Detecting Hub Genes Associated With Lauren Subtype of Gastric Cancer by Weighted Gene Co-Expression Network Analysis

XU LIU
Xi’an Jiaotong University  https://orcid.org/0000-0003-4050-9939

Li Yao
China XD Group Hospital

Jingkun Qu
Xi’an Jiaotong University

Lin Liu
Xi’an Jiaotong University

XU LIU
xi’an medical University

Jiansheng Wang
Xi’an Jiaotong University

George A. Calin
University of Texas MD Anderson Cancer Center

Jia Zhang (✉ zhangjiaxjtu@163.com)
Xi’an Jiaotong University

Research article

Keywords: gastric cancer, Lauren subtype, WGCNA, hub gene

DOI: https://doi.org/10.21203/rs.3.rs-61626/v1

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

Background Gastric cancer is a rather heterogeneous type of malignant tumor. Among the several classification system, Lauren classification can reflect biological and pathological differences of different gastric cancer.

Method to provide systematic biological perspectives, we employ weighted gene co-expression network analysis to reveal transcriptomic characteristics of gastric cancer. GSE15459 and TCGA STAD dataset were downloaded. Co-expressional network was constructed and gene modules were identified.

Result Two key modules blue and red were suggested to be associated with diffuse gastric cancer. Functional enrichment analysis of genes from the two modules was performed. Validating in TCGA STAD dataset, we propose 10 genes TNS1, PGM5, CPXM2, LIMS2, AOC3, CRYAB, ANGPTL1, BOC and TOP2A to be hub-genes for diffuse gastric cancer. Finally these ten genes were associated with gastric cancer survival.

Conclusion More attention need to be paid and further experimental study is required to elucidate the role of these genes.

Background

Gastric cancer (GC) is the fifth most common cancer globally. In 2018, gastric cancer-caused death ranks the third among all kinds of cancer (1). Multiple interactions of genetic, environmental and host factors brings tremendous complexity and heterogeneity to gastric cancer (2). With regard to medical management, several classification systems were developed, such as Lauren classification, WHO classification, Goseki & Ming classification. Based on multiple omics data, Asian Cancer Research Group (ACRG) and The Cancer Genome Atlas (TCGA) have also issued their own molecular subtype category system (2, 3). However, Lauren classification is still the most common used as it could reflect biological differences and individual subtypes within this classification do not transform into others (4).

In 1965, Lauren first described two distinct histologic types of gastric cancer: intestinal and diffuse(5). The intestinal type pathologically exhibits components of glandular or intestinal architecture which is thought to emerge from a sequence of multistep carcinogenesis often caused by chronic mucosal inflammation due to Helicobacter pylori (H. pylori) infection. This type is more common in male and older patients. The diffuse type manifests as poorly cohesive cells infiltrating the gastric wall, and can be a progressive disease ultimately leading to linitis plastica. This type is more common in female and younger patients and frequently associated with familial history (6, 7). A recent research reported Lauren classification as an indicator for survival and response to chemotherapy in advanced gastric cancer (8). Lauren subtype is even an important factor associated with pattern of recurrence following resection of gastric adenocarcinoma (8). Thus, uncovering the molecular mechanism of Lauren subtype may facilitate further understanding the pathogenesis of GC and providing potential therapy target (9). Although several typical oncogenes such as CDH1, KRAS and RHOA has been related with particular subtype, it is still unclear how the intestinal or diffused GCs are driven by particular genes (10-12).
Weighted gene co-expression network analysis (WGCNA) is a new systems biological method providing insights into gene networks that might be responsible for phenotypic traits of interest (13, 14). In this study, we perform an integrated bio-informatic analysis by using this algorism to identify hub genes for diffuse and intestinal GC based on transcriptomic data from GEO datasets and The Cancer Genome Atlas (TCGA). Furthermore, we found 10 genes that were expressed in diffuse gastric cancer at particular level and correlated with the overall survival of GC.

Materials And Methods

Data acquisition and study design

Two datasets GSE15459 and TCGA stomach adenocarcinoma gene expression data were used in this study. Gene expression matrix and clinical information of Dataset GSE15459 was download from Gene Expression Omnibus database of NCBI (https://www.ncbi.nlm.nih.gov/gds/). TCGA level 3 data and clinical phenotypic data were downloaded from UCSC portal (https://xenabrowser.net/datapages/). The data analysis flow is showed in Figure1. Briefly, GSE154559 data was normalized and genes of which expression variation in all samples rank top 50% were selected to construct gene module by WGCNA. Then, the relation between each module and clinical characteristics was analyzed. Hub gene from modules that has the strongest correlation with diffuse subtype were picked as candidate genes for later survival analysis and validated in TCGA dataset.

Construction of Co-expression Networks

The gene co-expression networks were constructed with WGCNA package in R program(15). After constructing a sample cluster tree to detect the outlying samples, we used the pickSoftThereshold function to calculate the best power (β) value to get a scale-free topology network in which the best power value was selected to generate a weighted adjacency matrix. After all, automatic module detection via dynamic tree cutting was performed to create a cluster tree, and define modules as branches of the tree.

Hub gene identification

In the network, module eigengen (ME) represents the first principal of a module and is an optimal summary of the gene expression profiles of a given module. To measure how a gene is close to the identified module, WGCNA algorithm issued module membership (MM) conception counted by the correlation of each gene to ME. Gene significance (GS) is determined as the absolute value of the correlation of a gene and sample external trait information. Module significance is defined as the average absolute gene significance for all genes in a given module.

Functional enrichment analysis of module genes
After the key module gene list was obtained, gene ontology and KEGG pathway analysis were performed by using “clusterProfiler” package of R language(16). P value<0.01 and q value<0.05 were chosen as the threshold to define the significantly enriched term.

**Differential expression (DEGs) analysis**

DEGs between diffuse and intestinal gastric cancer in TCGA dataset were analyzed by "limma" package in R software. Genes with |logFC|>1 and adj.P.Val<0.01 were regarded as DEGs. Afterward, DEGs were intersected with hub genes identified by WGCNA in GSE15459 dataset to get a final candidate gene list for survival analysis.

**Survival analysis**

R program package “survival”, “survminer”(17, 18) were utilized to perform survival analysis and draw survival curve. Log-rank test were used to evaluate the association between hub-genes expression and patient survival. Those genes with P value<0.05 were regarded as prognostic gene.

**Results**

**Construction of co-expressional net-work and recognition of co-expression module**

We included 10243 genes probes with top 50% expression variance in 192 samples from GSE15459 dataset which has complete clinical information to construct the co-expression network. First of all, sample cluster tree was built to identify outlying samples. No significant outliers were detected among these 192 samples (Figure2A). To receive a scale-free net, we selected 3 as the best power value provided by the `pickSoftThereshold` function with the scale-free R² to be 0.89 (Figure2B, FigureS1). Calculated with this power value, the co-expression matrix and the topological overlap matrix (TOM) were established (FiguresS2). We get a network consisted of 12 gene modules (Figure3), the number of gene within each module was showed in Table1. Inter-module correlation analysis suggested black, blue, and green modules were closed to each other, while red, green and yellow were closed (Figure4B).
Correlation of gene modules with Lauren subtype

To identify modules that are significantly associated with Lauren subtype, we correlate the module eigengene to clinical characteristic. In the cohort, there were 76 diffuse gastric cancer sample, 100 intestinal subtype and 18 mixed type. According to the result, blue, green and red module were revealed to be the most significant correlated with diffuse subtype (Figure4A), and they are inversely correlated with intestinal subtype. Meanwhile, no module associated with mixed subtype was detected.

Intramodular analysis

To measure how a given genes is connected to biologically interesting modules, WGCNA defines module membership (MM) by correlating the individual gene expression profile with the module eigengene of a given module. Meanwhile, the concept intramodular connectivity is introduced to reflect the connection of a gene to other member genes within a module. The gene significance (GS) and intramodular membership (MM) correlation analysis showed GS correlated well with MM in blue and red module (Figure4C, FigureS3) which means the more important a gene is to the module, the higher significance it posseses in particular subtype. The intramodular connectivity and module membership correlation analysis revealed these connectivity of genes in this two modules linked the best to module membership (FigureS4) and genes with high MM intend to have high connectivity indicating they are hub genes in the net. Thus we considered taking blue and red module as the key module for Lauren subtype.

Functional enrichment analysis of key modules

To elucidate critical biological process and pathways this two modules are involved in, gene functional enrichment analysis was performed. As a result, genes in blue module were mainly enriched in GO:0003779 (actin binding), GO:0005261 (cation channel activity), GO:0008201 (heparin binding), while those in red module were mainly enriched in GO:0140097 (catalytic activity, acting on DNA), GO:0008017 (microtubule binding), GO:0003678 (DNA helicase activity), GO:0016538 (cyclin-dependent protein serine/threonine kinase regulator.

### Table 1

<table>
<thead>
<tr>
<th>Module</th>
<th>Gene number</th>
<th>Module</th>
<th>Gene number</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>346</td>
<td>magenta</td>
<td>134</td>
</tr>
<tr>
<td>blue</td>
<td>2063</td>
<td>pink</td>
<td>342</td>
</tr>
<tr>
<td>brown</td>
<td>541</td>
<td>purple</td>
<td>97</td>
</tr>
<tr>
<td>green</td>
<td>388</td>
<td>red</td>
<td>352</td>
</tr>
<tr>
<td>greenyellow</td>
<td>43</td>
<td>turquoise</td>
<td>3081</td>
</tr>
<tr>
<td>grey</td>
<td>2346</td>
<td>yellow</td>
<td>510</td>
</tr>
</tbody>
</table>
activity) (Figure 5A). For KEGG pathway analysis, genes in red module were involved in “cell cycle”, “DNA replication”, “base excision repair”, “oocyte meiosis” and “p53 signaling pathway”, while blue module genes were involved in “Axon guidance”, “cardiomyopathy” and “vascular smooth muscle contraction” (Figure S5).

**Hub gene identification**

To identify hub genes in red and blue module. We chose genes that meet following criteria as hub gene candidates: ((abs(GS.diffuse)>0.3)&(abs(datKME$kMEblue/red)>0.9)&(NS1$q.Weighted <0.01)) In this screening, 23 genes in red and 32 genes in blue modules were obtained (Table 2). To validate the performance of these 55 genes in external dataset, we analyzed the differential expression in TCGA stomach adenocarcinoma comparing diffuse with intestinal subtype. (Figure 5B). Fortunately, all these genes were differentially expressed between diffuse and intestinal subtype (P<0.01). There was a higher expression level for the 32 blue module candidate genes in diffuse gastric cancer than those in intestinal subtype, consistent with the result that blue module is positive correlated with diffuse subtype (Figure 3A). This is on the contrary for the 23 red module genes, as they intend to get higher level in intestinal subtype while red module intend to be positive associated with intestinal trait. Although all these 55 genes were differentially expressed among these two subtype only ten genes including TNS1, PGM5, CPXM2, LIMS2, AOC3, CRYAB, ANGPTL1, BOC and TOP2A were with an absolute LogFC>=1. Thus, we regarded these 10 genes as the most possible hub genes in diffuse gastric cancer.
Survival analysis

To explore prognostic role of these ten hub genes, we then performed survival analysis. In GSE15459 dataset, high expression of HSPB6, TNS1, PGM5, CPXM2, LIMS2, AOC3, CRYAB, ANGPTL1, and BOC were associated with poor overall survival (Table 3). We further validated this in TCGA dataset (Figure 6). As expected, high expression of hub genes except for TOP2A were suggested to be significantly correlated with worse overall survival. At the same time AOC3, CRYAB, BOC were showed to be inversely associated with Relapse-free survival (RFS), whereas higher TOP2A was linked to a better RFS (Figure 7).
Table 3
prognostic role of ten hub genes in GSE15459 dataset

<table>
<thead>
<tr>
<th>Gene</th>
<th>HR (95% CI for HR)</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOC3</td>
<td>1.7 (1.1-2.6)</td>
<td>0.01</td>
</tr>
<tr>
<td>CRYAB</td>
<td>1.9 (1.3-2.9)</td>
<td>0.0028</td>
</tr>
<tr>
<td>LIMS2</td>
<td>1.6 (1-2.4)</td>
<td>0.029</td>
</tr>
<tr>
<td>BOC</td>
<td>1.5 (1-2.3)</td>
<td>0.041</td>
</tr>
<tr>
<td>PGM5</td>
<td>1.3 (0.88-2)</td>
<td>0.18</td>
</tr>
<tr>
<td>ANGPTL1</td>
<td>1.3 (0.85-1.9)</td>
<td>0.23</td>
</tr>
<tr>
<td>HSPB6</td>
<td>1.9 (1.3-3)</td>
<td>0.0024</td>
</tr>
<tr>
<td>CPXM2</td>
<td>1.7 (1.1-2.5)</td>
<td>0.017</td>
</tr>
<tr>
<td>TNS1</td>
<td>2.1 (1.3-3.2)</td>
<td>0.0011</td>
</tr>
<tr>
<td>TOP2A</td>
<td>0.8(0.5-1.2)</td>
<td>0.1939</td>
</tr>
</tbody>
</table>

Association of TNS1, PGM5, CPXM2, LIMS2, AOC3,CRYAB,ANGPTL1,BOC and TOP2A expression with overall survival in GSE15459 datasets.

**Discussion**

WGCNA is an informative method for detecting biologically relevant patterns using high dimensional data sets and had been widely applied in transcriptome analysis. It identifies gene modules by assign genes with strongly covarying patterns into groups and allows for the assessment of module relation to sample trait (19). The advantage of WGCNA lies in its focusing on the association between co-expression modules and clinical traits providing highly reliable and biologically significant results (20).

Lauren classification mainly distinguishes gastric cancer into two subtypes including diffuse and intestinal gastric cancer and the greatest advantage of this category system is it can perceive histological and biological trait of gastric cancer easily (21).

To our knowledge, the current study is the first one to utilize WGCNA method for identification of hub genes in Lauren gastric cancer subtype. We identified 12 gene modules in gastric cancer tissue samples and among these modules, blue and red were the most correlated with diffuse or intestinal subtype otherwise. Intramodular analysis showed MM correlated strongly with genes significance regarding to diffuse phenotype suggesting key genes in these two modules may make sense in diffuse subtype.

Involved in cytoskeleton remodeling, Actin is found to be critical for focal adhesion which is deficient in diffuse gastric cancer (7, 22, 23). In our study, result of gene functional enrichment analysis showed actin binding proteins were enriched in blue module that was positive with diffuse phenotype implied abnormal acting binding may occur in diffuse gastric cancer. In blue module, genes link to cation channel activity were
also found to be enriched. Recent several publications has reported calcium channel protein such as TRPV4 and TRPM2 were involved in gastric cancer cell survival, invasion and disease progression (24–26). Taken our result into consideration, we propose this case should be confined in particular gastric cancer subtype.

As regarding to red module which is positive associated with intestinal gastric cancer, GO term including GO: 0140097 (catalytic activity, acting on DNA), GO: 0008017 (microtubule binding), GO: 0003678 (DNA helicase activity), GO: 0016538 (cyclin-dependent protein serine/threonine kinase regulator activity) enriched among this module genes. Lei and colleagues identified 3 subtypes of gastric adenocarcinoma: proliferative, metabolic, and mesenchymal. Proliferative subtype is mainly consisted of intestinal gastric cancer (2, 27). As gene in these terms are involved in DNA replication and cell proliferation, this result provide a similar scenario for intestinal gastric cancer biology.

Hopefully, we identified 10 hub-genes in blue and red module. Nine of these genes including HSPB6, TNS1, PGM5, CPXM2, LIMS2, AOC3, CRYAB, ANGPTL1, and BOC were from blue module and associated with gastric cancer survival. As a member of tensin family, TNS1 is revealed to function as a scaffold for adhesion related signaling by binding to actin cytoskeleton and β1-integrin (28), and it had been studied in acute myeloid leukaemia and colorectal cancer (29, 30). LIMS2, a member of PINCH proteins, that contain 5 LIM domain also plays important roles in cytoskeletal organization and cell–extracellular matrix adhesion, migration, proliferation, and survival (31).

HSPB6 and CRYAB are two members of small heat shock protein famlily (sHSPs) that confer cells the ability to survive under stress conditions (32, 33). Previous study discovered that mesenchymal stem cells overexpression HSPB6 rendered cell resistance to oxidative stress via increasing secretion of growth factors including VEGF, FGF-2, and IGF-1 (34). On other side, CRYAB, also as a chaperone can interact with apoptosis-associated protein such as caspase-3, Bax, and Bcl-xS to inhibit cell apoptosis (35). Comprehensively, higher expression of these two gene means diffuse gastric cancer is more tolerant to stress.

In the present study, we also recognize CPXM2 to be a survival-related hub gene in diffuse gastric cancer. CPXM2 is a carboxypeptidases for C-terminal cleavage of peptide or protein have recently been validated to contribute the progression of liver cancer and osteosarcoma (36, 37). AOC3 that is an amine oxidase with copper as a cofactor is an inflammation-inducible endothelial cell molecule mediating leucocyte interactions with the blood vessels for its adhesive property (38). To date, AOC3 is has been showed to be a prognostic marker in breast cancer and astrocytomas (39, 40).

**Conclusion**

Overall, Lauren subtype of gastric cancer have different expression pattern which can be identified by WGCNA. HSPB6, TNS1, PGM5, CPXM2, LIMS2, AOC3, CRYAB, ANGPTL1, and BOC might be potential hub gene in diffuse gastric cancer, a subtype with a relatively poorer survival. These nine genes were also prognostic marker for gastric cancer thus more attention need to be paid and further experimental study is required to elucidate the role of these genes.

**Abbreviations**
WGCNA  Weighted Gene Co-Expression Network Analysis

ME  module eigengen

MM  module membership

TOM  topological overlap matrix

GS  gene significance

Declarations

Ethics approval and consent to participate

Not applied

Consent for publication

Not applied

Availability of data and materials

All the datasets involved in this study are publicly available. Gene expression matrix and clinical information of Dataset GSE15459 was download from Gene Expression Omnibus database of NCBI (https://www.ncbi.nlm.nih.gov/gds/). TCGA level 3 data and clinical phenotypic data were downloaded from UCSC portal (https://xenabrowser.net/datapages/).

Competing interests

All the authors in this manuscript declared no conflict of interest.

Funding

This study was supported by grants from National Natural Scientific Foundation of China\$81702430\$

Authors' contributions

Xu Liu and Li Yao drafted the manuscript, contributed to the conception and design of the study. Jingkun Qu, Ning Lu and Lin Liu were accountable for the data analyzing procedure. Jiansheng Wang, George A. Calin and Jia Zhang revised the manuscript. The work have been approved by all the author to be published.

References


Figures
Figure 1

work flow of this study

GSE15459

TCGA STAD data

Gene network construction

Differential expression analysis

Gene module recognition

Top 10 DEGs

Relating modules to clinical

((abs(GS.diffuse)>0.3)&(abs(datKME$kMERed)>0.9)&(NS1$q.Weight<0.01))

55 genes

Hub genes

Gene enrichment analysis

Survival analysis
Figure 2

detection of outlying sample and soft threshold. (A) sample cluster tree of GSE15459 dataset, (B) to pick up the best soft threshold. Left figure: soft threshold is the weight of the network, vertical coordinate is the scaled-free R2, representing the correlation of connectivity k and p(k), R2 of the best soft threshold(β value) should be closed to or over 0.9; right figure, correlation of soft threshold to mean connectivity.
Figure 3

Construction of co-expressional network. cluster dendrogram of the co-expression gene when setting soft threshold as three. Gene modules were attributed into different color.
Figure 4

identification of key modules. (A) module-trait correlation visualized by heatmap; (B) correlation among modules visualized by module eigengene cluster tree and heatmap; (C) correlation of module membership to gene significance for diffuse phenotype within blue, red, black and green module.
Figure 5

A. Gene ontology analysis of blue and red module genes; B. differential expression of the 55 hub gene candidates in blue and red module.

Figure 6
overall survival analysis of 10 hub genes in TCGA dataset

Figure 7

recurrence-free survival analysis of 10 hub genes in TCGA dataset

Supplementary Files

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

- FigureS5.tif
- FigureS4.tif
- FigureS3.tif
- FigureS2.tif
- FigureS1.tif