SPANDIDOS PUBLICATIONS

ONCOLOGY LETTERS 18: 5043-5054, 2019

Four targeted genes for predicting the prognosis of colorectal cancer: A bioinformatics analysis case

QINGLAI BIAN[1], JIAXU CHEN[1,2], WENQI QIU[1], CHENXI PENG[1], MEIFANG SONG[1], XUEBIN SUN[1], YUEYUN LIU[1], FENGMIN DING[3], JIANBEI CHEN[1] and LIQING ZHANG[4]

[1]School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100029; [2]Formula-Pattern Research Center, School of Traditional Chinese Medicine, Jinan University, Guangzhou, Guangdong 510632; [3]School of Basic Medical Science, Hubei University of Chinese Medicine, Wuhan, Hubei 430065, P.R. China; [4]Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA

Received February 1, 2019; Accepted August 13, 2019

DOI: 10.3892/ol.2019.10866

Abstract. The molecular mechanisms underlying the development and progression of colorectal cancer (CRC) have not been clarified. The purpose of the present study was to identify key genes that may serve as novel therapeutic targets or prognostic predictors in patients with CRC using bioinformatics analysis. Four gene expression datasets were downloaded from the Gene Expression Omnibus database, which revealed 19 upregulated and 34 downregulated differentially expressed genes (DEGs). The downregulated DEGs were significantly enriched in eight pathways according to Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis. A protein-protein interaction network was constructed with 52 DEGs and 458 edges. Ten key genes were identified according to the degree value, betweenness centrality and closeness centrality. Survival analysis revealed that low expression of four of the ten genes, carcinoembryonic antigen related cell adhesion molecule 7 (CEACAM7), solute carrier family 4 member 4 (SLC4A4), glucagon (GCG) and chloride channel accessory 1 (CLCA1) genes, were associated with unfavorable prognosis in CRC. Furthermore, gene set enrichment analysis revealed that two pathways were significantly enriched in the CEACAM7 low-expression group. Thus, CEACAM7, SLC4A4, GCG and CLCA1 may be prognostic markers or therapeutic targets of CRC. Low CEACAM7 expression may be associated with the activation of glycosaminoglycan biosynthesis-chondroitin sulfate and extracellular matrix receptor interaction pathways and may affect the prognosis of CRC.

Correspondence to: Professor Jiaxu Chen, School of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Yifu Science and Technology Building, 11 Bei San Huan Dong Lu, Beijing 100029, P.R. China
E-mail: chenjiaxu@hotmail.com

Dr Liqing Zhang, Department of Computer Science, Virginia Tech, 2160K Torgersen Hall, Blacksburg, VA 24060, USA
E-mail: lqzhang@cs.vt.edu

Key words: bioinformatics, genes, prognosis, colorectal cancer

Introduction

Colorectal cancer (CRC), which includes colon cancer and rectum cancer, is the third most common type of cancer and was the second leading cause of cancer-associated mortality for men and women worldwide in 2018 (1). It was estimated that there were more than 1.8 million new cases and over 0.8 million mortalities worldwide due to CRC in 2018 (1). In the USA, the number of newly diagnosed CRC cases in 2019 was ~145,600, accounting for 8.3% of all new cancer cases (2). Furthermore, the number of CRC-associated deaths in 2019 was ~51,020 in the USA, accounting for 8.4% of all cancer-associated deaths (2). The survival and prognosis of patients with CRC are closely associated with the staging of the tumor. If the tumors are diagnosed early and removed, the disease may be curable. The 5-year survival rate of patients with localized CRC was ~90% in the USA between 2001 and 2007 (3). However, in CRC cases at regional and distant stages, the 5-year survival rates are only ~70 and ~11%, respectively (3). Unfortunately, in developed countries including the USA, only ~40% of patients with CRC are diagnosed at early stages (4). Therefore, the identification of diagnostic and prognostic molecular markers for early detection and the prediction of prognosis for patients with CRC is clinically important.

The development of CRC involves interconnections between environmental and genetic factors. In recent decades, great progress has been made in understanding the molecular pathogenesis of CRC, which includes four main mechanisms of molecular pathogenesis: Adenoma-carcinoma sequence, inherited forms, mismatch repair deficiency, and high-level microsatellite instability (5). However, the precise molecular mechanisms underlying the development of CRC have not been fully elucidated. With the rapid development of bioinformatics and high-throughput platforms for detecting gene expression, screening key genes for CRC based on publicly available databases provides a strategy to clarify the molecular mechanisms of CRC. Large numbers of gene expression datasets for CRC are available in the Gene Expression Omnibus (GEO)

database, and numerous studies have used these datasets for the identification of differentially expressed genes (DEGs) in CRC (6-9). Previous studies have revealed the prognostic value of certain DEGs in CRC (10-13). However, the results of these individual studies varied and the studies demonstrated differences in sample collection, platform types and analysis methods. Furthermore, large-scale studies on the prognostic value of the DEGs in CRC are lacking. In addition, the enrichment pathways, gene set enrichment analysis, Gene Ontology (GO) functions and the interaction network involved in the DEGs remain to be clarified.

In order to overcome these shortcomings, the present study integrated and reanalyzed four online GEO datasets of CRC using bioinformatics analysis methods. DEGs between CRC samples and noncancerous samples were determined, and the interaction network among these DEGs was constructed. Enrichment analysis for these DEGs was conducted using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and GO functions. Through network analysis, gene expression confirmation and overall survival analysis, four key genes that were associated with the prognosis of CRC were identified. These four genes may provide valuable information for the identification of potential prognostic markers of CRC and to elucidate the molecular mechanism of CRC.

Materials and methods

Data source and identification of DEGs. In order to compensate for the limitation of small sample size and result offset in a single cohort study, four gene expression profiles (GSE113513, GSE87211, GSE35279 and GSE24551) were acquired from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) between January 1 2010 and August 31 2018. These profiles included both CRC samples and noncancerous samples, and all datasets contained at least five samples in each group (Table I). GSE113513 (unpublished, 2018; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113513) was based on the platform GPL15207 [(PrimeView) Affymetrix Human Gene Expression Array], GSE87211 was based on the platform GPL13497 (Agilent-026652 Whole Human Genome Microarray 4x44K v2) (6), GSE35279 was based on the platform GPL6480 (Agilent-014850 Whole Human Genome Microarray 4x44K G4112F) (7), and GSE24551 was based on the platform GPL5175 [(HuEx-1_0-st) Affymetrix Human Exon 1.0 ST Array] [transcript (gene) version] (8).

The DEGs between CRC samples and noncancerous samples in the four GEO series were filtered using the online GEO2R tool (https://www.ncbi.nlm.nih.gov/geo/geo2r/) (14,15). Genes that satisfied the threshold [$|\log_2$fold change (FC)$| \geq 2.0$; adjusted $P<0.05$] were classified as DEGs. Statistical analysis was performed for each dataset. The overlapping genes among the four profiles were determined and presented by the online tool Venny 2.1.0 (http://bioinfogp.cnb.csic.es/tools/venny/index.html) (16). The expression of the upregulated and downregulated DEGs in each dataset was visualized using a volcano plot generated by Sanger Box (version 0.0.9; http://sangerbox.com/).

GO and KEGG enrichment analysis. GO analysis is commonly performed for functional enrichment analysis, in which gene function is classified into biological process (BP), molecular function (MF) and cellular component (CC) terms (17,18). KEGG is frequently used for exploring the advanced functions and mechanisms involved in the biological system at the molecular level (19). In the present study, GO and KEGG analysis was completed with the Database for Annotation, Visualization and Integrated Discovery (DAVID) platform (version 6.8; http://david.ncifcrf.gov/) (20,21). Statistical significance was set as $P<0.05$.

Construction of a protein-protein interaction (PPI) network and identification of key genes. The GeneMANIA (http://genemania.org/) prediction server was designed to assess the PPI network (22). The network was analyzed and visualized by Cytoscape 3.6.1 (http://www.cytoscape.org/). In the network, a high degree value indicated a more essential role for that gene. The degree value of each gene was calculated by the network analyzer tool that was built in the Cytoscape software. The genes whose degree value, closeness centrality and betweenness centrality were greater than the median value were identified as key genes.

Confirmation of key genes and overall survival analysis. The expression of key genes in CRC samples and noncancerous samples were further examined using the Gene Expression Profiling Interactive Analysis (GEPIA) platform (http://gepia.cancer-pku.cn) (23). The expression profiles of key genes between tumor samples and adjacent normal samples of colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) were obtained from The Cancer Genome Atlas (TCGA) database by GEPIA (http://gepia.cancer-pku.cn) (23). Student's unpaired t-test was used to determine the statistical significance of the calculated differential expression. The fold change was defined as 2 and the P-value of significance was set at 0.01.

Overall survival analysis was performed in GEPIA by log-rank test based on gene expression. The overall survival analysis plot also contained the Cox proportional hazard ratio and the 95% confidence interval. The patients with CRC were divided into low-expression and high-expression groups, based on the median value of mRNA expression of the ten key genes. The differences between the two groups were evaluated separately for each of the ten key genes. $P<0.05$ was considered to indicate a statistically significant result.

Gene set enrichment analysis (GSEA) of prognosis-associated key genes. The pre-processed level 3 RNA-seq data and corresponding clinical information of patients with CRC were collected from the TCGA-COAD and TCGA-READ datasets (https://cancergenome.nih.gov/, updated in September 2018). GSEA (http://www.broadinstitute.org/gsea/index.jsp) (24,25) for the prognosis-associated key genes was performed on the TCGA datasets. The c2.cp.kegg.v6.0.symbols.gmt dataset was obtained from the molecular signatures database (MSigDB) v6.0 on the GSEA website. The CRC samples obtained from the TCGA database were divided into high- and low-expression groups according to the median expression level of prognosis-associated genes. The samples were analyzed by default weighted enrichment statistics using the GSEA 3.0 software. In the present study, the gene sets satisfying nominal $P<0.05$ and false discovery rate (FDR)<0.25 were considered

Table I. Sample numbers in the four GSE datasets.

Dataset ID	CRC	Noncancerous tissues	Total
GSE113513	14	14	28
GSE87211	203	160	363
GSE35279	74	5	79
GSE24551	160	13	173

CRC, colorectal cancer.

to be significantly enriched. The enrichment analysis was carried out by default weighted enrichment statistics, and the number of random permutations was set to 1,000 times.

Results

Identification of DEGs. A flow chart of the present study design is presented in Fig. 1. Four GEO datasets (GSE113513, GSE87211, GSE35279 and GSE24551) were downloaded. The numbers of CRC and noncancerous samples in each dataset are presented in Table I. GSE113513 consisted of 14 CRC samples and 14 noncancerous tissues samples; GSE87211 included 203 CRC samples and 160 noncancerous tissues samples; GSE35279 contained 74 CRC samples and 5 noncancerous tissues samples; and GSE24551 included 160 CRC samples and 13 noncancerous tissues samples. GSE113513 comprised 340 DEGs, 258 of which were upregulated genes and 82 were downregulated genes. GSE87211 comprised 971 DEGs, including 573 upregulated genes and 398 downregulated genes. GSE35279 included 1371 DEGs, with 222 upregulated genes and 1149 downregulated genes. GSE24551 comprised 213 DEGs, with 109 upregulated and 104 downregulated genes. The intersection of the DEGs is presented in Venn diagrams (Fig. 2). The DEGs in each dataset are presented in volcano plots (Fig. 3). In total, 53 common DEGs (19 upregulated and 34 downregulated) were identified among all four datasets.

GO and KEGG enrichment analysis. GO and KEGG analyses of the DEGs were performed using DAVID and the results are summarized in Table II. BP analysis indicated that upregulated genes were enriched in proteolysis, response to tumor necrosis factor, and positive regulation of gene expression, whereas the downregulated genes were enriched in seven terms comprised 'bicarbonate transport', 'one-carbon metabolic process', 'regulation of chloride transport', 'positive regulation of cellular pH reduction', 'chloride transmembrane transport', 'ethanol oxidation' and 'positive regulation of synaptic transmission'. Analysis of the CC function indicated that the main enriched functions of upregulated genes involved the proteinaceous extracellular matrix and extracellular space, whereas the downregulated genes involved the seven terms 'apical plasma membrane', 'basolateral plasma membrane', 'extracellular exosome', 'anchored component of membrane', 'plasma membrane', 'zymogen granule membrane' and 'integral component of membrane'. Analysis of the MF, identified calcium ion binding as a significantly enriched term of upregulated genes, whereas nine terms were significantly enriched in downregulated genes, including 'carbonate dehydratase activity', 'chloride channel activity', 'zinc-dependent alcohol dehydrogenase activity', 'arylesterase activity', 'alcohol dehydrogenase (NAD) activity', 'zinc ion binding', 'intracellular calcium activated chloride channel activity', 'retinol dehydrogenase activity' and 'carbohydrate binding'. Furthermore, KEGG analysis revealed that downregulated genes were significantly enriched in nitrogen metabolism, proximal tubule bicarbonate reclamation and six other pathways.

PPI network and key genes. The PPI network of the 53 DEGs was generated with the GeneMANIA platform (22) and visualized by Cytoscape (26) (Fig. 4). Following the removal of the single epoxide hydrolase 4 gene that had no connections in the network, the network contained 52 DEGs and 458 edges. The connections between the DEGs included physical interactions, co-expression and co-localization. Among the DEGs, ten genes were selected as key genes associated with CRC. The carbonic anhydrase 2 (CA2) gene was the most connected gene (degree, 44), followed by the guanylate cyclase activator 2A gene (GUCA2A; degree, 35), and carcinoembryonic antigen-related cell adhesion molecule 7 gene (CEACAM7; degree, 34), among others (Table III). Only one gene, matrix metalloprotease 7 (MMP7), was upregulated in the CRC samples; the other nine genes were downregulated.

Confirmation and survival analysis of ten key genes. GEPIA (23) was used to compare the expression levels of the ten key genes between CRC tumor samples and adjacent normal samples in COAD and READ obtained from the TCGA database. A total of 275 tumor samples and 41 adjacent normal samples in COAD and 92 tumor samples and ten adjacent normal samples in READ were available for analyses. Among the ten genes, only MMP7 was significantly upregulated in tumor samples compared with control samples, whereas the other nine genes were significantly downregulated in tumor samples (Fig. 5). The differential expression of the ten key genes in the present study was confirmed in GEPIA.

The overall survival analysis of the ten genes was also obtained from GEPIA. Among these key genes, low expression of CEACAM7, solute carrier family 4 member 4 (SLC4A4), glucagon (GCG) and chloride channel accessory 1 (CLCA1) genes were significantly associated with an unfavorable outcome of CRC (Fig. 6).

GSEA of the four prognosis-associated key genes. The mechanism of the four prognosis-associated genes (CEACAM7, SLC4A4, GCG and CLCA1) was further investigated by examining the associated pathways via GSEA. The expression matrix of CRC from the TCGA database was divided into high-expression (323 samples) and low-expression (324 samples) groups, according to the median expression level of CEACAM7, SLC4A4, GCG and CLCA1. In the low CEACAM7 expression group, two significantly enriched KEGG pathways at nominal $P<0.05$ and FDR<0.25 were identified, including glycosaminoglycan biosynthesis-chondroitin sulfate (nominal $P=0.002$; FDR, 0.101) and extracellular matrix-receptor interaction (nominal $P=0.022$; FDR, 0.151) (Fig. 7). However, no pathway was significantly enriched in the low SLC4A4, GCG and CLCA1 expression groups.

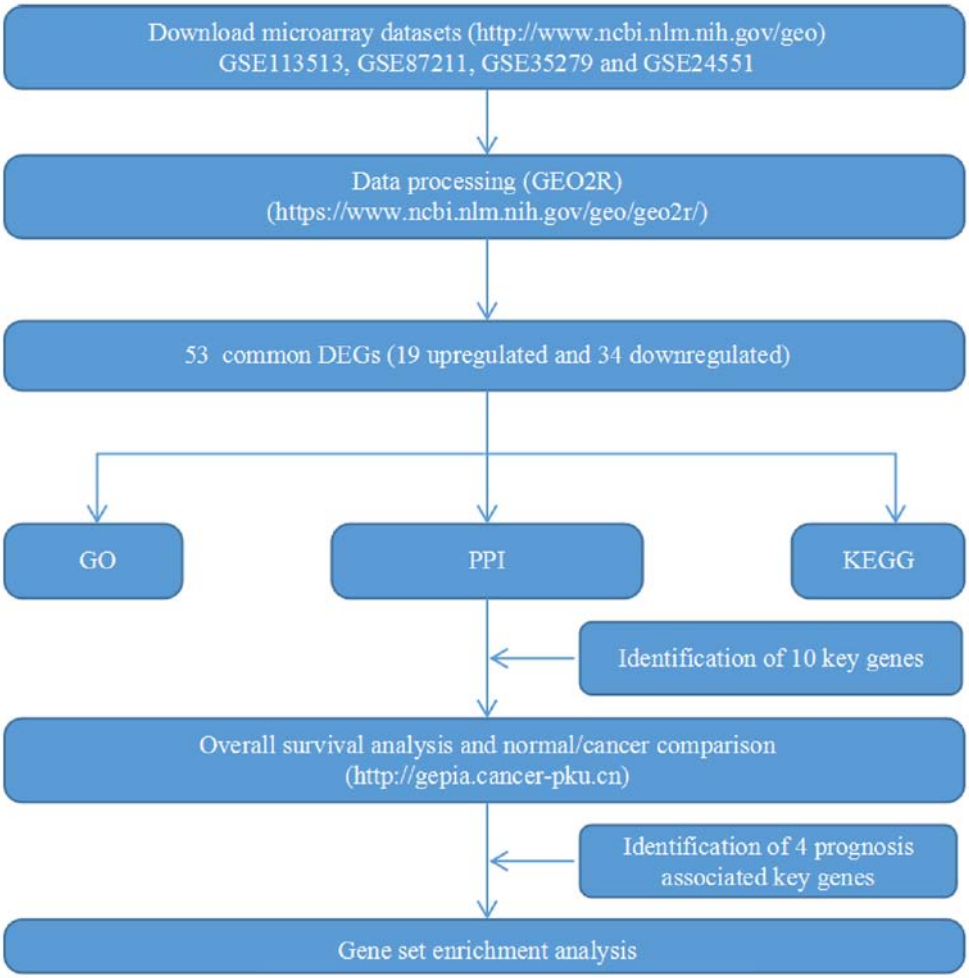

Figure 1. Flow chart of the present study. GEO, Gene Expression Omnibus; GO, Gene Ontology; PPI, protein-protein interaction; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes.

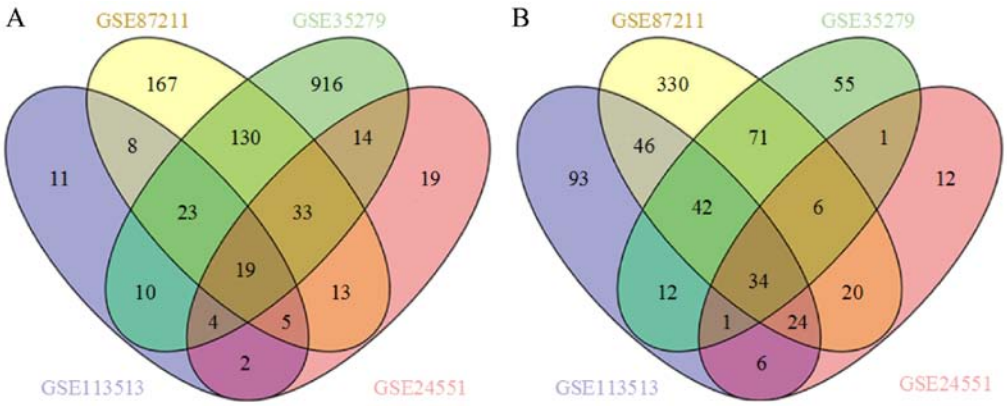

Figure 2. Venn diagram of DEGs in datasets GSE113513, GSE87211, GSE35279 and GSE24551. (A) A total of 19 upregulated common DEGs were extracted from datasets GSE113513, GSE87211, GSE35279 and GSE24551 with a threshold of $|\log_2 FC| \geq 2.0$ and adjusted $P<0.05$. (B) In total, 34 downregulated common DEGs were extracted from datasets GSE113513, GSE87211, GSE35279 and GSE24551 with a threshold of $|\log_2 FC| \geq 2.0$ and adjusted $P<0.05$. FC, fold change; DEGs, differentially expressed genes.

Discussion

In the present study, 53 DEGs were identified in CRC samples compared with noncancerous tissues samples, including 19 upregulated genes and 34 downregulated genes. The downregulated DEGs were significantly enriched in eight KEGG pathways, including nitrogen metabolism, proximal tubule bicarbonate reclamation and six other pathways. The upregulated DEGs were associated with six GO terms: Proteolysis, response to tumor necrosis factor, positive regulation of gene expression, proteinaceous extracellular matrix, extracellular space and calcium ion binding. The downregulated DEGs

Table II. Enriched GO terms and KEGG pathways of DEGs.

A, Upregulated expression

Category	Description	P-value
BP	GO:0006508-proteolysis	1.10×10^{-2}
BP	GO:0034612-response to tumor necrosis factor	2.36×10^{-2}
BP	GO:0010628-positive regulation of gene expression	2.52×10^{-2}
CC	GO:0005578-proteinaceous extracellular matrix	2.18×10^{-3}
CC	GO:0005615-extracellular space	8.32×10^{-3}
MF	GO:0005509-calcium ion binding	3.87×10^{-2}

B, Downregulated expression

Category	Description	P-value
BP	GO:0015701-bicarbonate transport	1.11×10^{-8}
BP	GO:0006730-one-carbon metabolic process	1.91×10^{-7}
BP	GO:2001225-regulation of chloride transport	3.45×10^{-3}
BP	GO:0032849-positive regulation of cellular pH reduction	6.89×10^{-3}
BP	GO:1902476-chloride transmembrane transport	1.12×10^{-2}
BP	GO:0006069-ethanol oxidation	2.05×10^{-2}
BP	GO:0032230-positive regulation of synaptic transmission, GABAergic	2.05×10^{-2}
CC	GO:0016324-apical plasma membrane	1.61×10^{-3}
CC	GO:0016323-basolateral plasma membrane	3.81×10^{-3}
CC	GO:0070062-extracellular exosome	6.42×10^{-3}
CC	GO:0031225-anchored component of membrane	1.67×10^{-2}
CC	GO:0005886-plasma membrane	1.71×10^{-2}
CC	GO:0042589-zymogen granule membrane	1.92×10^{-2}
CC	GO:0016021-integral component of membrane	4.47×10^{-2}
MF	GO:0004089-carbonate dehydratase activity	8.01×10^{-9}
MF	GO:0005254-chloride channel activity	4.13×10^{-3}
MF	GO:0004024-alcohol dehydrogenase activity, zinc-dependent	1.06×10^{-2}
MF	GO:0004064-arylesterase activity	1.06×10^{-2}
MF	GO:0004022-alcohol dehydrogenase (NAD) activity	1.24×10^{-2}
MF	GO:0008270-zinc ion binding	1.54×10^{-2}
MF	GO:0005229-intracellular calcium activated chloride channel activity	2.81×10^{-2}
MF	GO:0004745-retinol dehydrogenase activity	3.15×10^{-2}
MF	GO:0030246-carbohydrate binding	4.72×10^{-2}
KEGG	hsa00910-nitrogen metabolism	1.49×10^{-7}
KEGG	hsa04964-proximal tubule bicarbonate reclamation	4.18×10^{-5}
KEGG	hsa04972-pancreatic secretion	2.67×10^{-3}
KEGG	hsa00830-retinol metabolism	1.60×10^{-2}
KEGG	hsa00010-glycolysis/Gluconeogenesis	1.74×10^{-2}
KEGG	hsa00982-drug metabolism-cytochrome P450	1.79×10^{-2}
KEGG	hsa00980-metabolism of xenobiotics by cytochrome P450	2.10×10^{-2}
KEGG	hsa05204-chemical carcinogenesis	2.43×10^{-2}

GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, differentially expressed genes; BP, biological process; CC, cellular component; MF, molecular function; GABA, gamma aminobutyric acid.

were associated with 23 GO terms such as bicarbonate transport, one-carbon metabolic process and regulation of chloride transport, among others. A PPI network was constructed, consisting of 52 nodes and 458 edges, to evaluate the interactions among these DEGs. The ten key genes identified were CA2, GUCA2A, CEACAM7, MMP7, SLC4A4,

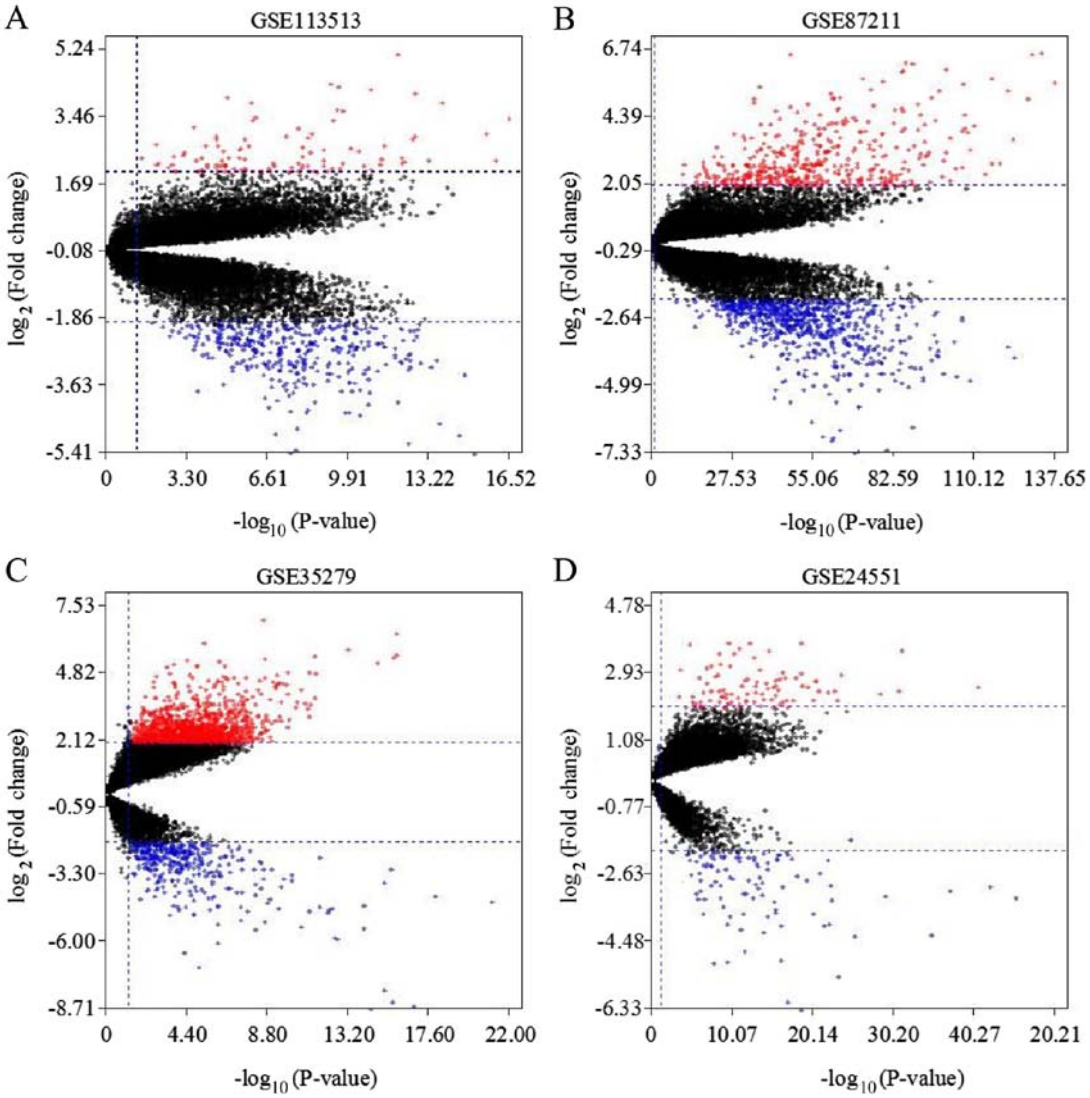

Figure 3. Volcano plot of DEGs in datasets GSE113513, GSE87211, GSE35279 and GSE24551. Red dots represent upregulated DEGs and blue dots represent downregulated DEGs with a threshold of $|\log_2 FC| \geq 2.0$ and adjusted $P<0.05$ in datasets (A) GSE113513, (B) GSE87211, (C) GSE35279 and (D) GSE24551. DEGs, differentially expressed genes; FC, fold change.

CA12, GCG, membrane spanning 4-domains A12, CA1 and CLCA1. Only the MMP7 gene was upregulated in patients with CRC, whereas the other nine genes were downregulated. Confirmation and survival analyses of these genes were performed using GEPIA. Survival analysis revealed that low expressions of CEACAM7, SLC4A4, GCG and CLCA1 genes were significantly associated with unfavorable prognosis in patients with CRC. Furthermore, GSEA results showed that the glycosaminoglycan biosynthesis-chondroitin sulfate and extracellular matrix-receptor interaction pathways were significantly enriched in the low CEACAM7 expression group of patients with CRC.

Genetic factors serve a critical role in the pathogenesis of CRC (27). CEACAM7, also termed CGM2, is a member of the carcinoembryonic antigen family and is expressed on highly differentiated colorectal epithelial cells and within ducts of pancreas epithelial cells (28). CEACAM7 sequences were detected only in human cDNA libraries of pancreas, pancreatic islets, colonic tumors and colon (29). The very narrow expression spectrum of CEACAM7 in pancreatic and colonic epithelial cells indicated a highly specialized function. Thompson *et al* (30) reported that CEACAM7 was downregulated in colorectal carcinoma. Messick *et al* (31) showed that CEACAM7 was significantly decreased in rectal cancer and considered a predictor for the recurrence of rectal cancer. Schölzel *et al* (28) identified that CEACAM7 was downregulated in hyperplastic polyps as well as early adenomas, which indicated early detected subtleties at the molecular level that lead to CRC.

The present study demonstrated that the glycosaminoglycan biosynthesis-chondroitin sulfate and extracellular matrix receptor interaction pathways were significantly enriched in the group with low expression of CEACAM7. The extracellular matrix components closely interact with cell surface receptors, growth factors and cytokines, supporting a substantial role for the extracellular matrix in the morphogenesis of tissues and organs and in maintaining the structure and function of cells and tissues. In addition, the functional macromolecules of extracellular matrix are involved in regulating the properties and function of cells. Notably,

Table III. Degree value of ten key genes.

Gene symbol	Description	Degree	Betweenness centrality	Closeness centrality
CA2	Carbonic anhydrase 2	44	7.61×10^{-2}	6.80×10^{-1}
GUCA2A	Guanylate cyclase activator 2A	35	8.12×10^{-2}	6.80×10^{-1}
CEACAM7	Carcinoembryonic antigen-related cell adhesion molecule 7	34	4.07×10^{-2}	6.30×10^{-1}
MMP7	Matrix metalloproteinase-7	33	4.37×10^{-2}	6.07×10^{-1}
SLC4A4	Solute carrier family 4 member 4	26	2.90×10^{-2}	5.86×10^{-1}
CA12	Carbonic anhydrase 12	25	3.27×10^{-2}	6.14×10^{-1}
GCG	Glucagon	25	1.79×10^{-2}	6.00×10^{-1}
MS4A12	Membrane-spanning 4-domains subfamily A member 12	24	1.04×10^{-2}	5.93×10^{-1}
CA1	Carbonic anhydrase 1	24	9.63×10^{-3}	5.93×10^{-1}
CLCA1	Calcium-activated chloride channel regulator 1	24	8.05×10^{-3}	5.73×10^{-1}

CA2, carbonic anhydrase 2; GUCA2A, guanylate cyclase activator 2A; CEACAM7, carcinoembryonic antigen-related cell adhesion molecule 7; MMP7, matrix metalloproteinase-7; SLC4A4, solute carrier family 4 member 4; CA, carbonic anhydrase 12; GCG, glucagon; MS4A12, membrane-spanning 4-domains subfamily A member 12; CLCA1, calcium-activated chloride channel regulator 1.

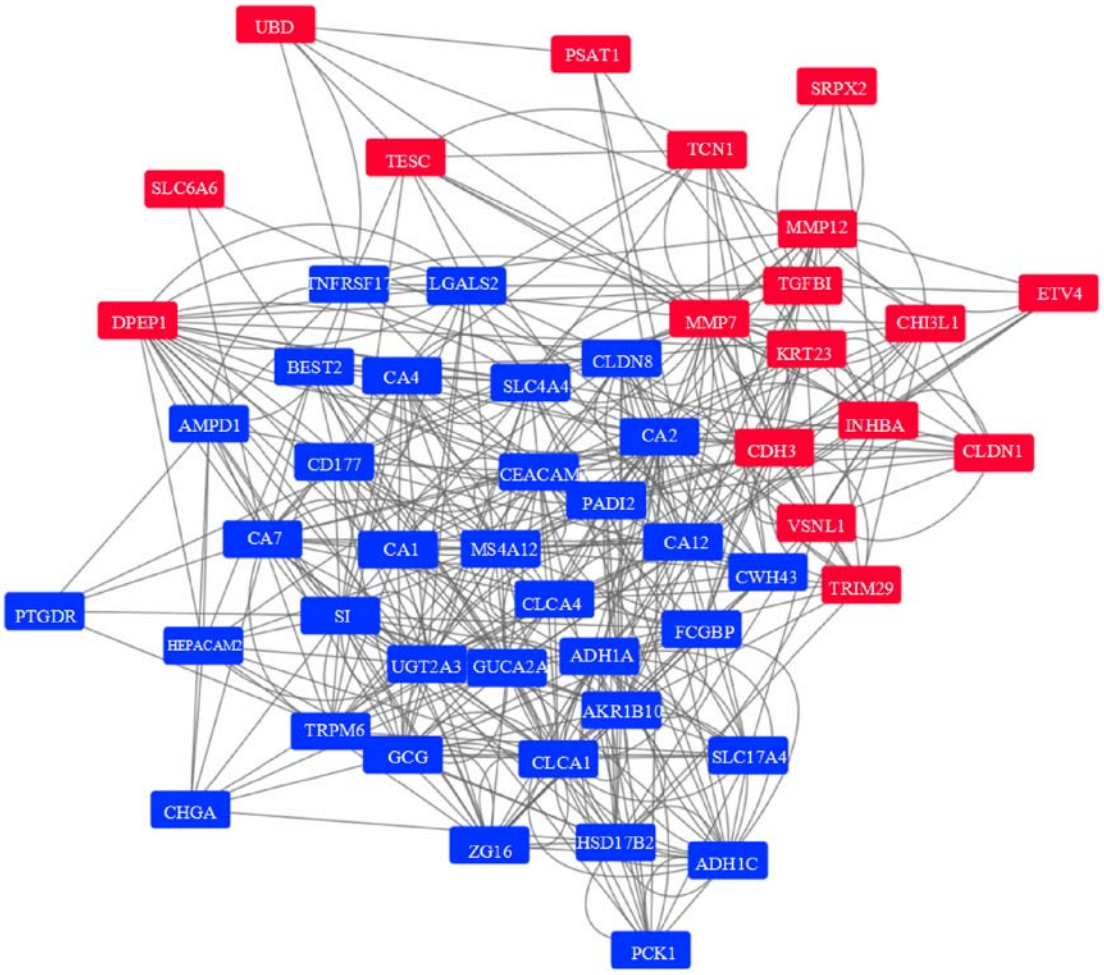

Figure 4. Protein-protein interaction network among the DEGs. Red nodes indicate upregulated DEGs while blue nodes indicate downregulated DEGs. The grey lines represent the connections between the DEGs. DEGs, differentially expressed genes.

surface molecules of matrix such as synaptophysin itself can also act as cell receptors or co-receptors. Consequently, the components of the extracellular matrix were closely associated with the cellular and molecular mechanisms of

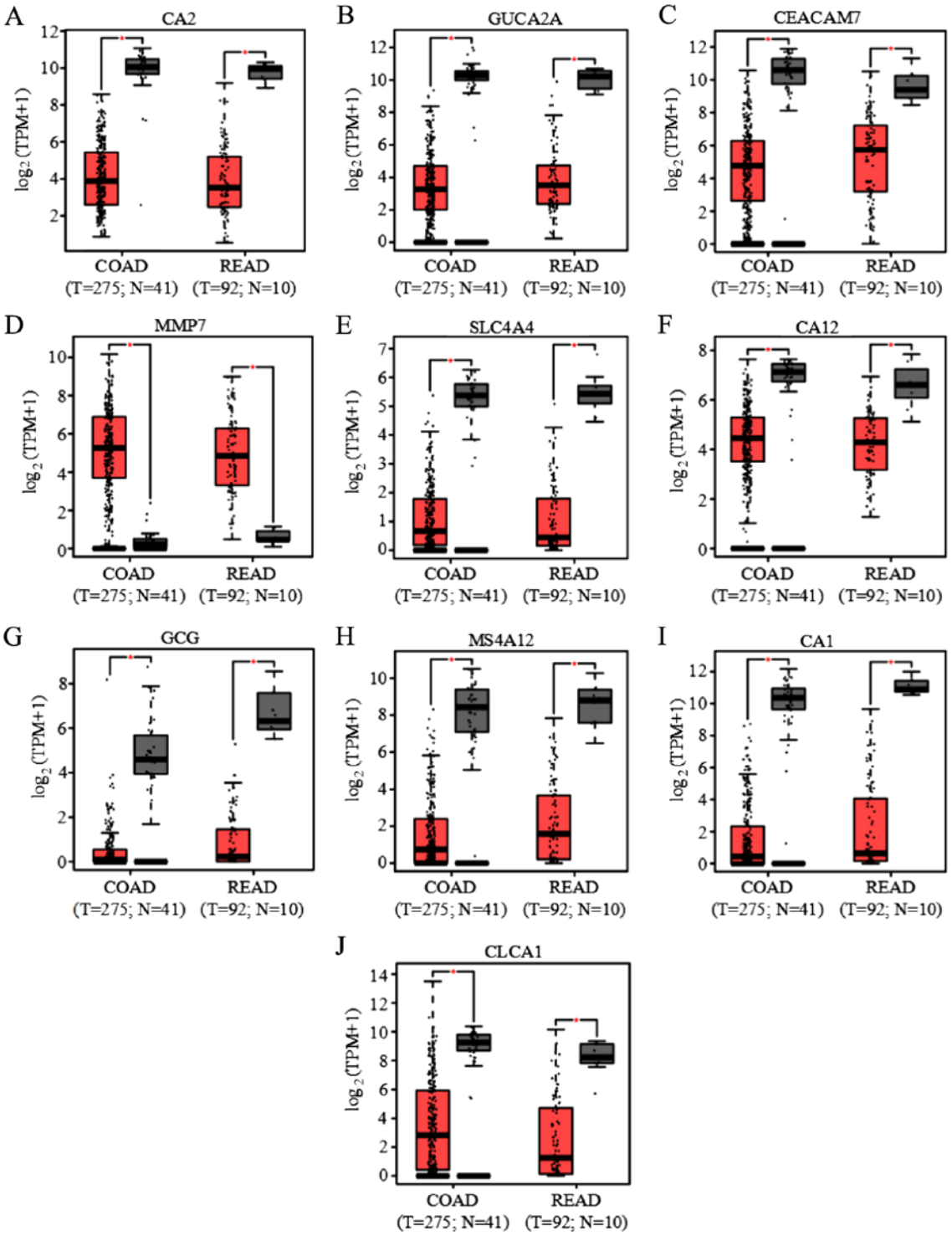

Figure 5. Boxplots of the expression of ten key genes in COAD and READ. Gene expression in COAD and READ of (A) CA2, (B) GUCA2A, (C) CEACAM7, (D) MMP7, (E) SLC4A4, (F) CA12, (G) GCG, (H) MS4A12, (I) CA1 and (J) CLCA1. *P<0.01 vs. the noncancerous group. COAD, colon adenocarcinoma; READ, rectum adenocarcinoma; T, tumor; N, normal; CA2, carbonic anhydrase 2; GUCA2A, guanylate cyclase activator 2A; CEACAM7, carcinoembryonic antigen-related cell adhesion molecule 7; MMP7, matrix metalloproteinase-7; SLC4A4, solute carrier family 4 member 4; CA2, carbonic anhydrase 2; GCG, glucagon; MS4A12, membrane-spanning 4-domains subfamily A member 12; CLCA1, calcium-activated chloride channel regulator 1; TPM, transcripts per million.

malignant cells (32). Recent studies have reported various roles of matrix molecules in tissue development, homeostasis and pathological processes (33-35). Glycosaminoglycans, a type of matrix molecule, affect the growth and progression of tumors by interacting with growth factors, cytokines and growth factor receptors (36). Furthermore, chondroitin sulfate is also a critical molecule in cancer progression (37). However, the association between CEACAM7 and glycosaminoglycan biosynthesis-chondroitin sulfate and extracellular matrix receptor interaction pathways has not yet been investigated. Further reports are required to clarify this potential connection.

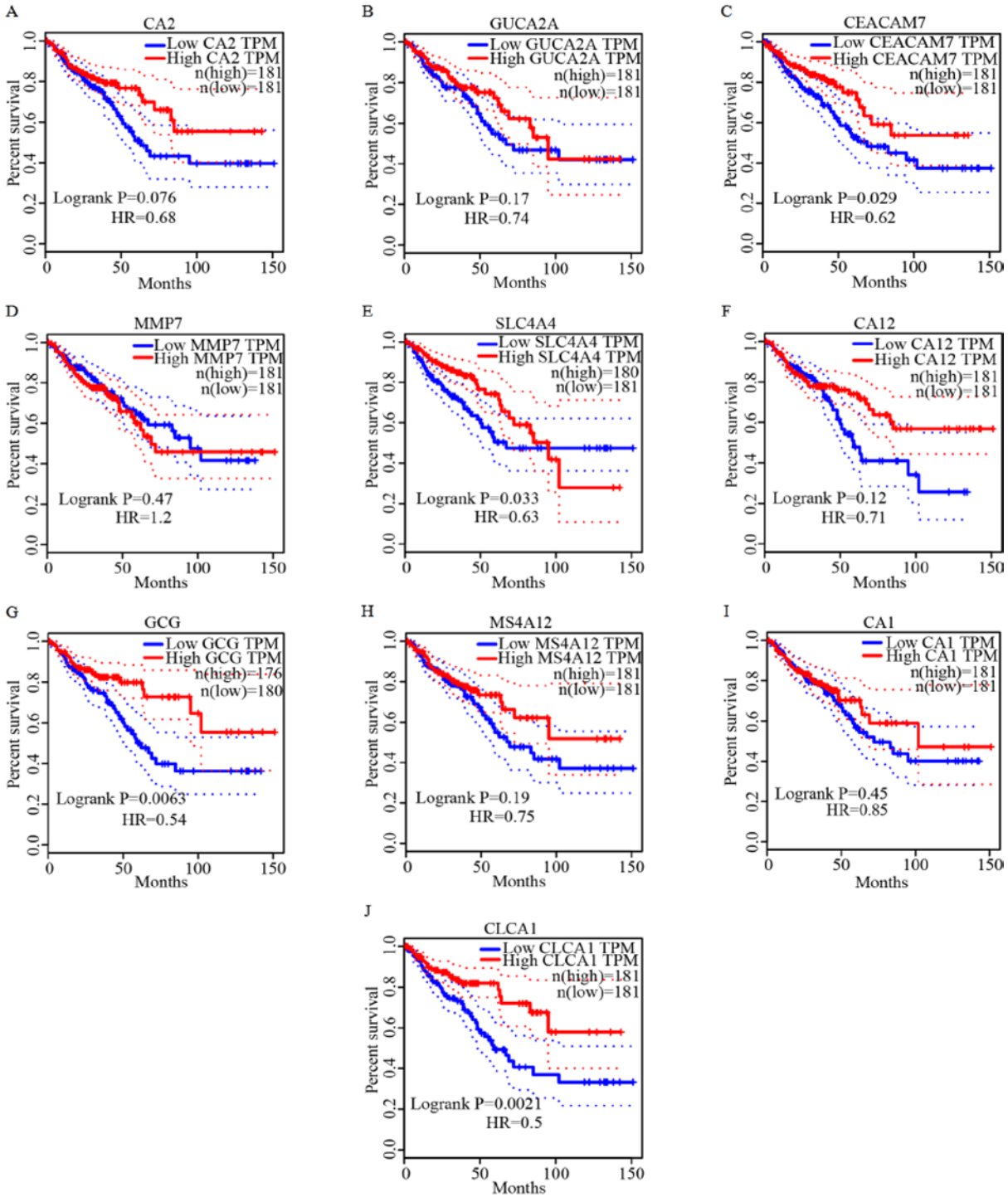

Figure 6. Survival curves of overall survival analysis of ten key genes. Survival curves of overall survival time of (A) CA2, (B) GUCA2A, (C) CEACAM7, (D) MM7, (E) SLC4A4, (F) CA12, (G) GCG, (H) MS4A12, (I) CA1 and (J) CLCA1. CA2, carbonic anhydrase 2; GUCA2A, guanylate cyclase activator 2A; CEACAM7, carcinoembryonic antigen-related cell adhesion molecule 7; MMP7, matrix metalloproteinase-7; SLC4A4, solute carrier family 4 member 4; GCG, glucagon; MS4A12, membrane-spanning 4-domains subfamily A member 12; CA2, carbonic anhydrase 2; CLCA1, calcium-activated chloride channel regulator 1; TPM, transcripts per million.

SLC4A4 may regulate bicarbonate influx and efflux in the basolateral membrane of cells and regulate intracellular pH (38-42). Certain studies have shown that SLC4A4 is significantly downregulated in CRC (43-45). However, the mechanism of SLC4A4 in affecting the prognosis of CRC is not well studied, and future studies should examine the potential function of SLC4A4 in CRC.

GCG serves a key role in glucose metabolism and homeostasis. Much attention has been drawn to its low expression in CRC tissues (44,46-48). GCG is cleaved into glucagon-like peptide (GLP)-1, GLP-2 and other small peptides in intestinal endocrine cells and brain neurons (49). Moreover, GLP-1 and its analogs have become an effective therapeutic strategy for numerous patients with type 2 diabetes (50). Notably, CRC

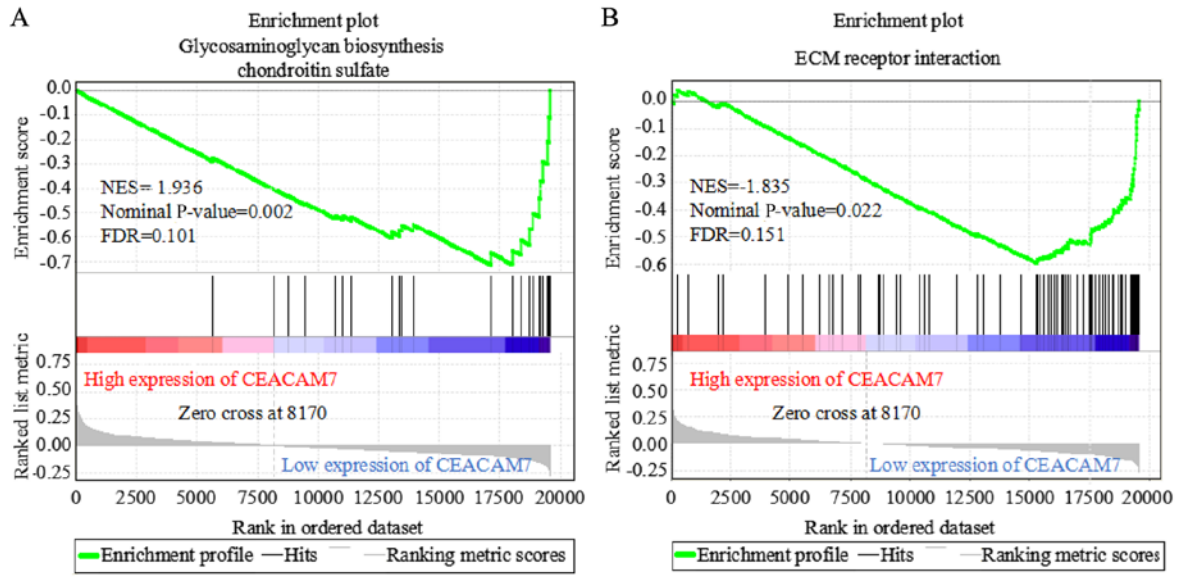

Figure 7. Gene set enrichment analysis of CEACAM7. Two significantly enriched pathways were identified in the low CEACAM7 expression group of patients with CRC at nominal P<0.05 and FDR<0.25, including (A) glycosaminoglycan biosynthesis-chondroitin sulfate (nominal P=0.002; FDR, 0.101) and (B) ECM receptor interaction (nominal P=0.022; FDR, 0.151). CEACAM7, carcinoembryonic antigen-related cell adhesion molecule 7; CRC, colorectal cancer; NES, normalized enrichment score; FDR, false discovery rate; ECM, extracellular matrix.

is more common in diabetic patients than in the non-diabetic population (51-54). Furthermore, Zanders *et al* (55) demonstrated that diabetes affects the presentation, treatment and outcome of CRC. Patients with both CRC and diabetes are likely to have a lower survival rate compared with patients with CRC without diabetes. In a study by Koehler *et al* (27), GLP-1 receptor activation decreased proliferation and survival of CT26 colon cancer cells that expressed the endogenous classical GLP-1 receptor. Hence, more studies on the associations between GLP-1 and CRC are required, particularly for patients with both diabetes and CRC. GLP-2, a nutrient-responsive neuropeptide and intestinal hormone, functions in promoting cell proliferation and survival (56,57). Previous studies have demonstrated the therapeutic potential of GLP-2 in surgical resections and ulcerative colitis (58,59). However, the function of GLP2 was shown to be controversial. A histopathological analysis in one study showed a significant increase in tumor load of mice treated with Gly2-GLP-2, which indicated that GLP2 promoted the development of CRC (60). These findings appear to be inconsistent with the present study, which revealed shorter overall survival time associated with low expression of GCG in patients with CRC. Therefore, further investigation is required to elucidate whether GLP2 promotes intestinal healing or accelerates the development of CRC.

CLCA1 is the first reported member of the CLCA family and is mainly expressed in the colon, small intestine and appendix (61). Yang *et al* (62) revealed that CLCA1 is expressed in differentiated, growth-arrested mammalian epithelial cells but is downregulated during tumor progression. CLCA1 has been identified as a regulator of the transition from proliferation to differentiation in Caco-2 cells. Further investigations demonstrated that low expression levels of CLCA1 predicted lower survival in patients with CRC (63). Li *et al* (64) demonstrated that increased expression levels of CLCA1 could suppress the aggressiveness of CRC via inhibiting the epithelial-mesenchymal transition process and the Wnt/β-catenin signaling pathway. Hence, CLCA1 is associated with CRC prognosis and may be a tumor suppressor in CRC.

Overall, the present study identified four prognosis-associated key genes, CEACAM7, SLC4A4, GCG and CLCA1, in CRC using bioinformatics analysis. All four genes were downregulated in patients with CRC. Differential expressions of these genes were also observed in CRC tumor samples. Low expression of these genes appeared to be associated with adverse clinical outcome in patients with CRC. These four genes may be potential prognosis markers or therapeutic targets of CRC. Low expression of CEACAM7 may affect the prognosis of patients with CRC via activating glycosaminoglycan biosynthesis-chondroitin sulfate and extracellular matrix receptor interaction pathways. It is speculated that CLCA1 is a potential prognosis predictor and therapeutic target of CRC. Further study is required to verify and investigate the molecular mechanisms of these genes in CRC *in vitro* and *in vivo*.

Acknowledgements

The authors would like to thank Dr Shaohua Lei, Dr Min Oh, Dr Mohammad Shabbir Hasan and Dr Saima Sultana Tithi (Zhang Lab, Virginia Tech) for their helpful discussions. The authors would also like to thank to Dr Gustavo Arango and Dr Dhoha Abid (Virginia Tech) for their guidance to improve the manuscript.

Funding

This study was funded by the National Natural Science Foundation of China (grant nos. 81630104 and 81973748), the China National Funds for Distinguished Young Scientists

(grant no. 30825046) and the China Scholarship Council (grant no. 201706550003).

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

QB drafted the manuscript. JXC and LZ designed the study. WQ, CP and FD collected the data. MS and XS performed data analyses. YL designed the figures and tables, and helped QB to interpret the results. JBC helped with statistical analysis. QB and LZ revised the manuscript. JXC contributed to funding acquisition. JXC and LZ supervised the study.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68: 394-424, 2018.
2. SEER Cancer Statistics Factsheets: Colon and Rectum Cancer. National Cancer Institute, Bethesda, MD, 2018. http://seer.cancer.gov/statfacts/html/colorect.html. Accessed February 2018.
3. Siegel R, DeSantis C, Virgo K, Stein K, Mariotto A, Smith T, Cooper D, Gansler T, Lerro C, Fedewa S, *et al*: Cancer treatment and survivorship statistics, 2012. CA Cancer J Clin 62: 220-241, 2012.
4. American Cancer Society, Colorectal Cancer Facts & Figures 2014-2016, American Cancer Society, Atlanta, Ga, USA, pp7, 2014.
5. Brenner H, Kloor M and Pox CP: Colorectal cancer. Lancet 383: 1490-1502, 2014.
6. Hu Y, Gaedcke J, Emons G, Beissbarth T, Grade M, Jo P, Yeager M, Chanock SJ, Wolff H, Camps J, *et al*: Colorectal cancer susceptibility loci as predictive markers of rectal cancer prognosis after surgery. Genes Chromosomes Cancer 57: 140-149, 2018.
7. Kagawa Y, Matsumoto S, Kamioka Y, Mimori K, Naito Y, Ishii T, Okuzaki D, Nishida N, Maeda S, Naito A, *et al*: Cell cycle-dependent Rho GTPase activity dynamically regulates cancer cell motility and invasion in vivo. PLoS One 8: e83629, 2013.
8. Sveen A, Agesen TH, Nesbakken A, Rognum TO, Lothe RA and Skotheim RI: Transcriptome instability in colorectal cancer identified by exon microarray analyses: Associations with splicing factor expression levels and patient survival. Genome Med 3: 32, 2011.
9. Agesen TH, Sveen A, Merok MA, Lind GE, Nesbakken A, Skotheim RI and Lothe RA: ColoGuideEx: A robust gene classifier specific for stage II colorectal cancer prognosis. Gut 61: 1560-1567, 2012.
10. Liu YR, Hu Y, Zeng Y, Li ZX, Zhang HB, Deng JL and Wang G: Neurexophilin and PC-esterase domain family member 4 (NXPE4) and prostate androgen-regulated mucin-like protein 1 (PARM1) as prognostic biomarkers for colorectal cancer. J Cell Biochem 120: 18041-18052, 2019.
11. Pan F, Chen T, Sun X, Li K, Jiang X, Försti A, Zhu Y and Lai M: Prognosis prediction of colorectal cancer using gene expression profiles. Front Oncol 9: 252, 2019.
12. Song X, Tang T, Li C, Liu X and Zhou L: CBX8 and CD96 are important prognostic biomarkers of colorectal cancer. Med Sci Monit 24: 7820-7827, 2018.
13. Yong L, YuFeng Z and Guang B: Association between PPP2CA expression and colorectal cancer prognosis tumor marker prognostic study. Int J Surg 59: 80-89, 2018.
14. Li Y, Jiang Q, Ding Z, Liu G, Yu P, Jiang G, Yu Z, Yang C, Qian J, Jiang H and Zou Y: Identification of a common different gene expression signature in ischemic cardiomyopathy. Genes (Basel) 9: pii: E56, 2018.
15. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3, 2004.
16. Oliveros JC: VENNY. An interactive tool for comparing lists with Venn's diagrams, 2007-2015. http://bioinfogp.cnb.csic.es/tools/venny/index.html.
17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat Genet 25: 25-29, 2000.
18. The Gene Ontology Consortium: The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res 47 (D1): D330-D338, 2019.
19. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H and Kanehisa M: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27: 29-34, 1999.
20. Huang da W, Sherman BT and Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57, 2009.
21. Huang da W, Sherman BT and Lempicki RA: Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13, 2009.
22. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, *et al*: The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38 (Web Server Issue): W214-W220, 2010.
23. Tang Z, Li C, Kang B, Gao G, Li C and Zhang Z: GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res 45 (W1): W98-W102, 2017.
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102: 15545-15550, 2005.
25. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P and Mesirov JP: Molecular signatures database (MSigDB) 3.0. Bioinformatics 27: 1739-1740, 2011.
26. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T: Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504, 2003.
27. Koehler JA, Kain T and Drucker DJ: Glucagon-like peptide-1 receptor activation inhibits growth and augments apoptosis in murine CT26 colon cancer cells. Endocrinology 152: 3362-3372, 2011.
28. Schölzel S, Zimmermann W, Schwarzkopf G, Grunert F, Rogaczewski B and Thompson J: Carcinoembryonic antigen family members CEACAM6 and CEACAM7 are differentially expressed in normal tissues and oppositely deregulated in hyperplastic colorectal polyps and early adenomas. Am J Pathol 156: 595-605, 2000.
29. Boguski MS: The turning point in genome research. Trends Biochem Sci 20: 295-296, 1995.
30. Thompson J, Zimmermann W, Nollau P, Neumaier M, Weber-Arden J, Schrewe H, Craig I and Willcocks T: CGM2, a member of the carcinoembryonic antigen gene family is down-regulated in colorectal carcinomas. J Biol Chem 269: 32924-32931, 1994.
31. Messick CA, Sanchez J, Dejulius KL, Hammel J, Ishwaran H and Kalady MF: CEACAM-7: A predictive marker for rectal cancer recurrence. Surgery 147: 713-719, 2010.
32. Afratis N, Gialeli C, Nikitovic D, Tsegenidis T, Karousou E, Theocharis AD, Pavão MS, Tzanakakis GN and Karamanos NK: Glycosaminoglycans: Key players in cancer cell biology and treatment. FEBS J 279: 1177-1197, 2012.

33. Mariman EC and Wang P: Adipocyte extracellular matrix composition, dynamics and role in obesity. Cell Mol Life Sci 67: 1277-1292, 2010.
34. Järveläinen H, Sainio A, Koulu M, Wight TN and Penttinen R: Extracellular matrix molecules: Potential targets in pharmacotherapy. Pharmacol Rev 61: 198-223, 2009.
35. Llacua LA, Faas MM and de Vos P: Extracellular matrix molecules and their potential contribution to the function of transplanted pancreatic islets. Diabetologia 61: 1261-1272, 2018.
36. Karamanos NK and Tzanakakis GN: Glycosaminoglycans: from 'cellular glue' to novel therapeutical agents. Curr Opin Pharmacol 12: 220-222, 2012.
37. Theocharis AD, Tsolakis I, Tzanakakis GN and Karamanos NK: Chondroitin sulfate as a key molecule in the development of atherosclerosis and cancer progression. Adv Pharmacol 53: 281-295, 2006.
38. Burnham CE, Amlal H, Wang Z, Shull GE and Soleimani M: Cloning and functional expression of a human kidney Na+:HCO3- cotransporter. J Biol Chem 272: 19111-19114, 1997.
39. Abuladze N, Lee I, Newman D, Hwang J, Boorer K, Pushkin A and Kurtz I: Molecular cloning, chromosomal localization, tissue distribution, and functional expression of the human pancreatic sodium bicarbonate cotransporter. J Biol Chem 273: 17689-17695, 1998.
40. Choi I, Romero MF, Khandoudi N, Bril A and Boron WF: Cloning and characterization of a human electrogenic Na+-HCO-3 cotransporter isoform (hhNBC). Am J Physiol 276: C576-C584, 1999.
41. Sun XC and Bonanno JA: Identification and cloning of the Na/HCO(3-) cotransporter (NBC) in human corneal endothelium. Exp Eye Res 77: 287-295, 2003.
42. Demirci FY, Chang MH, Mah TS, Romero MF and Gorin MB: Proximal renal tubular acidosis and ocular pathology: A novel missense mutation in the gene (SLC4A4) for sodium bicarbonate cotransporter protein (NBCe1). Mol Vis 12: 324-330, 2006.
43. Hong Y, Downey T, Eu KW, Koh PK and Cheah PY: A 'metastasis-prone' signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. Clin Exp Metastasis 27: 83-90, 2010.
44. Skrzypczak M, Goryca K, Rubel T, Paziewska A, Mikula M, Jarosz D, Pachlewski J, Oledzki J and Ostrowski J: Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. PLoS One 5: pii: e13091, 2010.
45. Gaedcke J, Grade M, Jung K, Camps J, Jo P, Emons G, Gehoff A, Sax U, Schirmer M, Becker H, *et al*: Mutated KRAS results in overexpression of DUSP4, a MAP-kinase phosphatase, and SMYD3, a histone methyltransferase, in rectal carcinomas. Genes Chromosomes Cancer 49: 1024-1034, 2010.
46. Notterman DA, Alon U, Sierk AJ and Levine AJ: Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res 61: 3124-3130, 2001.
47. Kaiser S, Park YK, Franklin JL, Halberg RB, Yu M, Jessen WJ, Freudenberg J, Chen X, Haigis K, Jegga AG, *et al*: Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumor models and human colon cancer. Genome Biol 8: R131, 2007.
48. Sabates-Bellver J, Van der Flier LG, de Palo M, Cattaneo E, Maake C, Rehrauer H, Laczko E, Kurowski MA, Bujnicki JM, Menigatti M, *et al*: Transcriptome profile of human colorectal adenomas. Mol Cancer Res 5: 1263-1275, 2007.
49. Janssen P, Rotondo A, Mulé F and Tack J: Review article: A comparison of glucagon-like peptides 1 and 2. Aliment Pharmacol Ther 37: 18-36, 2013.
50. Cho YM, Fujita Y and Kieffer TJ: Glucagon-like peptide-1: Glucose homeostasis and beyond. Annu Rev Physiol 76: 535-559, 2014.
51. Larsson SC, Orsini N and Wolk A: Diabetes mellitus and risk of colorectal cancer: A meta-analysis. J Natl Cancer Inst 97: 1679-1687, 2005.
52. Jiang Y, Ben Q, Shen H, Lu W, Zhang Y and Zhu J: Diabetes mellitus and incidence and mortality of colorectal cancer: A systematic review and meta-analysis of cohort studies. Eur J Epidemiol 26: 863-876, 2011.
53. Deng L, Gui Z, Zhao L, Wang J and Shen L: Diabetes mellitus and the incidence of colorectal cancer: An updated systematic review and meta-analysis. Dig Dis Sci 57: 1576-1585, 2012.
54. Krämer HU, Schöttker B, Raum E and Brenner H: Type 2 diabetes mellitus and colorectal cancer: Meta-analysis on sex-specific differences. Eur J Cancer 48: 1269-1282, 2012.
55. Zanders MM, Vissers PA, Haak HR and van de Poll-Franse LV: Colorectal cancer, diabetes and survival: Epidemiological insights. Diabetes Metab 40: 120-127, 2014.
56. Rowland KJ, Trivedi S, Lee D, Wan K, Kulkarni RN, Holzenberger M and Brubaker PL: Loss of glucagon-like peptide-2-induced proliferation following intestinal epithelial insulin-like growth factor-1-receptor deletion. Gastroenterology 141: 2166-2175.e7, 2011.
57. Shi X, Li X, Wang Y, Zhang K, Zhou F, Chan L, Li D and Guan X: Glucagon-like peptide-2-stimulated protein synthesis through the PI 3-kinase-dependent Akt-mTOR signaling pathway. Am J Physiol Endocrinol Metab 300: E554-E563, 2011.
58. L'Heureux MC and Brubaker PL: Glucagon-like peptide-2 and common therapeutics in a murine model of ulcerative colitis. J Pharmacol Exp Ther 306: 347-354, 2003.
59. Rowland KJ and Brubaker PL: Life in the crypt: A role for glucagon-like peptide-2? Mol Cell Endocrinol 288: 63-70, 2008.
60. Thulesen J, Hartmann B, Hare KJ, Kissow H, Ørskov C, Holst JJ and Poulsen SS: Glucagon-like peptide 2 (GLP-2) accelerates the growth of colonic neoplasms in mice. Gut 53: 1145-1150, 2004.
61. Gruber AD, Elble RC, Ji HL, Schreur KD, Fuller CM and Pauli BU: Genomic cloning, molecular characterization, and functional analysis of human CLCA1, the first human member of the family of Ca2+-activated Cl- channel proteins. Genomics 54: 200-214, 1998.
62. Yang B, Cao L, Liu B, McCaig CD and Pu J: The transition from proliferation to differentiation in colorectal cancer is regulated by the calcium activated chloride channel A1. PLoS One 8: e60861, 2013.
63. Yang B, Cao L, Liu J, Xu Y, Milne G, Chan W, Heys SD, McCaig CD and Pu J: Low expression of chloride channel accessory 1 predicts a poor prognosis in colorectal cancer. Cancer 121: 1570-1580, 2015.
64. Li X, Hu W, Zhou J, Huang Y, Peng J, Yuan Y, Yu J and Zheng S: CLCA1 suppresses colorectal cancer aggressiveness via inhibition of the Wnt/beta-catenin signaling pathway. Cell Commun Signal 15: 38, 2017.