COL1A1 and PTGS2 as potential core genes in gastric cancer

Yang Li1; Lipeng Niu2*

1Shihua Residential District Community Health Service Center, Jinshan District, Shanghai, 201500, China.
2Affiliated Hospital of NCO School, Army Medical University, Shijiazhuang, 050000, China.

Abstract

Background: Gastric Cancer (GC) is highly prevalent in China, where its mortality rate accounts for 50% of global GC deaths. Currently, GC pathogenesis is not fully understood.

Method: In this study, three gene expression profile datasets were retrieved from Gene Expression Omnibus (GEO) and examined to identify core GC biomarkers. First, DEGs in datasets GSE54129, GSE13911, and GSE118916 were identified using NCBI GEO2R and Venn diagram analyses. To determine functional enrichment, Gene Ontology (GO) and Kyoto Encyclopedia of Genes And Genomes (KEGG) analyses were conducted. The DEGs were identified from the Protein-Protein Interaction (PPI) network from which hub genes and gene modules were selected. Finally, we determined the impact of the hub genes on survival of GC using the GEPIA online platform.

Results: Venn diagram analysis identified 508 DEGs (117 upregulated and 331 downregulated) from the 3 gene expression datasets. The top-10 hub genes (FN1, MMP9, CXCL8, TIMP1, COL1A1, MYC, SPP1, COL3A1, COL1A2, PTGS2) were chosen from the DEG-based PPI network.

Conclusion: Using bioinformatics, GC-associate core genes with diagnostic, preventive, and therapeutic target potential were identified.

Keywords: Gastric cancer; Gene expression omnibus; Differently expressed genes; Hub genes.

Abbreviations: BP: Biological Process; CC: Cellular Component; GC: Gastric Cancer; DAVID: Database for Annotation Visualization and Integrated Discovery; DEG: Differentially Expressed Gene; GO: Gene Ontology; GEO: Gene Expression Omnibus; KEGG: Kyoto Encyclopedia of Genes and Genomes; MF: molecular function; PPI: Protein-Protein Interaction.

Introduction

GC is a commonly diagnosed cancer globally [1]. Increased GC screening has led to the detection of increasing GC rates [2]. Although the incidence of GC has been falling worldwide, deaths due to GC are among the highest [3]. Moreover, among these patients, the survival rate within 5 years of diagnosis is currently 20% due to the high rate of recurrence and metastasis [4]. Because of low endoscopy rates, more than 35% of Chinese GC patients have metastatic disease by the time they are diagnosed [5]. Here, we used bioinformatics to identify core genes in GC patients. First, to DEGs between GC versus normal tissues, GEO datasets were analyzed using GEO2R. The DEGs were then analyzed for pathway enrichment using GO and KEGG. Then, using STRING, we built a PPI network and identified GC-related hub genes and the GEPIA survival analysis tool was applied to determine the impact of the hub genes on survival of patients with GC.

Materials and methods

Data source

GC gene expression datasets were downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/). A total of 11096 series associated with human GC were retrieved from the database. After careful review, datasets GSE54129, GSE13911, and GSE118916 were selected for analyses.
Selection of DEGs

Identification of DEGs in GC tissues versus normal gastric tissues was performed using the GEO2R online analysis tool (https://www.ncbi.nlm.nih.gov/geo/geo2r/). After removing duplicate and invalid entries, and DEGs were chosen based on p<0.05 and |logFC|>1. To perform statistical analysis on each dataset, the Venn diagram webtool (bioinformatics.psb.ugent.be/webtools/Venn/) was used, and a Venn diagram of DEGs was created and saved in SVG format.

GO and KEGG pathway analysis of DEGs

GO analysis is frequently used for functional enrichment analyses of large-scale datasets [6]. Gene functions are categorized into Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). The KEGG analysis is a popular data repository for genomes, diseases, biological pathways, and chemicals [7]. GO and KEGG enrichment analyses of the DEGs were performed using DAVID website (https://david.ncifcrf.gov/). The text files of the selected results were downloaded and P<0.01 was deemed statistically significant as were gene counts ≥10.

The establishment of a PPI network and identification of a hub gene

By entering the DEGs into the search box and selecting Homo sapiens as the organism, we were able to create a Protein-Protein Interaction (PPI) network of the DEGs through STRING (http://string-db.org). Cytoscape and cytoHubba, a Cytoscape plugin that calculates protein node degrees and selects the ten top genes as hub genes, were used to build and visualize the PPI network.

Hub gene survival analysis

GEPIA, a web server, allows for the profiling of cancer, normal gene expression in addition to interactive analyses [8]. On GEPIA (http://gepia.cancer-pku.cn/), patient survival analyses were conducted. Based on each hub gene’s median mRNA expression level, cancer patients were assigned into two groups. P<=0.05 was taken as the threshold for statistical significance.

Results

DEG identification

Of the 3 gene expression datasets (GSE54129, GSE13911, and GSE118916) chosen for analysis, GSE54129 contained 111 and 21, GSE13911 had 38 and 31, and GSE118916 had 15 and 15 tumor and normal gastric specimens, respectively (Table 1). Using the p=0.05 and |logFC|>1 criterion, 3905 DEGs (1778 upregulated, 2127 downregulated), 3285 DEGs (1022 upregulated, 2263 downregulated), and 1763 DEGs (918 upregulated, 845 downregulated), were identified in datasets GSE54129, GSE13911, and GSE118916, respectively. Venn analysis identified 508 DEGs (177 significantly upregulated and 331 significantly downregulated) as being common between the 3 datasets (Figure 1).

Functional enrichment analyses

The DEGs were mostly enriched in BPs, according to GO analysis, which included digestion, the collagen catabolic process, the extracellular matrix organization, flavonoid glucuronidation, and the organization of collagen fibrils (Table 2). For CC, DEGs were found to be most abundant in extracellular spaces, exosomes, respectively, among others. DEGs involved in MF were primarily enriched in glucuronosyltransferase activity, heparin-binding, extracellular matrix structural constituent, integrin binding, and calcium ion binding. Moreover, the DEGs were found to be enriched in several pathways, including xenobiotic metabolism by cytochrome P450, drug metabolism, and chemical carcinogenesis.

Table 1: Gene expression data extracted from the 3 GEO datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Tumor</th>
<th>Normal</th>
<th>Total number</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE54129</td>
<td>111</td>
<td>21</td>
<td>132</td>
</tr>
<tr>
<td>GSE13911</td>
<td>38</td>
<td>31</td>
<td>69</td>
</tr>
<tr>
<td>GSE118916</td>
<td>15</td>
<td>15</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 2: DEGs regulate biological processes and pathways.

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>GO:0007586</td>
<td>digestion</td>
<td>18</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0030574</td>
<td>collagen catabolic process</td>
<td>18</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0030198</td>
<td>extracellular matrix organization</td>
<td>26</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0052696</td>
<td>flavonoid glucuronidation</td>
<td>10</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0030199</td>
<td>collagen fibril organization</td>
<td>11</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0006805</td>
<td>xenobiotic metabolic process</td>
<td>14</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0001503</td>
<td>ossification</td>
<td>14</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0008202</td>
<td>steroid metabolic process</td>
<td>10</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0007584</td>
<td>response to nutrient</td>
<td>12</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0045766</td>
<td>positive regulation of angiogenesis</td>
<td>14</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0042493</td>
<td>response to drug</td>
<td>23</td>
</tr>
<tr>
<td>BP</td>
<td>GO:0042127</td>
<td>regulation of cell proliferation</td>
<td>17</td>
</tr>
</tbody>
</table>
Construction of a PPI network and hub gene identification

In the PPI network, 494 nodes and 1834 edges were revealed by STRING analysis of protein interaction among the DEGs (Figure 2). The top ten genes identified in the PPI network based on connectivity degree were FN1, MMP9, CXCL8, TIMP1, COL1A1, MYC, SPP1, COL3A1, COL1A2, and PTGS2 (Table 3). GEPIA survival analysis of COL1A1 and PTGS2 prognostic value in GC revealed that their high expression correlated with poor GC overall and disease-free survival (Figure 3).

Table 3: The Top 10 in network string_interactions.tsv ranked by Degree.

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene function</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>FN1</td>
<td>Fibronectin 1</td>
<td>75</td>
</tr>
<tr>
<td>MMP9</td>
<td>Matrix metallopeptidase 9</td>
<td>68</td>
</tr>
<tr>
<td>CXCL8</td>
<td>C-X-C motif chemokine ligand 8</td>
<td>53</td>
</tr>
<tr>
<td>TIMP1</td>
<td>TIMP metallopeptidase inhibitor 1</td>
<td>46</td>
</tr>
<tr>
<td>COL1A1</td>
<td>Collagen type I alpha 1 chain</td>
<td>45</td>
</tr>
<tr>
<td>MYC</td>
<td>MYC proto-oncogene</td>
<td>44</td>
</tr>
<tr>
<td>SPP1</td>
<td>Secreted phosphoprotein 1</td>
<td>43</td>
</tr>
<tr>
<td>COL3A1</td>
<td>Collagen type III alpha 1 chain</td>
<td>43</td>
</tr>
<tr>
<td>COL1A2</td>
<td>Collagen type I alpha 2 chain</td>
<td>41</td>
</tr>
<tr>
<td>PTGS2</td>
<td>Prostaglandin-endoperoxide synthase 2</td>
<td>40</td>
</tr>
</tbody>
</table>

Figure 2: A network of protein-protein interactions encoding proteins encoded by differentially expressed genes. Genes that are upregulated are represented by red nodes, while genes that are downregulated are represented by yellow nodes.

Figure 3: The top ten hub genes in patients and their prognostic significance in patients with gastric cancer who have metastasis to the liver. Gene descriptions are present in Table 2.
Here, we sought to identify core GC-associated genes and uncover the mechanisms underlying their role in GC. To this end, we evaluated differential gene expression in GC tissue vs. normal tissue in 3 gene expression datasets using GEO2R and identified 508 common DEGs in the 3 datasets (177 upregulated, 331 downregulated). In this study, he top ten hub genes were found to be FN1, MMP9, CXCL8, TIMP1, COL1A1, MYC, SPP1, COL3A1, COL1A2, and PTGS2. Of these, COL1A1, COL1A2, and COL3A1 belong to the Collagen family. Collagen type I alpha 1 chain (also referred to as CAFYD, EDSC, Q1, O12, O13, O14, and EDSSRTH1) encodes type I collagen pro-alpha1 chains, which has a triple helix made up of one alpha2 chain and two alpha1 chains. When compared to normal tissues, mRNA expression levels of COL1A1 were found to be elevated in premalignant as well as malignant tissues, while mRNA expression levels of COL1A2 were significantly elevated in malignant tissues when compared to normal and premalignant tissues. Furthermore, expression levels of COL1A1 were not correlated with clinicopathological parameters, whereas expression of COL1A2 was correlated with tumor size and invasion depth. Additionally, COL1A1 and COL1A2 upregulation correlate with lower overall survival [9]. Silencing COL1A1 expression in BGC-823 cells significantly suppressed their cell proliferation and migration capacity [10]. 

**Discussion**

MiR-129-5p, by targeting COL1A1, inhibits GC cell proliferation, migration, and invasion [11].

Collagen type I alpha 2 chains (also referred to as O14, ED-SARTH2 or EDSCV) encode type I collagen's pro-alpha2 chain, whose triple helix is made up of one alpha2 chain and two alpha1 chains. Of the 7q21-22 candidate genes, two (SHFM1 and MCM7) were found to be expressed in intestinal-type gastric tumors, while COL1A2 was found to be expressed in diffuse-type gastric tumors.

MiR-25 has been reported to be co-expressed with SHFM1 and MCM7 in gastric tumors, and there is an inverse relationship between COL1A2 and miR-25 expression [12]. Through the PI3K- Akt signaling pathway, silencing COL1A2, THBS2 and COL6A3 suppresses GC cell proliferation, migration, as well as invasion while enhancing apoptosis [13].

Collagen type III alpha 1 chain (also referred to as EDS4A, PMGEDSV and EDSVASC) encodes type III collagens' pro-alpha1 chains. It is a fibrillar collagen that is found in extensible connective tissues including those of the skin, uterus, lungs, intestines, and the vascular system. Mostly, it is found in association with type I collagen. The co-expression of LUM and COL3A1 indicates the importance of LUM in collagen fiber assembly [14]. COL3A1 promotes the migration of iCCA cells and is a component of tumor-associated aligned collagen [15]. KCN15 expression levels in ESCC cell lines vary greatly and are associated with COL3A1, JAG1, and F11R [16].

TIMP metalloproteinase inhibitor 1 is a member of the TIMP gene family. It is also referred to as EPO, EPA, HCl, CLGI, TIMP-1, or TIMP. This gene family of proteins naturally inhibits matrix metalloproteinases (MMPs), a class of peptidases that degrade the extracellular matrix. Indeed, clinical data show that clinical-stage III patients with positive clinical lymph nodes had higher levels of CK18, TIMP1 and MMP-9 than those in clinical-stage II with negative clinical lymph nodes [17]. Previous research has identified tissue inhibitor of metalloproteinase-1, serum matrix metalloproteinase-8, and the ratio of matrix metalloproteinase-8/tissue inhibitor of metalloproteinase-1 as potential GC prognosis factors [18]. In a GC cell line, IL12A suppressed tissue inhibitors of metalloproteinase 1 expression while elevating insulin-like growth factor- binding protein 1 and plasminogen activator inhibitor 1 expression levels [19]. Matrix metalloproteinase 9 is also known as GELB, CLG4B, MMP-9, or MANDP2. During normal physiological processes, including embryonic development, tissue remodeling and reproduction, as well as in disease processes such as metastasis and arthritis, matrix metalloproteinases degrade the extracellular matrix. According to previous research, SETDB1 overexpression in GC upregulates CCND1 and MMP9 to influence gastric carcinogenesis and metastasis [20]. MMP 9, KLK6, and VEGF were significantly elevated in GC cell lines (SGC-7901, MKN-28, BGC-823, and MGC-803) when compared to the normal gastric epithelial cell line (GES-1) [21]. VGLL1 phosphorylation via TGF-β/ERK/RSK2 signaling has a crucial role in MMP9-driven GC [22].

Fibronectin 1 (also referred to as FN, FNZ, CIG, MSF, FNC, ED-B, GFND, GFND2, LET5, or SMDCF) encodes fibronectin, a glycoprotein. In plasma, fibronectin occurs in a soluble dimeric form while in the extracellular matrix and at the cell surface, it occurs in a dimeric or multimeric form. Impaired FN1 expression significantly suppressed proliferation, migration, invasion, and EMT in GC cells, while enhancing apoptosis [23]. CDH1 expression in CRML and non-tumor liver tissues were comparable, whereas FN1 and VIM expression in metastatic tissue was significantly lower [24]. Furthermore, in cell lines and GC tissues, miR-200c was reported to be significantly downregulated, whereas FN1 showed the opposite pattern [25].

C-X-C motif chemokine ligand 8 (also referred to as NAF, IL8, GCP1, LUCT, LECT, NAP1, LYNAP, GCP-1, MDNCF, NAP-1, MONAP, or SCYB8) encodes a CXC chemokine family member that is important in mediating inflammatory responses. CXCL8 is primarily secreted by macrophages and helps to create an immunosuppressive microenvironment by inducing PD-L1+CXCL8 [26]. CXCL7 and CXCL1 are CXCR2 ligands that may influence GC malignant progression via CXCR2 signaling [27]. The PTPRD-CXCL8 axis may have therapeutic potential, particularly for metastasis suppression in PTPRD-inactivated GCs [28].

MYC is a BHLH transcription factor that has an impact on cell cycle apoptosis, progression, and cellular transformation. It is also known as MRTL, MYCC, c-Myc, or bHLHc39. MYC activity and CCNE1 amplification may be associated with immune-deficient CIN GEAs, to allow for the study of immune evasion mechanisms, ultimately inspiring new therapeutic modalities [29].

Gastric carcinogenesis is initiated by YAP/TAZ activation, with MYC acting as a key downstream mediator [30]. KLFS, MYC/ LINCO0346/miR-34a-5p axis is an effector of GC tumorigenesis and may have therapeutic target potential against GC [31].

Secretem phosphophrotein 1 (BNSP, OPN, ETA-1 or BSPI) helps osteoclasts adhere to the mineralized bone matrix. Through PI3K/AKT signaling, MIR-340 targeting of SPP1 may suppress GC cell proliferation, migration, invasion, as well as EMT. SPP1 silencing by MiR-340 suppresses GC cell proliferation, migration, and EMT while promoting GC cell apoptosis [32]. Multivariate analysis revealed that high epithelial PDGFBR levels (p=0.036 and p=0.02) and SPP1 (p=0.003 and p<0.001) are independent prognostic factors for GC DFS and OS [33].

Prostaglandin-endoperoxide synthase 2 (also referred to as COX-2, COX2, PHS-2, PGHS-2, PGG/HS, GRIPGHS or hCox-2) is a dioxygenase and peroxidase enzyme involved in prostaglan-
din biosynthesis. Clinical specimen analysis revealed that PTGS2 and BCL2 were positively correlated with human GC. Furthermore, xenograft models revealed that inhibiting PTGS2 with celecoxib significantly increased cisplatin cytotoxicity in resistant GC by suppressing PTGS2 and BCL2 expression, which is regulated by the ERK1/2 and P38 signaling axis, implying that PTGS2 may have an adjunctive therapeutic potential for chemoresistance reversal in a subset of cisplatin-resistant GC [34]. These findings indicate that lycopene induces apoptosis and inhibits proliferation in AGS cells by inhibiting the ROS-activated EGFR/Ras/ERK as well as the p38 MAPK pathways, as well as the NF-B-mediated COX-2 gene expression. It has been proposed that lycopene-rich foods may reduce the incidence of GC [35]. Overexpressed LMP1 and LMP2A suppresses COX-2 via a decrease in TRAF2. In GC, P-ERK is involved in COX-2 inhibition by LMP1 [36]. GC patients with low COL1A1 and PTGS2 expression had significantly better overall survival, according to survival analysis.

**Conclusion**

Here, 508 DEGs (177 upregulated, 331 downregulated) were identified in GC vs normal gastric tissues using bioinformatic analysis of GEO gene expression datasets. Of these, 10 hub genes (FN1, MMP9, CXCL8, TIMP1, COL1A1, MYC, SPP1, COL3A1, COL1A2, and PTGS2) that may be core GC genes were selected. COL1A1 and PTGS2 are potential core genes in gastric cancer. Further studies are needed to validate our findings.

**Declarations**

**Acknowledgments**: We acknowledge GEO database for providing their platforms and contributors for uploading their meaningful datasets.

**Authors’ contributions**: YL collected and analyzed the data. LN designed and supervised the study. YL drafted the first version of the manuscript. YL, LN reviewed and revised the manuscript. LN constructed the diagrams and tables of the article. All authors read and approved the final manuscript.

**Funding**: None.

**Conflict of interest**: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Ethics approval**: Our study did not require an ethical board approval because it did not contain human or animal trials.

**Consent to participate**: Not applicable.

**Consent for publication**: Not applicable.

**Availability of data and materials**: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

**Code availability**: Not applicable.

**References**


