Original Article
Identification of potential biomarker genes in ulcerative colitis-associated colorectal cancer using a bioinformatics approach

Dejun Cui1,2, Shujiro Okuda2, Yiwei Ling2

1Department of Gastroenterology, Guizhou Provincial People’s Hospital, Medical College of Guizhou University, Guiyang, Guizhou Province, China; 2Division of Bioinformatics, Niigata University Graduate School of Medical and Dental Sciences, Japan

Received April 7, 2020; Accepted August 21, 2020; Epub November 15, 2020; Published November 30, 2020

Abstract: Background: Ulcerative colitis-associated colorectal cancer (UC-CRC) is the main cause of death in UC patients. It lacks specific non-invasive molecular markers for early diagnosis and treatment, so we explored these in the present study using bioinformatics analysis of genomic data. Methods: Gene expression datasets (GSE3629 and GSE37283) were obtained from the GEO database. Differentially expressed genes (DEGs) between UC-CRC and UC samples were identified and analyzed using the GEO2R online analysis tool. GO and KEGG enrichment analysis were performed. The STRING database and Cytoscape software were used for protein-protein interactions visualization. The CytoHubba plugin was employed to calculate the degree of each protein node and identify ‘hub’ genes. Finally, we selected colon and rectum adenocarcinoma of the TCGA dataset to perform expression and survival analysis for each ‘hub’ gene using UALCAN. Results: We identified a total of 163 up-regulated DEGs and 266 down-regulated DEGs in UC-CRC, among which 20 ‘hub’ genes with a higher degree of connectivity were selected. The low expression of CDH1 and B3GNT7 hub genes was associated with a poor prognosis of UC-CRC. Conclusion: Our findings indicate that CDH1 and B3GNT7 may be novel biomarkers for the early diagnosis and prognosis of UC-CRC.

Keywords: Colitis-associated cancer, hub genes, ulcerative colitis, expression profiling data, bioinformatics

Introduction
Ulcerative colitis-associated colorectal cancer (UC-CRC) is one of the major complications in UC patients, accounting for 1-2% of cases of colorectal cancer, and is an important cause of death in these patients [1]. Studies have shown that patients with UC at 10, 20, and 30 years are 1.15%, 3.56%, and 14.36% at risk of developing UC-CRC [2]. Compared with sporadic colorectal cancer, UC-CRC has an earlier age of onset and a worse prognosis [3, 4].

The mechanism of UC-CRC has not been fully defined and specific and effective treatment methods are currently lacking. Current monitoring and treatment of UC-CRC and its precancerous lesions rely largely on endoscopic endoscopy and pathological biopsy [3, 5]. Unfortunately, endoscopic screening is associated with misdiagnosis, the risks of an invasive procedure, poor patient compliance, and a financial burden, which affects the early diagnosis and treatment of UC-CRC [3]. However, there is currently no specific, sensitive, non-invasive biomarker for the early diagnosis and prognosis of UC-CRC [6].

In this study, we tested new UC-CRC prognostic indicators and UC-CRC susceptibility genes to identify UC-CRC treatment targets and predict molecular marker genes involved in the progression of UC to CRC. To detect genes that were differentially expressed (DEGs) between UC-CRC and UC tissue, a bioinformatics approach was adopted to evaluate gene expression profiling data retrieved from the Gene Expression Omnibus (GEO). Obtained DEGs then underwent functional clustering, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, and Gene Ontology (GO) functional annotation analysis. A protein-protein interaction (PPI) network was constructed to select ‘hub’ genes associ-
Potential biomarker genes in UC-CRC

Materials and methods

High throughput data

Gene expression datasets were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). A total of 2679 series on UC were extracted from the database. Search term: ulcerative colitis. After careful scrutiny, two gene expression profiles (GSE3629 and GSE37283) were selected [7, 8]. GSE3629 was based on the GPL570 platform (HG-U133_Plus_2, Affymetrix Human Genome U133 Plus 2.0 Array), and GSE37283 was based on platform GPL13158 (HT_HG-U133_Plus_PM, Affymetrix HT HG-U133 + PM Array Plate). All data are freely available online. This study did not involve human or animal experiments and therefore had no ethical issues.

Screening for DEGs

DEGs between UC-CRC and UC samples in each chip were identified and analyzed using the GEO2R online analysis tool (https://www.ncbi.nlm.nih.gov/geo/geo2r/), and the adjusted P-value (Benjamini-Hochberg method) and |logFC|, as well as the absolute value of the logarithm of fold change, were calculated. The cutoff criterion of DEG was adjusted to \( P < 0.05 \) and \(|\text{logFC}| \geq 1.0\). Volcano plots of DEGs were drawn using SangerBox software (http://sangerbox.com/). Datasets were statistically processed and intersecting DEGs were identified using an online Venn diagram tool (http://bioinformatics-psb.ugent.be/webtools/Venn/).

Functional clustering

GO enrichment analysis, mainly including the three functional annotations of biological process (BP), molecular function (MF), and cellular component (CC), is one of the most important methods to quickly understand the functional tendency of target genes in biological information analysis. The KEGG database integrates genomic, chemical, and system function information. This study used the Database for Annotation, Visualization and Integrated Discovery tools (v.6.8) for GO and KEGG analysis (https://david.ncifcrf.gov/) [9, 10], with \( P < 0.01 \) and gene counts \( \geq 5 \) regarded as statistically significant.

Interacting network

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING v.11.0, https://string-db.org/) was used to identify protein-protein interactions (PPI). Previously obtained DEGs were imported into the STRING database, selecting those with a combined score > 0.4. Next, Cytoscape software (v.3.6.0) was used for PPI visualization [11]. Finally, the CytoHubba plugin in Cytoscape was employed to calculate the degree of each protein node and identify ‘hub’ genes [12]. The top 20 genes were defined as ‘hub’ genes.

Clinical significance

UALCAN is a comprehensive and interactive web resource for analyzing cancer OMICS data (http://ualcan.path.uab.edu/index.html) [13]. This online tool provides graphs depicting gene expression or patient survival information. We selected UC-CRC subtypes, colon adenocarcinoma and rectum adenocarcinoma of the TCGA dataset to perform expression and survival analysis for each ‘hub’ gene using UALCAN. \( P < 0.05 \) represented a significant difference.

Results

DEGs

We identified two gene expression profiles (GSE3629 and GSE37283). GSE3629 contained six UC-CRC and 43 UC-nonCa samples, while GSE37283 included 11 UC-CRC and four UC-nonCa specimens (Table 1). According to \( P < 0.05 \) and \(|\text{logFC}| \geq 1\), 18,870 DEGs, including 10,202 up-regulated genes and 8,668 down-regulated genes, were found in

Table 1. Two groups of the two microarray databases derived from the GEO database

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Number of UC-CRC cases</th>
<th>Number of UC-nonCa cases</th>
<th>Total number of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE3629</td>
<td>6</td>
<td>43</td>
<td>49</td>
</tr>
<tr>
<td>GSE37283</td>
<td>11</td>
<td>4</td>
<td>15</td>
</tr>
</tbody>
</table>

UC-CRC: ulcerative colitis-associated colorectal cancer. UC-nonCa: ulcerative colitis-non cancer.
Potential biomarker genes in UC-CRC

GSE3629. A total of 1,019 DEGs were identified in GSE37283, of which 550 were up-regulated and 469 were down-regulated. Figure 1 shows volcano plots of the data. The intersection of the DEG profiles is shown in Figure 2; 429 DEGs were obtained, of which 163 were up-regulated and 266 were down-regulated.

**GO and KEGG enrichment analysis**

Table 2 shows the findings of GO and KEGG enrichment analyses. BP and MF enrichment analysis showed no statistical significance, and DEGs were only found to be enriched in CC, including extracellular space, lateral plasma membrane, and extracellular exosome. Subsequent KEGG pathway enrichment analysis indicated that DEGs were enriched in pathways associated with proximal tubule bicarbonate reclamation and choline metabolism in cancer and melanoma.

**PPI network and ‘hub’ genes**

Using the STRING tool, we found 56 nodes and 174 edges in the PPI network, as shown in Figure 3A. The top 20 ‘hub’ genes for the connectivity degree are shown in Table 3. Eight of these were up-regulated and 12 were down-regulated in UC-CRC. The ‘hub’ gene interaction network is shown in Figure 3B.

**Hub genes expression and survival analysis**

To explore the prognostic values of the 20 potential ‘hub’ genes, UALCAN analysis was performed on TCGA samples. We found that low expression of the cadherin 1 ‘hub’ gene (CDH1) was associated with the unfavorable overall survival of rectum adenocarcinoma (READ) patients ($P=0.031$, Figure 4A). How-
Table 2. Significantly enriched GO terms and KEGG pathways of differentially expressed genes

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Count</th>
<th>Percentage</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>GO:0005615: extracellular space</td>
<td>58</td>
<td>13.679</td>
<td>7.74E-07</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>GO:0016328: lateral plasma membrane</td>
<td>10</td>
<td>2.358</td>
<td>1.80E-06</td>
</tr>
<tr>
<td>GOTERM_CC_DIRECT</td>
<td>GO:0070062: extracellular exosome</td>
<td>96</td>
<td>22.642</td>
<td>3.63E-06</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa04964: Proximal tubule bicarbonate reclamation</td>
<td>5</td>
<td>1.179</td>
<td>0.002</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa05231: Choline metabolism in cancer</td>
<td>9</td>
<td>2.123</td>
<td>0.003</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa05218: Melanoma</td>
<td>7</td>
<td>1.651</td>
<td>0.007</td>
</tr>
</tbody>
</table>

Figure 3. A. PPI network constructed with DEGs. Note: Red nodes indicate up-regulated genes, and blue nodes indicate down-regulated genes, the nodes with yellow borders are potential Hub genes. B. The ‘hub’ gene interaction network. Red and orange nodes indicate ‘hub’ genes.
ever, the expression of CDH1 had no statistical significance in READ ($P=8.031\times10^{-1}$). Additionally, low expression of B3GNT7 was an unfavorable prognostic factor of survival in colon adenocarcinoma (COAD) patients ($P=0.013$, Figure 4B), and the expression of B3GNT7 had statistical significance in COAD and READ patients ($P<0.05$, Figure 4C and 4D). Moreover, promoter methylation levels of B3GNT7 were significantly increased in COAD and READ patients ($P<0.05$, Figure 4E and 4F).

Discussion

We found that 163 up-regulated and 266 down-regulated DEGs were associated with GO CC terms such as extracellular space, lateral plasma membrane, and extracellular exosome, and were significantly enriched in the KEGG terms proximal tubule bicarbonate reclamation and choline metabolism in cancer and melanoma. Our interaction network and clinical significance analyses found that low expression of CDH1 and B3GNT7 ‘hub’ genes was an unfavorable prognostic factor in UC-CRC patients.

Different subtypes of CRC have diverse markers that correlate with patient survival time [14, 15]. The pathogenesis of UC-CRC differs from that of sporadic CRC [6, 16]. Long-term chronic inflammatory stimulation appears to be the main factor of UC oncogenesis, and has a clear genetic component [17]. However, because of the time and spatial heterogeneity of tumors, it remains unclear which UC-CRC genetic mutations and epigenetic factors are associated with chronic inflammation.

CDH1 is located on human chromosome 16q22.1 and encodes E-cadherin, which is involved in regulating cell adhesion, migration, and epithelial cell proliferation. CDH1 mutations are associated with gastric, breast, colorectal, thyroid, and ovarian cancers. Moreover, CDH1 deficiency is closely related to poor prognosis, metastasis, and tumor progression in a variety of human tumors [18-23]. A meta-analysis suggested that CDH1 promoter methylation plays an important role in colorectal carcinogenesis [24]. However, one study found that the CDH1-160C > A polymorphism does not contribute to the genetic susceptibility of CRC and may not directly affect progression of the disease in Turkish patients [25]. Another study documented hypermethylation of the CDH1 promoter region in 46% of colorectal cancers, but found no difference in

<p>| Table 3. Top twenty hub genes with higher degrees |</p>
<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Degree</th>
<th>Up- or down-regulation</th>
<th>Description of gene symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR</td>
<td>11</td>
<td>down</td>
<td>epidermal growth factor receptor</td>
</tr>
<tr>
<td>GCG</td>
<td>8</td>
<td>down</td>
<td>glucagon</td>
</tr>
<tr>
<td>ANXA1</td>
<td>7</td>
<td>up</td>
<td>annexin A1</td>
</tr>
<tr>
<td>STOM</td>
<td>7</td>
<td>up</td>
<td>stomatin</td>
</tr>
<tr>
<td>EDN1</td>
<td>7</td>
<td>down</td>
<td>endothelin 1</td>
</tr>
<tr>
<td>EDNRA</td>
<td>7</td>
<td>up</td>
<td>endothelin receptor type A</td>
</tr>
<tr>
<td>CDH1</td>
<td>7</td>
<td>down</td>
<td>cadherin 1</td>
</tr>
<tr>
<td>HGF</td>
<td>7</td>
<td>up</td>
<td>hepatocyte growth factor</td>
</tr>
<tr>
<td>PLCB1</td>
<td>7</td>
<td>up</td>
<td>phospholipase C beta 1</td>
</tr>
<tr>
<td>EDN3</td>
<td>6</td>
<td>down</td>
<td>endothelin 3</td>
</tr>
<tr>
<td>FOS</td>
<td>6</td>
<td>up</td>
<td>Fos proto-oncogene</td>
</tr>
<tr>
<td>AVPR1A</td>
<td>6</td>
<td>up</td>
<td>arginine vasopressin receptor 1A</td>
</tr>
<tr>
<td>LILRB2</td>
<td>6</td>
<td>up</td>
<td>leukocyte immunoglobulin like receptor B2</td>
</tr>
<tr>
<td>TMEM63A</td>
<td>5</td>
<td>down</td>
<td>transmembrane protein 63A</td>
</tr>
<tr>
<td>CD177</td>
<td>5</td>
<td>down</td>
<td>CD177 molecule</td>
</tr>
<tr>
<td>MUC20</td>
<td>5</td>
<td>down</td>
<td>mucin 20, cell surface associated</td>
</tr>
<tr>
<td>LYZ</td>
<td>5</td>
<td>down</td>
<td>lysozyme</td>
</tr>
<tr>
<td>DYNC1H1</td>
<td>4</td>
<td>down</td>
<td>dynein cytoplasmic 1 heavy chain 1</td>
</tr>
<tr>
<td>B3GNT7</td>
<td>4</td>
<td>down</td>
<td>UDP-GlcNAc: betaGal beta-1, 3-N-acetylglucosaminyltransferase 7</td>
</tr>
<tr>
<td>WASL</td>
<td>4</td>
<td>down</td>
<td>WASP like actin nucleation promoting factor</td>
</tr>
</tbody>
</table>
expression between UC-CRC and sporadic CRC groups [26]. In our study, CDH1 was expressed at lower levels in UC-CRC compared with UC tissue, and correlated with the unfavorable survival of rectal cancer patients. However, we found no significant difference in colon cancer survival analysis.

As a down-regulated ‘hub’ gene in GEO, CDH1 was verified in the TCGA database. Its expression was shown to be reduced in CRC, but the difference was not significant. This may be associated with the different disease subtypes included in different databases. Although Saito et al reported that active inflammation was an independent factor of methylation for CDH1 and GDNF in UC, further studies are needed to expand the role of CDH1 in UC-CRC and UC-related dysplasia, and the effects of non-coding RNA and exosomes on CDH1 expression [27].

B3GNT7 was first reported by Kataoka and Huh in 2002 and shown to be involved in tissue invasion [28]. In humans, B3GNT7 is expressed in the brain, thymus, esophagus, pancreas, intestine, connective tissue, lung, muscle, ovary, spleen, testis, trachea, and vascular system. It appears to affect the adhesiveness and motility of certain cancer cells [28-31], while the expression of B3GNT7 was markedly down-regulated during colon cancer tumorigenesis [32]. We found that B3GNT7 was down-regulated in UC-CRC, which is meaningful for colon cancer survival analysis, but not in rectal cancer. Additionally, we showed that B3GNT7 expression was decreased in both colon and rectal cancer, and that this difference was statistically significant. B3GNT7 was also significantly methylated in colorectal adenocarcinoma.

The present study has some limitations, including the failure to analyze selective bias factors caused by UC medication, lesion sites, course, race, and dysplasia. Nevertheless, based on our combined GO and KEGG enrichment analyses, B3GNT7 may be a prognostic factor and potential therapeutic target for UC-CRC. Insights from our gene function enrichment analysis suggest that it will be necessary to further study the role of B3GNT7 in the choline metabolism pathway of the intestinal nervous system and exosomes in UC-CRC.

In conclusion, the present study was designed to identify aberrantly DEGs that may be involved in the carcinogenesis of UC. A total of 20 ‘hub’ genes were selected, of which CDH1 and B3GNT7 were identified as potential novel biomarkers for the early and accurate diagnosis and prognosis of UC-CRC. Further research should pay close attention to the epigenetic mutations involved in driving tumorigenesis in UC patients.

Acknowledgements

This work was financially supported by Cultivation Fund of National Natural Science Foundation (Grant No. qiankehe2018-5764-11), Beijing Medical and Health Foundation [NO. YWKJHKKJYJ-B184053] and Doctor Foundation of Guizhou Provincial People's Hospital [NO. GZSYBS(2017)09]. We thank Chinese Medical Association and Takeda Science Foundation for providing scholarship support.

Disclosure of conflict of interest

None.

Address correspondence to: Dejun Cui, Department of Gastroenterology, Guizhou Provincial People’s Hospital, Medical College of Guizhou University, No. 83, East Zhongshan Road, Guiyang 550002, Guizhou Province, China. Tel: +86-851-85937074; Fax: +86-851-85937194; E-mail: hxuczij@163.com; Shujiro Okuda, Division of Bioinformatics, Niigata University Graduate School of Medical and Dental Sciences, Japan. E-mail: okd@med.niigata-u.ac.jp
References


Potential biomarker genes in UC-CRC


