Functional investigation and two-sample Mendelian randomization study of Early gastric cancer (EGC) hub genes obtained by WGCNA analysis

Xiao-Jun Ren
rxiaojun2016@hotmail.com

Hubei University of Chinese Medicine

Man-Ling Zhang
Hubei University of Chinese Medicine

Hui-Hui Zhao
Hubei University of Chinese Medicine

Zhao-Hong Shi
Hubei University of Chinese Medicine

Research Article

Keywords: Early gastric cancer, Mendelian randomization analysis, CLU

Posted Date: February 26th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3974859/v1

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

Additional Declarations: No competing interests reported.
Abstract

Objective: This study aimed to identify differentially expressed genes associated with early gastric cancer and analyze their potential functions through bioinformatics analysis. Additionally, we sought to validate the genetic causality of identified differential molecules related to early gastric cancer using Mendelian randomization.

Methods: We analyzed the expression profiles of 19 gastritis specimens and 20 early gastric cancer specimens (high-grade intraepithelial neoplasia) from the GEO database using combined differential gene expression analysis and weighted gene co-expression network analysis (WGCNA). Hub genes most relevant to early gastric cancer were selected and subjected to functional enrichment and signaling pathway analysis using GO and KEGG enrichment. The diagnostic efficacy of the predictive model based on the top five ranked hub genes was evaluated using ROC curves. Finally, Mendelian randomization analysis was employed to assess genetic causality between hub genes and early gastric cancer.

Results: A total of 755 hub genes were identified after intersecting the most relevant modular genes from WGCNA with genes exhibiting significant expression differences. GO and KEGG pathway enrichment analyses revealed significant differential expression of hub genes in responses to xenobiotic stimuli, hormone transport, apical cell parts, and oxidoreductase activity targeting the CH-OH group of donors, as well as involvement in pluripotency of stem cells, hepatocellular carcinoma, and axon guidance. The top five core genes—IL6, CLU, UGT2B15, NANOG, and NFE2L2—were analyzed using Cytoscape software. The column-linear graph prediction model demonstrated excellent performance in predicting the risk of early gastric cancer, as evidenced by ROC analysis. In the inverse variance weighting (IVW) method, we found that the core gene CLU was associated with an increased risk of early gastric cancer (OR = 1.157, 95% CI = 1.043-1.283, p = 0.0057).

Conclusion: Our bioinformatics analysis identified the CLU gene as genetically causally associated with early gastric cancer, suggesting its potential as a diagnostic or therapeutic target for this condition.

1. Introduction

Gastric cancer presents a significant global health burden, characterized by elevated morbidity and mortality rates, posing substantial threats to individuals' well-being[1]. The absence of specific early diagnostic markers often results in gastric cancers being diagnosed at advanced stages with distant metastases[2]. Particularly in East Asian nations, notably China, gastric cancer exhibits alarming incidence rates, high mortality, and inadequate early detection rates[3]. The pathogenesis of gastric cancer is complex, involving multiple stages from chronic gastritis to the development of differentiated gastric cancer[4–7]. Pre-cancerous lesions such as atrophy, intestinal metaplasia, and dysplasia serve as pivotal transitions from inflammatory states to carcinogenesis[8], with early gastric cancer representing a progression of these precancerous states. Notably, early gastric cancer confined to the mucosal or submucosal layers offers a higher curative potential and lower propensity for lymph node metastasis[9].
Therefore, effective early diagnosis and treatment are imperative to mitigate the progression to advanced stages, enhance cure rates, and improve patient survival outcomes.

Current screening strategies for early gastric cancer primarily rely on Helicobacter pylori testing and gastroscopy, yet their implementation, especially in populous developing countries like China, poses logistical challenges\(^\text{[10,11]}\). While serological markers like gastrin 17 combined with pepsinogen show promise in detecting gastric precancerous lesions and cancer, their lack of specificity limits their utility in early gastric cancer diagnosis\(^\text{[12]}\). Consequently, there is a pressing need to identify specific markers pertinent to the diagnosis and treatment of early gastric cancer.

Mendelian randomization (MR) analysis emerges as a robust statistical method to ascertain causal relationships between exposures and outcomes, utilizing single nucleotide polymorphisms (SNPs) as instrumental variables to mitigate biases arising from confounding and reverse causation\(^\text{[13,14]}\). Leveraging the Gene Expression Omnibus (GEO) database, renowned for its extensive, accurate, and well-curated gene expression datasets\(^\text{[15]}\), this study aims to conduct bioinformatics analyses on high-quality gene expression profiles of early gastric cancer. Through Mendelian randomization analysis, we seek to evaluate potential genetic causal relationships between identified markers and early gastric cancer, thereby elucidating valuable molecular markers for early gastric cancer diagnosis and treatment.

2. Materials and methods

2.1. Data source

The dataset, sourced from the Gene Expression Omnibus (GEO) database, includes transcriptome microarray assays of chronic gastritis tissue specimens (N = 19) and high-grade intraepithelial neoplasia specimens (N = 20). Diagnosis of high-grade intraepithelial neoplasia was based on specimens obtained through surgical procedures or Endoscopic Submucosal Dissection (ESD).

2.2. Differentially expressed genes identification

We initially analyzed differential genes in the GSE55696 dataset using the "limma" package in R (version 3.6.1). Subsequently, we conducted pre-processing, applying batch correction and normalization to expression levels. Visualization employed the "pheatmap" and "ggplot2" R packages, creating volcano and heat maps to highlight significantly different genes. Selection criteria for significance included a > 2-fold expression difference between sample groups, with a corrected P Value < 0.05.

2.3. Weighted gene co-expression network analysis

WGCNA is a powerful bioinformatics tool that allows for the construction of gene co-expression networks, identifying modules of genes with correlated expression profiles across different samples. Utilizing the "WGCNA" R package, we constructed a gene co-expression network to identify modules of genes with correlated expression profiles in gastric mucosal high-grade intraepithelial neoplasia. We assessed the
correlation of these gene modules with the occurrence of high-grade intraepithelial neoplasia and identified the most relevant module as the focal point for subsequent WGCNA screening.

2.4. Screening of candidate pivotal genes and Go/KEGG analysis

Intersecting differential genes from chronic gastritis and high-grade intraepithelial neoplasia with WGCNA-derived modules identified candidate core genes relevant to gastric mucosal high-grade intraepithelial neoplasia. Subsequent GO and KEGG analyses using "clusterProfiler" provided insights into the associated biological processes and pathways, enhancing our understanding of the disease mechanism.

2.5. Protein–protein interaction network hub genes

We utilized the STRING online tool to assemble a protein-protein interaction (PPI) network for the identified pivotal genes. Subsequently, we employed Cytoscape software to assess the significance of the candidate pivotal genes based on their Degree scores. To further refine our analysis, we employed the CytoHubba plugin to identify and score the top 10 genes. Finally, the visual representation of these top-ranked genes was generated using the CytoHubba plugin.

2.6. Nomogram model construction

We employed the "rms" R package to construct a nomogram model incorporating the top five core genes. This model was designed for predicting the risk of developing high-grade intraepithelial neoplasia in gastric mucosa specimens. The effectiveness of the nomogram was assessed through the calculation of the Harrell consistency index. Additionally, the diagnostic accuracy of the core genes was evaluated by generating ROC curves using the "pROC" R package.

2.7. Immune cell analysis

To investigate the distribution of diverse immune cells in high-grade intraepithelial neoplasia within the gastric mucosa, we performed an analysis to gauge the infiltration levels of 22 distinct immune cell types. This assessment was conducted utilizing CIBERSORT analysis.

2.8. Mendelian randomization

The study utilized data sourced from publicly available databases. A two-sample Mendelian randomization analysis was employed to investigate potential causal relationships between core genes and high-grade intraepithelial neoplasia in the gastric mucosa. Single nucleotide polymorphisms (SNPs) were designated as instrumental variables (IVs). Core genes associated with IL-6, CLU, and gastric cancer were extracted from the Genome-Wide Association Study (GWAS) database for Mendelian randomization analysis, primarily utilizing the "TwoSampleMR" R package. The MR analysis was conducted through the "TwoSampleMR" R software package, employing inverse variance weighting (IVW) to assess the risk association between hub genes and the occurrence of high-grade intraepithelial neoplasia in the gastric mucosa. Additionally, sensitivity analysis was carried out using MR-Egger.
3. Results

3.1. DEGs screening

The dataset for high-grade intraepithelial neoplasia in gastric mucosa (GSE55696) was initially retrieved from the GEO database, followed by conducting a differential gene expression analysis. The results revealed 18,352 differentially expressed genes in high-grade intraepithelial neoplasia compared to the chronic gastritis group. Subsequent filtering, employing criteria of LogFC = 1 and Adj P Value < 0.05, identified 1777 genes with significant differences, including 776 up-regulated and 1001 down-regulated genes. Further analysis focused on the subset of significantly different genes. Visualization was carried out by generating heat maps (Fig. 1A) representing the top 50 genes with the most significant up- and down-regulation, while volcano maps (Fig. 2B) depicted genes with notable differences.

3.2. Construction of WGCNA network and identification of high-grade intraepithelial neoplasia module

To assess the potential correlation of gene modules with gastric mucosal high-grade intraepithelial neoplasia, WGCNA analysis was conducted on the entire set of differentially expressed genes within the GSE55696 dataset. A total of 24 gene modules were identified through WGCNA analysis. Following an examination of positive correlation coefficients, the turquoise module genes emerged as the most pertinent genes associated with high-grade gastric intraepithelial neoplasia.

3.3. Go/KEGG analyses

Upon intersecting WGCNA-derived hub genes with Differentially Expressed Genes (DEGs), the analysis yielded 775 candidate genes. Subsequent Gene Ontology (GO) enrichment analysis of the intersected genes revealed predominant associations with responses to xenobiotic stimuli, hormone transport, apical cell parts, and oxidoreductase activity targeting the CH-OH group of donors. Additionally, Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis highlighted significant impacts on signaling pathways regulating pluripotency of stem cells, hepatocellular carcinoma, and axon guidance for the overlapping genes.

3.4. PPI network analysis for hub genes

Protein-Protein Interaction (PPI) networks for overlapping genes were constructed using the String online web tool. Subsequently, the CytoHubba plugin facilitated the visualization of the top ten core genes, identified based on their degree score. Notably, IL6, CLU, NANOG, UGT2B15, NFE2L2, CXCL8, APOA2, LIF, HGF, and PER2 were discerningly chosen as the 10 core genes with the highest composite scores. Additionally, it is noteworthy that the darker coloration of the genes within the network corresponds to higher scores.
3.5. Construction of nomogram model for High-grade gastric intraepithelial neoplasia

A nomogram model was constructed to predict the risk of developing high-grade gastric intraepithelial neoplasia. The results of the nomogram model demonstrated strong predictive performance. Subsequently, we assessed the diagnostic efficacy of the top five core genes (IL6, CLU, NANOG, UGT2B15, NFE2L2) using ROC curves. The AUC values for IL6, CLU, NANOG, UGT2B15, and NFE2L2 were 0.928, 0.848, 0.850, 0.975, and 0.922, respectively, indicating excellent diagnostic efficacy.

3.6. CLU was causally associated with the risk of High-grade gastric intraepithelial neoplasia

In Fig. 6A and 6B, the impact of each genetic variant on high-grade gastric intraepithelial neoplasia is depicted. Specifically, we investigated the relationship between CLU and high-grade intraepithelial neoplasia. Employing the IVW method, we identified a causal link between CLU and the development of high-grade intraepithelial neoplasia (OR = 1.157, 95% CI = 1.043–1.283, p = 0.005). The mR-Egger method yielded non-significant results (OR = 1.042, 95% CI = 0.753–1.443, p = 0.807). Assessment of the funnel plot (Fig. 6C) indicated symmetrical distribution, and the MR Egger regression intercept did not reveal horizontal pleiotropy (p = 0.529), affirming the absence of bias from pleiotropic effects. Systematic MR analysis, removing each SNP individually (Fig. 6D), consistently upheld the causal significance, suggesting no dominant SNP influencing CLU levels and high-grade intraepithelial neoplasia. These findings validate the robustness of the previous MR results.

4. Discussion

Gastric cancer ranks as a major cancer worldwide, with the fifth highest incidence and the fourth highest mortality rate[16]. The classical theory proposed by Correa suggests that the formation of gastric cancer is a multi-stage and multi-step process[6]. Early gastric cancer, being confined to the mucosal layer or submucosal superficial layer, has a lower risk of lymph node metastasis and, thus, a higher cure rate[17]. However, early gastric cancer usually does not cause symptoms[18] and lacks specific diagnostic biomarkers[19], leading to most patients being diagnosed at an advanced stage of gastric cancer[20]. Currently, screening for early gastric cancer relies primarily on detailed gastroscopic examination under high-definition endoscopy and targeted biopsy[21]. The reduction in the incidence and mortality rates of gastric cancer in developed countries such as Japan and South Korea has proven the effectiveness of this screening method[22, 23]. However, for populous developing countries like China, the strategy of using high-definition endoscopy and biopsy for early gastric cancer screening faces challenges due to high medical costs and low population coverage[24]. Therefore, screening for biomarkers with specific diagnostic value has significant clinical implications.
In this study, we downloaded and analyzed GSE55696 mRNA-seq data, identifying 18,352 differentially expressed genes between the chronic gastritis group and the early gastric cancer (high-grade gastric mucosal intraepithelial neoplasia) group. By filtering these genes with a Log Fold Change of 1 and an Adj P value < 0.05, we ultimately selected 1,777 genes with significant differential expression, including 776 upregulated and 1,001 downregulated genes. Weighted correlation network analysis (WGCNA) is a systems biology method used to describe gene association patterns between different samples, enabling the identification of highly co-variated gene sets\cite{25}. Through WGCNA, we identified 2,520 core genes related to early gastric cancer. We then intersected these 1,777 significantly differentially expressed genes with the 2,520 core genes identified by WGCNA, obtaining 755 genes strongly related to early gastric cancer. GO and KEGG analyses were employed to further perform functional clustering analysis and pathway analysis on the intersected genes. GO analysis indicated that the intersected genes are mainly involved in biological processes such as response to xenobiotic stimulus, hormone transport, hormone secretion, and peptide transport; primarily targeting the apical part of the cell and apical plasma membrane; and participating in molecular functions including oxidoreductase activity acting on the CH-OH group of donors, with NAD or NADP as acceptor, and monooxygenase activity. KEGG results showed that the intersected genes are mainly related to signaling pathways regulating the pluri-potency of stem cells, hepatocellular carcinoma, and retinol metabolism.

Mendelian randomization analysis is an important method for verifying causal relationships. By conducting Mendelian randomization analysis on the ten most relevant core genes selected through the CytoHubba plugin, we discovered a genetic causal relationship between the core gene CLU and early gastric cancer. Clusterin protein, first identified and isolated in rat testes in 1979\cite{26}, has been found to be ubiquitously present in almost all bodily fluids and the intracellular matrix, performing a variety of biological functions\cite{27}. Named CLU for its cell aggregation function in vitro\cite{28}, it has been found that humans have at least three isoforms of CLU\cite{29}. Under certain stress conditions, immature Clusterin can be converted into a mature or nuclear form of about 55kDa and relocate to the nucleus, exerting pro-apoptotic effects through a caspase 3-dependent pathway\cite{30}. The function of cytoplasmic Clusterin is more complex; under certain stress conditions, pre-secreted clusterin can bind with GRP87 (Bip) in the endoplasmic reticulum to increase its stability, then enter mitochondria to inhibit the formation of the Bax-Bak complex, thereby exerting anti-apoptotic effects\cite{31}. Secreted clusterin (sCLU) has been proven to be an 80kDa glycoprotein\cite{32}, an important extracellular chaperone molecule. It plays protective roles in cells and tissues by clearing cell debris and misfolded proteins, inducing cell survival and proliferation pathways\cite{33}, and is implicated in the development of diseases such as Alzheimer’s\cite{34}, atherosclerosis, and malignant tumors\cite{35}.

Previous studies have shown that Clusterin acts as either an oncogene or a tumor suppressor gene in different cancers, indicating that the CLU gene plays a dual role in tumorigenesis\cite{36}. Reports suggest that overexpression of CLU in gastric cancer is associated with lymph node metastasis, tumor invasion, and high tumor stage\cite{37}. However, literature on the value of blood-secreted Clusterin as a biomarker for gastric cancer has yielded contradictory results\cite{38}. In our study, we found that CLU is underexpressed in
early gastric cancer. We hypothesize that this underexpression indicates a reduction in its function to clear cancer-related proteins, leading to tumor development. However, Mendelian randomization analysis revealed a genetic causal relationship between CLU and gastric cancer, identifying it as a risk factor for gastric cancer. This finding, which seems contradictory to the underexpression of CLU in early gastric cancer, further highlights the multifunctionality of CLU. It prompts further investigation into whether CLU exhibits differential expression and functionality between early and advanced stages of gastric cancer, or whether different isoforms of CLU perform different functions at these stages. This will likely be a focus of our research team's future studies.

Regardless, this is the first study to explore the causal relationship between CLU levels and gastric cancer using GWAS data for CLU (exposure) and early gastric cancer (outcome) through bidirectional Mendelian randomization analysis. This analysis indicates a genetic causal relationship between CLU and gastric cancer, identifying CLU as a risk factor for the development of gastric cancer. Despite the significance of this finding for the diagnosis of early gastric cancer, the study has its limitations. Firstly, although we found CLU to be underexpressed in early gastric cancer with significant weight, the Mendelian randomization analysis was ultimately conducted using gastric cancer GWAS data due to the lack of early gastric cancer GWAS data. Secondly, this study only utilized bioinformatics analysis to examine hub genes and potential functions related to the onset of early gastric cancer. Further research will require validation of CLU's specific mechanisms through cellular and animal model functional experiments, as well as exploration of whether differences exist in the expression and function of CLU at different stages of gastric cancer.

**Declarations**

**Author Contribution**

Author contributions: Xiao-Jun Ren and Zhao-Hong Shi designed the study and revised the manuscript. Xiao-Jun Ren and Man-Ling Zhang acquired the raw data and analyzed the data. Xiao-Jun Ren and Hui-Hui Zhao wrote the manuscript. All authors contributed to the article and approved the submitted version.

**Ethics approval and consent to participate:** Not applicable.

**Consent for publication:** Not applicable.

**Availability of data and materials:** The datasets generated and/or analysed during the current study are available in the Gene Expression Omnibus (GEO).

**Conflict of interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Acknowledgements:** The authors acknowledge all individuals who contributed to this study.
References


Figures

Figure 1

depicts the differential expression analysis results between high-grade intraepithelial neoplasia and chronic gastritis in GSE55696. (A) illustrates the volcanic map, showcasing genes with significant differential expression. (B) the heat map visualizes the differential expression analysis, where blue indicates down-regulated genes, red denotes up-regulated genes, and black signifies no differentially expressed genes.

3.2. Construction of WGCNA network and identification of high-grade intraepithelial neoplasia module
Figure 2 illustrates the WGCNA identification of gene modules associated with gastric mucosal high-grade intraepithelial neoplasia in the GSE55696 dataset. (A) presents a clustering tree of all differential genes, with each branch representing a gene and various modules of co-expressed genes indicated by distinct colors. In Panel (B), a heatmap displays the correlation between clustered modules of different genes and gastric mucosal high-grade intraepithelial neoplasia, with each module accompanied by its corresponding correlation coefficient and p-value. (C) features a scatterplot highlighting the turquoise modules, demonstrating the strongest positive correlation with gastric mucosal high-grade intraepithelial neoplasia in the dataset.
Figure 3 illustrates the identification and validation of candidate hub genes. Panel (A) displays a Venn diagram showcasing 775 overlapping candidate hub genes. Panel (B) presents the results of Gene Ontology (GO) enrichment analysis for these candidate hub genes, while Panel (C) outlines the findings of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis for the same set of candidate hub genes.

3.4. PPI network analysis for hub genes
Figure 4

depicts the construction of the Protein-Protein Interaction (PPI) network. Subfigure (A) illustrates the PPI network of overlapping hub genes, while subfigure (B) highlights the core genes obtained through the degree algorithm within the interaction network.
Figure 5

illustrates the prediction of high-grade intraepithelial neoplasia risk in the gastric mucosa through nomograms. Subfigure (A) displays the nomogram model incorporating hub genes, while Subfigure (B) showcases ROC curves evaluating the diagnostic efficacy of both the nomogram model and individual hub genes.

Figure 6

presents the results of the Mendelian randomization study. Subfigure (A) features a scatter plot illustrating the causal effect of CLU on the risk of high-grade gastric intraepithelial neoplasia. In Subfigure (B), a forest plot displays the individual causal effects of each SNP on the risk of trigeminal neuralgia. Subfigure (C) showcases funnel plots, offering insight into the overall heterogeneity of Mendelian randomization estimates for the effect of CLU on high-grade gastric intraepithelial neoplasia. Finally, Subfigure (D) presents a leave-one-out plot, providing a visual representation of the causal effect of CLU on the risk of high-grade gastric intraepithelial neoplasia when each SNP is sequentially omitted.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- table.MRresult.csv
- table.SNP.csv
- table.heterogeneity.csv
- table.pleiotropy.csv