Collagen family genes and related genes might be associated with prognosis of patients with gastric cancer: an integrated bioinformatics analysis and experimental validation

Kongyan Weng¹,², Yinger Huang¹,², Hao Deng¹,², Ruixue Wang³, Shuhong Luo³, Hongfeng Wu¹,², Jialing Chen¹,², Mingjian Long⁴, Wenbo Hao¹,²

¹Institute of Antibody Engineering, School of Laboratory Medicine and Biotechnology, Southern Medical University, Guangzhou, China; ²Guangdong Provincial Key Laboratory of Construction and Detection in Tissue Engineering, Southern Medical University, Guangzhou, China; ³Department of Laboratory Medicine, School of Stomatology and Medicine, Foshan University, Foshan, China; ⁴Department of Laboratory Medicine, The Fifth Affiliated Hospital, Southern Medical University, Guangzhou, China

Contributions: (I) Conception and design: K Weng, W Hao; (II) Administrative support: W Hao; (III) Provision of study materials: K Weng, Y Huang; (IV) Collection and assembly of data: K Weng; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Wenbo Hao. Institute of Antibody Engineering, School of Laboratory Medicine and Biotechnology, Southern Medical University, Guangzhou, China. Email: haowa@126.com.

**Background:** Gastric cancer (GC) is disease with a high morbidity. The purpose of this study was to identify genes essential to GC development in patients and to reveal the underlying mechanisms of progression.

**Methods:** Bioinformatics analysis is an effective tool for discovering essential genes of different disease states. We used the Gene Expression Omnibus (GEO) database to identify differentially expressed genes (DEGs), the DAVID online tool to perform Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of DEGs, the STRING database to construct the protein-protein interaction (PPI) network of DEGs, the Oncomine and the Cancer Genome Atlas-Stomach Adenocarcinoma (TCGA-STAD) databases to analyze the gene expression differences, the Human pan-Cancer Methylation database (MethHC) to compare the DNA methylation of genes, and the Kaplan-Meier plotter to show the survival analysis of DEGs. We performed Real-Time quantitative PCR (RT-qPCR) experiment to confirm our analysis results.

**Results:** After the integration of four Gene Expression Series (GSEs), we identified 407 DEGs. GO and KEGG pathway analysis indicated that the upregulated DEGs were significantly enriched in Extracellular Matrix (ECM) related functions and pathways. The main DEGs were collagens (COLs). Moreover, the downregulated DEGs were enriched in ethanol oxidation. Several groups of DEGs, such as insulin-like growth factor binding protein (IGFBP), collagen (COL) and serpin peptidase inhibitors (SERPIN) gene families, constituted several PPI networks. In the Oncomine database, all of the collagen genes were highly expressed in breast cancer, esophageal cancer, GC, head and neck cancer and pancreatic cancer, compared with normal tissues. Consistently, from the TCGA-STAD database, most of the collagens (COLs) were highly expressed and exhibited methylated variation in GC patients. In GC patients, some of these collagen (COL) genes related to worse prognosis, as evidenced by the results from the Kaplan-Meier plotter database analysis. Our RT-qPCR results showed that collagen type III α1 chain (COL3A1) was highly expressed in GC cells. Collagen type V α1 chain (COL5A1) was highly expressed, except in AGS cells, which was consistent with our analysis.

**Conclusions:** Collagen (COL) family genes might serve as progression and prognosis markers of GC.

**Keywords:** Gastric cancer (GC); bioinformatics analysis; collagens; prognosis; experimental validation
Introduction

Data from the GLOBOCAN database indicates that, globally, there are more than 1,000,000 new cases of gastric cancer (GC) each year, causing an estimated 783,000 deaths in 2018, making it the fifth most frequently diagnosed cancer and the third leading cause of cancer deaths (1). While new treatment strategies and drug developments have made significant progress, due to the low early detection rate of GC, the survival rate of GC patients remains low (2,3). In addition to the existing primary treatments, targeted therapy is expected to be an essential supplementary treatment for advanced GCs (4). Therefore, it is necessary to explore new molecular targets as well as new, highly sensitive and specific biomarkers to elucidate the molecular mechanisms of GC and improve the prognosis of patients with GC.

Recently, bioinformatics analyses (5) have become increasingly popular for analyzing gene expression changes of the in the progression and development of diseases. For example, the online GEO database (http://www.ncbi.nlm.nih.gov/geo) is a public functional genomics tool that can be utilized to analyze experimental gene expression data uploaded by researchers to identify differentially expressed genes (DEGs) of importance to disease. The DAVID online database (http://david.ncifcrf.gov) holds information related to proteins and genes, and can be used to mine data for Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses of these genes. Similarly, differences in gene expression between tumor and normal tissues can be obtained from the TCGA database. STRING (http://string-db.org) is an online database for use in analyzing PPI networks. These online databases assist in experimental data integration and identification of important genes. In the present study, using GO enrichment analysis, we found several DEGs in GC patients, including collagens (COLs), alcohol dehydrogenases (ADHs), N-acetyl galactosyltransferases (GALNTs). Combining the KEGG, GO, and PPI network analysis results, we selected COLs for more in-depth analysis.

From the HUGO Gene Nomenclature Committee database (https://www.genenames.org/), we know that collagen-encoded proteins contain one or more collagen-like domains. Found in vertebrates, this fibrin is a significant component of skin, bones, tendons, cartilage, blood vessels and teeth. Moreover, it is a substantial component of the tumor microenvironment and is involved in cancer fibrosis (6,7). Cancer cells can regulate collagen biosynthesis through mutant genes (8), transcription factors (9,10), signaling pathways and receptors (11,12). Furthermore, collagen can affect tumor cell behavior through tyrosine kinase receptors, integrins, domain receptors, discoidin and some signaling pathways. In GC, collagen type IV α3 chain (COL4A3) has been identified as a potential prognostic factor (13), but few articles have discussed the relationship between collagen genes and GC (14). Therefore, we performed an in-depth study of the COL gene family's role in GC in order to expose progression mechanisms and to identify prognostic and progression markers.

We present the following article in accordance with the MDAR checklist (available at http://dx.doi.org/10.21037/tcr-20-1726).

Methods

Microarray data and Identification of DEGs.

We downloaded four gene expression series (GSE79973, GSE26899, GSE54129 and GSE29272) from the GEO database and screened the DEGs of each series between GC and normal samples by GEO2R (http://www.ncbi.nlm.nih.gov/geo2r). Genes with more than one probe set or probe sets without corresponding gene symbols were removed or averaged, respectively. Adjusted P value <0.01 and |log2 Fold Change| >1 were considered statistically significant. Venn diagram of the differentially upregulated and downregulated genes were created (http://bioinfogp.cnb.csic.es/tools/venny/index.html). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

KEGG and GO enrichment analyses of DEGs

We used the DAVID (http://david.ncifcrf.gov) online database (version 6.8) to analyze the function of identified DEGs and P<0.05 was considered statistically significant.
PPI network construction and module analysis

In the present study, we used STRING (http://string-db.org) (version 11.0) to construct the PPI network of the DEGs, where a combined score >0.9 was considered statistically significant. We utilized Cytoscape (version 3.7.2) to analyze the molecular interaction networks and MCODE, a Cytoscape app for finding densely connected regions in a given network, was used to identify the most significant modules in the PPI networks. The criteria for selection were as follows: node score cut-off =0.1, degree cut-off =2, k-score =2 and Max depth =100. The genes in the module were analyzed by GO and KEGG using DAVID.

COLs Gene Expression between normal and tumor samples

We utilized Oncomine (https://www.oncomine.org/) to investigate the mRNA levels of COLs in normal and tumor tissues. We retrieved twelve members of COL family genes from the Oncomine database. In our study, the P values of comparison were generated from the student's t-test. The fold change and cut-off P value were defined as 2 and 0.01, respectively. The expression of COL genes in normal and gastric tumor tissues was also studied using the TCGA-STAD database (http://ualcan.path.uab.edu/index.html).

COL gene methylation between normal versus tumor tissues

We compared the DNA methylation of COL genes between normal and GC tissues using the Human Pan-cancer Methylation database, MethHC (http://methhc.mbc.nctu.edu.tw/). The correlation between COL mRNA expression and the methylation in GC patients was analyzed. In our study, the average value was used as a method for evaluating methylation levels and promoter regions selected for analysis.

Prognostic values of COL members in GC patients

The Kaplan-Meier plotter online database (http://kmplot.com) was used to analyze the relationship between COL expression and the overall survival (OS), first progression (FP), and post-progression survival (PPS) in GC patients. The median COL expression was used as the cut-off. Log-rank P value and hazard ratios, with 95% CI, were calculated.

Cell culture, RNA extraction and real-time quantitative PCR

Human GC cells (AGS, MKN45, HGC27, SGC7901) and human gastric mucosal epithelial cells (GES-1) brought from ATCC were cultured in DMEM supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin. Cells were maintained at 37 °C in a 5% CO2 atmosphere. Total RNA was extracted from cell samples using an Animal Total RNA Isolation Kit (Foregene, China). After quality control, total RNA was reverse transcribed into cDNA using a RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific, Waltham, MA, USA). A SYBR Green ITM premix Ex TaqTM II reagent kit (Takara Biomedical Technology, Guangzhou, China) was employed to amplify and quantify the cDNA templates. All PCR reaction systems and conditions were conducted according to the manufacturer’s instructions. Primers for COL3A1 were 5’-GAAAGAGGATCTGAGGGCTCC-3’ (forward) and 5’-AAACCGCCAGCTTTTTCACC-3’ (reverse) and those for COL5A1 were 5’-CTGACAAGAAGTCCGAAGGG-3’ (forward) and 5’-CGTCCACATAGGAGAGCAGTTT-3’ (reverse). Primers for β-actin were 5’-AACTGGGACGACATGGAGAAAA-3’ (forward) and 5’-GGATAGCACAGCCTGGATAGCA-3’ (reverse). The 2−∆∆Ct method was used to calculate expression levels of target genes.

Statistical analysis

Normally distributed data were expressed as mean ± standard deviation (x ± SD). To examine statistical differences between mRNA expression levels and DNA methylation levels of normal and tumor tissues in GC patients, a two-tailed unpaired Student’s t-test was used, P<0.05 was considered to indicate a statistically significant difference. The RT-qPCR analysis was made by GraphPad Prism 7 software and the t-test was used.

Results

Identification of DEGs and GO enrichment and KEGG analyses

After integrating microarray results according to our standards, we identified several DEGs (3160 in GSE54129, 1581 in GSE79973, 428 in GSE26899 and 445 in GSE29272). The overlap among the four gene
expression series contained 407 genes, as shown in the Venn diagram (Figure 1), consisting of 275 downregulated genes (Figure 1A) and 132 upregulated genes (Figure 1B). Among the 407 overlapping genes, we used the DAVID online analysis tool to upload all genes that were upregulated and downregulated, thereby determining statistically rich GO terms and KEGG pathways. GO analysis results showed that upregulated DEGs were involved mainly in extracellular matrix (ECM), organization in biological processes (BP), the ECM in cell component (CC), and ECM structural constituent in molecular function (MF). Moreover, downregulated DEGs were involved mainly with ethanol oxidation in BP and ADH activity in MF (Table 1). The significantly enriched pathways of the DEGs analyzed by the KEGG database are shown in Table 2. Upregulated genes were enriched mainly in chemical carcinogenesis, retinol metabolism, glycolysis/gluconeogenesis, metabolism of xenobiotics by cytochrome P450 and drug metabolism-cytochrome P450.

**DEG PPI network analyses**

PPI networks involving 150 DEGs (consisting of 75 downregulated genes and 75 upregulated genes) were constructed (Figure 2A), excluding the DEGs which could not constitute a part of a network. With the cut-off criterion set as degrees ≥12, there were 26 genes selected as hub genes, including Quiescin sulfhydryl oxidase-1 (QSOX1), Fibronectin-1 (FN1), Tissue inhibitor of metalloproteinases-1 (TIMP1), C3 complement, Collagen 18A1 (COL18A1), Mesothelin (MSLN) and Collagen 1A1 (COL1A1). Cytoscape was used to obtain the 5 most significant submodules (Figure 2B,C,D,E,F). In these submodules, we found several members of the insulin-like growth factor binding protein (IGFBP) gene family (first submodule), collagen (COL) gene family members (second submodule), and serpin peptidase inhibitors (SERPIN) gene family members (fourth submodule). These results suggested that IGFBP, COL and SERPIN family members play an essential role in the development of GC. Functional enrichment results of the second submodule, which involved collagen gene family members, revealed that the development of GC was associated with ECM organization in a biological process, similar to the GO analysis, platelet-derived growth factor binding in MF, and collagen trimer in the cellular component. Other submodules are detailed in Table S1.

**Up-regulation of COLs in GC patients**

In the GO and KEGG enrichment analysis, several members of the collagen gene family frequently were enriched and, in the PPI network analysis, several COL genes were involved in the second significant submodule. Therefore, COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL5A2, COL6A2, COL6A3, COL8A1, COL17A1 and COL18A1, which were in the
Table 1 The enriched Gene Ontology terms of up-regulated and down-regulated genes

| Ontology ID | Description                                      | P value     | Amount | Main gene                              |
|-------------|--------------------------------------------------|-------------|--------|----------------------------------------|
| GO:0030198  | Extracellular matrix organization                | 7.39E-20    | 23     | COL1A1, COL4A2, COL4A1, OLFML2B, COL3A1|
| GO:0030574  | Collagen catabolic process                       | 4.86E-14    | 13     | COL1A1, COL4A2, COL4A1, COL3A1         |
| GO:0030199  | Collagen fibril organization                     | 1.51E-08    | 8      | COL3A1, COL1A2, FOXC1, COL1A1, GREM1   |
| GO:0071230  | Cellular response to amino acid stimulus         | 0.00043     | 5      | COL4A1, COL3A1, COL1A2, COL1A1, COL5A2 |
| GO:0030168  | Platelet activation                              | 0.0018      | 6      | PDPN, COL3A1, COL1A2, FCER1G, COL1A1   |
| GO:0001568  | Blood vessel development                         | 0.0029      | 4      | SPHK1, COL1A2, COL1A1, COL5A1          |
| GO:0031012  | Extracellular matrix                             | 3.53E-20    | 26     | ASPN, IGFBP7, COL3A1, SERPINE2, APOE   |
| GO:0005581  | Collagen trimer                                   | 4.41E-11    | 12     | COL1A1, CTHRC1, C1QB, COL3A1, COL6A3   |
| GO:0005788  | Endoplasmic reticulum lumen                      | 1.08E-08    | 13     | COL1A1, COL4A2, COL4A1, COL3A1, COL6A3 |
| GO:0005201  | Extracellular matrix structural constituent      | 1.035E-09   | 10     | COL4A2, BGN, COL4A1, COL3A1, COL1A2    |
| GO:0048407  | Platelet-derived growth factor binding           | 7.71E-07    | 5      | COL4A1, COL3A1, COL1A2, COL1A1, COL5A1 |
| GO:0046332  | SMAD binding                                     | 0.0035      | 4      | COL3A1, COL1A2, COL5A2, FLNA           |
| GO:0004252  | Serine-type endopeptidase activity               | 0.0023      | 8      | CTSL, C1QB, BMP1, C3, FAP              |

GO, gene ontology; BP, biological process; CC, cell component; MF, molecular function.

COL gene family and involved in the DEGs, were chosen for more in-depth analysis. To understand better the potential relationship between GC and collagen genes, we used the Oncomine and TCGA-STAD databases to examine the mRNA expression levels of COL genes in normal and gastric tumor tissue. We assessed the expression differences of COLs in 20 cancer samples and their paired normal tissues in the Oncomine database. In these tumor datasets, COL isoforms were significantly upregulated in breast cancer, esophageal cancer, GC, head and neck cancer and pancreatic cancer (Figure 3) compared to matched normal tissues. As the Oncomine and TCGA-STAD databases showed, other COLs were significantly upregulated in tumor tissues (Figures 3, 4), except for
Table 2 The enriched KEGG pathways of up-regulated and down-regulated genes

| Ontology                                         | ID       | Description                              | P value    | Counts | Gene                                      |
|--------------------------------------------------|----------|------------------------------------------|------------|--------|-------------------------------------------|
| The enriched KEGG pathway of upregulated genes    |          |                                          |            |        |                                           |
| KEGGPATHWAY hsa04512                             |          | ECM-receptor interaction                 | 8.11E-13   | 14     | COL4A2, COL4A1, COL3A1, COL5A2, COL5A1   |
| KEGGPATHWAY hsa04510                             |          | Focal adhesion                           | 4.65E-10   | 16     | COL4A2, COL4A1, COL3A1, COL5A2, FLNA     |
| KEGGPATHWAY hsa04974                             |          | Protein digestion and absorption         | 1.08E-07   | 10     | COL18A1, COL4A2, COL4A1, COL3A1, COL6A3  |
| KEGGPATHWAY hsa05146                             |          | Amoebiasis                               | 5.42E-07   | 10     | COL4A2, COL4A1, COL3A1, COL1A2, CXCL8    |
| KEGGPATHWAY hsa04151                             |          | PI3K-Akt signaling pathway               | 1.55E-05   | 14     | COL4A2, COL4A1, COL3A1, COL5A2, COL5A1   |
| KEGGPATHWAY hsa04611                             |          | Platelet activation                      | 1.35E-03   | 7      | COL3A1, COL1A2, FCER1G, FCGR2A, COL1A1   |
| KEGGPATHWAY hsa05133                             |          | Pertussis                                | 6.81E-05   | 7      | C1QB, ITGA5, C3, CXCL8, SERPING1         |
| KEGGPATHWAY hsa04610                             |          | Complement and coagulation cascades      | 4.55E-04   | 6      | C1QB, C3, SERPINE1, SERPING1, C2         |
| KEGGPATHWAY hsa04610                             |          | Complement and coagulation cascades      | 4.55E-04   | 6      | C1QB, C3, SERPINE1, SERPING1, C2         |
| The enriched KEGG pathway of downregulated genes  |          |                                          |            |        |                                           |
| KEGGPATHWAY hsa05204                             |          | Chemical carcinogenesis                   | 1.36E-05   | 9      | CYP2C9, CYP2C18, SULT1A1, ADH1C, ADH1B    |
| KEGGPATHWAY hsa00830                             |          | Retinol metabolism                       | 2.68E-05   | 8      | CYP2C9, CYP2C18, ADH1C, ADH1B, ADH1A     |
| KEGGPATHWAY hsa00010                             |          | Glycolysis / Gluconeogenesis             | 3.63E-05   | 8      | LDHB, ALDOB, ADH1C, ADH1B, ADH1A         |
| KEGGPATHWAY hsa00980                             |          | Metabolism of xenobiotics by cytochrome P450 | 6.94E-05   | 8      | CYP2C9, ADH1C, ADH1B, ADH1A, ADH7        |
| KEGGPATHWAY hsa00982                             |          | Drug metabolism - cytochrome P450         | 3.44E-04   | 7      | CYP2C9, ADH1C, ADH1B, ADH1A, ADH7        |
| KEGGPATHWAY hsa00350                             |          | Tyrosine metabolism                      | 0.00138831 | 5      | ADH1C, ADH1B, ADH1A, ADH7, ADH3A1        |
| KEGGPATHWAY hsa00512                             |          | Mucin type O-Glycan biosynthesis         | 8.40E-04   | 5      | GALNT10, GALNT6, GALNT5, GCNT1, GALNT12  |

hsa, Homo sapiens.

COL6A2 and COL17A1 (data not shown). The details of COL gene expression in all GC datasets in the Oncomine database are shown in Table S2.

**DNA methylation of COL genes in GC patients**

In order to explore the role of methylation in the regulation of COL expression in GC patients, the MethHC method was used to analyze the methylation level of the COL genes promoter regions, and the relationship between DNA methylation level and mRNA expression level. Among the COL members, the methylation levels between normal and cancer samples were statistically different (P<0.05, Figure 5),
except for COL8A1. Notably, DNA methylation of most COLs (10/11) in GC was higher than in the matched normal tissue, except for COL5A2, which was lower than the normal tissue (Figure 5). The relationship between DNA methylation and mRNA expression of COL members in GC are listed in Table S3, although the R values did not prove the relationship between mRNA level and DNA methylation.

Figure 2 Module (A) and 5 submodules (B-F) of protein-protein interaction (PPI) network. Line color indicates the type of interaction evidence. The number of lines between two genes indicate the level of interaction between the two genes.
Weng et al. Collagen genes may affect the prognosis of GC

Table S4

Figure 6C.

Transl Cancer Res 3  7

be used to determine the prognostic value of COL gene
mixed. Therefore, the Kaplan-Meier plotter online tool can
GC is divided into three categories: diffuse, intestinal and
significantly correlated only with increased FP, but was
expression levels led to reduced OS, FP and PPS in GC
High COL1A1, COL1A2, COL4A1, COL4A2, COL5A1, COL6A2, COL6A3, and
COL18A1 led to reduced OS, FP and PPS in intestinal
Prognostic characteristics of COLs in GC patients
Prognostic characteristics of GC patients, including OS,
and diffuse-type GC patients. Additionally, in the mixed-
levels of COL17A1 had no effect on the three subtypes,
levels of COL17A1 significantly correlated with increased
result where the COL17A1 mRNA expression level showed
GC subtype survival time (OS, FP, PPS) with the COLmRNA
isoforms in different GC subtypes. The data showed that
high expression levels of COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL6A2, COL6A3, and COL18A1 led to reduced OS, FP and PPS in intestinal and diffuse-type GC patients. Additionally, in the mixed-type GC patients, most of the COLs were with no-significance because the number of the cases were too small for statistical analysis (Table S4). The different transcript levels of COL17A1 had no effect on the three subtypes, except the OS in intestinal type, which corresponded to the result where the COL17A1 mRNA expression level showed no difference between the normal and tumor tissues. The complex relationship of these GC subtype survival time (OS, FP, PPS) with the COLmRNA expression was shown in the supplementary materials (Figures S1-S3).

mRNA expression of COL3A1 and COL5A1 in different GC cells

Except for COL6A2 and COL17A1, COLs were highly expressed according to the TCGA-STAD databases. We
Figure 4 U aclan database showed that mRNA expression of collagen family genes differed between primary tumor and corresponding normal tissues in gastric cancer patients using (A-J). The blue box represents normal tissue; red box represents tumor tissue. Only P<0.05 was shown.

chose the COL3A1 and COL5A1 genes, which lacked experimental verification in GC but have already been shown to play roles in other cancers (15,16), for RT-qPCR experiments to validate our analysis results. As shown in Figure 7, the COL3A1 level was 1.310×10^6 folds higher in HGC27, 185 folds higher in SGC7901, 96 folds higher in
MKN45 and six folds higher in AGS human GC cell lines, compared with GES-1 normal human gastric mucosal epithelial cell line, and these results were consistent with the analyses from the TCGA-STAD databases. Interestingly, COL5A1 was highly expressed in HGC27, MKN45 and SGC7901, at about 3-7 folds, which were also consistent with the above results. However, in the AGS cell lines, COL5A1 was 400 folds lower than in GES-1. These results require additional in-depth exploration.

**Discussion**

In recent years, significant efforts have been made in order to understand better the early diagnosis, targeted therapy
Collagen genes may affect the prognosis of GC

© Translational Cancer Research. All rights reserved.

Transl Cancer Res. 2020;9(10):6246-6262 | http://dx.doi.org/10.21037/tcr-20-1726
Figure 6 Different mRNA levels of collagen genes prognostic values in gastric cancer patients (Kaplan-Meier plotter). Kaplan-Meier plots show the relationship between OS (A), FP (B) and PPS (C) and the expression of collagens in gastric cancer patients, with hazard ratio (HR) and statistical significance.
and prognosis of GC (17,18). However, the OS of patients with GC remains unimproved, particularly in developing countries (19,20). Our study aimed to identify nuclear genes with similar functions that are highly expressed in GC, compared to normal controls, and to reveal their underlying mechanisms. In the present study, we downloaded the gene expression series of GSE79973, GSE26899, GSE54129 and GSE29272 from the GEO database and found 132 upregulated and 275 downregulated overlap DEGs between GC and normal controls. GO term analysis showed that upregulated DEGs were related primarily with ECM. As reported in previous studies, several ECM-related genes had impacts on the development of GC (21-23). Increased deposition of matrix proteins favors tumor progression by interfering with cell polarity, cell-cell adhesion and, ultimately, amplifying growth factor signaling. As the most significant ECM component (24), collagen determines the functional properties of the matrix and changes in the deposition or degradation of collagen can lead to a decline of ECM homeostasis. It has been reported that increased collagen cross-linking and deposition leads to tumor progression via increased integrin signaling (25). PPI network analysis showed that the IGFBP, SERPIN, and COL gene families were enriched in several submodules. Previous studies have shown that IGFBPs play a protective role in the process of GC development (26-28). However, in our meta-analysis, we found that IGFBP3, IGFBP4, and IGFBP7 were upregulated in GC patients, which was opposed to normal tissues. This might be a self-protection mechanism in GC patients, and additional experiments and analyses are required to investigate this unusual situation. Wang et al. (29), Ju et al. (30), and Yang et al. (31) found that SERPINs can be used as a novel prognostic factor in GC. Additionally, Tian et al. found that SERPINH1 was overexpressed in GC patients and took part in the regulation of EMT (32), which supported the results of our analysis.

Among the identified DEGs, 12 collagen genes were found. Most of these collagen genes with high mRNA and DNA methylation levels excepting COL6A2, COL8A1, COL17A1 and COL5A2, these collagen genes were found to have high mRNA and DNA methylation levels. DNA methylation causes gene silencing. Our results showed, however, high DNA methylation in the promoter region (except for COL5A2), similar to the mRNA levels, in the GC cells. The results showed that methylation in the promoter region did not influence mRNA expression levels COL genes and suggested that methylation may
exist in another region or some other mechanism may have affected mRNA levels. Kaplan-Meier analysis revealed that most of the COLs showed a positive relationship between high expression and significantly worse prognoses in GC patients, which supported the idea that COLs could be prognostic markers in GC patients. Previously, only a few isoforms of COLs involved in GC were reported. Previous studies have demonstrated that upregulated expression of COL1A1 (33), COL1A2 and COL6A3 (34) enhanced the invasive properties of GC cells. COL4A3 was confirmed as a prognostic factor in GC (13). The role of other COLs in GC has not been published (14). Accordingly, additional experimental verification is required to confirm our results and evaluate their meaning. It has been shown that COL3A1 and COL5A1 can be a diagnostic marker in breast cancer and plays a role in non-small cell lung cancer (15,16,35). Therefore, we chose these two COLs for RT-qPCR experiments. After our repeated experiments, data showed that COL3A1 was highly expressed in the four cell lines, and that COL5A1 was highly expressed, in except AGS cells. The differing expression levels between GC cell lines suggested to determine the differences between the cell lines. We found that between these four GC cell lines, HGC27 had the highest degree of malignancy, while AGS was the lowest (36-39).These results are consistent with the expression levels of COL3A1 and COL5A1 in each of the cell lines, which provided a basis for COL3A1 and COL5A1 as markers for the progression and prognosis of GC.

Conclusions

Additional experimentation is required in order to determine whether the COL gene family can be utilized as markers of GC progression and prognosis. Our analysis provides a feasible basis for the idea that COLs may be used as progression and prognosis markers of GC.

Acknowledgments

We thank the contributions of the staff of the Institute of Antibody Engineering, School of Laboratory Medicine and Biotechnology, Southern Medical University in Guangzhou, China.

Funding: This study was supported by the National Natural Science Foundation of China (Grant no. 31672536 and no. 81773271), Guangdong Provincial Education Department of Education of Guangdong Province (2017KZDXM088 and 2018KQNCX284). The Joint Fund of Basic and Applied Basic Research Fund of Guangdong Province (2019A1515110689) and Guangdong Provincial Department of Science and Technology (2019B110233003). The funders had no role in study design, data collection, analysis, decision to publish or preparation of the manuscript.

Footnote

Reporting Checklist: The authors have completed the MDAR checklist. Available at http://dx.doi.org/10.21037/tcr-20-1726

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/tcr-20-1726). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. Zhang H, Wang X, Huang H, et al. Hsa_circ_0067997 promotes the progression of gastric cancer by inhibition of miR-515-5p and activation of X chromosome-linked inhibitor of apoptosis (XIAP). Artif Cells Nanomed Biotechnol 2019;47:308-18.
3. Wang YN, Xu F, Zhang P, et al. MicroRNA-575 regulates development of gastric cancer by targeting PTEN. Biomed
Pharmacother 2019;113:108716.

4. Shah MA, Xu RH, Bang YJ, et al. HELOISE: Phase IIIb Randomized Multicenter Study Comparing Standard-of-Care and Higher-Dose Trastuzumab Regimens Combined With Chemotherapy as First-Line Therapy in Patients With Human Epidermal Growth Factor Receptor 2-Positive Metastatic Gastric or Gastroesophageal Junction Adenocarcinoma. J Clin Oncol 2017;35:2558-67.

5. Sedlazeck FJ, Lee H, Darby CA, et al. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet 2018;19:329-46.

6. Discher DE, Smith L, Cho S, et al. Matrix Mehanosensing: From Scaling Concepts in 'Omithics Data to Mechanisms in the Nucleus, Regeneration, and Cancer. Annu Rev Biophys 2017;46:295-315.

7. Yamauchi M, Barker TH, Gibbons DL, et al. The fibrotic tumor stroma. J Clin Invest 2018;128:16-25.

8. Wörmann SM, Song L, Ai J, et al. Loss of P53 Function Activates JAK2-STAT3 Signaling to Promote Pancreatic Tumor Growth, Stroma Modification, and Gemicatibin Resistance in Mice and Is Associated With Patient Survival. Gastroenterology 2016;151:180-93.e12.

9. Yoshida T, Hashimura M, Kuwata T, et al. Transcriptional regulation of the alpha-1 type II collagen gene by nuclear factor B/p65 and Sox9 in the chondrocytic phenotype of uterine carcinosarcomas. Hum Pathol 2013;44:1780-8.

10. Nagathihalli NS, Castellanos JA, Shi C, et al. Signal Transducer and Activator of Transcription 3, Mediated Remodeling of the Tumor Microenvironment Results in Enhanced Tumor Drug Delivery in a Mouse Model of Pancreatic Cancer. Gastroenterology 2015;149:1932-43.e9.

11. Laklai H, Miroshnikova YA, Pickup MW, et al. Genotype tunes pancreatic ductal adenocarcinoma tissue tension to induce matricellular fibrosis and tumor progression. Nat Med 2016;22:497-505.

12. Miskolczi Z, Smith MP, Rowling EJ, et al. Collagen abundance controls melanoma phenotypes through lineage-specific microenvironment sensing. Oncogene 2018;37:3166-82.

13. Nie XC, Wang JP, Zhu W, et al. COL4A3 expression correlates with pathogenesis, pathologic behaviors, and prognosis of gastric carcinomas. Hum Pathol 2013;44:77-86.

14. Xu S, Xu H, Wang W, et al. The role of collagen in cancer: from bench to bedside. J Transl Med 2019;17:309.

15. Wang Y, Resnick MB, Lu S, et al. Collagen type III alpha1 as a useful diagnostic immunohistochemical marker for fibroepithelial lesions of the breast. Hum Pathol 2016;57:176-81.

16. Souza P, Rizzardi F, Noleto G, et al. Refractory remodeling of the microenvironment by abnormal type V collagen, apoptosis, and immune response in non-small cell lung cancer. Hum Pathol 2010;41:239-48.

17. Badiyan SN, Hallemeier CL, Lin SH, et al. Proton beam therapy for gastrointestinal cancers: past, present, and future. J Gastrointest Oncol 2018;9:962-71.

18. Tan AC, Chan DL, Faisal W, et al. New drug developments in metastatic gastric cancer. Therap Adv Gastroenterol 2018;11:1756284818808072.

19. Goetze OT, Al-Batran SE, Chevallay M, et al. Multimodal treatment in locally advanced gastric cancer. Updates Surg 2018;70:173-9.

20. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. CA Cancer J Clin 2016;66:115-32.

21. Wang K, Wang B, Xing AY, et al. Prognostic significance of SERPINE2 in gastric cancer and its biological function in SGC7901 cells. J Cancer Res Clin
30. Ju H, Lim B, Kim M, et al. SERPINE1 intron polymorphisms affecting gene expression are associated with diffuse-type gastric cancer susceptibility. Cancer 2010;116:4248-55.
31. Yang J, Xiong X, Wang X, et al. Identification of peptide regions of SERPNA1 and ENOSF1 and their protein expression as potential serum biomarkers for gastric cancer. Tumour Biol 2015;36:5109-18.
32. Tian S, Peng P, Li J, et al. SERPINH1 regulates EMT and gastric cancer metastasis via the Wnt/beta-catenin signaling pathway. Aging (Albany NY) 2020;12:3574-93.
33. Shi Y, Duan Z, Zhang X, et al. Down-regulation of the let-7i facilitates gastric cancer invasion and metastasis by targeting COL1A1. Protein Cell 2019;10:143-8.
34. Ao R, Guan L, Wang Y, et al. Silencing of COL1A2, COL6A3, and THBS2 inhibits gastric cancer cell proliferation, migration, and invasion while promoting apoptosis through the PI3k-Akt signaling pathway. J Cell Biochem 2018;119:4420-34.
35. Pan J, Mor G, Ju W, et al. Viral Infection-Induced Differential Expression of LncRNAs Associated with Collagen in Mouse Placentas and Amniotic Sacs. Am J Reprod Immunol 2015;74:237-57.
36. Barranco SC, Townsend CM Jr, Casartelli C, et al. Establishment and characterization of an in vitro model system for human adenocarcinoma of the stomach. Cancer Res 1983;43:1703-9.
37. Lin C, Fu Z, Liu Y, et al. The establishment of human gastric carcinoma cell line (SGC7901). China Academic Journal Electronic Publishing House 1981;1:1-03.
38. Akagi T, Kimoto T. Human cell line (HGC-27) derived from the metastatic lymph node of gastric cancer. Acta Med Okayama 1976;30:215-9.
39. Naito Y, Kino I, Horiuchi K, et al. Promotion of collagen production by human fibroblasts with gastric cancer cells in vitro. Virchows Arch B Cell Pathol Incl Mol Pathol 1984;46:145-54.

Cite this article as: Weng K, Huang Y, Deng H, Wang R, Luo S, Wu H, Chen J, Long M, Hao W. Collagen family genes and related genes might be associated with prognosis of patients with gastric cancer: an integrated bioinformatics analysis and experimental validation. Transl Cancer Res 2020;9(10):6246-6262. doi: 10.21037/tcr-20-1726
### Table S1: Features of module and five submodules of protein-protein interaction (PPI) networks

| Characteristics | Nodes | Edges | Average node degree | Average local clustering coefficient | PPI enrichment P value | Key genes | Functional enrichment |
|-----------------|-------|-------|----------------------|--------------------------------------|------------------------|-----------|----------------------|
| Module          | 365   | 371   | 2.03                 | 0.36                                 | <1.0e-16               | QSOX1, FN1, TIMP1, C3, MSLN | Extracellular structure organization Platelet-derived growth factor binding Endoplasmic reticulum lumen |

**Submodules**

|   |       |     |                     |                        |                       |                        |                                      |
|---|-------|-----|----------------------|------------------------|-----------------------|------------------------|--------------------------------------|
| 1 | 13    | 78  | 12                   | 1                      | <1.0e-16              | IGFBP7, IGFBP3, QSOX1, VCAN, TIMP1 | Post-translational protein modification Insulin-like growth factor binding Endoplasmic reticulum lumen |
| 2 | 13    | 78  | 12                   | 1                      | <1.0e-16              | COL1A1, COL1A2, SERPINH1, COL17A1, COL4A2 | Extracellular matrix organization Platelet-derived growth factor binding Collagen trimer |
| 3 | 6     | 15  | 5                    | 1                      | <1.0e-16              | MGAM, PLAU, SIRPA, FCER1G, CYSTM1 | Neutrophil degranulation |                                      |
| 4 | 5     | 10  | 4                    | 1                      | 2.03E-13              | SERPINE1, SERPING1, THBS1, ISLR, SPARC | Platelet degranulation Extracellular matrix binding Platelet alpha granule lumen |
| 5 | 7     | 15  | 4.29                 | 0.886                  | <1.0e-16              | METTL7A, CHI3L1, LTF, TCN1, TNFAIP6 | Neutrophil degranulation Carbohydrate derivative binding Tertiary granule lumen |

Nodes, the gene numbers in the modules. Edges, the interaction numbers in the modules. PPI enrichment P value indicate that the nodes are not random and that the observed number of edges is significant.
| Genes | Type of Gastric Cancer | Log2FoldChange | P Value       | Log2FoldChange | P Value       | Log2FoldChange | P Value       | Log2FoldChange | P Value       |
|-------|-----------------------|----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 2.29E-07 | 0.002        | 4.619        | 8.902        | 9.017        | 0.002        | 9.076        | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
|      | Diffuse Gastric Adenocarcinoma vs. Normal | 8.89        | -3.494       | 1.65E-07     | 7.514        | 2.57E-11     | 2.118        | 7.68         | 0.002        |
### Table S3 The relationship between DNA methylation and mRNA expression in the collagen gene members of gastric cancer patients (MethHC)

| Gene name | COL1A1 | COL1A2 | COL3A1 | COL4A1 | COL4A2 | COL5A1 | COL5A2 | COL6A2 | COL6A3 | COL8A1 | COL17A1 | COL18A1 |
|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| R value   | -0.0634 | -0.162 | 0.0552 | -0.114 | -0.24  | -0.157 | 0.0779 | -0.128 | 0.0166 | -0.259 | -0.262 | -0.00749 |
| P value   | 0      | 0      | 3.33E-16 | 0      | 0      | 0      | 0.341  | 0      | 3.33E-16 | 0.259  | 0      | 0      |

Notes: P value was analyzed using the survival analysis test. OS, overall survival; PPS, post-progression survival; FP, first progression; HR, hazard ratio.

### Table S4 The prognostic values of collagen isoforms in different subtypes of gastric cancer patients (Kaplan-Meier plotter)

| Collagen family | Lauren classification | OS | P value | PPS | P value |
|-----------------|-----------------------|----|---------|-----|---------|
| COL1A1          | Intestinal            | 320| 2.08    | 1.43–3.02 | 8.70E-05 | 192| 2.27 | 1.39–3.71 | 0.0007 |
|                 | Diffuse               | 241| 1.52    | 1.05–2.21 | 0.026 | 176| 1.86 | 1.26–2.74 | 0.0016 |
|                 | Mixed                 | 32 | 0.62    | 0.2–1.97 | 0.417 | 16 | —   | —       | —      |
| COL1A2          | Intestinal            | 320| 1.79    | 1.31–2.46 | 0.00026 | 192| 1.6   | 1.04–2.47 | 0.032  |
|                 | Diffuse               | 241| 1.71    | 1.21–2.42 | 0.002 | 176| 2.13 | 1.44–3.16 | 0.00011 |
|                 | Mixed                 | 32 | 0.4     | 0.11–1.47 | 0.154 | 16 | —   | —       | —      |
| COL3A1          | Intestinal            | 320| 1.49    | 1.09–2.05 | 0.0123 | 192| 2.02 | 1.34–3.06 | 0.0006 |
|                 | Diffuse               | 241| 1.42    | 1.01–1.99 | 0.045 | 176| 2.02 | 1.38–2.96 | 0.00022 |
|                 | Mixed                 | 32 | 2.82    | 1–7.99   | 0.042 | 16 | —   | —       | —      |
| COL4A1          | Intestinal            | 320| 1.63    | 1.15–2.32 | 0.0056 | 192| 1.76 | 1.15–2.69 | 0.0082 |
|                 | Diffuse               | 241| 2.08    | 1.4–3.1   | 0.00022 | 176| 1.99 | 1.26–3.12 | 0.0024 |
|                 | Mixed                 | 32 | 1.97    | 0.55–7.09 | 0.2922 | 16 | —   | —       | —      |
| COL4A2          | Intestinal            | 320| 2.58    | 1.81–3.67 | 5.60E-08 | 192| 2.69 | 1.91–4.37 | 1.60E-07 |
|                 | Diffuse               | 241| 2.45    | 1.72–3.49 | 3.40E-07 | 176| 2.63 | 1.77–3.89 | 5.60E-07 |
|                 | Mixed                 | 32 | 2.2     | 0.49–9.93 | 0.29 | 16 | —   | —       | —      |
| COL5A1          | Intestinal            | 320| 2.55    | 1.85–3.5  | 2.50E-09 | 192| 2.79 | 1.85–4.22 | 3.80E-07 |
|                 | Diffuse               | 241| 1.86    | 1.31–2.64 | 0.00043 | 176| 2.46 | 1.67–3.67 | 2.60E-06 |
|                 | Mixed                 | 32 | 3.83    | 1.35–10.87 | 0.0067 | 16 | —   | —       | —      |
| COL5A2          | Intestinal            | 320| 1.49    | 1.07–2.08 | 0.0167 | 192| 1.34 | 0.89–2.02 | 0.1596 |
|                 | Diffuse               | 241| 1.26    | 0.89–1.79 | 0.1984 | 176| 1.54 | 1.03–2.32 | 0.036  |
|                 | Mixed                 | 32 | 2.45    | 0.84–7.12 | 0.0886 | 16 | —   | —       | —      |
| COL6A2          | Intestinal            | 320| 2.78    | 1.98–3.91 | 8.10E-10 | 192| 3.54 | 2.33–5.39 | 3.50E-10 |
|                 | Diffuse               | 241| 2.13    | 1.48–3.06 | 3.10E-05 | 176| 2.51 | 1.7–3.72  | 2.00E-06 |
|                 | Mixed                 | 32 | 3.58    | 1.26–10.22 | 0.0111 | 16 | —   | —       | —      |
| COL6A3          | Intestinal            | 320| 2.26    | 1.65–3.1  | 1.90E-07 | 192| 2.51 | 1.67–3.79 | 5.40E-06 |
|                 | Diffuse               | 241| 1.43    | 1–2.03    | 0.048 | 176| 1.92 | 1.29–2.86 | 0.0011 |
|                 | Mixed                 | 32 | 1.77    | 0.62–5.03 | 0.279 | 16 | —   | —       | —      |
| COL8A1          | Intestinal            | 320| 1.94    | 1.38–2.72 | 8.80E-05 | 192| 2.19 | 1.43–3.35 | 0.0002 |
|                 | Diffuse               | 241| 1.31    | 0.93–1.86 | 0.13 | 176| 2.21 | 1.5–3.25  | 3.90E-05 |
|                 | Mixed                 | 32 | 0.39    | 0.13–1.25 | 0.1012 | 16 | —   | —       | —      |
| COL17A1         | Intestinal            | 320| 0.69    | 0.5–0.94  | 0.0181 | 192| 1.27 | 0.83–1.95 | 0.2725 |
|                 | Diffuse               | 241| 0.72    | 0.47–1.09 | 0.1189 | 176| 1.28 | 0.83–1.98 | 0.26   |
|                 | Mixed                 | 32 | 2.84    | 0.94–8.53 | 0.053 | 16 | —   | —       | —      |
| COL18A1         | Intestinal            | 320| 2.76    | 2.01–3.8  | 7.90E-11 | 192| 3.14 | 2.08–4.76 | 1.20E-08 |
|                 | Diffuse               | 241| 1.76    | 1.22–2.53 | 0.002 | 176| 1.97 | 1.34–2.89 | 0.00042 |
|                 | Mixed                 | 32 | 2.59    | 0.81–8.31 | 0.0967 | 16 | —   | —       | —      |

Notes: P value was analyzed using the survival analysis test. OS, overall survival; PPS, post-progression survival; FP, first progression; HR, hazard ratio.
Figure S1: Different mRNA levels of collagens' prognostic values in diffuse subtype gastric cancer patients (Kaplan-Meier plots). Notes: Kaplan-Meier plots show the relationship between OS (A), PFS (B), and DFS (C) and the expression of collagen in gastric cancer patients, respectively, with hazard ratio (HR) and statistical significance.
Figure S2 Different mRNA level of collagens' prognostic values in intestinal subtype gastric cancer patients (Kaplan-Meier plots). Notes: Kaplan-Meier plots show the relationship between OS (A), EFS (B) and FPS (C) and the expression of collagens in gastric cancer patients, respectively, with hazard ratio (HR) and statistical significance.
Figure S3 Different mRNA level of collagens' prognostic values in mixed subtype gastric cancer patients (Kaplan-Meier plotter). Notes: Kaplan-Meier plots show the relationship between OS (A), FP (B) and PPS (C) and the expression of collagens in gastric cancer patients, respectively, with hazard ratio (HR) and statistical significance.