Contents lists available at ScienceDirect

Pathology - Research and Practice

journal homepage: www.elsevier.com/locate/prp

Integrated bioinformatics analysis reveals novel key biomarkers and potential candidate small molecule drugs in gastric cancer

Qiong Wu[a,b,1], Bo Zhang[a,b,1], Ziheng Wang[c,1], Xinyi Hu[c], Yidan Sun[d], Ran Xu[a], Xinming Chen[e], Qiuhong Wang[b], Fei Ju[b], Shiqi Ren[c], Chenlin Zhang[f], Fuwei Qi[g], Qianqian Ma[h], Qun Xue[e,*], You Lang Zhou[b,*]

[a] Medical School of Nantong University, Nantong 226001, PR China
[b] The Hand Surgery Research Center, Department of Hand Surgery, Affiliated Hospital of Nantong University, Nantong 226001, PR China
[c] Department of Medicine, Nantong University Xinling College, Nantong, Jiangsu, 226001, PR China
[d] Department of Oncology, First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin 300193, PR China
[e] Department of Thoracic Surgery, Affiliated Hospital of Nantong University, Nantong, 226001, PR China
[f] Department of Spine, Chinese Medicine Hospital,Wuxi Hospital Affiliated to Nanjing University of Chinese Medicine, Wuxi 214000, PR China
[g] Department of anesthesiology, The First people's Hospital of Taicang City, Taicang Affiliated Hospital of Soochow University, Suzhou 215400, P.R. China
[h] Emergency Office, Wuxi Center for Disease Control and Prevention, Wuxi, Jiangsu, 214023, PR China

ARTICLE INFO

Keywords:
Gastric cancer
Differentially expressed genes
Novel biomarkers
Candidate small molecules
Bioinformatics analysis

ABSTRACT

Background and objective: The underlying molecular mechanisms of gastric cancer (GC) have yet not been investigated clearly. In this study, we aimed to identify hub genes involved in the pathogenesis and prognosis of GC.

Methods: We integrated five microarray datasets from Gene Expression Omnibus (GEO) database. The differentially expressed genes (DEGs) between GC and normal samples were analyzed with limma package. Gene ontology (GO) and KEGG enrichment analysis were performed using DAVID. Then we established the protein-protein interaction (PPI) network of DEGs by the Search Tool for the Retrieval of Interacting Genes database (STRING). The prognostic analysis of hub genes were performed through Gene Expression Profiling Interactive Analysis (GEPIA). Additionally, we used real-time quantitative PCR to validate the expression of hub genes in 5 pairs of tumor tissues and corresponding adjacent tissues. Finally, the candidate small molecules as potential drugs to treat GC were predicted in CMap database.

Results: Through integrating five microarray datasets, a total of 172 overlap DEGs were detected including 79 up-regulated and 93 down-regulated genes. Biological process analysis of functional enrichment showed these DEGs were mainly enriched in digestion, collagen fibril organization and cell adhesion. Signaling pathway analysis indicated that these DEGs played an vital in ECM-receptor interaction, focal adhesion and metabolism of xenobiotics by cytochrome P450. Protein-protein interaction network among the overlap DEGs was established with 124 nodes and 365 interactions. Three DEGs with high degree of connectivity (NID2, COL4A1 and COL4A2) were selected as hub genes. The GEPIA database confirmed that overexpression levels of hub genes were significantly associated with worse survival of patients. Finally, the 20 most significant small molecules were obtained based on CMap database and spiradoline was the most promising small molecule to reverse the GC gene expression.

Conclusions: Our results indicated that NID2, COL4A1 and COL4A2 could be the potential novel biomarkers for GC diagnosis prognosis and the promising therapeutic targets. The present study may be crucial to understanding the molecular mechanism of GC initiation and progression.

[*] Corresponding authors.
E-mail addresses: qunyuemi@sina.com (Q. Xue), zhouyoulang@ntu.edu.cn (Y.L. Zhou).
[1] These authors contributed equally to this study.

https://doi.org/10.1016/j.prp.2019.02.012
Received 3 December 2018; Received in revised form 16 February 2019; Accepted 26 February 2019

1. Introduction

Gastric cancer (GC) is the fifth most frequently cancer and the third leading cause of cancer death, with an estimated 1,000,000 new cases and 783,000 deaths in 2018. Although several therapeutic strategies have been developed for gastric cancer including chemotherapy, targeted therapy, surgery and radiotherapy, the patients' prognosis still remains far from ideal, with 5-year survival below 20%. The main reason for poor 5-year survival of GC is that most patients are initially diagnosed at advanced stage [1–3]. Thus, identifying the promising novel diagnostic and prognostic biomarkers of GC is urgently demanded. In recent years, the rapid development of high throughput sequencing and microarray technologies have made a great contribution to discovering novel biomarkers associated with tumor initiation, progression, diagnosis and prognosis, and has resulted in the development of a variety of new targeted drugs [4,5]. Trastuzumab, a human antiepidermal growth factor receptor 2 (HER2) antibody, combined with chemotherapy has demonstrated a significant survival benefit in HER2 + GC patients compared to chemotherapy monotherapy [6] Cetuximab and ramucirumab were also the targeted drugs for GC approved by Food and Drug Administration (FDA) [6–8]. Gene Expression Omnibus (GEO) is a public repository for archiving high-throughput microarray experimental data. GEO database has provided a powerful platform for bioinformatics to explore novel biomarkers for cancer diagnosis, treatment and prognosis analysis [9,10]. In this study, we performed an integrated analysis of five microarray datasets (GSE13911, GSE19826, GSE29272, GSE54129 and GSE79973) from GEO database and identified the differentially expressed genes (DEGs) between GC and normal tissues. Then we conducted functional and pathway enrichment to further explore biological functions of these DEGs in GC. The hub genes correlated with the pathogenesis of GC were selected by PPI module analysis and prognosis analysis. Finally, the CMap database was used to predict small molecules targeting the gene expression of GC. Fig. 1 showed the workflow of our study.

2. Materials and methods

2.1. Microarray data

The microarray datasets of GSE13911, GSE19826, GSE29272, GSE54129 and GSE79973 database were downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/). These RNA profiles were based on GPL570 platforms (Affymetrix Human Genome U113 Plus 2.0 Array) and GPL6947 platforms (Illumina HumanHT-12 V3.0 expression beadchip), containing a total of 305 gastric cancer samples and 210 normal samples. The GSE13911 profile included 38 tumor samples and 31 normal samples, GSE19826 profile contained 12 tumor samples and 14 normal samples, GSE29272 profile provided 134 tumor samples and 134 normal samples, GSE54129 profile consisted of 111 tumor samples and 21 normal samples and GSE79973 profile included 10 tumor samples and 10 normal samples.

2.2. Identification of DEGs

After the data standardization, the limma package in R software was applied to screen DEGs between tumor and normal tissues in each datasets [11,12]. The genes with P value < 0.05 and $|\log FC| > 1$ were considered as DEGs. The hierarchical clustering of DEGs was established by UCSC Cancer Genomics Browser (http://genome-cancer.ucsc.edu).

2.3. Gene ontology and pathway enrichment analysis

To depict the potential biological functions of the overlapping DEGs, we performed Gene ontology (GO) enrichment analysis based on three aspects including biological process (BP), molecular function (MF), and cellular component (CC) [13,14]. Then we conducted Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis to investigate the potential signaling pathways related to the overlapping DEGs. KEGG database stores a large number of information about genomes, biological pathways, diseases, chemical substances and drugs and is widely used for identifying functional and metabolic pathways [15]. GO and KEGG pathway enrichment analysis were carried out by DAVID (Database for Annotation Visualization and Integrated Discovery), a reliable program providing users with integrated functional annotation information of genes and proteins [16].

2.4. Protein-protein interaction (PPI) network construction and module analysis

The potential interactions of the overlapping DEGs were analyzed using the online tool STRING (Search Tool for the Retrieval of Interacting Genes, https://string-db.org/). The interactions with a combined score > 0.4 were extracted for the construction of PPI network. Subsequently, the PPI network was visualized and further analyzed using Cytoscape software (www.cytoscape.org/) [17]. The MCODE (Molecular Complex Detection) plug-in of the Cytoscape software was utilized to detect significant modules from the PPI network. The degree cutoff = 2, node score cutoff = 0.2, k-core = 2, and max. depth = 100 were set as cutoff criteria. The biological process of the module DEGs was analyzed and visualized by the Networks Gene Oncology tool (BiNGO) plugin of Cytoscape [18,19].

2.5. Analysis and validation of hub genes

The cBioPortal online platform (http://www.cbioportal.org) was used to construct a network of DEGs and their co-expression genes [20,21]. The DEGs with high degree of connectivity in the modules were selected as hub genes. In order to further confirm the reliability of hub genes from the results of bioinformatics analysis, we used the Gene Expression Profiling Interactive Analysis (GEPIA) database to analyze their expression and prognostic value [22]. GC patients were divided into high expression and low expression groups. The expression values of hub genes and patients' prognosis were visualized through the Kaplan-Meier curve and boxplot. We also analyzed the protein expression of the hub genes between GC and normal tissues using the human protein atlas (HPA, www.proteinatlas.org) database. HPA database is an online tool widely used for determining the protein level of genes in clinical samples.

2.6. Identification of small molecules

We queried the Connectivity Map (CMap, http://www.broadinstitute.org/cmap/) to detect the candidate small molecule drugs for use in patients based on the gene signature of GC. CMap is a collection of databases that stores thousands of gene transcription-expression profiles from cultured mammalian cells exposed to active small molecules [23]. First, these overlapping DEGs were divided into up-regulated and down-regulated groups. Then the probesets from each group were utilized to query the CMap database. Finally, the enrichment scores ranging from -1 to +1 were calculated, which represented the similarity. A positive connectivity value (closer to +1) demonstrated that a small molecule is able to induce the gene expression of GC cells, whereas a negative connectivity value (closer to -1) demonstrated that a small molecule is able to imitate the status of normal cells.

2.7. Real-time quantitative PCR

Total RNA was extracted from tumor tissues and non-tumorous tissues using Trizol (Invitrogen, Carlsbad, CA). 2 μg of total RNA was subjected to DNaseI digestion (Fermentas, MD, USA) at 37 °C for 0.5 h. cDNA was synthesized using a Omniscript Reverse Transcription kit

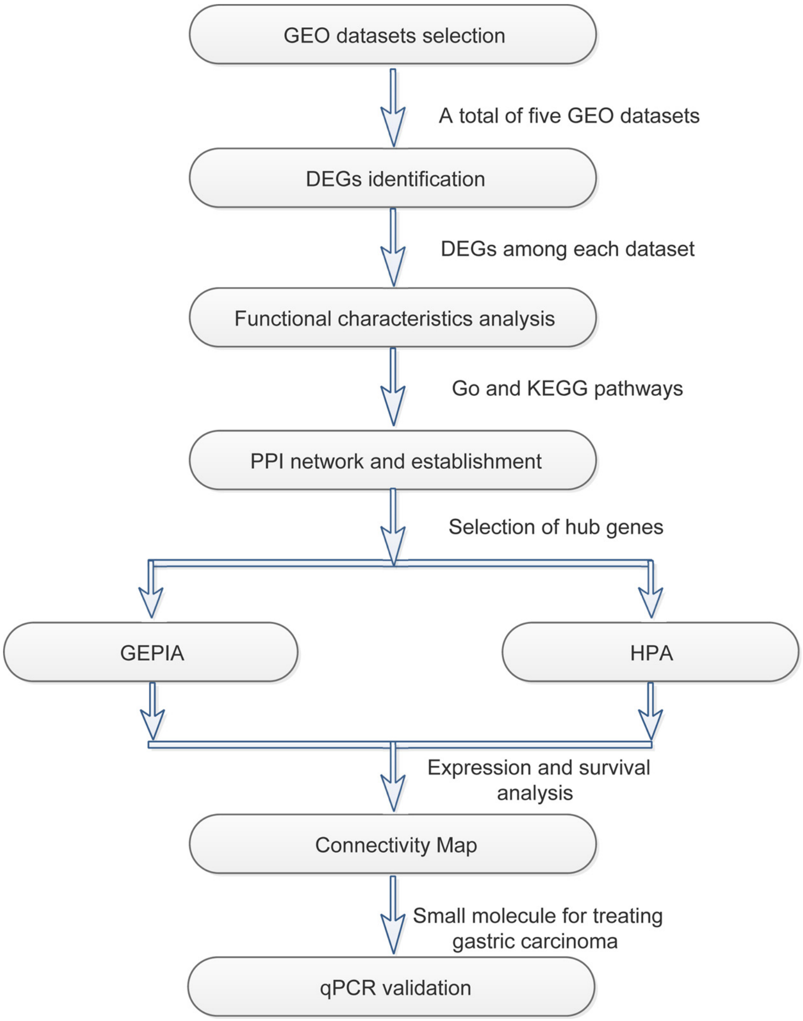

GEO datasets selection
A total of five GEO datasets
DEGs identification
DEGs among each dataset
Functional characteristics analysis
Go and KEGG pathways
PPI network and establishment
Selection of hub genes
GEPIA
HPA
Expression and survival analysis
Connectivity Map
Small molecule for treating gastric carcinoma
qPCR validation

Fig. 1. The workflow of the present study. Abbreviations:
GEO: Gene Expression Omnibus; DEGs: differentially expressed genes; GO: Gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; PPI: protein-protein interaction; GEPIA: Gene Expression Profiling Interactive Analysis; HPA: human protein atlas.

(Qiagen, Valencia, CA). mRNA level was tested using EvaGreen Master Mix (Biotium Inc.,Hayward, CA). Relative expression level for each target gene was normalized by the Ct value of GAPDH (internal control) using a $2^{-\Delta\Delta Ct}$ relative quantification method. The primers were as follows: NID2 gene 5'- CTACTGCCCAACAGGAAGAAA-3' (sense) and 5'-TGCAGTCACTGTTCTTTAGGG-3' (anti-sense). COL4A1 gene 5'- CTAA TGTCACAACATGGTGCTAC-3' (sense) and 5'- GCAGGGTGTGTTAGTT ACGC-3' (anti-sense). COL4A2 gene 5'- TGGGACAGACGAGACAACAG-3' (sense) and 5'- CAACGGTATTTGGGAGAACAT-3' (anti-sense). All reactions were performed on the Eppendorf Mastercycler ep realplex (2S; Eppendorf, Hamburg, Germany) using following cycling parameters, 95 °C for 2 min, followed by 40 cycles of 95 °C for 15 s, 60 °C for 45 s.

3. Results

3.1. Identification of DEGs in GC

Using P value < 0.05 and |logFC| > 1 as cut-off criteria, we detected a total of 172 overlapping DEGs after performing integrated analysis for five microarray datasets, of which 79 were markedly up-regulated and 93 were down-regulated. The volcano plot showed the up-regulated and down-regulated genes in each dataset. The detailed records of Venn diagram among the five datasets were presented in Fig. 2B. The heat map indicted that the genes exhibited significant difference between GC tissues and the noncancerous tissues (Fig. 2C). We used cBioPortal platform to construct a network of the overlapping DEGs and their co-expression genes (Fig. 2D).

3.2. GO term and KEGG pathway enrichment analyses of DEGs

Functional and pathway enrichment analyses for DEGs were performed using DAVID.

GO terms were divided into biological processes, cell component and molecular function. For biological processes, GO analysis results indicated that these overlap DEGs were significantly enriched in digestion, collagen fibril organization, cell adhesion, biological adhesion and skeletal system development. Molecular function analysis showed that the DEGs were particularly enriched in extracellular matrix structural constituent, structural molecule activity, platelet-derived growth factor binding, aldo-keto reductase activity and glycosaminoglycan binding. Similarly, changes in cell component of overlap DEGs were significantly enriched in extracellular region, extracellular region part, extracellular matrix, proteinaceous extracellular matrix and extracellular matrix part. Additionally, the results of KEGG pathway analysis revealed that these DEGs were mainly involved in ECM-receptor interaction, focal adhesion, metabolism of xenobiotics by cytochrome P450, drug metabolism and retinol metabolism (Fig. 3A &

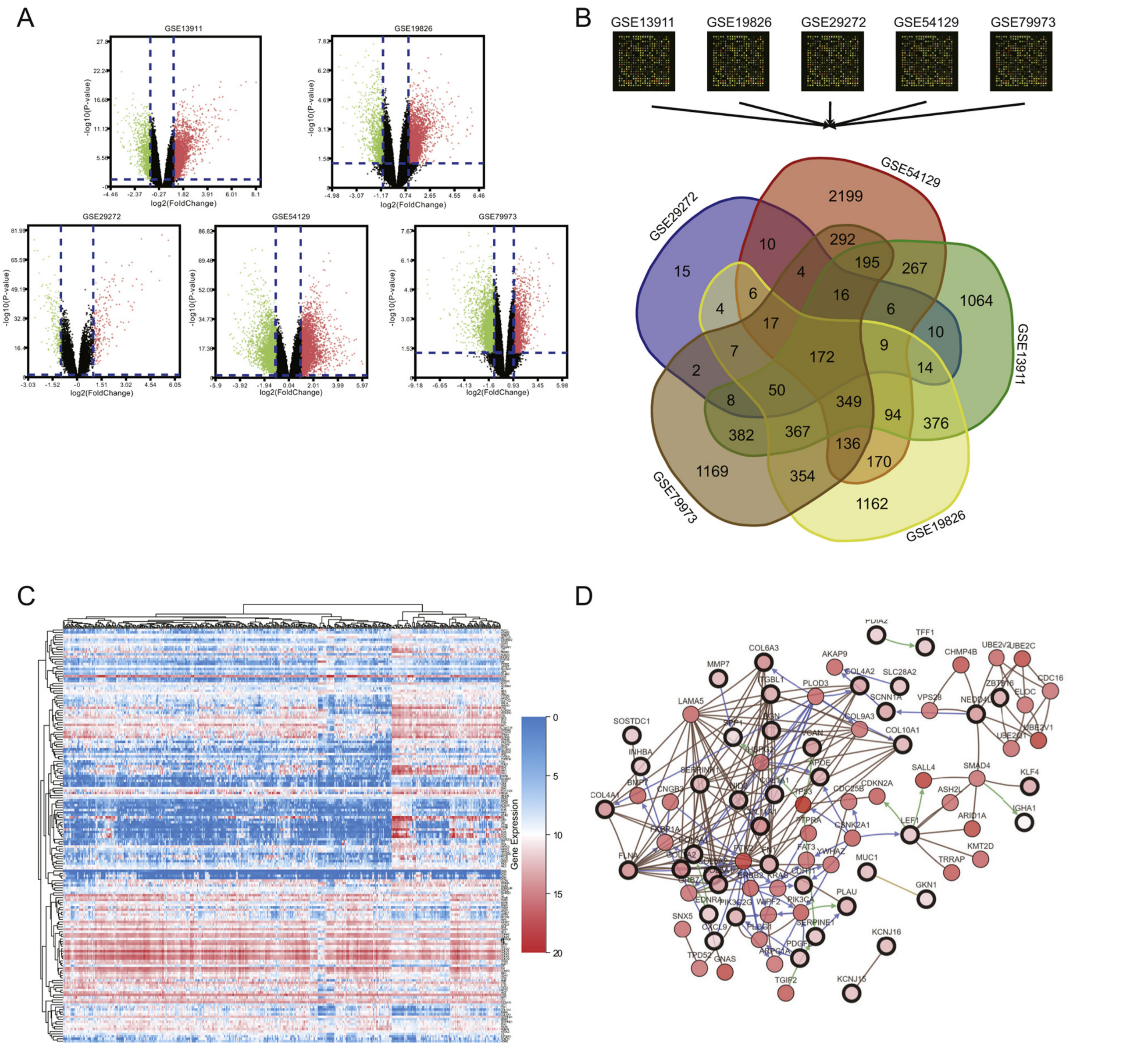


Fig. 2. **(A)** Volcano plot of gene expression profile data between gastric cancer and normal tissues in each dataset. Red dots: significantly up-regulated genes in GC; Green dots: significantly down-regulated genes in GC; Black dots: non-differentially expressed genes. $P < 0.05$ and $|\log2\ FC| > 1$ were considered as significant. **(B)** The Venn diagram of 172 overlapping DEGs among five datasets. **(C)** The expression heat map of overlapping DEGs. **(D)** A network of the overlap DEGs and their co-expression genes. Nodes with thick outline: hub genes; Nodes with thin outline: co-expression genes. Abbreviations:
GC: gastric cancer; DEGs: differentially expressed genes.

Table 1).

3.3. PPI network construction and selection of hub genes from significant modules

A PPI network from the STRING database was established to predict the potential interactions of these DEGs at the protein level. In total, 124 nodes and 365 interactions were obtained using the Cytoscape software (Fig. 3B). Then, the two most important modules were selected from PPI network using MCODE, one consisting of up-regulated genes and the other consisting of down-regulated genes. The biological process of the genes in these two modules was significantly associated with the process of response to stimulus, response to chemical stimulus and response to stress (Fig. 3C). Enrichment analysis suggested that the down-regulated module were mainly enriched in polyketide metabolic process, daunorubicin metabolic process and doxorubicin metabolic process, while the up-regulated module were enriched in laminin interactions, naba basement membranes and naba core matrisome (Table 2). NID2, COL4A1 and COl4A2, which existed a high degree of connectivity in the modules, were chosen as hub genes. The expression of NID2, COL4A1 and COL4A2 in GC tissues was significantly up-regulated compared to adjacent normal samples (Fig. 4). The detailed information of these hub genes was shown in Table S1.

3.4. Validation and survival analysis of hub genes

We used HPA database and the cBioPortal for Cancer Genomics database to confirm the important role of hub genes in the GC initiation and progression. The hub gene median expression of tumor and normal samples in bodymap was shown in Fig. 5A and Fig. S1&S2, respectively. When exploring the expression and survival value of hub genes in the GEPIA online database, NID2, COL4A1 and COL4A2 were significantly differentially expressed between tumor and normal tissues (Fig. 5B), and markedly correlated with the overall survival of GC patients

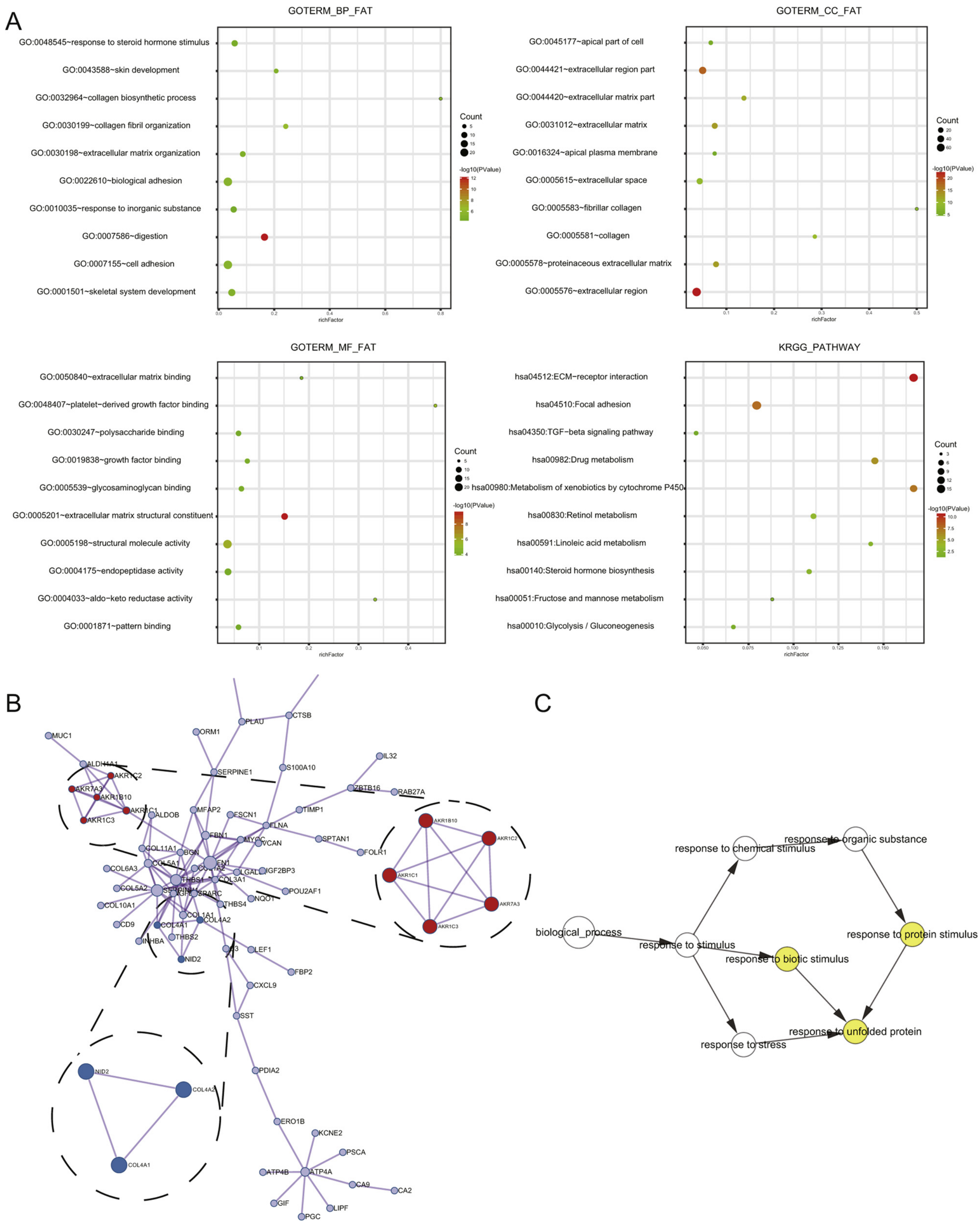


Fig. 3. (A) Functional and pathway enrichment analysis of 172 overlap DEGs. **(B)** The protein-protein interaction network of the 172 overlap genes. **(C)** The biological process analysis of DEGs constructed by BiNGO. The color depth of nodes represents the corrected P-value. The size of nodes represents the number of genes involved. Abbreviations:
DEGs: differentially expressed genes.

(Fig. 5C). Considering that gene expression was not always consistent with its protein level [24], we further analyzed the protein level of NID2, COL4A1 and COL4A2 in clinical GC tissues from HPA database. The results of immunohistochemical indicated that the protein expression level of NID2, COL4A1 and COL4A2 exhibited significant difference between GC tissues and normal tissues (Fig. 5D). We also

Table 1
Functional and pathway enrichment analysis of the overlap DEGs.

Category	Term	Count	PValue
GOTERM_BP_FAT	GO:0007586~digestion	15	6.70E-13
GOTERM_BP_FAT	GO:0030199~collagen fibril organization	7	5.09E-07
GOTERM_BP_FAT	GO:0007155~cell adhesion	23	4.64E-06
GOTERM_BP_FAT	GO:0022610~biological adhesion	23	4.75E-06
GOTERM_BP_FAT	GO:0001501~skeletal system development	15	7.71E-06
GOTERM_BP_FAT	GO:0032964~collagen biosynthetic process	4	1.16E-05
GOTERM_BP_FAT	GO:0043588~skin development	6	1.23E-05
GOTERM_BP_FAT	GO:0030198~extracellular matrix organization	9	1.48E-05
GOTERM_BP_FAT	GO:0048545~response to steroid hormone stimulus	11	3.88E-05
GOTERM_BP_FAT	GO:0010035~response to inorganic substance	11	6.75E-05
GOTERM_CC_FAT	GO:0005576~extracellular region	75	2.48E-23
GOTERM_CC_FAT	GO:0044421~extracellular region part	48	1.09E-18
GOTERM_CC_FAT	GO:0031012~extracellular matrix	26	1.04E-13
GOTERM_CC_FAT	GO:0005578~proteinaceous extracellular matrix	25	1.62E-13
GOTERM_CC_FAT	GO:0044420~extracellular matrix part	16	3.74E-12
GOTERM_CC_FAT	GO:0005581~collagen	10	1.33E-10
GOTERM_CC_FAT	GO:0005615~extracellular space	30	5.68E-10
GOTERM_CC_FAT	GO:0005583~fibrillar collagen	6	1.30E-07
GOTERM_CC_FAT	GO:0045177~apical part of cell	12	5.95E-06
GOTERM_CC_FAT	GO:0016324~apical plasma membrane	10	2.00E-05
GOTERM_MF_FAT	GO:0005201~extracellular matrix structural constituent	13	1.52E-10
GOTERM_MF_FAT	GO:0005198~structural molecule activity	23	1.59E-06
GOTERM_MF_FAT	GO:0048407~platelet-derived growth factor binding	5	4.39E-06
GOTERM_MF_FAT	GO:0004033~aldo-keto reductase activity	5	1.75E-05
GOTERM_MF_FAT	GO:0005539~glycosaminoglycan binding	9	1.57E-04
GOTERM_MF_FAT	GO:0019838~growth factor binding	8	1.57E-04
GOTERM_MF_FAT	GO:0050840~extracellular matrix binding	5	2.03E-04
GOTERM_MF_FAT	GO:0004175~endopeptidase activity	14	2.55E-04
GOTERM_MF_FAT	GO:0001871~pattern binding	9	3.00E-04
GOTERM_MF_FAT	GO:0030247~polysaccharide binding	9	3.00E-04
KEGG_PATHWAY	hsa04512:ECM-receptor interaction	14	2.16E-11
KEGG_PATHWAY	hsa04510:Focal adhesion	16	1.93E-08
KEGG_PATHWAY	hsa00980:Metabolism of xenobiotics by cytochrome P450	10	4.56E-08
KEGG_PATHWAY	hsa00982:Drug metabolism	9	9.00E-07
KEGG_PATHWAY	hsa00830:Retinol metabolism	6	5.70E-04
KEGG_PATHWAY	hsa00140:Steroid hormone biosynthesis	5	0.002653
KEGG_PATHWAY	hsa00591:Linoleic acid metabolism	4	0.005199
KEGG_PATHWAY	hsa00010:Glycolysis / Gluconeogenesis	4	0.040709
KEGG_PATHWAY	hsa00051:Fructose and mannose metabolism	3	0.069486
KEGG_PATHWAY	hsa04350:TGF-beta signaling pathway	4	0.099275

Table 2
Gene ontology and pathway enrichment analysis of Module 1 and Module 2.

MCODE	GO	Description	Log10(P)
MCODE_1	GO:0030638	polyketide metabolic process	−13.5
MCODE_1	GO:0044597	daunorubicin metabolic process	−13.5
MCODE_1	GO:0044598	doxorubicin metabolic process	−13.5
MCODE_2	R-HSA-3000157	laminin interactions	−8.7
MCODE_2	M5887	naba basement membranes	−8.3
MCODE_2	M5884	naba core matrisome	−5.8

predicted potential transcription associated with NID2, COL4A1 and COL4A2 (Fig. 5E), and established a regulatory network of lncRNA, miRNA and mRNA by GCBI analysis (Gene-Cloud Biotechnology Information, Fig. 5F).

3.5. Identification of related active small molecules

To screen out candidate small molecule drugs, CMap database was utilized to analyze consistent differently expressed probesets between GC tissues and normal tissues. Table 3 and Fig. 5G showed the predicted small molecules that could inhibit GC-associated gene expression. The enrichment scores and p value were also listed. Among these small molecules, spiradoline (enrichment score = −0.821) and loracarbef (enrichment score = −0.746) showed higher negative correlation between the gene expression of GC and the small molecules, which could be the most promising small molecules to reverse the tumoral status of GC. The prediction of small molecule drugs aim to exploit the potential drugs targeting GC and make existing drugs fully utilized. However, further studies were urgently needed to validate the above results.

3.6. Evaluation of NID2, COL4A1 and COL4A2 expression in GC

In order to verify the expression of NID2, COL4A1 and COL4A2 genes in tumor tissues and corresponding adjacent non-tumorous tissues, we selected 5 pairs of tumor tissues and corresponding adjacent tissues. The present study was performed with the approval of the institutional ethics board of the Affiliated Hospital of Nantong University. Relative expression of NID2, COL4A1 and COL4A2 mRNA in GC and adjacent non-tumorous tissues were quantified by qPCR. The results showed that the average NID2, COL4A1 and COL4A2 mRNA expression level was significantly higher in GC tissues compared with non-tumorous tissues ($p = 0.028$, $p = 0.02$ and $p = 0.005$, respectively, Fig. 5H).

4. Discussion

The failure to early screening and diagnosis results in the poor long-term survival of GC patients. Therefore, identifying sensitive and specific biomarkers for GC treatment is urgently necessary. The rapid development of high-throughput sequencing and bioinformatics technologies has brought hope to decipher the pathogenesis of GC. For example, Cao et al. identified hub genes and potential molecular mechanisms associated with progression of GC by integrated bioinformatics analysis including DEGs selection, enrichment analysis, PPI network construction, survival analysis and reverse transcription-quantitative polymerase chain reaction validation [25]. Ting et al integrated three microarray datasets to identify key genes associated with patients' prognosis using a bioinformatics approach consisting of DEGs selection, enrichment analysis, survival analysis and the prediction of transcription factors and potential stem loop miRNAs [26]. Furthermore, Ru et al. revealed COL4A1 may be involved in the trastuzumab resistance in gastric cancer based on bioinformatics analysis [27].

In the present study, we performed a comprehensive bioinformatics analysis of GC based on a larger sample size by mining the GEO database. A total of 305 gastric cancer samples and 210 normal samples were available for DEGs screen. Using limma package, the five gene expression profiles revealed a total of 172 DEGs, in which 79 genes were significantly up-regulated and 93 genes were down-regulated. These DEGs could play an important role in regulating the occurrence and development of GC and may be potential targets for treating GC. To get a more in-depth understanding of biological function related to these DEGs, we conducted the functional and pathway enrichment analysis by using DAVID. Digestion, collagen fibril organization and cell adhesion were the top three significantly major functions among the biological process of these DEGs. Molecular function enriched for the overlap DEGs were mainly within extracellular matrix structural constituent, structural molecule activity and platelet-derived growth factor binding. Changes in cell component of were mainly associated with extracellular region, extracellular region part and extracellular matrix. Additionally, the overlap DEGs were enriched in 10 KEGG pathways, mainly including ECM-receptor interaction, focal adhesion, metabolism

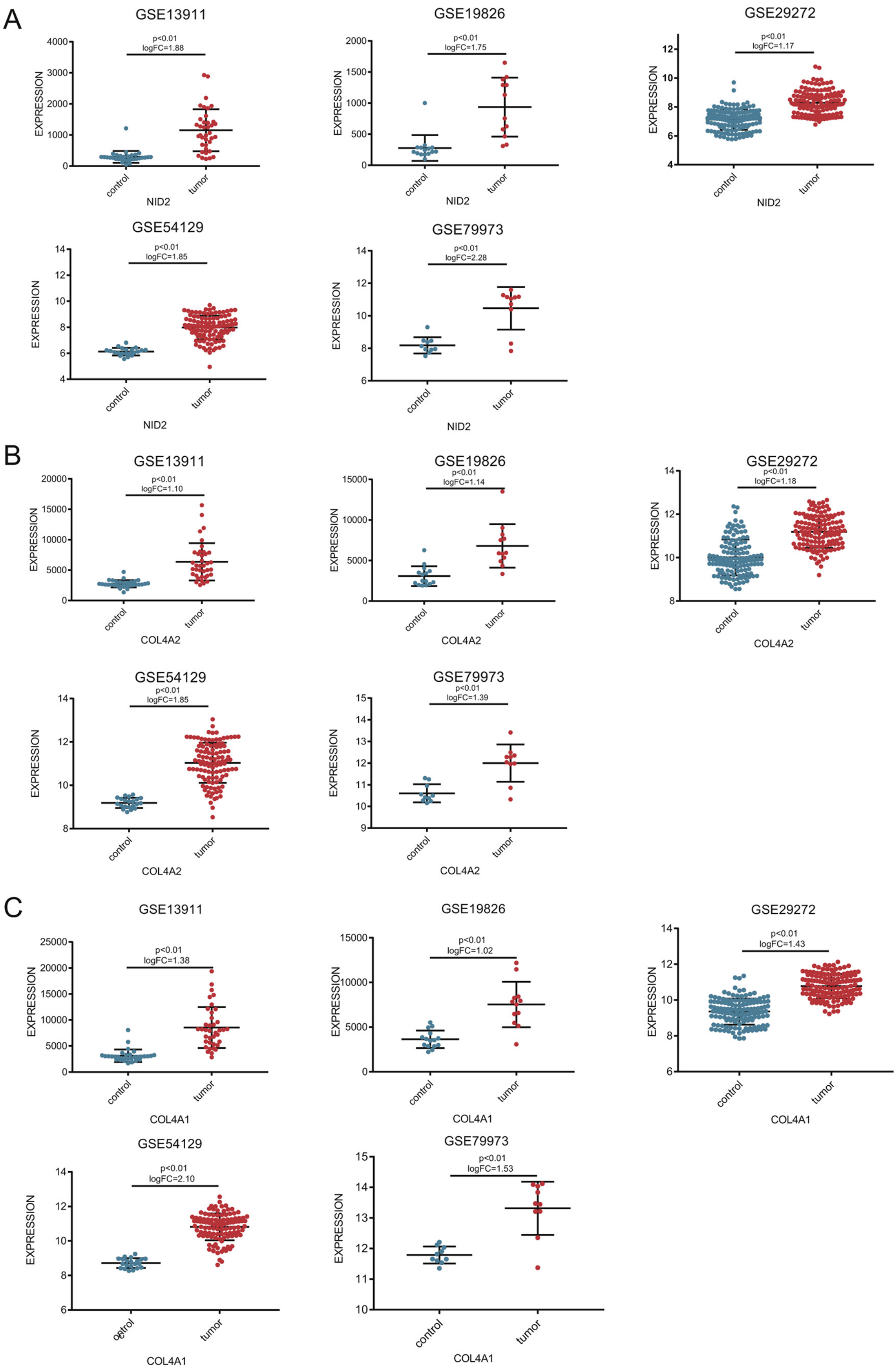


Fig. 4. The expression level of NID2, COL4A1, COL4A2 between GC and normal tissues in five datasets.

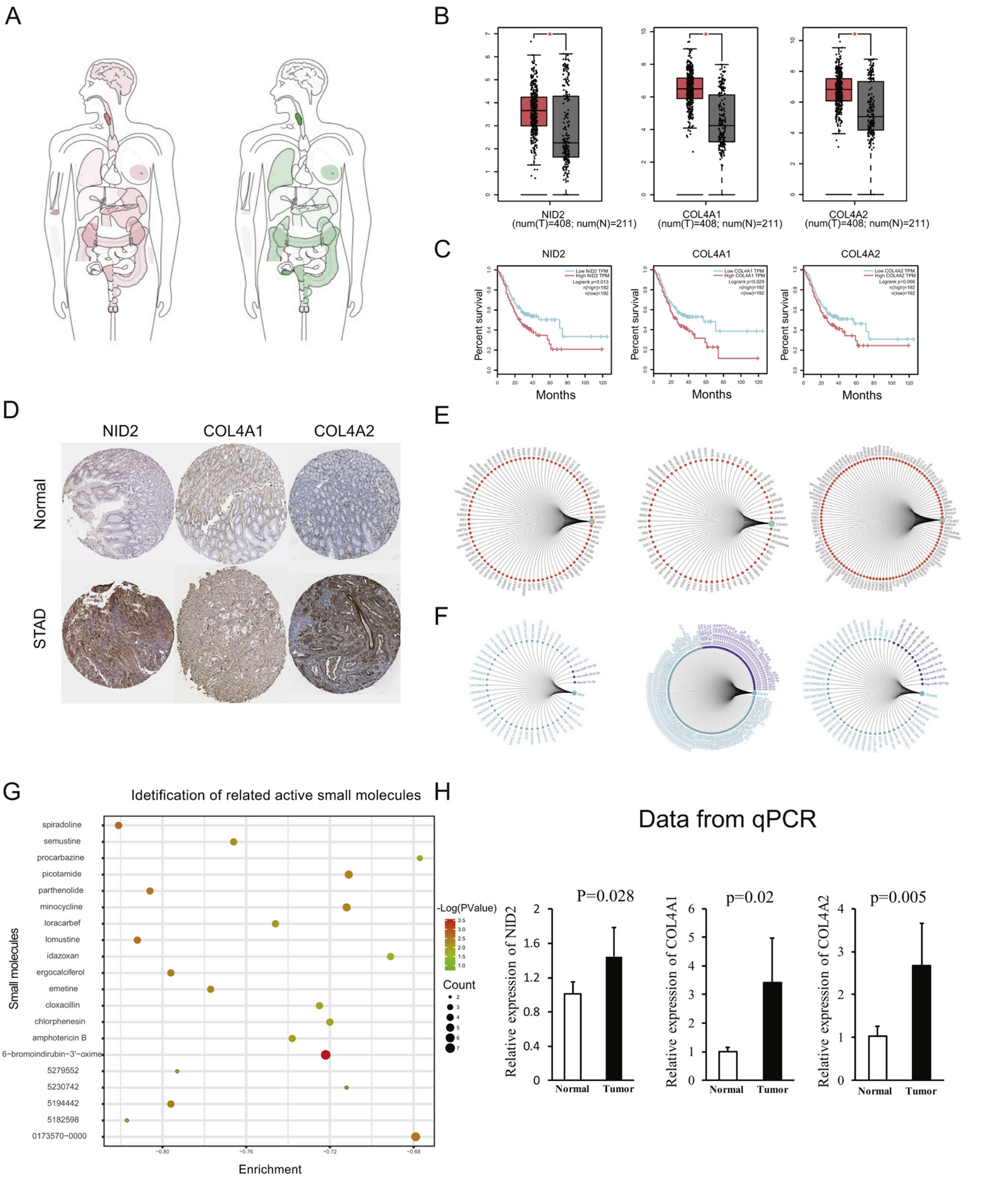
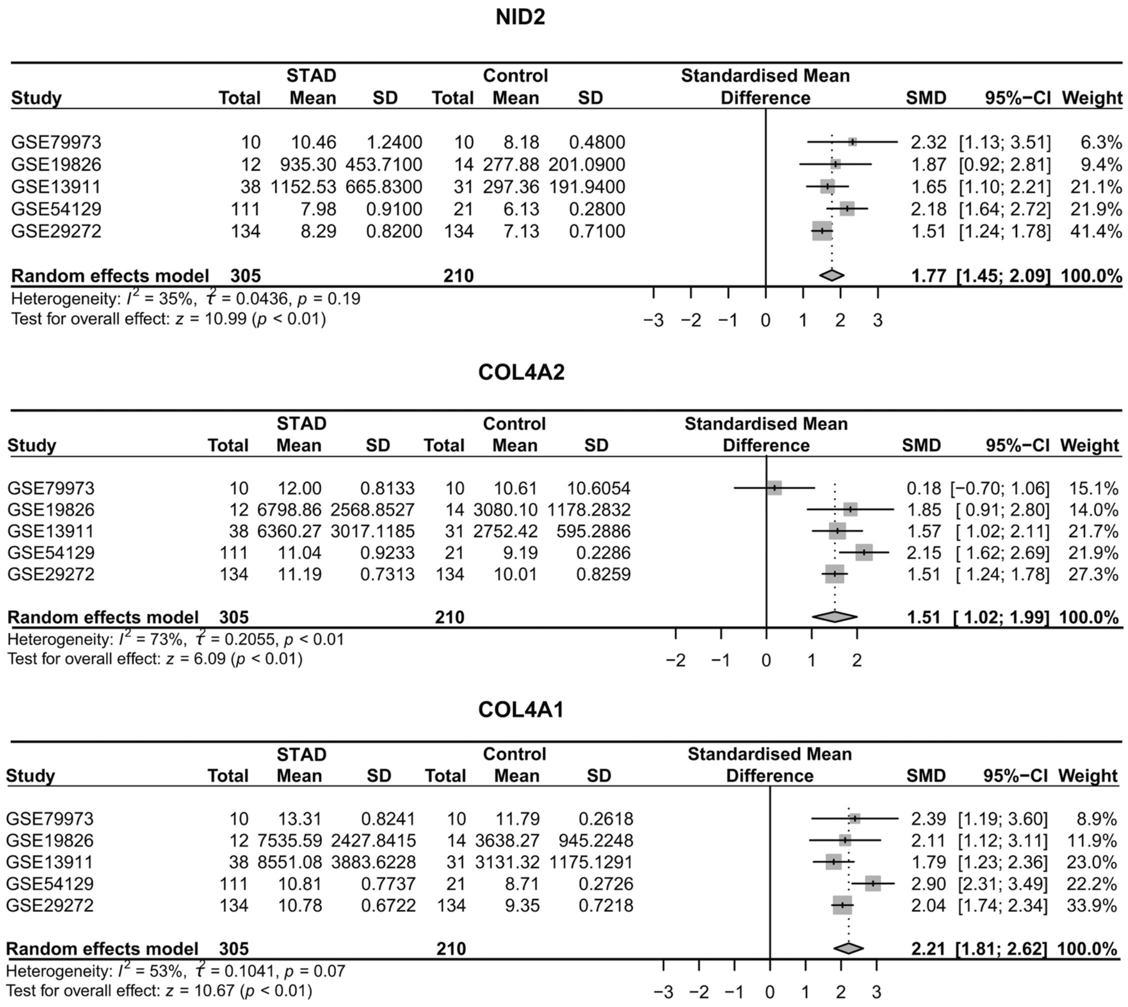


Fig. 5. **(A)** The median expression of tumor and normal samples in bodymap for NID2 **(B,C)** The expression level of hub genes and their prognostic value according to the GEPIA database. **(D)** Representative immunohistochemistry staining results reveal the protein level expression of NID2, COL4A1 and COL4A2 in gastric cancer and normal tissues. **(E)** The related transcription factors that may potentially regulated hub genes. **(F)** The regulatory network of lncRNA, miRNA and mRNA constructed by GCBI. Purple nodes: Related lncRNA; Blue nodes: Targeted miRNA. **(G)** The top 20 candidate small molecules targeting the gene expression of GC. **(H)** The average NID2, COL4A1 and COL4A2 mRNA expression in five paired gastric cancer tissues measured by qPCR. Abbreviations:
GEPIA: Gene Expression Profiling Interactive Analysis; GCBI: Gene-Cloud Biotechnology Information; STAD: stomach adenocarcinoma.

Table 3
List of the 20 most significant small molecule drugs that can reverse the tumoral status of gastric cancer.

rank	cmap name	n	enrichment	p
4	spiradoline	4	−0.821	0.00193
43	5182598	2	−0.817	0.06672
5	lomustine	4	−0.812	0.00249
6	parthenolide	4	−0.806	0.00275
8	5194442	4	−0.796	0.00338
9	ergocalciferol	4	−0.796	0.00346
51	5279552	2	−0.793	0.08388
13	emetine	4	−0.777	0.00511
14	semustine	4	−0.766	0.00613
16	loracarbef	4	−0.746	0.00822
17	amphotericin B	4	−0.738	0.00949
21	cloxacillin	4	−0.725	0.01166
2	6-bromoindirubin-3'-oxime	7	−0.722	0.0003
22	chlorphenesin	4	−0.72	0.01247
10	minocycline	5	−0.712	0.00423
80	5230742	2	−0.712	0.16526
11	picotamide	5	−0.711	0.00433
25	idazoxan	4	−0.691	0.01967
7	0173570-0000	6	−0.679	0.00298
45	procarbazine	3	−0.677	0.06794

of xenobiotics by cytochrome P450 drug metabolism and retinol metabolism. The ECM-receptor interaction pathway is involved in the various cancer cells proliferation and invasion. A previous study showed that Twist2 promoted the proliferation and invasion of kidney cancer cell by regulating the expression of ITGA6 and CD44 in the ECM-receptor interaction pathway [28]. High expression of focal adhesion kinase activity (FAK) has been reported to be associated with increased fibrosis and poor infiltration of CD8 + T cells. Inhibiting FAK can significantly limit tumor progression and prolong the overall survival of patients [29]. Tumor invasion and metastasis involve a series of complex molecular regulatory mechanisms. We established a PPI network with 124 nodes and 365 interactions on the basis of 172 DEGs. To determine the hug genes that played vital role in the progression of GC, we used the MCODE to perform module analysis in PPI network. Three DEGs with high degree of connectivity (NID2, COL4A1 and COL4A2) in these modules were selected as hub genes. GO enrichment analysis for hub genes showed that these genes were mainly related to laminin interactions, naba basement membranes and naba core matrisome. Nidogen-2 (NID2) is a major component of the basement membrane and involved in the stabilization of the extracellular matrix (ECM) network. Previous studies revealed the abnormal methylation of NID2 could be used as biomarkers for the early noninvasive detection of bladder cancer [30–32]. The highly expressed COL4A1 plays a vital role in promoting the proliferation and migration of multiple tumors. Ru et al. indicated that the overexpression of COL4A1 may increase resistance to trastuzumab in GC patients. COL4A2, one of the six subunits of encoding type IV collagen, suppressed by siRNA could significantly inhibit the migration and proliferation of triple-negative breast cancer cells [33–36]. To better understand the role of these hub genes in the GC



NID2

Study	STAD Total	STAD Mean	STAD SD	Control Total	Control Mean	Control SD	SMD	95%−CI	Weight
GSE79973	10	10.46	1.2400	10	8.18	0.4800	2.32	[1.13; 3.51]	6.3%
GSE19826	12	935.30	453.7100	14	277.88	201.0900	1.87	[0.92; 2.81]	9.4%
GSE13911	38	1152.53	665.8300	31	297.36	191.9400	1.65	[1.10; 2.21]	21.1%
GSE54129	111	7.98	0.9100	21	6.13	0.2800	2.18	[1.64; 2.72]	21.9%
GSE29272	134	8.29	0.8200	134	7.13	0.7100	1.51	[1.24; 1.78]	41.4%
Random effects model	**305**			**210**			**1.77**	**[1.45; 2.09]**	**100.0%**

Heterogeneity: $I^2 = 35\%$, $\tau^2 = 0.0436$, $p = 0.19$
Test for overall effect: $z = 10.99$ ($p < 0.01$)

COL4A2

Study	STAD Total	STAD Mean	STAD SD	Control Total	Control Mean	Control SD	SMD	95%−CI	Weight
GSE79973	10	12.00	0.8133	10	10.61	10.6054	0.18	[−0.70; 1.06]	15.1%
GSE19826	12	6798.86	2568.8527	14	3080.10	1178.2832	1.85	[0.91; 2.80]	14.0%
GSE13911	38	6360.27	3017.1185	31	2752.42	595.2886	1.57	[1.02; 2.11]	21.7%
GSE54129	111	11.04	0.9233	21	9.19	0.2286	2.15	[1.62; 2.69]	21.9%
GSE29272	134	11.19	0.7313	134	10.01	0.8259	1.51	[1.24; 1.78]	27.3%
Random effects model	**305**			**210**			**1.51**	**[1.02; 1.99]**	**100.0%**

Heterogeneity: $I^2 = 73\%$, $\tau^2 = 0.2055$, $p < 0.01$
Test for overall effect: $z = 6.09$ ($p < 0.01$)

COL4A1

Study	STAD Total	STAD Mean	STAD SD	Control Total	Control Mean	Control SD	SMD	95%−CI	Weight
GSE79973	10	13.31	0.8241	10	11.79	0.2618	2.39	[1.19; 3.60]	8.9%
GSE19826	12	7535.59	2427.8415	14	3638.27	945.2248	2.11	[1.12; 3.11]	11.9%
GSE13911	38	8551.08	3883.6228	31	3131.32	1175.1291	1.79	[1.23; 2.36]	23.0%
GSE54129	111	10.81	0.7737	21	8.71	0.2726	2.90	[2.31; 3.49]	22.2%
GSE29272	134	10.78	0.6722	134	9.35	0.7218	2.04	[1.74; 2.34]	33.9%
Random effects model	**305**			**210**			**2.21**	**[1.81; 2.62]**	**100.0%**

Heterogeneity: $I^2 = 53\%$, $\tau^2 = 0.1041$, $p = 0.07$
Test for overall effect: $z = 10.67$ ($p < 0.01$)

Fig. 6. Meta-analysis of the expression amount of NID2, COL4A1 and COL4A2 among the five datasets. Abbreviations: STAD: stomach adenocarcinoma.

initiation and progression, we constructed a molecular regulatory network of lncRNA - miRNA - mRNA and predicted potential transcription factors related to NID2, COL4A1 and COL4A2. Meanwhile, we performed a meta-analysis to pool the hub gene expression levels in different datasets. The meta-analysis also demonstrated that the expression of NID2 COL4A1 and COL4A2 in GC tissues was significantly higher than that in normal tissues. (Fig. 6) The GEPIA database was used to further validate our findings from bioinformatics analysis, and we found that patients with high expression levels of NID2 COL4A1 and COL4A2 had a worse survival compared to those with low expression. Similarly, the RT-qPCR analysis in five paired GC tissues showed the same trend in expression, thus verifying the accuracy of our findings. In addition, based on the overlapping DEGs, we obtained several potential small molecule drugs for the treatment of GC from the CMap database. These small molecules had the potential to alter the gene expression of GC, thereby controlling the progression of tumors. For example, procarbazine (enrichment score = − 0.677) has been clinically used for the treatment of glioma and Hodgkin's lymphoma. However, its efficacy and safety on GC are still not investigated [37–38]. Therefore, further studies are urgently demanded to confirm the huge potential of these candidate small molecules in treating GC.

Similar bioinformatics analysis for identifying key genes associated with the molecular mechanism of GC have been conducted before. As mentioned above, Cao et al. [24] and Ting et al. [25] revealed some key genes involved in GC by integrated bioinformatics analysis. Nevertheless, when compared with our study, their analysis was only based on three gene expression profiles and the selected key genes were validated only by the Kaplan–Meier plotter database. They did not analyze the expression of hub genes at the protein level especially considering that gene expression was not always consistent with its protein. In additionally, we constructed a molecular regulatory network of lncRNA - miRNA - mRNA, which contributed to better understanding the important role of hub genes in the tumor occurrence and development. More importantly, the sample size including in our study was much higher than that of their studies, thus ensuring the higher reliability of our results.

In conclusion, with integrated bioinformatics analysis for microarray datasets, we identified three key genes which could play a critical role in the pathogenesis of GC. Our study firstly uncovered that NID2 and COL4A2 could be involved in the initiation and progression of GC. These three key genes could act as the promising diagnostic and prognostic biomarkers in patients with GC. We also found spiradoline was the most promising small molecule to reverse the gene expression of GC. These findings could provide new sights about future genomic individualized treatment of GC and survival prediction.

Conflict of interest

None.

Funding

This study was funded by the National Natural Science Foundation of China (No. 81401796 and 81572267), Jiangsu youth medical talent project (QNRC2016702) and the Social Development Foundation of Nantong City (MS12017006-2).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.prp.2019.02.012.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (6) (2018) 394–424.
[2] L. Zong, M. Abe, Y. Seto, J. Ji, The challenge of screening for early gastric cancer in China, Lancet (London, England) 388 (10060) (2016) 2606.
[3] E. Van Cutsem, X. Sagaert, B. Topal, K. Haustermans, H. Prenen, Gastric cancer, Lancet (London, England) 388 (10060) (2016) 2654–2664.
[4] V. Kulasingam, E.P. Diamandis, Strategies for discovering novel cancer biomarkers through utilization of emerging technologies, Nat. Clin. Pract. Oncol. 5 (10) (2008) 588–599.
[5] Comprehensive molecular characterization of gastric adenocarcinoma, Nature 513 (7517) (2014) 202–209.
[6] B. Xiong, L. Ma, W. Huang, H. Luo, Y. Zeng, Y. Tian, The efficiency and safety of trastuzumab for advanced gastric and gastroesophageal cancer: a meta-analysis of five randomized controlled trials, Growth Factors (Chur, Switzerland) 34 (5-6) (2016) 187–195.
[7] H. Wilke, K. Muro, E. Van Cutsem, et al., Ramucirumab plus paclitaxel versus placebo plus paclitaxel in patients with previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (RAINBOW): a double-blind, randomised phase 3 trial, Lancet Oncol. 15 (11) (2014) 1224–1235.
[8] E. Van Cutsem, S. de Haas, Y.K. Kang, et al., Bevacizumab in combination with chemotherapy as first-line therapy in advanced gastric cancer: a biomarker evaluation from the AVAGAST randomized phase III trial, J. Clin. Oncol. 30 (17) (2012) 2119–2127.
[9] P. Jiang, X.S. Liu, Big data mining yields novel insights on cancer, Nat. Genet. 47 (2) (2015) 103–104.
[10] P. Yan, Y. He, K. Xie, S. Kong, W. Zhao, In silico analyses for potential key genes associated with gastric cancer, PeerJ 6 (2018) 6092.
[11] L. Gautier, L. Cope, B.M. Bolstad, R.A. Irizarry, Affy–analysis of Affymetrix GeneChip data at the probe level, Bioinformatics 20 (3) (2004) 307–315.
[12] M.E. Ritchie, B. Phipson, D. Wu, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47.
[13] The Gene Ontology (GO) project in 2006, Nucleic Acids Res. 34 (2006) D322–326 (Database issue).
[14] M. Ashburner, C.A. Ball, J.A. Blake, et al., Gene ontology: tool for the unification of biology, Gene Ontol. Consortium. Nat Genet. 25 (1) (2000) 25–29.
[15] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[16] G. Dennis Jr, B.T. Sherman, D.A. Hosack, et al., DAVID: database for annotation, visualization, and integrated discovery, Genome Biol. 4 (5) (2003) P3.
[17] M.E. Smoot, K. Ono, J. Ruscheinski, P.L. Wang, T. Ideker, Cytoscape 2.8: new features for data integration and network visualization, Bioinformatics 27 (3) (2011) 431–432.
[18] S. Maere, K. Heymans, M. Kuiper, BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks, Bioinformatics 21 (16) (2005) 3448–3449.
[19] W.P. Bandettini, P. Kellman, C. Mancini, et al., MultiContrast delayed Enhancement (MCODE) improves detection of subendocardial myocardial infarction by late gadolinium enhancement cardiovascular magnetic resonance: a clinical validation study, J. Cardiovasc. Magn. Reson. 14 (2012) 83.
[20] E. Cerami, J. Gao, U. Dogrusoz, et al., The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Cancer Discov. 2 (5) (2012) 401–404.
[21] J. Gao, B.A. Aksoy, U. Dogrusoz, et al., Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal, Sci. Signal. 6 (269) (2013) pl1.
[22] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, Z. Zhang, GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses, Nucleic Acids Res. 45 (W1) (2017) W98–w102.
[23] J. Lamb, E.D. Crawford, D. Peck, et al., The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease, Science 313 (5795) (2006) 1929–1935.
[24] T. Maier, M. Guell, L. Serrano, Correlation of mRNA and protein in complex biological samples, FEBS Lett. 583 (24) (2009) 3966–3973.
[25] L. Cao, Y. Chen, M. Zhang, et al., Identification of hub genes and potential molecular mechanisms in gastric cancer by integrated bioinformatics analysis, PeerJ 6 (2018) e5180.
[26] T. Li, X. Gao, L. Han, J. Yu, H. Li, Identification of hub genes with prognostic values in gastric cancer by bioinformatics analysis, World J. Surg. Oncol. 16 (1) (2018) 114.
[27] R. Huang, W. Gu, B. Sun, L. Gao, Identification of COL4A1 as a potential gene conferring trastuzumab resistance in gastric cancer based on bioinformatics analysis, Mol. Med. Rep. 17 (5) (2018) 6387–6396.
[28] H.J. Zhang, J. Tao, L. Sheng, et al., Twist2 promotes kidney cancer cell proliferation and invasion by regulating ITGA6 and CD44 expression in the ECM-receptor interaction pathway, Oncol. Targets Ther. 9 (2016) 1801–1812.
[29] H. Jiang, S. Hegde, B.L. Knolhoff, et al., Targeting focal adhesion kinase renders pancreatic cancers responsive to checkpoint immunotherapy, Nat. Med. 22 (2016) 851.
[30] A.W. Chai, A.K. Cheung, W. Dai, et al., Metastasis-suppressing NID2, an epigenetically-silenced gene, in the pathogenesis of nasopharyngeal carcinoma and esophageal squamous cell carcinoma, Oncotarget 7 (48) (2016) 78859–78871.
[31] Z. Yegin, S. Gunes, R. Buyukalpelli, Hypermethylation of TWIST1 and NID2 in tumor tissues and voided urine in urinary bladder cancer patients, DNA Cell Biol. 32 (7) (2013) 386–392.
[32] J.J. Fantony, T.A. Longo, A. Gopalakrishna, et al., Urinary NID2 and TWIST1 methylation to augment conventional urine cytology for the detection of bladder cancer, Cancer Biomarkers: Sect. A Dis. Markers 18 (4) (2017) 381–387.

[33] M. Miyake, S. Hori, Y. Morizawa, et al., Collagen type IV alpha 1 (COL4A1) and collagen type XIII alpha 1 (COL13A1) produced in cancer cells promote tumor budding at the invasion front in human urothelial carcinoma of the bladder, Oncotarget 8 (22) (2017) 36099–36114.

[34] R. Jin, J. Shen, T. Zhang, et al., The highly expressed COL4A1 genes contributes to the proliferation and migration of the invasive ductal carcinomas, Oncotarget 8 (35) (2017) 58172–58183.

[35] C.W. Brown, A.S. Brodsky, R.N. Freiman, Notch3 overexpression promotes anoikis resistance in epithelial ovarian cancer via upregulation of COL4A2, Mol. Cancer Res. 13 (1) (2015) 78–85.

[36] H. JingSong, G. Hong, J. Yang, et al., siRNA-mediated suppression of collagen type iv alpha 2 (COL4A2) mRNA inhibits triple-negative breast cancer cell proliferation and migration, Oncotarget 8 (2) (2017) 2585–2593.

[37] C. Mauz-Korholz, D. Hasenclever, W. Dorffel, et al., Procarbazine-free OEPA-COPDAC chemotherapy in boys and standard OPPA-COPP in girls have comparable effectiveness in pediatric Hodgkin's lymphoma: the GPOH-HD-2002 study, J. Clin. Oncol. 28 (23) (2010) 3680–3686.

[38] S. Parasramka, G. Talari, M. Rosenfeld, J. Guo, J.L. Villano, Procarbazine, lomustine and vincristine for recurrent high-grade glioma, Cochrane Database Syst. Rev. 7 (2017) Cd011773.