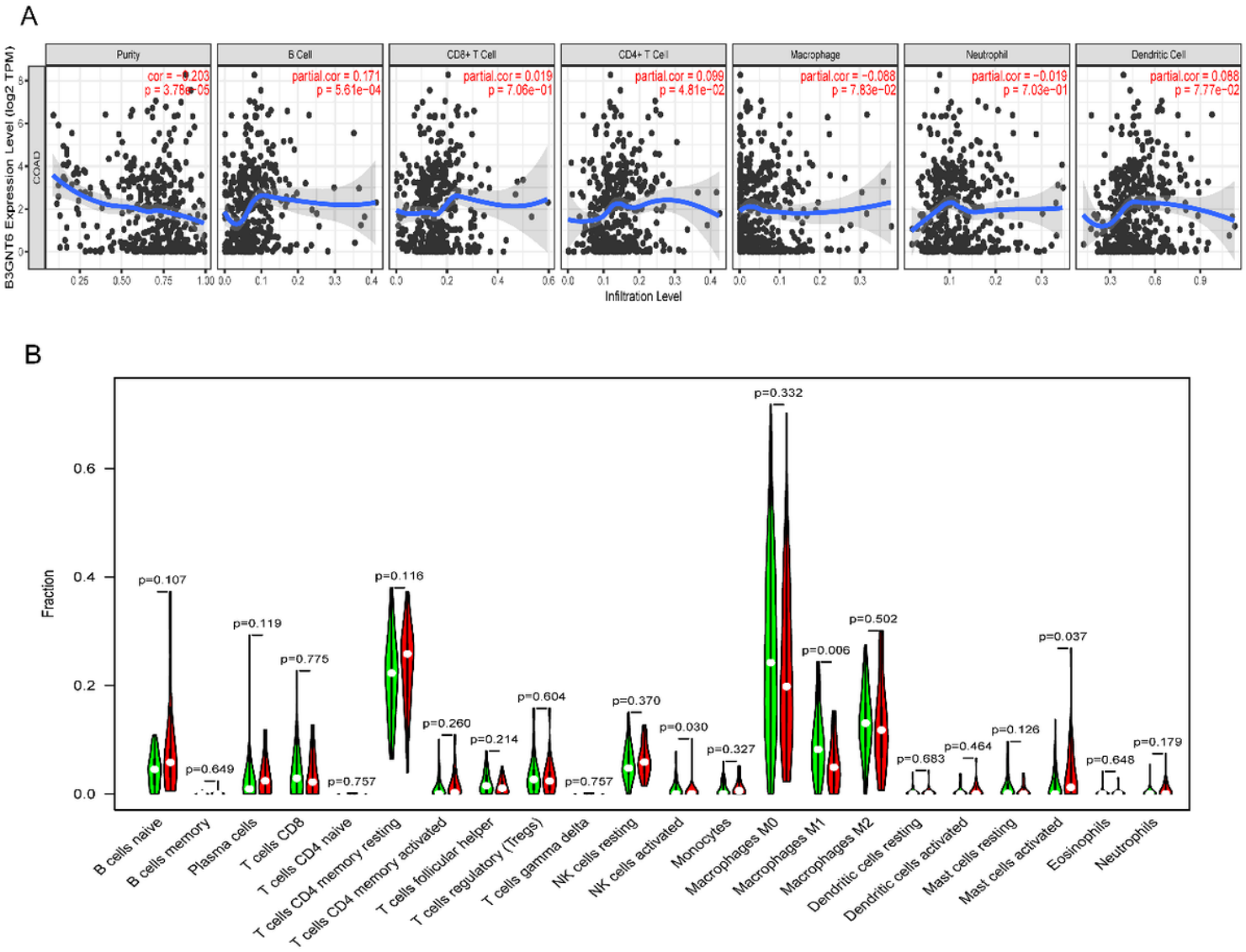

Figure 10

Correlation between B3GNT6 and tumor immune infiltrating cells. (A) Correlation between B3GNT6 expression and immune infiltrating cells in COAD. (B) Changes of 22 immune cell subtypes in high and low expression groups of B3GNT6 in COAD.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement.docx