Identification of potential hub genes of gastric cancer

Xu-dong Zhou  
Affiliated Clinical College of the Armed Police General Hospital of Anhui Medical University

Ya-Wei Qu  
Third Medical Center of PLA General Hospital

Li Wang  
Graduate School of Jinzhou Medical University

Fu-hua Jia  
Huamei Hospital of China National University of Science and Technology

Peng Chen  
Graduate School of Jinzhou Medical University

Yin-pu Wang  
Baoji Hospital Affiliated to Xi’an Jiaotong University

Hai-Feng Liu (✉ liuhaifeng247@163.com)  
Affiliated Clinical College of the Armed Police General Hospital of Anhui Medical University

Research Article

Keywords: Gastric cancer, differentially expressed genes, protein-protein interaction, bioinformatic analysis

Posted Date: May 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1628077/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

Gastric cancer (GC) is a malignant tumor originated from gastric mucosa epithelium. It is the third leading cause of cancer mortality in China. The early symptoms are not obvious. When it is discovered, it has developed to the advanced stage, and the prognosis is poor. In order to screen for potential genes for GC development, this study obtained GSE118916 and GSE109476 from the GEO database for bioinformatics analysis.

Methods

First, GEO2R was used to identify differentially expressed genes (DEG) and the functional annotation of DEGs was performed by GO and KEGG analysis. The STRING tool was used to construct protein-protein interaction network and the most important modules and hub genes were mined. RT-qPCR assay was performed to verify the expression level of hub genes.

Results

A total of 139 DEGs were identified. The functional changes of DEGs are mainly concentrated in the cytoskeleton, extracellular matrix and collagen synthesis. Eleven genes were identified as core genes. Bioinformatics analysis shows that the core genes are mainly enriched in many processes related to cell adhesion and collagen.

Conclusions

In summary, the DEGs and hub genes found in this study may be potential diagnostic and therapeutic targets.

Introduction

Gastric cancer (GC) is a malignant tumor originated from the gastric mucosal epithelium, mainly gastric adenocarcinoma. GC accounts for more than 95% of malignant tumors in the stomach and is one of the malignant tumors that seriously endanger human health. According to the results of the National Cancer Center of China in 2015, GC accounts for the third place in the mortality rate of malignant tumors in China[1]. The occurrence of GC is closely related to the adverse environment, lifestyle, dietary structure changes and Helicobacter pylori infection. Early GC symptoms are not obvious, some patients may have dyspepsia symptoms, and advanced GC may have upper abdominal pain, postprandial aggravation, poor appetite, anorexia, fatigue and weight loss. The common examination methods are gastroscopy and computed tomography, which are invasive and expensive[2]. When the patient has obvious symptoms, he
is admitted to the hospital. The disease has developed to the advanced stage of GC, and the best surgical
treatment time is lost. Exception for Japan and South Korea, the 5-year survival rate of advanced GC in other
countries and regions in the world is even less than 10%[3]. However, if GC can be diagnosed early, its 5-
year survival rate will rise to 95%[4], which means that the fundamental method for providing GC
prognosis is early diagnosis and timely treatment. Currently, some serum biomarkers are used for
screening early GC, such as CA19-9 and CEA, but these tumor markers are less sensitive and specific[5].
Therefore, to find a new effective biomarker for early GC, to further explore the pathogenesis of GC, to find
potential diagnostic and therapeutic targets, to achieve early detection, early diagnosis and targeted
therapy, with significant clinical value and market Application prospect.

Bioinformatics is an emerging interdisciplinary subject that combines life sciences with computer
science. It focuses on the collection, storage, processing, dissemination, analysis, and interpretation of
biological information. The ability to process large amounts of complex biological data can be processed
through the use of biological and informatics techniques. Microarray data information analysis
technology has been widely used in the study of diseases such as tumors to explore the genetic
correlation[6, 7]. Microarray analysis technology can simultaneously acquire the expression information
of tens of thousands of genes, and then explore the genomic changes related to the development of
diseases. A large number of research and scholars[8, 9] have used bioinformatics techniques to analyze
differentially expressed genes in tumor progression, and to study their roles in biological processes,
molecular functions and signaling pathways, and to elucidate the pathogenesis of diseases, so as to
provide theoretical basis for early diagnosis and treatment.

In this study, bioinformatics technology was used to find the gene sequencing data of GC patients and
normal people from Gene Expression Omnibus (GEO). Two high-quality genetic data sets were extracted
and analyzed for further analysis. Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and
Genomes (KEGG) analysis were performed by Gene Set Enrichment Analysis (GSEA), and then important
modules of the PPI network were screened. Using the genetic data of tumor patients and normal people in
the sample, 73 gene sets and 11 significantly differentially expressed gene molecules were found to be
differentially expressed. These findings will enhance our understanding of the underlying mechanisms of
GC and provide the basis for finding new diagnostic markers and targeted therapies.

Materials And Methods

Access to Public Data

GEO (http://www.ncbi.nlm.nih.gov/geo)[10] is an open high-throughput genomic database that includes
microarrays, gene expression data and chips.

On November 20, 2019, the key words “(gastric cancer) AND gene expression” were set to detect the
datasets, using a filter of “expression profiling by array” and “recent two years”. There were five inclusion
criteria: a sample number of more than ten per dataset (samples of less than ten were excluded), data
from Homo sapiens (data from other species were excluded), a series entry type, expression profiling by
array (data using methylation profiling by array were excluded), and a diagnosis of gastric cancer (data from other cancer diagnoses were excluded).

Two expression profile data sets—GSE118916 and GSE109476—were downloaded from the GEO database. The annotation platform for GSE118916 is GPL15207 platform, [PrimeView] Affymetrix Human Gene Expression Array. The GSE118916 data set is composed of 15 GC tissues and 15 stomach normal tissues. The annotation platform for GSE109476 is GPL24530 platform, Arraystar Human LncRNA microarray V2.0 (Agilent-033010; custom-annotation; probe name version). The GSE109476 date set is composed of 5 GC tissues and 5 stomach normal tissues. All probe numbers are converted to gene symbols based on the annotation information in the platform.

Screening of DEGs via GEO2R

GEO2R (http://www.ncbi.nlm.nih.gov/geo/geo2r)[11] is a system for online analysis of data in GEO. This tool system runs in the R language. It is accurate to say that it uses two R packages: GEOquery and limma. The former is used for data reading and the latter is used for calculation. The best thing about GEO2R is that it is an online tool, easy and efficient to operate. GEO2R can perform a command to compare gene expression profiles between groups in order to identify DEGs between GC and stomach normal groups. In general, when the probe set has a corresponding gene symbol, the probe is considered valuable and will be retained. Statistically significant measure is P-value<0.01 and Fold change (FC) >1.

Functional Annotation of DEGs via GO and KEGG Analysis

DAVID(https://david.ncifcrf.gov/home.jsp) (version 6.8) is a bioinformatics database that integrates biological data and analytical tools[12]. Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.kegg.jp/) could help researcher to understand advanced functions and biological systems[13]. Gene ontology (GO) is an ontology widely used in bioinformatics, which covers three aspects of biology, including cellular components, molecular function and biological process[14]. In order to analyze the GO and pathway enrichment information of DEGs, the DAVID online tool was executed. Statistically significant measure is P<0.05.

Construction and Analysis of Protein-Protein Interaction Network

Search Tool for the Retrieval of Interacting Genes (STRING; http://string-db.org) (version 10.5)[15] is a network that can be used to predict and track protein-protein interactions. Introducing DEGs into the tool makes intermolecular network analysis. The analysis of the interactions between different proteins can provide insights into the mechanisms of generation or development of GC. In this study, STRING database was used to construct PPI network with DEGs. The minimum required interaction score is that medium confidence > 0.4. Cytoscape (version 3.6.1) is an open visualization software that can be used to visualize protein-protein interaction network[16]. Based on topological principles, the Molecular Complex Detection (MCODE) (version 1.5.1), a plug-in for Cytoscape, can mine tightly coupled regions from PPI. First, Cytoscape software plots the PPI network. Secondly, MCODE identifies the most important
modules in the PPI network graph. The criteria of MCODE analysis is that node score cut-off=0.2, degree cut-off=2, Max depth=100, MCODE scores >5, and k-score=2.

**Mining and Screening of Core Gene**

The hub genes were selected with degrees≥10. A network of the genes and their co-expression genes was analyzed using cBioPortal (http://www.cbioportal.org)[17, 18] online platform. Hierarchical clustering of hub genes was constructed using UCSC Cancer Genomics Browser (http://genome-cancer.ucsc.edu)[19]. The overall survival and disease-free survival analyses of hub genes were performed using Kaplan-Meier curve in cBioPortal.

**RT-qPCR assay**

A total of 10 GC participates were recruited. After surgery, 10 GC tumor samples from GC patients and 10 adjacent normal stomach tissues samples were obtained. The research conformed to the Declaration of Helsinki and was authorized by the Human Ethics and Research Ethics Committees of Third Medical Center of PLA General Hospital. The written informed consents were obtained from all participates.

Total RNA was extracted from 10 GC tumor samples and 10 adjacent normal stomach tissues samples by the RNAiso Plus (Trizol) kit (Thermofisher, Massachusetts, America), and reverse transcribed to cDNA. RT-qPCR was performed using a Light Cycler® 4800 System with specific primers for genes. Table 1 presents the primer sequences used in the experiments. The RQ values (2^{ΔΔCt}, where Ct is the threshold cycle) of each sample were calculated, and are presented as fold change in gene expression relative to the control group. GAPDH was used as an endogenous control.

**The verification of hub genes expression and role on the overall survival of GC patients using The Cancer Genome Atlas (TCGA) data**

The gene expression dataset of GC in the TCGA was downloaded. There were a total of 580 samples including 478 GC samples and 102 normal gastric samples. The IlluminaHiSeq UNC was selected as gene expression RNAseq in the research. In addition, the gene expression levels of hub genes between GC and normal gastric samples were compared using the one-way Anova.

Furthermore, effect of gene expression of hub genes on overall survival was analyzed by using the TCGA data.

**Statistical Analysis**

Student’s t test was used to determine the statistical significance when comparing the two groups. Statistical analysis was carried out using SPSS software version 21.0(IBM Corp. Armonk, NY, USA). Value of P<0.05 were considered statistically significant.

**Results**
Identification of DEGs in GC

One volcano plot presents the DEGs in the GSE118916 (Fig. 1A), and another volcano plot presents the DEGs in the GSE109476 (Fig. 1B). After standardization of the microarray results, DEGs (1,768 in GSE118916, and 564 in GSE109476) were identified. The overlap among the 2 datasets contained 139 genes as shown in the Venn diagram (Fig. 1C), consisting of 189 downregulated genes and 84 upregulated genes between GC tissues and non-cancerous tissues.

KEGG and GO enrichment analyses of DEGs

To analyze the biological classification of DEGs, functional and pathway enrichment analyses were performed using DAVID. GO analysis results showed that changes in biological processes (BP) of DEGs were significantly enriched in collagen catabolic process, collagen fibril organization, extracellular matrix organization, integrin-mediated signaling pathway, cell adhesion and so on. Changes in molecular function (MF) were mainly enriched in collagen binding, growth factor binding, heparin binding, extracellular matrix structural constituent and so on (Table I). Changes in cell component (CC) of DEGs were mainly enriched in extracellular matrix, proteinaceous extracellular matrix, collagen trimer, extracellular region and so on. The KEGG pathway analysis showed that all DEGs are mainly concentrated in ECM-receptor interaction, PI3K-Akt signaling pathway, Metabolism of xenobiotics by cytochrome P450, platelet activation, Gap junction, Protein digestion and absorption and Phagosome (Table 2).

PPI network construction and module analysis

The PPI network of DEGs was constructed (Fig. 2) and the most significant module was obtained using Cytoscape (Fig. 3). The functional analyses of genes involved in this module were analyzed using DAVID.

Hub gene selection and analysis

A total of 11 genes were identified as hub genes with degrees ≥10. The names, abbreviations and functions for these hub genes are shown in Table 3. A network of the hub genes and their co-expression genes was analyzed using cBioPortal online platform (Fig. 4A). Hierarchical clustering showed that the hub genes could basically differentiate the GC samples from the non-cancerous samples (Fig. 4B). Subsequently, the overall survival analysis of the hub genes was performed using Kaplan-Meier curve. GC patients with COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, SERPINH1, COL8A2, COL6A3, LAMA4, LOXL1 and COL5A2 alteration showed worse overall survival (Fig. 5, and 6).

Results of RT-qPCR analysis

According to the above expression analysis, COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, SERPINH1, COL8A2, COL6A3, LAMA4, LOXL1, and COL5A2 were markedly up-regulated in GC tumor samples. As presented in Fig. 7, the relative expression levels of COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, SERPINH1, COL8A2, COL6A3, LAMA4, LOXL1, and COL5A2 were significantly higher in GC samples,
compared with the normal stomach tissues groups. The result demonstrated that COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, SERPINH1, COL8A2, COL6A3, LAMA4, LOXL1, and COL5A2 might be considered as biomarkers for GC.

**The verification by TCGA**

According to the above expression analysis, COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, SERPINH1, COL8A2, COL6A3, LAMA4, LOXL1, and COL5A2 were significantly up-regulated in GC tumor samples compared with the normal gastric samples. After confirmation using TCGA data, these genes expression levels in GC samples were also significantly higher than the normal gastric samples (Fig. 8).

Overall survival analysis showed that GC patients with high expression levels of COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, SERPINH1, COL8A2, COL6A3, LAMA4, LOXL1, and COL5A2 had poorer overall survival times than those with low expression levels (P<0.05, Fig.9).

**Discussion**

In 2018, there were more than 1 million new cases of GC in the world, and 783,000 deaths[20]. The most common sites of GC were gastric antrum (58%), cardia (20%), corpus (15%), whole stomach or most stomach (7%). GC can spread through direct spread, lymph node metastasis, hematogenous dissemination, and plant metastasis. At present, the treatment of GC is often treated by multiple means. The treatment may include partial gastrectomy or total gastrectomy, lymph node dissection and perioperative chemotherapy or postoperative radiotherapy and chemotherapy.[21–23]. Patients may experience malnutrition, reduced immunity, and decreased quality of life during treatment. And it will bring a series of adverse reactions to patients, so that patients with GC not only suffer from physiologically great pain, but also psychologically bear tremendous pressure. After gastrectomy, the physiological function of patients will be seriously disturbed, and the body will also suffer from malnutrition, reflux esophagitis, absorption disorders and other adverse consequences[24, 25]. On the other hand, since medicinal chemotherapy kills cancer cells and kills normal cells of the patient, it causes toxic effects and a series of adverse reactions, which cause serious damage to the patient's body and mind. The prognosis of patients is often associated with timely diagnosis and treatment, but there are large clinical heterogeneities in different individuals and tumor types. Therefore, it is of great clinical significance to further explore the pathogenesis of GC, to find early diagnostic markers, targeted therapeutic genes and molecules, and to achieve early diagnosis and individualized treatment according to different individuals and pathological types.

Bioinformatics technology has been widely used to find genetic molecules related to tumorigenesis and development, and to find genes and molecules that can be used as therapeutic targets. Cao et al. found the PLEKHG1 molecule related to GC through this technology, and further confirmed the correlation between the gene and GC, suggesting that the molecule is a biomarker for diagnosis and prediction of outcome[26]. Wang et al. found a molecule related to colorectal cancer proliferation and metastasis through bioinformatics technology, suggesting that it may serve as a potential therapeutic target[27]. In
this study, DEGs between GC tissues and non-cancer tissues were obtained by analyzing two mRNA microarray data sets. A total of 139 DEGs were identified in 2 data sets. Bioinformatics analysis revealed high expression of COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, SERPINH1, COL8A2, COL6A3, LAMA4, LOXL1, and COL5A2 in GC patients. At the same time, multiple gene sets that were significantly up-regulated and down-regulated were found by Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis.

COL1A2 (Collagen Type I Alpha 2 Chain) is a member of the fibrocollagen family and encodes a pro-alpha 2 chain of type I collagen[28]. It acts to support the matrix structure, forming the interstitial part of most solid tumors, and regulates cell movement through interaction with the cytoskeleton. Studies have found that COL1A2 gene mainly affects cell proliferation, differentiation, adhesion and metastasis through extracellular matrix receptor interaction pathway and local adhesion pathway, mainly related to tumor invasion and metastasis[29]. Li et al found that the expression of COL1A2 in GC tissues was higher than that in adjacent normal tissues[30], which was the same as the bioinformatics analysis in this study. Ponticos et al. suggest that low expression of COL1A2 can inhibit the expression of TGF-B in cancer cells[31]. Since TGF-B contributes to the activation of PI3K signaling pathway, it is hypothesized that low expression of COL1A2 may inhibit the activation of PI3K signaling pathway by down-regulating the expression of TGF-B in cancer cells, and promote the apoptosis of GC cells[28]. The high expression of COL1A2 can promote the proliferation, invasion and migration of GC, while the low expression of COL1A2 can inhibit the proliferation of GC cells, delay cell migration, and promote the apoptosis of GC cells. Therefore, COL1A2 can be a potential biomarker and therapeutic target.

SPARC (secreted protein acid and cysteine rich) is located in 5q33.1. It is a relative molecular mass of 32,000 nonstructural secreted extracellular matrix glycoprotein, it consists of a single polypeptide (285 amino acids), with the first 1981 U.S. TERMINE equal separation and purification of fetal bovine bone in humans[32]. It mediates the interaction of cell-microenvironment and has a wide range of biological effects in tumorigenesis, invasion, metastasis, angiogenesis and inflammation[33]. The study found that in some tumors with high metastatic characteristics, such as glioblastoma, melanoma, breast cancer and prostate cancer, SPARC can promote bone metastasis and epithelial-mesenchymal transition (EMT) and promote tumor development, but as an anti-angiogenesis pancreatic cancer, colorectal cancer, gastric low metastatic tumors, pro-apoptotic, inhibition of cell proliferation and inhibition of cell cycle antitumor factor[34, 35]. Its role in GC cells is highly controversial. Tsutomu et al found that the expression of SPARC mRNA in GC tissues was higher than that in the normal control group, and the prognosis of high SPARC expression was poor compared with low SPARC expression[36]. Chen et al also showed that in 140 ovarian cancer patients, high SPARC expression had a worse prognosis than low SPARC expression[37]. Chew et al. and Liang et al reported that low SPARC expression was associated with poor long-term survival in 120 patients and 114 patients with colorectal cancer[38, 39]. SPARC may play different roles in different stages of cancer and different stages of development of the same cancer. This study found that SPARC expression in GC tissues was higher than that in adjacent tissues, and the prognosis was poor.
SERPINH1 (Serpin Family H Member 1) is a member of the serine protease inhibitor H subfamily, also known as HPS47, a heat shock protein 47, and the coding gene is located in the 11q13.5 region of human chromosome 11. It is involved in biological processes such as collagen synthesis and endopeptidase activity, and can be used as a partner in the biosynthesis pathway of collagen[40]. SERPINH1 is closely associated with collagen-related diseases, including osteogenesis imperfecta (OI), keloids, and fibrosis[41, 42]. Qi et al found that SERPINH1 is highly expressed in renal clear cell carcinoma with poor prognosis[43]. Studies have reported that SERPINH1 is associated with the occurrence and development of glioma and cervical cancer, and is a possible therapeutic target[44, 45]. Zhang et al found that SERPINH1 is up-regulated in GC[46], and it is possible to promote tumor growth and invasion by regulating the extracellular matrix (ECM) network. This study found that high expression of SERPINH1 in GC tissues, poor prognosis in patients with low expression, can be a potential biomarker.

Our study identified 139 DEGs and 11 Hub genes that may be associated with the occurrence and development of GC. There are corresponding literatures indicating that COL1A2[47, 48], COL3A1[49], SPARC[50], SERPINH1[51], COL6A3[26]. These genes are highly expressed in gastric cancer tissues, and the expression of LOXL1[52] is also related to distant metastasis of gastric cancer. However, the PCOLCE, COL8A2, COL8A1, and LAMA4 genes have not yet been documented to indicate their role in gastric cancer, and we subsequently recruited some patients. Relevant RT-qPCR experimental verification of these Hub genes is more indicative of the role of these genes in the development of gastric cancer than other studies.

Although the study conducted a rigorous bioinformatics analysis, a large number of clinical samples, animal experiments should be comprehensively verified to better understand the pathogenesis of primary colorectal cancer.

In summary, we identified 20 gene sets and 10 distinct differentially expressed genes from genetic samples from patients with GC and normal subjects through bioinformatics analysis. Hub genes in DEGs may provide new ideas and evidence for the diagnosis and targeted therapy of GC.

**Conclusion**

In conclusion, the present research aimed to identify DEGs which might be contained in the occurrence or development of GC. Finally, 139 DEGs and 11 hub genes were confirmed between GC tissues and normal tissues, which could be used as diagnostic and therapeutic biomarkers for GC. However, the biological functions of the all hub genes in GC require further researches.

**Abbreviations**

Gastric cancer (GC), differentially expressed genes (DEG), Gene Expression Omnibus (GEO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Set Enrichment Analysis (GSEA), Gene ontology (GO), The Cancer Genome Atlas (TCGA), biological processes (BP), molecular function (MF), cell component
(CC), secreted protein acid and cysteine rich (SPARC), epithelial-mesenchymal transition (EMT), Serpin Family H Member 1 (SERPINH1), osteogenesis imperfecta (OI), extracellular matrix (ECM)

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication

All the authors agreed to be published.

Availability of data and materials

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

Competing interests

The authors declare that there is no conflict of interest.

Funding

This project was funded by the Capital Clinical Application Research-Research on the Diagnosis Value of Early Gastric Cancer and Precancerous Lesions by High-resolution Microendoscopy, No. Z141107002514099.

Authors' contributions

Xu-dong Zhou participated in the topic selection and design of the paper. Ya-Wei Qu and Li Wang participated in data analysis and interpretation. Fu-hua Jia and Peng Chen drafted and revised key theories in the paper, Yin-pu Wang and Hai-Feng Liu answered academic questions.

Acknowledgements

Not Applicable

References


Tables
<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5′–3′)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COL1A2-hF</td>
<td>AGGGAAGGTAGTAACAGTAG</td>
</tr>
<tr>
<td>COL1A2-hR</td>
<td>CCAGGATTACCCTATGAG</td>
</tr>
<tr>
<td>COL3A1-hF</td>
<td>TGAGCCTGGTAAGAATGG</td>
</tr>
<tr>
<td>COL3A1-hR</td>
<td>CCTGGAACACCTGGAATA</td>
</tr>
<tr>
<td>SPARC-hF</td>
<td>GGCTGGTCACATAGGTAC</td>
</tr>
<tr>
<td>SPARC-hR</td>
<td>GAGGGTTAAGCAAGGAAT</td>
</tr>
<tr>
<td>PCOLCE-hF</td>
<td>TCCTCCGTGCTGTGGTGT</td>
</tr>
<tr>
<td>PCOLCE-hR</td>
<td>GTTCAGATCCCGTCCCT</td>
</tr>
<tr>
<td>COL8A1-hF</td>
<td>TGAACCAATCTGGCCTCC</td>
</tr>
<tr>
<td>COL8A1-hR</td>
<td>TTTGCTGCTAAGCCGTGA</td>
</tr>
<tr>
<td>SERPINH1-hF</td>
<td>CCTGAAGAATGGAGCAA</td>
</tr>
<tr>
<td>SERPINH1-hR</td>
<td>AGGAGCGGAAAAGGACACT</td>
</tr>
<tr>
<td>COL8A2-hF</td>
<td>GAAACAAGAGCGATGACG</td>
</tr>
<tr>
<td>COL8A2-hR</td>
<td>CAGCGGTGAGAAGGTTGT</td>
</tr>
<tr>
<td>COL6A3-hF</td>
<td>CCCAGGAGTTCAAGACCA</td>
</tr>
<tr>
<td>COL6A3-hR</td>
<td>GAGGAGGACAACACCATC</td>
</tr>
<tr>
<td>LAMA4-hF</td>
<td>CTGGACCTAACTGTGAAA</td>
</tr>
<tr>
<td>LAMA4-hR</td>
<td>GTATAAAGAATGGGCGAAA</td>
</tr>
<tr>
<td>LOXL1-hF</td>
<td>CGTTCACCTGAGCGTGTTT</td>
</tr>
<tr>
<td>LOXL1-hR</td>
<td>GTGCACTCCTCTATGTCCTC</td>
</tr>
<tr>
<td>COL5A2-hF</td>
<td>AGCCAGGTTTGAGGAGCA</td>
</tr>
<tr>
<td>COL5A2-hR</td>
<td>GCAGCAATTAGTTGAGCC</td>
</tr>
<tr>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>GO:0030574</td>
<td>collagen catabolic process</td>
</tr>
<tr>
<td>GO:0030199</td>
<td>collagen fibril organization</td>
</tr>
<tr>
<td>GO:0030198</td>
<td>extracellular matrix organization</td>
</tr>
<tr>
<td>GO:0007229</td>
<td>integrin-mediated signaling pathway</td>
</tr>
<tr>
<td>GO:004593</td>
<td>camera-type eye morphogenesis</td>
</tr>
<tr>
<td>GO:0007263</td>
<td>nitric oxide mediated signal transduction</td>
</tr>
<tr>
<td>GO:0045926</td>
<td>negative regulation of growth</td>
</tr>
<tr>
<td>GO:0071294</td>
<td>cellular response to zinc ion</td>
</tr>
<tr>
<td>GO:0030199</td>
<td>extracellular matrix</td>
</tr>
<tr>
<td>GO:0005615</td>
<td>extracellular space</td>
</tr>
<tr>
<td>GO:0005578</td>
<td>proteinaceous extracellular matrix</td>
</tr>
<tr>
<td>GO:0005581</td>
<td>collagen trimer</td>
</tr>
<tr>
<td>GO:0005576</td>
<td>extracellular region</td>
</tr>
<tr>
<td>GO:0005788</td>
<td>endoplasmic reticulum lumen</td>
</tr>
<tr>
<td>hsa04512</td>
<td>ECM-receptor interaction</td>
</tr>
<tr>
<td>hsa04151</td>
<td>PI3K-Akt signaling pathway</td>
</tr>
<tr>
<td>hsa00980</td>
<td>Metabolism of xenobiotics by cytochrome P450</td>
</tr>
<tr>
<td>hsa04611</td>
<td>Platelet activation</td>
</tr>
<tr>
<td>hsa04540</td>
<td>Gap junction</td>
</tr>
<tr>
<td>hsa04974</td>
<td>Protein digestion and absorption</td>
</tr>
<tr>
<td>hsa04145</td>
<td>Phagosome</td>
</tr>
</tbody>
</table>

GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; DEGs: differentially expressed genes.
Table 3. Summaries for the function of 11 hub genes.

<table>
<thead>
<tr>
<th>No.</th>
<th>Gene symbol</th>
<th>Full name</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>COL1A2</td>
<td>Collagen Type I Alpha 2 Chain</td>
<td>Type I collagen is a member of group I collagen (fibrillar forming collagen). collagen type I, alpha 2, fibril forming, putative down-regulated c-Myc target gene, COL1A2.</td>
</tr>
<tr>
<td>2</td>
<td>COL3A1</td>
<td>Collagen Type III Alpha 1 Chain</td>
<td>Involved in regulation of cortical development. Is the major ligand of ADGRG1 in the developing brain and binding to ADGRG1 inhibits neuronal migration and activates the RhoA pathway by coupling ADGRG1 to GNA13 and possibly GNA12.</td>
</tr>
<tr>
<td>3</td>
<td>SPARC</td>
<td>Secreted Protein Acidic And Cysteine Rich</td>
<td>Appears to regulate cell growth through interactions with the extracellular matrix and cytokines. Binds calcium and copper, several types of collagen, albumin, thrombospondin, PDGF and cell membranes.</td>
</tr>
<tr>
<td>4</td>
<td>PCOLCE</td>
<td>Procollagen C-Endopeptidase Enhancer</td>
<td>Binds to the C-terminal propeptide of type I procollagen and enhances procollagen C-proteinase activity. C-terminal processed part of PCPE (CT-PCPE) may have an metalloproteinase inhibitory activity.</td>
</tr>
<tr>
<td>5</td>
<td>COL8A1</td>
<td>Collagen Type VIII Alpha 1 Chain</td>
<td>Macromolecular component of the subendothelium. Major component of the Descemet's membrane (basement membrane) of corneal endothelial cells. Also component of the endothelia of blood vessels.</td>
</tr>
<tr>
<td>6</td>
<td>SERPINH1</td>
<td>Serpin Family H Member 1</td>
<td>Binds specifically to collagen. Could be involved as a chaperone in the biosynthetic pathway of collagen.</td>
</tr>
<tr>
<td>7</td>
<td>COL8A2</td>
<td>Collagen Type VIII Alpha 2 Chain</td>
<td>Necessary for migration and proliferation of vascular smooth muscle cells and thus, has a potential role in the maintenance of vessel wall integrity and structure, in particular in atherogenesis.</td>
</tr>
<tr>
<td>8</td>
<td>COL6A3</td>
<td>Collagen Type VI Alpha 3 Chain</td>
<td>Collagen VI acts as a cell-binding protein. collagen type VI, alpha 3 (300kDa), microfibrillar, putative down-regulated c-Myc target gene, COL6A3</td>
</tr>
<tr>
<td>9</td>
<td>LAMA4</td>
<td>Laminin Subunit Alpha 4</td>
<td>Binding to cells via a high affinity receptor.</td>
</tr>
<tr>
<td>10</td>
<td>LOXL1</td>
<td>Lysyl Oxidase Like 1</td>
<td>Active on elastin and collagen substrates.</td>
</tr>
<tr>
<td>11</td>
<td>COL5A2</td>
<td>Collagen Type V Alpha 2 Chain</td>
<td>Type V collagen binds to DNA, heparan sulfate, thrombospondin, heparin, and insulin. Type V collagen is a key determinant in the assembly of tissue-specific matrices.</td>
</tr>
</tbody>
</table>

**Figures**
**Figure 1**

DEGs in GC. (A) One volcano plot presents the DEGs in the GSE118916. (B) another volcano plot presents the DEGs in the GSE109476. (C) Venn diagram, PPI network and the most significant module of DEGs. (A) DEGs were selected with a fold change >1 and P-value <0.01 among the mRNA expression profiling sets GSE118916 and GSE109476. The 2 datasets showed an overlap of 139 genes.
Figure 2

The PPI network of DEGs was constructed using Cytoscape.
Figure 3

The most significant module was obtained from PPI network with 11 nodes.
Figure 4

Interaction network and biological process analysis of the hub genes. (A) Hub genes and their co-expression genes were analyzed using cBioPortal. Nodes with bold black outline represent hub genes. Nodes with thin black outline represent the co-expression genes. (B) Hierarchical clustering of hub genes was constructed using UCSC. The samples under the pink bar are non-cancerous samples and the
samples under the blue bar are GC samples. Upregulation of genes is marked in red; downregulation of genes is marked in blue.

**Figure 5**

Overall survival analyses of hub genes (COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, and SERPINH1). P<0.05 was considered statistically significant.
Figure 6

Overall survival analyses of hub genes (COL8A2, COL6A3, LAMA4, LOXL1 and COL5A2). P<0.05 was considered statistically significant.
Figure 7

Relative expression of COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, SERPINH1, COL8A2, COL6A3, LAMA4, LOXL1, and COL5A2 by RT-qPCR analysis. *p < 0.05, compared with normal stomach tissues.
Figure 8

The confirmation of gene expression level using The Cancer Genome Atlas (TCGA) data. The genes expression levels of COL1A2, COL3A1, SPARC, PCOLCE, COL8A1, SERPINH1, COL8A2, COL6A3, LAMA4, LOXL1, and COL5A2 in GC samples were significantly higher than the normal gastric samples.
Figure 9

The effect of gene expression on overall survival by using the TCGA data.