RESEARCH ARTICLE

Identification of hub genes and construction of an mRNA-miRNA-lncRNA network of gastric carcinoma using integrated bioinformatics analysis

Gang Wei[1], Youhong Dong[2], Zhongshi He[2], Hu Qiu[1], Yong Wu[1], Yongshun Chen[1]*

1 Department of Clinical Oncology, Renmin Hospital of Wuhan University, Wuhan, China, **2** Department of Clinical Oncology, The First People's Hospital of Xiangyang, Xiangyang, China

* yongshunchen2020@163.com

OPEN ACCESS

Citation: Wei G, Dong Y, He Z, Qiu H, Wu Y, Chen Y (2021) Identification of hub genes and construction of an mRNA-miRNA-lncRNA network of gastric carcinoma using integrated bioinformatics analysis. PLoS ONE 16(12): e0261728. https://doi.org/10.1371/journal.pone.0261728

Editor: Qi Zhao, University of Science and Technology Liaoning, CHINA

Received: July 23, 2021

Accepted: December 8, 2021

Published: December 30, 2021

Data Availability Statement: The GEO database and TCGA database are open, no account registration is required, anyone can download and use the data inside. To further explain the source of the data, we will repeat the main steps we used to obtain the data: Obtaining gastric cancer transcriptome data in the TCGA database: 1. Enter the data selection interface: https://portal.gdc.cancer.gov/repository 2. In the "Cases" interface: Select "stomach" in the "Primary Site" item; Select " TCGA " in the " Program " item; Select "

Abstract

Background

Gastric carcinoma (GC) is one of the most common cancer globally. Despite its worldwide decline in incidence and mortality over the past decades, gastric cancer still has a poor prognosis. However, the key regulators driving this process and their exact mechanisms have not been thoroughly studied. This study aimed to identify hub genes to improve the prognostic prediction of GC and construct a messenger RNA-microRNA-long non-coding RNA (mRNA-miRNA-lncRNA) regulatory network.

Methods

The GSE66229 dataset, from the Gene Expression Omnibus (GEO) database, and The Cancer Genome Atlas (TCGA) database were used for the bioinformatic analysis. Differential gene expression analysis methods and Weighted Gene Co-expression Network Analysis (WGCNA) were used to identify a common set of differentially co-expressed genes in GC. The genes were validated using samples from TCGA database and further validation using the online tools GEPIA database and Kaplan-Meier(KM) plotter database. Gene set enrichment analysis(GSEA) was used to identify hub genes related to signaling pathways in GC. The RNAInter database and Cytoscape software were used to construct an mRNA-miRNA-lncRNA network.

Results

A total of 12 genes were identified as the common set of differentially co-expressed genes in GC. After verification of these genes, 3 hub genes, namely *CTHRC1*, *FNDC1*, and *INHBA*, were found to be upregulated in tumor and associated with poor GC patient survival. In addition, an mRNA-miRNA-lncRNA regulatory network was established, which included 12 lncRNAs, 5 miRNAs, and the 3 hub genes.

TCGA-STAD " in the " Project " item; Continue to filter the data in the "Files" interface: Select " transcriptome profiling " in the “Data Category " item; Select " Gene Expression Quantification " in the “Data Type" item; Select " RNA-Seq " in the “Experimental Strategy " item; Select " HTSeq-Counts " in the “Workflow Type " item; Then click “Add All Files to Cart”. 3. Enter the "Cart" interface to download "Metadata", "Manifest" and "Cart", a total of three files; these three files contain the data needed in the TCGA database in our study. Obtaining the GSE66229 data set in the GEO database: 1. Enter the GSE66229 data set interface: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66229; 2. The data we downloaded is: GSE66229_RAW.tar, download address: https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE66229&format=file; 3. GPL570 platform comment file download address: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?mode=raw&is_datatable=true&acc=GPL570&id=55999&db=GeoDb_blob143; The above are the steps of data acquisition in our study. If there is anything that is not clear, please contact us.

Funding: This research was supported by the National Natural Science Foundation of China (No. U1604175), Central Leading Local Science and Technology Development Special Foundation (ZYYD2020000169). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Conclusions

In summary, the identification of these hub genes and the establishment of the mRNA-miRNA-lncRNA regulatory network provide new insights into the underlying mechanisms of gastric carcinogenesis. In addition, the identified hub genes, *CTHRC1*, *FNDC1*, *and INHBA*, may serve as novel prognostic biomarkers and therapeutic targets.

Introduction

Gastric carcinoma (GC) is a common malignant tumor originating from the gastric mucosal epithelium. Despite its worldwide decline in incidence and mortality rates over the past decades, GC has a poor prognosis [1]. In 2020, there were about 1,089,103 new gastric cancer cases, which resulted in 768,793 deaths, making it the fifth-most commonly diagnosed cancer type and the fourth leading cause of cancer-related deaths after lung, colorectal, and liver cancers [2]. Although the pathogenesis of gastric cancer remains unclear to date, gastric cancer is widely considered to be a highly heterogeneous disease caused by multiple factors, including chronic infection with Helicobacter pylori [3, 4], Epstein–Barr virus [5], unhealthy diet [6], smoking [7], *etc*., which interact with genes and ultimately lead to tumor development [8, 9]. A gene mutation can be an early indicator of the risk of cancer development and even its future aggressiveness, and genes whose expression is correlated with the progression and prognosis of GC need to be identified. In recent years, biomarkers and therapeutic targets for GC have greatly contributed to improving the diagnosis and treatment of GC. For example, IDO1 and COL12A1, which were found to synergistically promote gastric cancer metastasis, appear to be promising targets for the treatment of gastric cancer [10]. Sha *et al.* [11] found that ORAI2 promotes gastric cancer cell migration and tumor metastasis through MAPK-dependent focal adhesion disassembly and PI3K/Akt signaling, which suggests the possibility of developing potential therapies for GC by targeting the ORAI2 signaling pathway. However, identifying novel diagnostic and prognostic biomarkers remains urgently necessary in view of the biological complexity, poor prognosis, and high reoccurrence of GC.

In the past few decades, microarray technology and bioinformatics analysis have been widely used in cancer functional genomics research to identify genes closely related to tumor development, progression and prognosis through genomics and clinical data analysis [12, 13]. Accordingly, an approach integrating technologies is helpful to identify key genes associated with gastric cancer development and progression. In this study, two major transcriptome analysis methods were used to identify GC-associated genes. One method is the analysis of differentially expressed genes (DEGs), which is used to determine quantitative changes in expression levels between different groups [14]. Studies to identify DEGs between groups under specific conditions, which are widely conducted using RNA-seq data analysis, are critical to understanding phenotypic variation. Differential gene expression analysis can provide in-depth insights into the genetic mechanisms of different phenotypes. For example, through the differential gene expression analysis of multiple data sets, a total of 31 hub genes were identified in colorectal cancer, and these hub genes were found to be significantly enriched in multiple pathways, mainly those related to the cell cycle process, and chemokines and G-protein coupled receptors [15]. The other method is Weighted Gene Co-expression Network Analysis (WGCNA) [16], a data mining-based method used to analyze biological networks, which is used to identify highly coordinated gene sets. Then, based on the interconnectivity of the identified gene sets, candidate biomarker genes or therapeutic targets as well as the association between gene sets and phenotypes can be identified. Compared with focusing only on DEGs, WGCNA can be used to analyze thousands, or nearly thousands, of the genes with the most altered expression, as well as the information about all genes, to identify the gene sets of interest and perform significant association analysis between the genes sets and the phenotype. For instance, Yin *et al*. [17] used WGCNA to identify 5 hub genes that may play a key role in the progression of hepatocellular carcinoma. Additional studies using WGCNA have shown that four genes (*RACGAP1*, *ZWINT*, *TKI*, and *LMNB1*) may serve as potential diagnostic and prognostic markers [18]. In this study, WGCNA was based on the correlation of variables to establish a gene interaction network within the biological system, using the transcriptome and clinical data to identify the modules of genes with characteristic co-expression pattern, and further examine the relationship between the gene modules and the clinical traits [19]. Therefore, we used two approaches, combining the results of the WGCNA and differential gene expression analysis, to enhance the identification of highly correlated genes, which could thus serve as candidate biomarkers for GC prognosis.

Materials and methods

Data sources and pre-processing

The gene expression profiles in dataset GSE66229, which consists of the GSE62254 and GSE66222 datasets, obtained on the GPL570 platform using the Affymetrix Human Genome U133 Plus 2.0 array, (Affymetrix Inc., Santa Clara, CA, USA), were downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/gds). The GSE66222 dataset contains data from 100 samples of normal gastric tissue, and the GSE62254 dataset contains data from 300 gastric cancer samples. The probes were converted into gene symbols according to the annotation file provided by the manufacturer, and probes corresponding to multiple genes were removed. If multiple probes corresponded to one gene, the median value was used. Ultimately, a total of 20,549 genes were subjected to further analysis.

The RNA-seq expression profiles (count format) and clinical data of GC patients were obtained from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). Genes with an average expression level value below 1 in all samples were removed. The mRNA expression matrix consists of 22,634 genes and 407 samples, of which 375 are tumor samples and 32 are normal tissue adjacent to the tumor. A total of 371 tumor samples were available for survival analysis.

Weighted gene co-expression network analysis

All data analysis was performed using the R software (Version 3.63, https://www.r-project.org/). The R package WGCNA was used to analyze the gene co-expression network of the two datasets. First, the genes with the absolute median difference in the top 5,000 were retained. Second, the samples were clustered and the outliers were removed. To construct a scale-free network, soft strengths of β = 3 and 4 were chosen for these two datasets with the function pickSoftThreshold, separately. In the co-expression network, genes with high absolute correlation were aggregated into modules with different colors by using the function blockwiseModules. Then, the correlations between modules and clinical feature information were calculated using the WGCNA package, and modules with high correlation with tumor traits were further analyzed.

Analysis of differentially expressed genes

The R package edgeR was used to identify DEGs in TCGA datasets. For the GEO data, the limma package was used to identify DEGs between the tumor samples and normal samples. False discovery rate (FDR) was used to adjust the P-value. Genes with |FC| (fold change)> = 2 and adjusted $P < 0.05$ were considered to be DEGs. The DEGs of TCGA datasets and the GSE66229 datasets were visualized as volcano plots using the R package ggplot2.

Validation of hub genes

In the two datasets, the overlapping genes were selected from upregulated genes in DEGs and the module genes in the co-expression network, which were visualized using the VennDiagram package [20]. To identify the true hub genes, Kaplan-Meier survival analysis was performed to evaluate the association of hub genes with overall survival (OS) in TCGA datasets using the survival package. Tumor samples with follow-up time were divided into two groups according to the median value of gene expression. Genes with p-values < = 0.05 are verified again in two online databases the GEPIA database (http://gepia.cancer-pku.cn/) and Kaplan-Meier (KM) plotter database (https://kmplot.com/analysis/). Then, the Human Protein Atlas (HPA) database (http://www.proteinatlas.org/) was used to validate the hub genes by immunohistochemistry (IHC).

Gene set enrichment analysis of real hub genes

In TCGA datasets, samples of GC were divided into two groups according to the expression level of the hub genes (high expression *vs*. low expression based on the median expression value of each hub gene). The gene set enrichment analysis (GSEA) software downloaded from http://www.gsea-msigdb.org/gsea/index.jsp was used to identify the potential function of the hub genes. FDR $P < 0.05$ was used as the criterion for significant enrichment.

Construction of lncRNA-miRNA-hub gene network

RNAInter database (https://www.rna-society.org/rnainter/), a complete resource of RNA interactome data from the literature and other databases containing over 41 million RNA-related interactions of RDI, RCI, RPI, RHI, and RRI [21], was used to investigate the relationship between mRNAs, long non-coding RNAs (lncRNAs) and microRNAs (miRNAs). We used the lncRNA–miRNA and mRNA–miRNA relationships with strong experimental evidence for further analysis. Then, the lncRNA-miRNA-mRNA regulatory network was visualized with the Cytoscape software.

Results

Identification of modules associated with tumor and normal tissues

The data were processed and analyzed as shown in the flowchart in Fig 1. By filtering the two gene expression matrices, genes in the top 5,000 with absolute median differences in TCGA dataset and GEO dataset were further screened for co-expression network analysis. Soft strengths of $\beta = 3$ in TCGA datasets and $\beta = 4$ in the GSE66229 datasets were chosen using the function pickSoftThreshold. Then, the co-expression networks were established, and gene modules were identified using the function blockwiseModules. A total of 9 and 14 modules were identified in TCGA datasets and the GSE datasets (Fig 2A), respectively. Each color represents an independent module that contains a set of highly-related genes. Eventually, the relationship between different co-expression modules and clinical features was visualized by heat

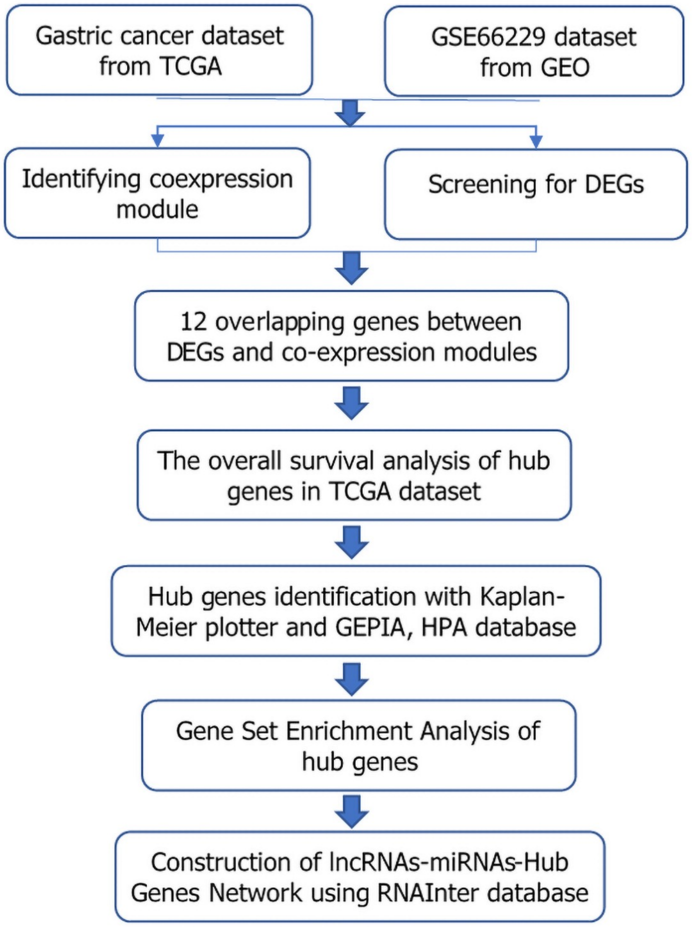
Gastric cancer dataset from TCGA
GSE66229 dataset from GEO
Identifying coexpression module
Screening for DEGs
12 overlapping genes between DEGs and co-expression modules
The overall survival analysis of hub genes in TCGA dataset
Hub genes identification with Kaplan-Meier plotter and GEPIA, HPA database
Gene Set Enrichment Analysis of hub genes
Construction of lncRNAs-miRNAs-Hub Genes Network using RNAInter database

Fig 1. Study design and workflow of this study.

https://doi.org/10.1371/journal.pone.0261728.g001

map (Fig 2B). The top number in each cell is the correlation coefficient, and the bottom one is the p-value. Modules with high correlation with tumor traits and p-value< 0.05were further analyzed, and the results were consistent with one module (pink, with 113 genes) in TCGA datasets, and with two modules (blue, purple, with 1,581 genes in all) in the GEO datasets.

Analysis of differential gene expression

Using |FC|(fold change)> = 2 and FDR-adjusted P < 0.05 as the cut-off criterion, we obtained 6,065 and 1,205 DEGs from TCGA datasets and the GSE66229 datasets, respectively, and visualized them with volcano plots (Fig 3A and 3B). A total of 857 DEGs (321 upregulated and 536 downregulated) were identified by gene integration analysis (Fig 3C). According to the calculations, there were 12 overlapping genes (*EVA1A*, *RARRES1*, *ADAM12*, *COL10A1*, *COL11A1*, *COL1A1*, *COL1A2*, *COL10A1*, *CTHRC1*, *FAP*, *FNDC1*, *and INHBA*) between the upregulated genes and co-expression Modules (Fig 3D).

Validation of the actual hub genes

We identified the prognostic value of the 12 genes in TCGA datasets using the survival package in R. Patients were divided into a high group and a low group based on the median expression

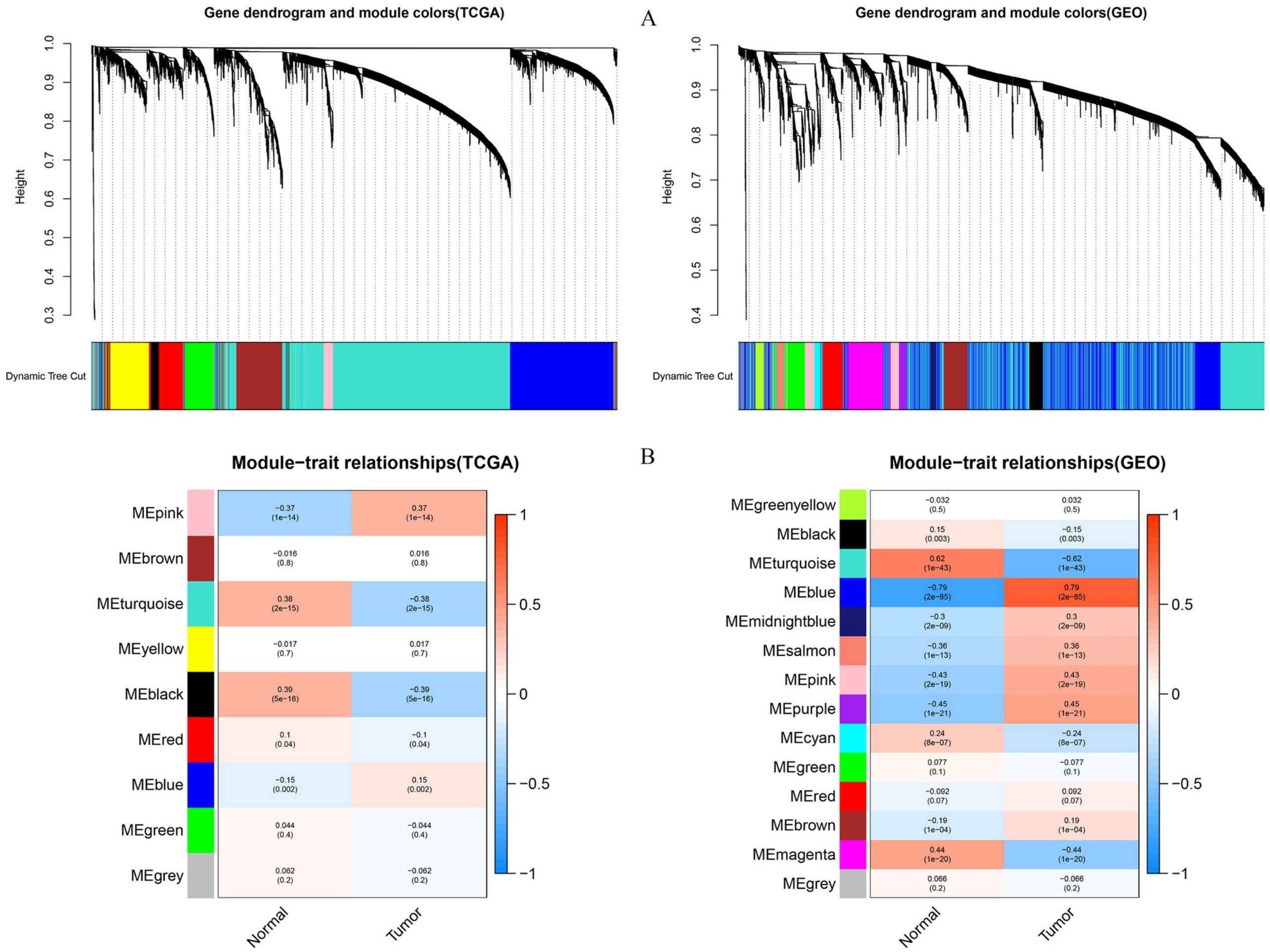

Fig 2. Identification of co-expression modules associated with clinical features in gastric cancer. The results on the left are from TCGA, and those on the right are from GSE66229 (A) Gene cluster dendrograms and module colors. The gene dendrogram is obtained by overlaying the topology with the corresponding module colors. Each color represents an independent module that contains a set of highly-related genes. (B) Heat map of the correlation between co-expression module genes and clinical features (tumor and normal), the top number in each cell is the correlation coefficient, and the bottom one is the p-value.

https://doi.org/10.1371/journal.pone.0261728.g002

of the genes. The results of the Kaplan–Meier curve analysis indicated that the higher the expression of C*OL10A1*, *CTHRC1*, *FAP*, *FNDC1*, *and INHBA*, the worse the prognosis of the GC patients (P < 0.05) (Fig 4). These five genes were further validated by survival analysis using the Kaplan-Meier plotter database (Fig 5A–5E) and GEPIA database (Fig 5F–5J). After the above validation, three candidate genes (*CTHRC1*, *FNDC1*, *and INHBA*) were ultimately determined to be the hub genes. Then, the expression level of these three genes was evaluated using the GEPIA database (Fig 6), which revealed that compared with normal gastric tissue samples, the expression of *CTHRC1*, *FNDC1 and INHBA* was elevated in GC samples. These findings are consistent with the results of our analysis. The IHC staining data obtained from the HPA database were used to determine the protein levels of these three candidate hub genes (Fig 7). The results also showed that the protein levels of *CTHRC1*, *FNDC1* were dysregulated in GC tissues (*INHBA* was not found in the HPA database).

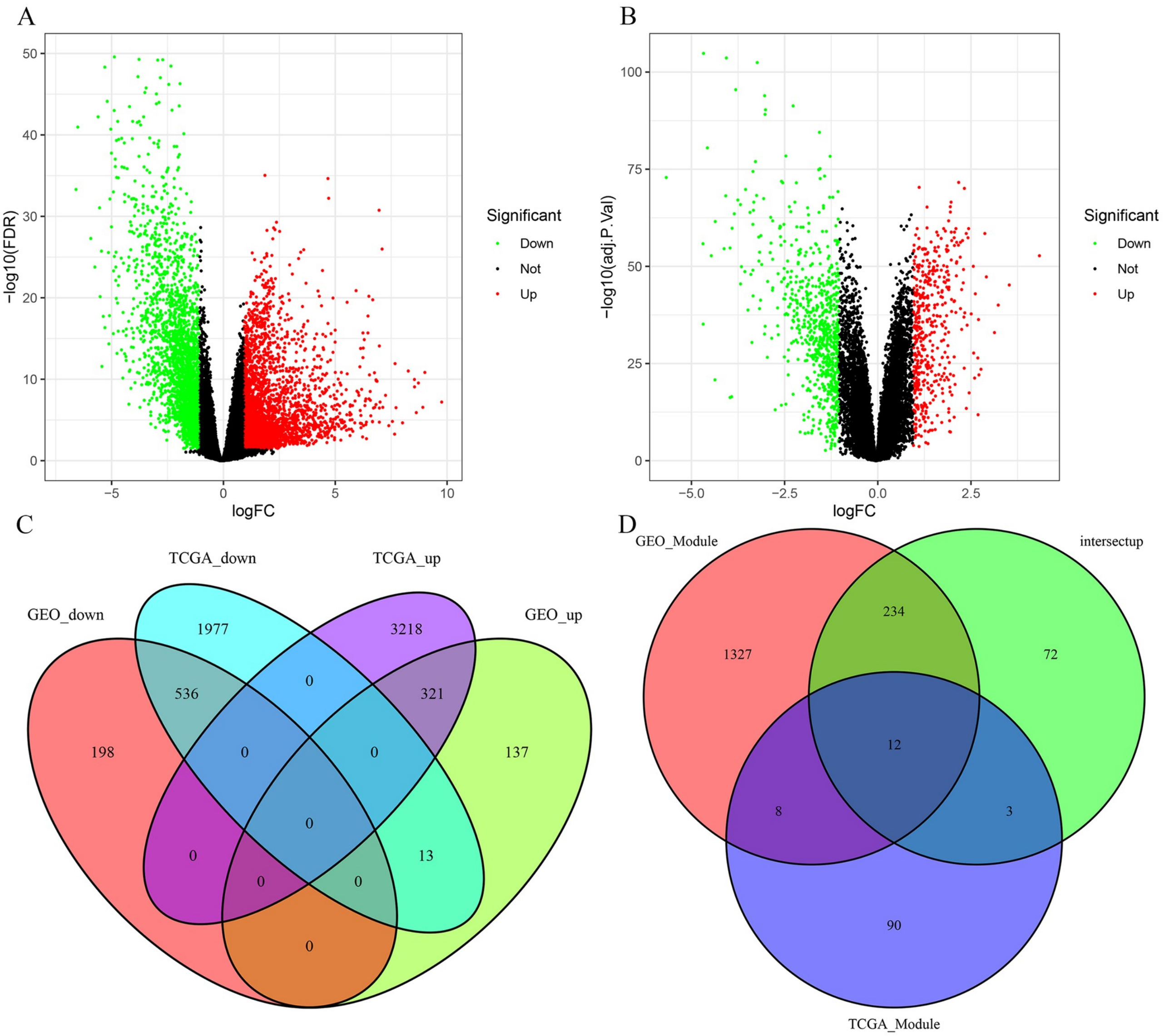

Fig 3. Analysis of differential gene expression with TCGA-GC datasets and the GSE66229 datasets. (A) Volcano plot of TCGA dataset. (B) Volcano plot of the GSE66229 dataset. (C) Venn diagram of genes between DEGs. (D) Venn diagram comparison of genes from the upregulated genes and co-expression modules. A total of 12 overlapping genes were identified.

https://doi.org/10.1371/journal.pone.0261728.g003

Gene set enrichment analysis revealed a close relationship between hub genes and tumor development

To further investigate the potential biological functions of *CTHRC1*, *FNDC1*, and *INHBA*, we performed Kyoto Encyclopedia of Genes and Genomes (KEGG)-GSEA analysis of the RNA-seq data of GC samples from TCGA. As shown in Fig 8, genes in higher-expression groups of *CTHRC1*, *FNDC1*, and *INHBA* were all involved in "BASAL CELL CARCINOMA", "FOCAL ADHESION", "HEDGEHOG SIGNALING PATHWAY", "MELANOMA", and "TGF BETA SIGNALING PATHWAY". In addition, the "PATHWAY IN CANCER" was enriched in the *FNDC1* and *INHBA* high-expression groups. The sets of genes with the highest enrichment scores are closely related to the occurrence and development of tumor.

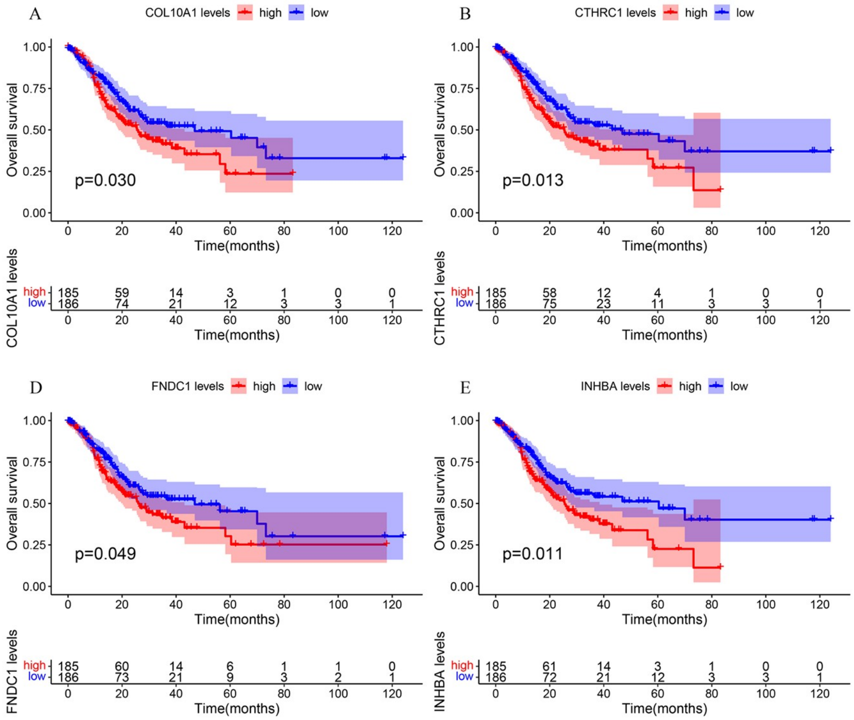

Fig 4. The Kaplan–Meier survival curves of the five genes in TCGA dataset. (A) *COL10A1* (B) *CTHRC1* (C) *FAP* (D) *FNDC1* (E) *INHBA*.

https://doi.org/10.1371/journal.pone.0261728.g004

Construction of lncRNA-miRNA–hub gene network

We also undertook to establish the transcriptional regulatory network of lncRNAs, miRNAs, and hub genes by selecting strong experimental evidence in the RNAInter database. As depicted in Fig 9, the network includes 12 lncRNAs, 5 miRNAs, and 3 hub genes. The scores of the correlations between lncRNA and miRNA and between miRNA and mRNA are shown in Table 1. MiRNAs regulate gene expression by interacting with their target genes [22]. LncRNAs function as competing endogenous RNAs (ceRNAs), competing for shared miRNAs and sequestering miRNAs from mRNAs [23]. This network reflects the regulatory relationships in the process of hub genes expression as well as the complex mechanisms of tumorigenesis.

Discussion

In this study, we first identified tumor-associated co-expression modules in two datasets using WGCNA, and then the DEGs between tumor and normal tissues, ultimately identifying a total of 12 overlapping genes between the upregulated genes and co-expression modules. These genes were not only upregulated in GC but also highly associated with GC. We additionally performed a prognostic analysis of these genes using TCGA database and further validated them using the Kaplan-Meier plotter database and GEPIA database. The GEPIA and Kaplan-Meier plotter databases are two web-based tools that deliver fast and customizable functionalities. The GEPIA database provides key interactive and customizable functions including profiling plotting, differential expression analysis, patient survival analysis base on TCGA and GTEx data [24]. The Kaplan-Meier Plotter database includes gene expression data and clinical data, which is a powerful tool that can be used to evaluate the effect of genes on the survival of patients with gastric cancer [25]. The use of these two databases can make our results more

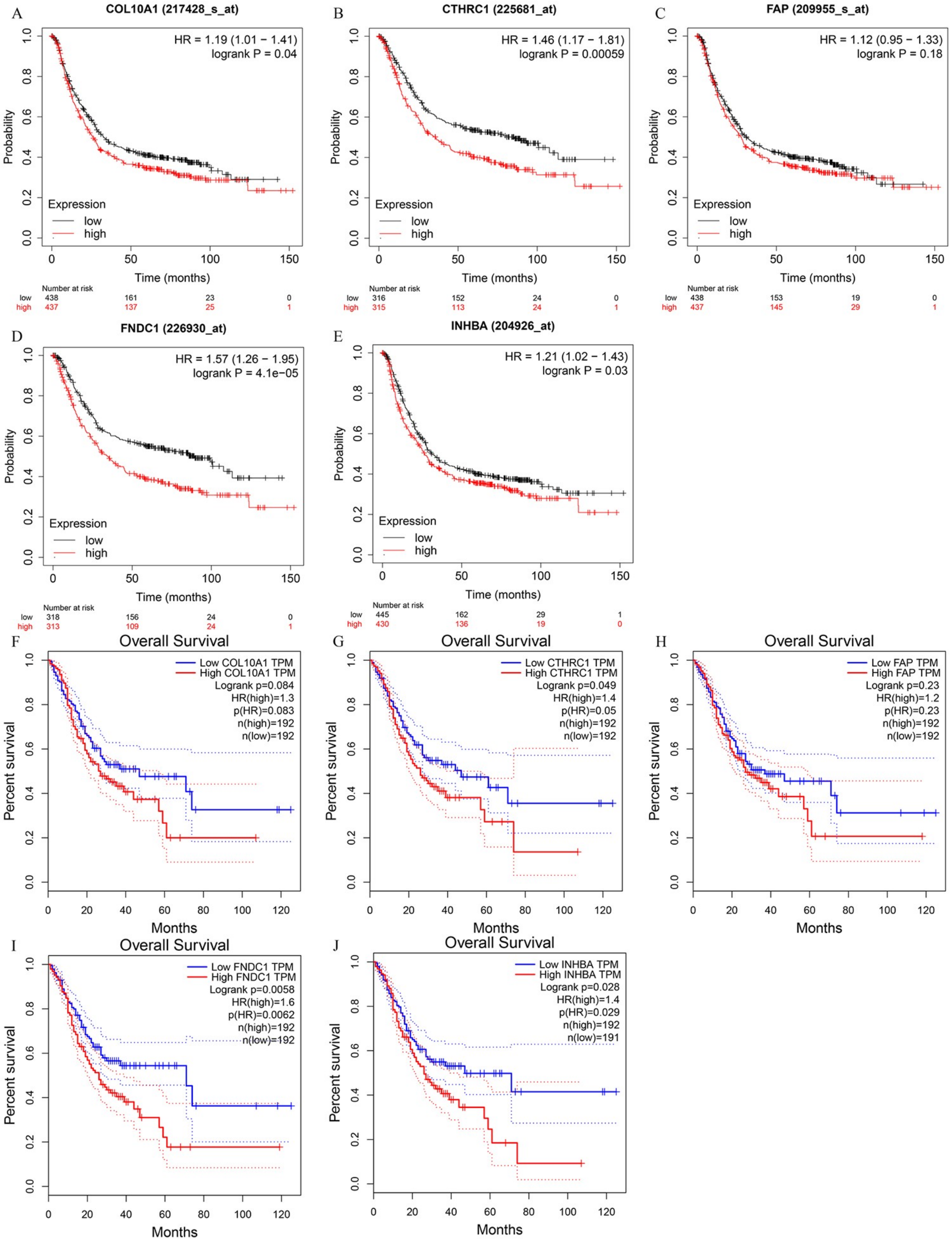

Fig 5. The Kaplan–Meier survival curves of the 5 genes.

https://doi.org/10.1371/journal.pone.0261728.g005

reliable and compelling. After the validation process, revealed that higher expression of *CTHRC1*, *FNDC1*, and *INHBA* indicated poorer survival in patients with GC, these 3 genes were considered to be the hub genes. The KEGG-GSEA analysis indicated that these mRNAs were significantly enriched in cancer-related pathways, including basal cell carcinoma, focal adhesion [26], hedgehog signaling pathway [27], melanoma, pathway in cancer, and TGF beta signaling pathway [28].

CTHRC1 (Collagen Triple Helix Repeat Containing 1) is a protein-encoding gene that appears to play a role in the cellular response to arterial injury through its involvement in vascular remodeling. It has been reported that, following injury, *CTHRC1* is transiently

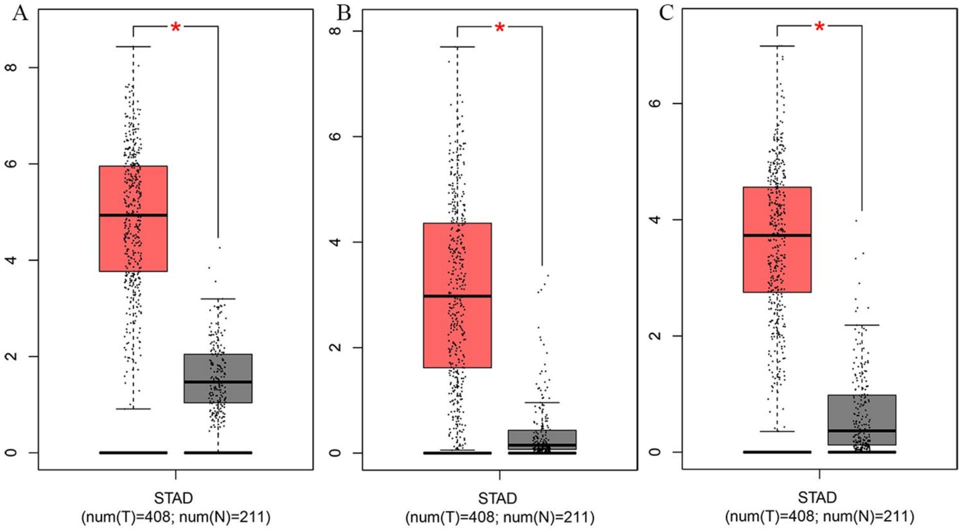

Fig 6. Verification of the expression levels of these 3 hub genes using GEPIA. (A) *CTHRC1* (B) *FNDC1* (C) *INHBA*.

https://doi.org/10.1371/journal.pone.0261728.g006

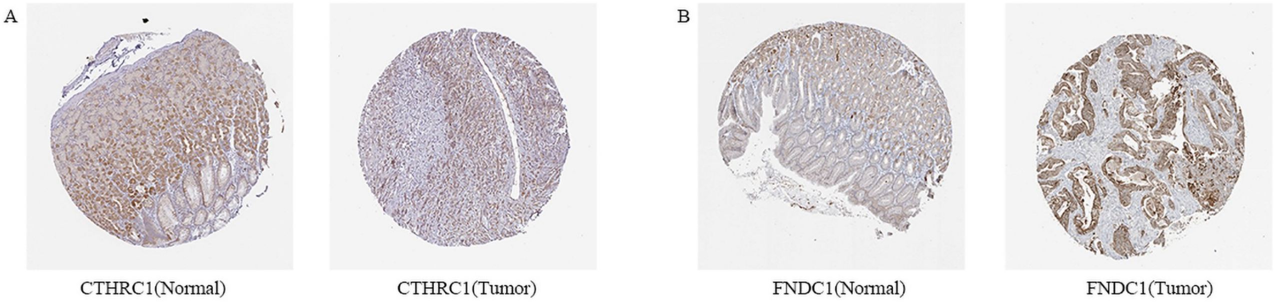

Fig 7. Immunohistochemistry (IHC) analysis of 3 hub genes on the HPA database. (A) *CTHRC1*, (B) *FNDC1*. (*INHBA* was not found in the HPA database).

https://doi.org/10.1371/journal.pone.0261728.g007

overexpressed in the adventitial and intimal smooth muscle of rat arteries [29]. Although physiologically *CTHRC1* plays an important role in wound healing, abnormal expression of *CTHRC1* also promotes the development of various human tumors. For instance, *CTHRC1* promotes cervical cancer metastasis and activates the Wnt/PCP pathway [30]. In addition,

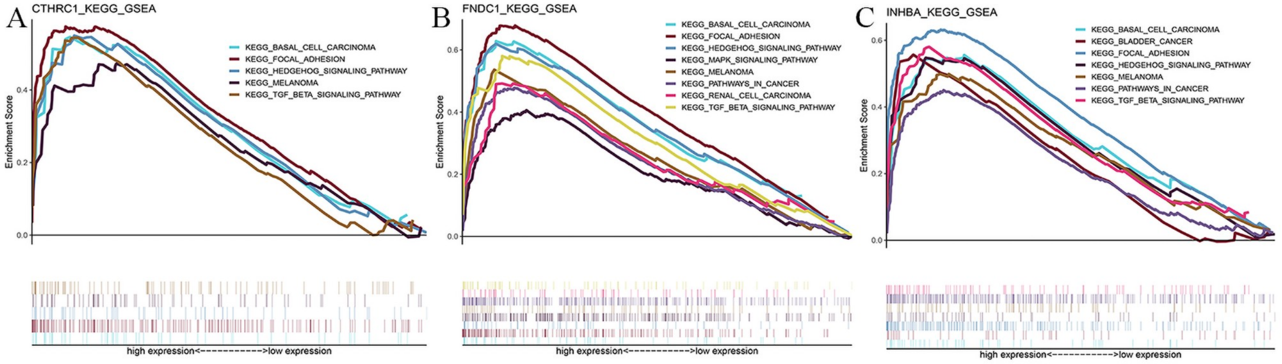

Fig 8. KEGG-GSEA analyses of the 3 identified hub genes. (A) CTHRC1 (B) FNDC1 (C) INHBA.

https://doi.org/10.1371/journal.pone.0261728.g008

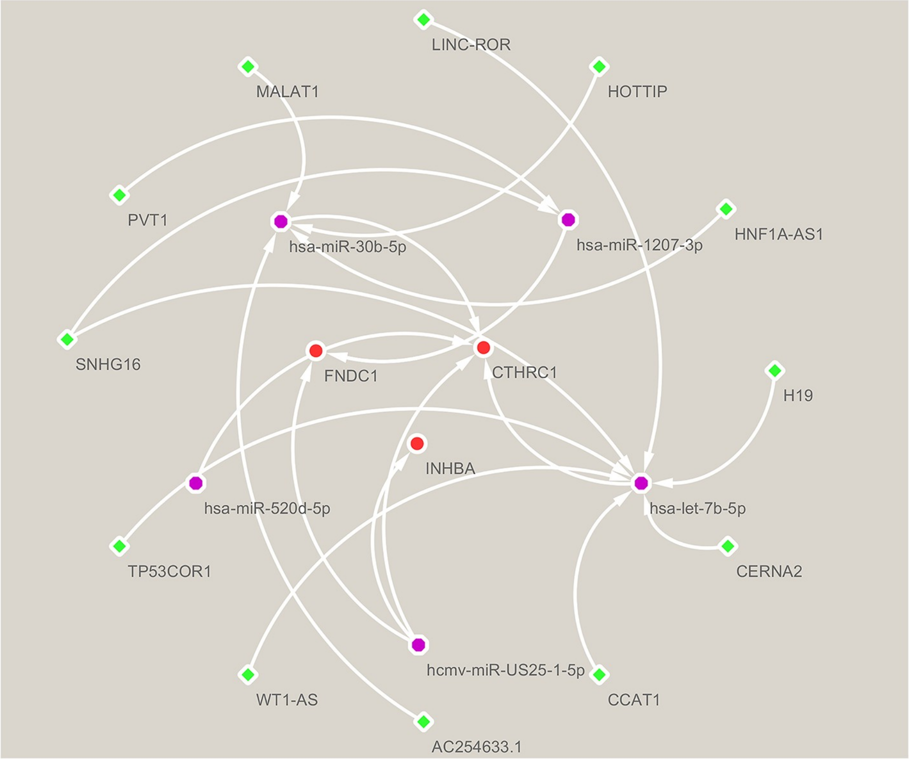

Fig 9. The lncRNA-miRNA-mRNA network was established using the cytoscape software.

https://doi.org/10.1371/journal.pone.0261728.g009

Table 1. The correlation between lncRNA, miRNA, and mRNA according to the RNAInter database.

RNAInter_ID	Interactor1	Category1	Interactor2	Category2	Score
RR00377685	hsa-let-7b-5p	miRNA	CTHRC1	mRNA	0.9943
RR00377701	hsa-miR-30b-5p	miRNA	CTHRC1	mRNA	0.9928
RR00377703	hsa-miR-520d-5p	miRNA	CTHRC1	mRNA	0.9609
RR00377694	hcmv-miR-US25-1-5p	miRNA	CTHRC1	mRNA	0.7311
RR00572052	hsa-miR-1207-3p	miRNA	FNDC1	mRNA	0.9526
RR00572047	hcmv-miR-US25-1-5p	miRNA	FNDC1	mRNA	0.7311
RR00751831	hcmv-miR-US25-1-5p	miRNA	INHBA	mRNA	0.7311
RR00694900	H19	lncRNA	hsa-let-7b-5p	miRNA	1
RR01338124	SNHG16	lncRNA	hsa-let-7b-5p	miRNA	0.9856
RR00287286	CCAT1	lncRNA	hsa-let-7b-5p	miRNA	0.982
RR01472766	TP53COR1	lncRNA	hsa-let-7b-5p	miRNA	0.9526
RR00845074	LINC-ROR	lncRNA	hsa-let-7b-5p	miRNA	0.9526
RR00320272	CERNA2	lncRNA	hsa-let-7b-5p	miRNA	0.9526
RR05191717	WT1-AS	lncRNA	hsa-let-7b-5p	miRNA	0.7704
RR00719481	HOTTIP	lncRNA	hsa-miR-30b-5p	miRNA	0.9975
RR00032111	AC254633.1	lncRNA	hsa-miR-30b-5p	miRNA	0.9912
RR00715424	HNF1A-AS1	lncRNA	hsa-miR-30b-5p	miRNA	0.982
RR00889541	MALAT1	lncRNA	hsa-miR-30b-5p	miRNA	0.7311
RR01338152	SNHG16	lncRNA	hsa-miR-1207-3p	miRNA	0.9818
RR01159638	PVT1	lncRNA	hsa-miR-1207-3p	miRNA	0.8808

https://doi.org/10.1371/journal.pone.0261728.t001

CTHRC1 induces non-small cell lung cancer invasion by upregulating MMP-7/MMP-9 [31]. Moreover, in gastric cancer, *CTHRC1* was reported to promote tumor metastasis through the HIF-1α/CXCR4 signaling pathway [32]. *FNDC1* encodes a protein containing a fibronectin type III structural domain. Fibronectin interaction with integrins is involved in cell proliferation, migration, and differentiation [33]. Recent studies have found that *FNDC1* also has a role in different diseases including cancer. *FNDC1* was shown to be involved in the pathological changes in inflammatory bowel disease [34]. The silencing of *FNDC1* inhibited the proliferation and migration of prostate cancer cells [35]. *FNDC1* was also shown to promote apoptosis through hypermethylation in human salivary-like cystic carcinoma cells [36]. Additionally, high *FNDC1* expression was also reported to be associated with poor prognosis in gastric cancer [37], which is consistent with our findings. *INHBA* encodes a member of the TGF-beta superfamily of proteins, which has been found to be associated with various types of human cancers. Previous studies have shown that besides being associated with cell proliferation and migration, the *INHBA* gene is overexpressed in various tumors, such as colorectal cancer [38], esophageal cancer [39], and nasopharyngeal cancer [40]. All the above studies similarly indicate that these three genes may serve as potential diagnostic and prognostic biomarkers for gastric cancer. However, few studies have investigated the important upstream mediators of these hub genes. In this study, we made predictions about the upstream regulatory mechanisms of these genes using the RNAinter database in combination with the strong experimental evidence and established a regulatory network of lncRNAs-miRNAs-hub genes of GC, involving 12 lncRNAs, 5 miRNAs, and 3 hub genes, such as models HOTTIP-miR30b-*CTHRC1*, H19-let7b-*CTHRC1*, and PVT1-miR1207-*FNDC1*.

Non-coding RNAs, such as lncRNAs, circRNAs, and miRNAs, which were considered as transcriptional noise in the past and are now known to account for over 90% of the human genome [41], have been shown to have regulatory roles in various biological processes and play a crucial role in the development of diseases [42]. In recent years, considerable attention has been devoted to some calculation methods for predicting the potential associations of miRNAs, lncRNAs, circRNAs, and diseases as they can provide the most promising reference for the experiment, greatly reducing the time and cost of the experiment [43–45]. For instance, Chen *et al.* proposed Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction (MDHGI) by combining the sparse learning method with the heterogeneous graph inference method to calculate and predict the association of potential miRNA and disease [46]. Additionally, Chen *et al.* also developed a model of Inductive Matrix Completion for MiRNA-Disease Association prediction (IMCMDA), which was successfully validated in five human tumors [47]. Advances in interaction prediction research in various fields of computational biology have also provided valuable insights for the development of mRNA-miRNA-lncRNA networks, such as miRNA-lncRNA interaction prediction. Zhang *et al.* constructed the LMI-INGI and NDALMA models to predict the interactions between lncRNAs and miRNAs, and obtained satisfactory results in five-fold cross-validation, showing good prediction performance [48, 49]. LMFNRLMI is another algorithm which can achieve a good prediction of the relationship between lncRNA and miRNA [50]. Liu et al. proposed the "IMBDANET" algorithm that can predict genes directly or indirectly related to the target gene [51]. In the future, we can try to use these computational models to identify non-coding RNA biomarkers for gastric cancer and explore potential regulatory networks.

However, this study has several limitations, including the following. First, some key genes may have been removed during the gene filtering performed before performing WGCNA analysis. Second, when analyzing DEGs, some factors were not considered, such as age, sex, tumor staging, and patient classification. Finally, although the upstream regulatory network of the three hub genes has been predicted, experiments are still needed for further verification.

Conclusions

In conclusion, through comprehensive bioinformatics analysis, we successfully identified three hub genes (*CTHRC1*, *FNDC1*, and *INHBA*) that are differentially expressed between tumor and normal tissues in both TCGA-STAD and GSE66229 datasets and highly correlated with GC. These genes may play key roles in the development of gastric cancer. Their upstream lncRNA and miRNA regulators may reveal the potential mechanism by which these hub genes modulate the progression of GC. Overall, this study provides a new perspective on the diagnosis, prognosis, and treatment strategies for this malignant disease.

Supporting information

S1 Checklist.
(PDF)

S2 Checklist.
(DOCX)

Acknowledgments

We highly value the support of oncologists from Renmin Hospital of Wuhan University and the First People's Hospital of Xiangyang.

Author Contributions

Conceptualization: Gang Wei, Yongshun Chen.

Data curation: Gang Wei, Youhong Dong, Yongshun Chen.

Formal analysis: Gang Wei, Youhong Dong.

Funding acquisition: Yongshun Chen.

Investigation: Gang Wei, Youhong Dong.

Methodology: Gang Wei, Zhongshi He, Hu Qiu, Yongshun Chen.

Project administration: Yongshun Chen.

Resources: Gang Wei, Youhong Dong, Zhongshi He.

Software: Zhongshi He, Hu Qiu.

Supervision: Yongshun Chen.

Validation: Gang Wei, Yong Wu.

Visualization: Youhong Dong, Hu Qiu, Yong Wu.

Writing – original draft: Youhong Dong, Zhongshi He, Hu Qiu, Yong Wu.

Writing – review & editing: Gang Wei.

References

1. Thrift AP, El-Serag HB. Burden of Gastric Cancer. Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association. 2020; 18(3):534–542. Epub 2019/07/31. https://doi.org/10.1016/j.cgh.2019.07.045 PMID: 31362118.

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.

CA: a cancer journal for clinicians. 2021; 71(3):209–249. Epub 2021/02/05. https://doi.org/10.3322/caac.21660 PMID: 33538338.

3. Leite M, Marques MS, Melo J, Pinto MT, Cavadas B, Aroso M, et al. Helicobacter Pylori Targets the EPHA2 Receptor Tyrosine Kinase in Gastric Cells Modulating Key Cellular Functions. Cells. 2020; 9(2). Epub 2020/02/28. https://doi.org/10.3390/cells9020513 PMID: 32102381.
4. Chang WL, Yeh YC, Sheu BS. The impacts of H. pylori virulence factors on the development of gastroduodenal diseases. Journal of biomedical science. 2018; 25(1):68. Epub 2018/09/13. https://doi.org/10.1186/s12929-018-0466-9 PMID: 30205817.
5. Camargo MC, Murphy G, Koriyama C, Pfeiffer RM, Kim WH, Herrera-Goepfert R, et al. Determinants of Epstein-Barr virus-positive gastric cancer: an international pooled analysis. British journal of cancer. 2011; 105(1):38–43. Epub 2011/06/10. https://doi.org/10.1038/bjc.2011.215 PMID: 21654677.
6. Kim J, Cho YA, Choi WJ, Jeong SH. Gene-diet interactions in gastric cancer risk: a systematic review. World journal of gastroenterology. 2014; 20(28):9600–10. Epub 2014/07/30. https://doi.org/10.3748/wjg.v20.i28.9600 PMID: 25071358.
7. Lai HT, Koriyama C, Tokudome S, Tran HH, Tran LT, Nandakumar A, et al. Waterpipe Tobacco Smoking and Gastric Cancer Risk among Vietnamese Men. PloS one. 2016; 11(11):e0165587. Epub 2016/11/02. https://doi.org/10.1371/journal.pone.0165587 PMID: 27802311.
8. Kang GH, Lee S, Kim JS, Jung HY. Profile of aberrant CpG island methylation along the multistep pathway of gastric carcinogenesis. Laboratory investigation; a journal of technical methods and pathology. 2003; 83(5):635–41. Epub 2003/05/15. https://doi.org/10.1097/01.lab.0000067481.08984.3f PMID: 12746473.
9. Yoon HH, Shi Q, Sukov WR, Wiktor AE, Khan M, Sattler CA, et al. Association of HER2/ErbB2 expression and gene amplification with pathologic features and prognosis in esophageal adenocarcinomas. Clinical cancer research: an official journal of the American Association for Cancer Research. 2012; 18(2):546–54. Epub 2012/01/19. https://doi.org/10.1158/1078-0432.CCR-11-2272 PMID: 22252257.
10. Xiang Z, Li J, Song S, Wang J, Cai W, Hu W, et al. A positive feedback between IDO1 metabolite and COL12A1 via MAPK pathway to promote gastric cancer metastasis. Journal of experimental & clinical cancer research: CR. 2019; 38(1):314. Epub 2019/07/19. https://doi.org/10.1186/s13046-019-1318-5 PMID: 31315643.
11. Wu S, Chen M, Huang J, Zhang F, Lv Z, Jia Y, et al. ORAI2 Promotes Gastric Cancer Tumorigenicity and Metastasis through PI3K/Akt Signaling and MAPK-Dependent Focal Adhesion Disassembly. Cancer research. 2021; 81(4):986–1000. Epub 2020/12/15. https://doi.org/10.1158/0008-5472.CAN-20-0049 PMID: 33310726.
12. Zhou Z, Cheng Y, Jiang Y, Liu S, Zhang M, Liu J, et al. Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. International journal of biological sciences. 2018; 14(2):124–136. Epub 2018/02/28. https://doi.org/10.7150/ijbs.22619 PMID: 29483831.
13. Zheng MJ, Li X, Hu YX, Dong H, Gou R, Nie X, et al. Identification of molecular marker associated with ovarian cancer prognosis using bioinformatics analysis and experiments. Journal of cellular physiology. 2019; 234(7):11023–11036. Epub 2019/01/12. https://doi.org/10.1002/jcp.27926 PMID: 30633343.
14. Segundo-Val IS, Sanz-Lozano CS. Introduction to the Gene Expression Analysis. Methods in molecular biology (Clifton, NJ). 2016; 1434:29–43. Epub 2016/06/15. https://doi.org/10.1007/978-1-4939-3652-6_3 PMID: 27300529.
15. Guo Y, Bao Y, Ma M, Yang W. Identification of Key Candidate Genes and Pathways in Colorectal Cancer by Integrated Bioinformatical Analysis. International journal of molecular sciences. 2017; 18(4). Epub 2017/03/30. https://doi.org/10.3390/ijms18040722 PMID: 28350360.
16. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008; 9:559. Epub 2008/12/31. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008.
17. Yin L, Cai Z, Zhu B, Xu C. Identification of Key Pathways and Genes in the Dynamic Progression of HCC Based on WGCNA. Genes. 2018; 9(2). Epub 2018/02/15. https://doi.org/10.3390/genes9020092 PMID: 29443924.
18. Song ZY, Chao F, Zhuo Z, Ma Z, Li W, Chen G. Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis. Aging. 2019; 11(13):4736–4756. Epub 2019/07/16. https://doi.org/10.18632/aging.102087 PMID: 31306099.
19. Niemira M, Collin F, Szalkowska A, Bielska A, Chwialkowska K, Reszec J, et al. Molecular Signature of Subtypes of Non-Small-Cell Lung Cancer by Large-Scale Transcriptional Profiling: Identification of Key Modules and Genes by Weighted Gene Co-Expression Network Analysis (WGCNA). Cancers. 2019; 12(1). Epub 2019/12/28. https://doi.org/10.3390/cancers12010037 PMID: 31877723.

20. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC bioinformatics. 2011; 12:35. Epub 2011/01/29. https://doi.org/10.1186/1471-2105-12-35 PMID: 21269502.

21. Lin Y, Liu T, Cui T, Wang Z, Zhang Y, Tan P, et al. RNAInter in 2020: RNA interactome repository with increased coverage and annotation. Nucleic acids research. 2020; 48(D1):D189–d97. Epub 2020/01/08. https://doi.org/10.1093/nar/gkz804 PMID: 31906603.

22. Esquela-Kerscher A, Slack FJ. Oncomirs—microRNAs with a role in cancer. Nature reviews Cancer. 2006; 6(4):259–269. Epub 2006/03/25. https://doi.org/10.1038/nrc1840 PMID: 16557279.

23. Thomson DW, Dinger ME. Endogenous microRNA sponges: evidence and controversy. Nature reviews Genetics. 2016; 17(5):272–83. Epub 2016/04/05. https://doi.org/10.1038/nrg.2016.20 PMID: 27040487.

24. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic acids research. 2017; 45(W1):W98–w102. Epub 2017/04/14. https://doi.org/10.1093/nar/gkx247 PMID: 28407145.

25. Szász AM, Lánczky A, Nagy Á, Förster S, Hark K, Green JE, et al. Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. Oncotarget. 2016; 7(31):49322–49333. Epub 2016/07/08. https://doi.org/10.18632/oncotarget.10337 PMID: 27384994.

26. Tilghman RW, Parsons JT. Focal adhesion kinase as a regulator of cell tension in the progression of cancer. Seminars in cancer biology. 2008; 18(1):45–52. Epub 2007/10/12. https://doi.org/10.1016/j.semcancer.2007.08.002 PMID: 17928235.

27. Zeng X, Ju D. Hedgehog Signaling Pathway and Autophagy in Cancer. International journal of molecular sciences. 2018; 19(8). Epub 2018/08/08. https://doi.org/10.3390/ijms19082279 PMID: 30081498.

28. Seoane J, Gomis RR. TGF-β Family Signaling in Tumor Suppression and Cancer Progression. Cold Spring Harbor perspectives in biology. 2017; 9(12). Epub 2017/03/02. https://doi.org/10.1101/cshperspect.a022277 PMID: 28246180.

29. Pyagay P, Heroult M, Wang Q, Lehnert W, Belden J, Liaw L, et al. Collagen triple helix repeat containing 1, a novel secreted protein in injured and diseased arteries, inhibits collagen expression and promotes cell migration. Circulation research. 2005; 96(2):261–268. Epub 2004/12/25. https://doi.org/10.1161/01.RES.0000154262.07264.12 PMID: 15618538.

30. Zhang R, Lu H, Lyu YY, Yang XM, Zhu LY, Yang GD, et al. E6/E7-P53-POU2F1-CTHRC1 axis promotes cervical cancer metastasis and activates Wnt/PCP pathway. Scientific reports. 2017; 7:44744. Epub 2017/03/18. https://doi.org/10.1038/srep44744 PMID: 28303973.

31. He W, Zhang H, Wang Y, Zhou Y, Luo Y, Cui Y, et al. CTHRC1 induces non-small cell lung cancer (NSCLC) invasion through upregulating MMP-7/MMP-9. BMC cancer. 2018; 18(1):400. Epub 2018/04/11. https://doi.org/10.1186/s12885-018-4317-6 PMID: 29631554.

32. Ding X, Huang R, Zhong Y, Cui N, Wang Y, Weng J, et al. CTHRC1 promotes gastric cancer metastasis via HIF-1α/CXCR4 signaling pathway. Biomedicine & pharmacotherapy=Biomedecine & pharmacotherapie. 2020; 123:109742. Epub 2019/12/20. https://doi.org/10.1016/j.biopha.2019.109742 PMID: 31855733.

33. Gao M, Craig D, Lequin O, Campbell ID, Vogel V, Schulten K. Structure and functional significance of mechanically unfolded fibronectin type III1 intermediates. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100(25):14784–14789. Epub 2003/12/06. https://doi.org/10.1073/pnas.2334390100 PMID: 14657397.

34. Wuensch T, Wizenty J, Quint J, Spitz W, Bosma M, Becker O, et al. Expression Analysis of Fibronectin Type III Domain-Containing (FNDC) Genes in Inflammatory Bowel Disease and Colorectal Cancer. Gastroenterology research and practice. 2019; 2019:3784172. Epub 2019/05/17. https://doi.org/10.1155/2019/3784172 PMID: 31093274.

35. Das DK, Naidoo M, Ilboudo A, Park JY, Ali T, Krampis K, et al. miR-1207-3p regulates the androgen receptor in prostate cancer via FNDC1/fibronectin. Experimental cell research. 2016; 348(2):190–200. Epub 2016/10/25. https://doi.org/10.1016/j.yexcr.2016.09.021 PMID: 27693493.

36. Bell A, Bell D, Weber RS, El-Naggar AK. CpG island methylation profiling in human salivary gland adenoid cystic carcinoma. Cancer. 2011; 117(13):2898–909. Epub 2011/06/22. https://doi.org/10.1002/cncr.25818 PMID: 21692051.

37. Zhong M, Zhang Y, Yuan F, Peng Y, Wu J, Yuan J, et al. High FNDC1 expression correlates with poor prognosis in gastric cancer. Experimental and therapeutic medicine. 2018; 16(5):3847–54. Epub 2018/11/08. https://doi.org/10.3892/etm.2018.6731 PMID: 30402143.

38. Okano M, Yamamoto H, Ohkuma H, Kano Y, Kim H, Nishikawa S, et al. Significance of INHBA expression in human colorectal cancer. Oncology reports. 2013; 30(6):2903–2908. Epub 2013/10/03. https://doi.org/10.3892/or.2013.2761 PMID: 24085226.

39. Seder CW, Hartojo W, Lin L, Silvers AL, Wang Z, Thomas DG, et al. INHBA overexpression promotes cell proliferation and may be epigenetically regulated in esophageal adenocarcinoma. Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer. 2009; 4(4):455–462. Epub 2009/02/26. https://doi.org/10.1097/JTO.0b013e31819c791a PMID: 19240652.

40. Peng S, Wang J, Hu P, Zhang W, Li H, Xu L. INHBA knockdown inhibits proliferation and invasion of nasopharyngeal carcinoma SUNE1 cells in vitro. International journal of clinical and experimental pathology. 2020; 13(5):854–868. Epub 2020/06/09. PMID: 32509056.

41. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447(7146):799–816. Epub 2007/06/16. https://doi.org/10.1038/nature05874 PMID: 17571346.

42. Adams BD, Parsons C, Walker L, Zhang WC, Slack FJ. Targeting noncoding RNAs in disease. The Journal of clinical investigation. 2017; 127(3):761–771. Epub 2017/03/02. https://doi.org/10.1172/JCI84424 PMID: 28248199 7,893,034) in, and is a consultant and advisory board member for, Mirna Therapeutics. B.D. Adams holds patent interests with, and consults with, AUM LifeTech.

43. Chen X, Yan CC, Zhang X, You ZH. Long non-coding RNAs and complex diseases: from experimental results to computational models. Briefings in bioinformatics. 2017; 18(4):558–576. Epub 2016/06/28. https://doi.org/10.1093/bib/bbw060 PMID: 27345524.

44. Chen X, Xie D, Zhao Q, You ZH. MicroRNAs and complex diseases: from experimental results to computational models. Briefings in bioinformatics. 2019; 20(2):515–539. Epub 2017/10/19. https://doi.org/10.1093/bib/bbx130 PMID: 29045685.

45. Wang CC, Han CD, Zhao Q, Chen X. Circular RNAs and complex diseases: from experimental results to computational models. Briefings in bioinformatics. 2021; 22(6). Epub 2021/07/31. https://doi.org/10.1093/bib/bbab286 PMID: 34329377.

46. Chen X, Yin J, Qu J, Huang L. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. PLoS computational biology. 2018; 14(8):e1006418. Epub 2018/08/25. https://doi.org/10.1371/journal.pcbi.1006418 PMID: 30142158.

47. Chen X, Wang L, Qu J, Guan NN, Li JQ. Predicting miRNA-disease association based on inductive matrix completion. Bioinformatics (Oxford, England). 2018; 34(24):4256–4265. Epub 2018/06/26. https://doi.org/10.1093/bioinformatics/bty503 PMID: 29939227.

48. Zhang L, Yang P, Feng H, Zhao Q, Liu H. Using Network Distance Analysis to Predict lncRNA-miRNA Interactions. Interdisciplinary sciences, computational life sciences. 2021; 13(3):535–545. Epub 2021/07/08. https://doi.org/10.1007/s12539-021-00458-z PMID: 34232474.

49. Zhang L, Liu T, Chen H, Zhao Q, Liu H. Predicting lncRNA-miRNA interactions based on interactome network and graphlet interaction. Genomics. 2021; 113(3):874–880. Epub 2021/02/16. https://doi.org/10.1016/j.ygeno.2021.02.002 PMID: 33588070.

50. Liu H, Ren G, Chen H, Liu Q, Yang Y, Zhao Q. Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized. Knowledge-Based Systems. 2020; 191. https://doi.org/10.1016/j.knosys.2019.105261

51. Liu W, Jiang Y, Peng L, Sun X, Gan W, Zhao Q, et al. Inferring Gene Regulatory Networks Using the Improved Markov Blanket Discovery Algorithm. Interdisciplinary sciences, computational life sciences. 2021. Epub 2021/09/09. https://doi.org/10.1007/s12539-021-00478-9 PMID: 34495484.