Identifying miRNA and gene modules of colon cancer associated with pathological stage by weighted gene co-expression network analysis

Xian-guo Zhou,1,2,* Xiao-liang Huang,1,2,* Si-yuan Liang,1,2 Shao-mei Tang,1,2 Si-kao Wu,1,2 Tong-tong Huang,1,2 Zeng-nan Mo,1,2,4 Qiu-yan Wang1,2,5

1Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning, Guangxi Zhuang Autonomous Region, People’s Republic of China; 2Guangxi Key Laboratory for Genomic and Personalized Medicine, Guangxi Collaborative Innovation Center for Genomic and Personalized Medicine, Nanning, Guangxi Zhuang Autonomous Region, People’s Republic of China; 3Department of Colorectal Surgery, First Affiliated Hospital of Guangxi Medical University, Nanning, Guangxi Zhuang Autonomous Region, People’s Republic of China; 4Department of Urology and Nephrology, The First Affiliated Hospital of Guangxi Medical University, Nanning, Guangxi Zhuang Autonomous Region, People’s Republic of China; 5Guangxi Colleges and Universities Key Laboratory of Biological Molecular Medicine Research, Guangxi Medical University, Nanning, Guangxi Zhuang Autonomous Region, People’s Republic of China

*These authors contributed equally to this work

Introduction: Colorectal cancer (CRC) is the fourth most common cause of cancer-related mortality worldwide. The tumor, node, metastasis (TNM) stage remains the standard for CRC prognostication. Identification of meaningful microRNA (miRNA) and gene modules or representative biomarkers related to the pathological stage of colon cancer helps to predict prognosis and reveal the mechanisms behind cancer progression.

Materials and methods: We applied a systems biology approach by combining differential expression analysis and weighted gene co-expression network analysis (WGCNA) to detect the pathological stage-related miRNA and gene modules and construct a miRNA–gene network. The Cancer Genome Atlas (TCGA) colon adenocarcinoma (CAC) RNA-sequencing data and miRNA-sequencing data were subjected to WGCNA analysis, and the GSE29623, GSE35602 and GSE39396 were utilized to validate and characterize the results of WGCNA.

Results: Two gene modules (Gmagenta and Ggreen) and one miRNA module were associated with the pathological stage. Six hub genes (COL1A2, THBS2, BGN, COL1A1, TAGLN and DACT3) were related to prognosis and validated to be associated with the pathological stage. Five hub miRNAs were identified to be related to prognosis (hsa-miR-125b-5p, hsa-miR-145-5p, hsa-let-7c-5p, hsa-miR-218-5p and hsa-miR-125b-2-3p). A total of 18 hub genes and seven hub miRNAs were predominantly expressed in tumor stroma. Proteoglycans in cancer, focal adhesion, extracellular matrix (ECM)–receptor interaction and so on were common pathways of the three modules. Hsa-let-7c-5p was located at the core of miRNA–gene network.

Conclusion: These findings help to advance the understanding of tumor stroma in the progression of CAC and provide prognostic biomarkers as well as therapeutic targets.

Keywords: colon adenocarcinoma, weighted gene co-expression network analysis, differentially expressed genes, differentially expressed miRNA, tumor stroma

Introduction: Colorectal cancer (CRC) ranks the third most frequently diagnosed cancer and the fourth leading cause of cancer-related deaths in the world. It is estimated that more than 2.2 million new cases and 1.1 million deaths will occur by 2030. Colon adenocarcinoma (CAC) is the main histological subtype of CRC and accounts for more than 90% colon cancer. Typically, CAC develops through a multistage process, including the transformation of normal mucosa to adenoma and then to carcinoma. With the progression of CAC, invasion of tumor into the intestinal wall, involvement of regional lymph nodes and distant metastasis may be present. The tumor, node, metastasis (TNM) staging system provides a method to assess the status of cancer and predict prognosis, which is widely used in clinical practice. Patients with CAC are grouped into different stage groups...
according to the TNM status, and patients in the same stage group have a similar prognosis. In most of the cases, patients with a higher pathological stage tend to have a worse prognosis. Thus, it is of great value to identify biomarkers associated with the TNM stage to reveal the molecular mechanisms related to the development of CAC and represent potential prognostic biomarkers as well as therapeutic targets for CAC.

MicroRNAs (miRNAs) are a class of small, noncoding regulatory RNAs. By binding to the target mRNAs, miRNAs repress the expression of target mRNAs. More importantly, one miRNA may target hundreds of mRNAs; thus, miRNA–mRNA constitute networks to participate in many cellular pathways, such as proliferation, apoptosis and differentiation. A significant number of studies search for deregulated miRNA–mRNA networks in cancers. For instance, Jacobsen et al inferred recurrent cancer-associated miRNA–target in diverse cancer types according to negative relationships of expression, sequence-based predictions and miRNA perturbation experiments. However, the utilization of some important clinical parameters such as the TNM stage and survival is limited, which may weaken the ability to discover molecular mechanisms and biomarkers associated with cancer progress and prognosis. In addition, although the miRNA–target relationships have been deeply studied, the relationships among genes or miRNAs are not well considered, since no genes or miRNAs act alone.

With the recent advancements in bioinformatics, new and effective methods have been invented to achieve a more comprehensive identification of new biomarkers. Based on the theory that genes with closely functional linkages or involved in similar pathways may have similar expression profiles, the weighted gene co-expression network analysis (WGCNA) approach provides a systems biology strategy to explore the system-level functionality of a transcriptome. WGCNA constructs a gene network considering the correlations among genes across samples. Genes with highly co-expressed relationships are grouped into the same module; thus, the genes in the same module may have similar functions or possibly share common biological regulatory roles.

The most central genes within a module are identified as hub genes, suggesting their important roles in the module. More importantly, WGCNA analyzes the relationships between modules and sample traits, which provides an effective way to explore the mechanisms behind certain traits. WGCNA has been widely applied to identify gene modules associated with clinical parameters in many cancers. As for colon cancer, WGCNA has been successfully applied to detect gene modules associated with recurrence of colon cancer. Liu et al detected a gene module associated with the recurrence-free survival. In addition, Zhai et al also detected a gene module associated with recurrence of colon cancer by using microarray dataset. However, there is no study applying WGCNA to systematically identify gene co-expression networks and miRNA co-expression networks associated with the pathological stage in CAC. To fulfill this gap, we conducted a differential expression analysis followed by WGCNA to detect the miRNA and gene modules associated with the TNM stage in CAC and predicted the targeted relationships between genes and miRNAs to depict a miRNA–gene network in the progress of CAC.

**Materials and methods**

**Data collection and preprocessing**

The level three RNA-sequencing data (RNA-Seq-HTSeq) and level three miRNA isoform expression quantification data (miRNA-Seq-BGCSC miRNA Profiling) of CAC patients were downloaded from The Cancer Genome Atlas (TCGA; http://cancergenome.nih.gov/) on October 5, 2017. The sequencing data were generated by the Illumina HiSeq platform. For miRNA isoform data, the expression of a mature miRNA was obtained by adding all the isoform expression of it. The genes or miRNAs estimated with counts in less than 20% of cases were removed. The outlier samples were detected by cluster diagram and then removed. The fragments per kilobase of exon per million fragments mapped (FPKM) and read per million (RPM) were used to represent the expression of genes and miRNAs, respectively. Clinical information of CAC patients was downloaded and extracted at the same time. Microarray-based mRNA and miRNA expression data were obtained from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) based on the following inclusion criteria: 1) the expression data were generated from CRC samples; 2) the clinicopathological information, especially the American Joint Committee on Cancer TNM stage, was included and 3) the sample size was >40. The microarray dataset GSE29623, which contained both mRNA and miRNA expression data, was used to estimate module preservation. The datasets GSE28364 (miRNA expression data) and GSE39582 (gene expression data) served as an independent cohort to validate module–clinical trait relationships. The normalized expression matrix files and platform annotations were downloaded. If multiple probes were applied to detect one miRNA or mRNA, the average value of the probes was used. All the miRNAs in the present study were renamed according to miRBase database v21 (www.mirbase.org). Data processing...
was performed using the R programming language (v3.1.0) and complied with the human subject protection and data access policies of GEO and TCGA.

Identification of differentially expressed genes (DEGs) and miRNAs
The R package, edgeR, was used to detect the DEGs and differentially expressed miRNAs (DEMs) between cancer and normal samples. The thresholds for both DEGs and DEMs were set at the values of \(|\log_2(\text{fold change [FC]})| > 1\) and false discovery rate (FDR) <0.05.

Weighted gene co-expression network construction and module detection
The gene co-expression networks and miRNA co-expression networks were constructed by the WGCNA package based on DEGs and DEMs. The processes for constructing gene co-expression networks and miRNA co-expression networks were similar except for some parameters. So, we took the construction of weighted gene co-expression networks as an example. First, a matrix of similarity was constructed by calculating the correlations of all pairs of genes. Second, an appropriate soft-thresholding power \(\beta\) was selected by using the integrated function (pickSoftThreshold) in the WGCNA package. With this soft-thresholding power, the co-expression similarity was raised to achieve scale-free topology. Third, the adjacency was transformed into a topological overlap matrix (TOM) by using TOM similarity. Then, the corresponding dissimilarity (dissTOM) was also calculated. Fourth, co-expression gene modules were identified by using dynamic tree cut with the following major parameters: maxBlockSize of 20000, minModuleSize of 30 and deepSplit of 2. As for the miRNA network, the minModuleSize was set at 15 due to the small number of DEMs. In order to distinguish miRNA modules from gene modules, we named miRNA modules as “M+color” and gene modules as “G+color” in the full text. The module eigen-gene (ME), which was the first principal component (PC) of each module’s gene expression matrix, was obtained by the WGCNA to represent the expression profiles of module genes. Some highly similar modules with the height of 0.25 were merged together. A clustering dendrogram was used to display the results of dynamic tree cut and merge.

Module preservation analysis
To test the stability of modules, we conducted the module preservation analysis. The theories for preservation analysis are, briefly, to test whether a reference network can be constructed in another test network. The datasets GSE29621 and GSE29622 were used to test the preservation of co-expressed gene modules and miRNA modules, respectively. An integrated function of WGCNA (module preservation) was used to calculate the preservation (nPermutations=100) and the Z summary score (Z score). A Z score of lower than 2 indicates that the modules have no preservation, that of 2–10 indicates low to moderate preservation and that of >10 indicates strong preservation. It must be noted that in both gene and miRNA modules, the gray module contains genes or miRNAs that do not belong to any of the modules and the gold module is generated by module preservation function for statistical aims. Thus, the two modules were not shown in results of preservation analysis.

Relating modules to clinical parameters and identifying hub genes
One of the advantages of WGCNA is that the correlation between modules and clinical parameters can be analyzed. Given that the ME can summarize the gene expression profiles, we calculated the correlation between MEs and external clinical parameters. This process was referred as the module–trait relationship (MTR) analysis. To validate the MTRs of gene modules, we utilized the microarray dataset GSE39582, which contained the gene expression data from 566 cases of colon cancers. We first categorized genes into the corresponding module referring to the modules constructed. Then, the ME of each module was calculated and related to clinical parameters. The MTRs of miRNA modules were validated by using GSE28364 in the same way. Furthermore, a \(q\)-value was introduced to control the false-positive rate and a \(q\)-value of \(\leq 5\%\) was considered to be statistically significant. Hub genes were identified by calculating the gene significance (GS) and module membership (MM). The GS of a gene was referred to as (the absolute value of) the correlation between the gene and a certain interested clinical parameter, and the MM of a gene was referred to as the correlation between the ME and the gene expression profile. Genes with a high MM and GS were defined as hub genes in a module. This method was also applicable to identify hub miRNAs.

Pathway enrichment analysis of genes and miRNAs
Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis for genes was conducted based on the Database for Annotation, Visualization and Integrated Discovery (DAVID; https://david.ncifcrf.gov/).
The miRNA pathway analysis was conducted using mirPath v.3 (http://snf-515788.vm.okeanos.grnet.gr), which was combined with miRNA target prediction based on DIANA-microT-CDS algorithm and information of experimentally validated miRNA interactions. A P-value of $\leq 5\%$ was considered to be statistically significant.

**Characterization and validation of hub genes and survival analysis**

The dataset GSE29621 served as an independent cohort to validate the association between hub genes and the pathological stage. The microarray dataset GSE35602, which contained miRNA and gene expression profiles from epithelium and stroma of CRC tissues dissected by a laser microdissection technology, was used to evaluate the expression levels of hub genes and hub miRNAs in the epithelium and stroma. Another microarray dataset, GSE39396, was used to assess the gene expression levels in four cell populations from CRC samples (Epcam$^+$: epithelial [Ep] cancer cells; CD45$^+$: leukocytes; CD31$^+$: endothelial cells; FAP$^+$: cancer-associated fibroblasts [CAF$^+$]). The one-way analysis of variance (ANOVA) test was used to compare the mean values of multiple groups with equal variances, and the Welch’s test was used when the variances were unequal. Moreover, we applied two web servers Gene Expression Profiling Interactive Analysis (http://gepia.cancer-pku.cn) and OncoLnc (http://www.oncolnc.org) to evaluate the prognostic value of hub genes and hub miRNA, respectively.

**Construction of the miRNA–gene interaction network**

The miRNA–gene interaction network was constructed based on the miRNAs and genes involved in the same pathway. Two web servers, microT-CDS (v5.0) and TargetScan, were applied to predicate the interaction between miRNAs and genes. Only the interactions with a score of microT-CDS $>0.6$ and a context++ score percentile of TargetScan $>70$ were used to construct the interaction network. Furthermore, we also used the STRING to analyze the interaction between proteins in the network. Finally, the network graph was visualized and analyzed using Cytoscape v3.4.0. The core region in the network, which possessed a dense connection, was detected by Molecular Complex Detection (MCODE), a Cytoscape plugin.

**Results**

**Data preprocessing**

A workflow of this study is shown in Figure 1. The level three RNA-sequencing data of CAC contained 477 CAC samples and 41 normal colon samples. After removing outlier samples and samples without clinical information, 450 CAC samples and 41 normal colon samples remained for WGCNA analysis. As for miRNA-sequencing data, 455 CAC samples and eight normal colon samples were downloaded. After outlier samples and samples without clinical information were eliminated, 442 CAC samples and eight normal colon samples were subjected to WGCNA analysis. As for the GEO dataset GSE29623, the subdatasets GSE29621 (mRNA data) and GSE29622 (miRNA data) were derived from the same 65 CAC patients. All the 65 samples remained after the same data processing methods were performed. The original clinical information of CAC patients in TCGA and GSE29623 is listed in Tables S1 and S2, respectively.

**DEGs and miRNAs**

The RNA-sequencing data and miRNA-sequencing data from TCGA were subjected to differential expression analysis. Based on the threshold of $|\log_2(FC)| > 1$ and FDR $< 0.05$, 6486 DEGs and 544 DEMs were identified. Among the 6486 DEGs, 3678 DEGs were significantly upregulated in cancer samples, while 2808 DEGs were significantly downregulated in cancer samples. Furthermore, among the 544 DEMs, 194 DEMs were significantly downregulated in cancer samples and 350 DEMs were significantly upregulated in cancer samples. The volcano plots of mRNAs and miRNAs are displayed in Figure S1, Table S3 lists the top 20 DEGs and DEMs ranked by FDR.

**Identification of co-expression networks and module preservation analysis**

The 6486 DEGs and 450 CAC samples were used to construct co-expressed gene networks. When the soft thresholding was set at five, the scale-free topology fit index reached 0.90 (Figure S2). In all, 21 co-expressed gene modules were identified by the dynamic tree cut method, and each of the modules was marked by a color. The Gcyan and Gred, as well as the Gpink and Gpurple, were respectively merged together due to the similarity between them (the height of ME in the clustering was $< 0.25$). Finally, 19 co-expressed gene modules were detected. The Glightyellow and Groyal-blue were the smallest modules with 38 genes. The Ggrey, which contained genes not attributed to any modules, was the largest module with 1432 genes. For miRNA co-expression network, 544 DEMs and 442 CAC samples were subjected to WGCNA analysis. When the soft thresholding was set at five, the scale-free topology fit index reached 0.90 (Figure S3). Eight co-expression miRNA modules were detected, and no module was highly similar that needed to be merged.
Identifying miRNA and gene modules of colon cancer

The Mturquoise contained the most miRNA with a number of 136, while the Mblack contained the least miRNA with a number of 36. The clustering dendrograms of genes and miRNAs are shown in Figure 2. All the modules and their contents as well as the GS and MM of each gene and each miRNA are listed in the Supplement material Table S4.

To test the stability of gene modules, we used the dataset GSE29621 as an independent validation cohort. Six gene
modules (Gblue, Gbrown, Gblack, Gmagenta, Ggreen and Gturquoise) were found to be strongly conservative with a Z summary score (Z score) >10. The Gblue was the most stable module with a Z score of 26 and a median rank statistic of 12. However, the preservation of the remaining modules was not strong enough since the Z scores were <10. The preservation of miRNA modules was tested by using the dataset GSE29622. However, none of the module was found to have strong preservation. The Mblack was the most stable module with a Z score of 3.4 and a median rank statistic of 3, which reflected that the Mblack had low to moderate stability.

The reason for this phenomenon might be that the size of miRNA module was generally small (the maximum size was 136), which reduced the stability of modules (Figure S4).

### Module–clinical trait relationships and validation

Identifying genes associated with a certain clinical trait is of great value to explore the molecular mechanisms behind the trait. In the present study, the clinical parameters of CRC patients, including age at diagnosis, pathological stage, pathological T (pT) category, pathological N (pN) category, M category, lymphatic invasion, venous invasion, perineural invasion, person neoplasm cancer status, vital status and microsatellite instability were extracted for analysis. In addition, some potential prognostic biomarkers of CRC, such as IGFBP2, PTGS2 and MIP3, were involved in the MTR analysis.

As for the correlation between miRNA modules and clinical traits, the Mblack, which possessed the strongest preservation among the miRNA modules, was significantly associated with the pathological stage (r=0.19, q=3.0×10^-4), pT category (r=0.11, q=0.033), pN category (r=0.14, q=8.1×10^-5), M category (r=0.18, q=3.2×10^-4), lymphatic invasion (r=0.17, q=8.8×10^-4), venous invasion (r=0.14, q=7.5×10^-4), perineural invasion (r=0.13, q=0.031), microsatellite instability (r=0.26, q=1.6×10^-5), IGFBP2 (r=0.16, q=2.0×10^-3) and IMP3 (r=0.25, q=8.2×10^-7; Figure 4). The dataset GSE28364 was utilized to validate module–clinical trait relationships. We found that the Mblack was positively correlated with the pathological stage, but did not reach the significant level (r=0.016, q=0.89; Figure S6). Table S5 lists the genes in the magenta and green modules and the miRNAs in the black module.

### Pathway enrichment analysis of genes and miRNAs

To explore the biological functions of the Gmagenta and Ggreen, we performed pathway analysis to the genes in the Gmagenta and Ggreen, respectively. For the Gmagenta, the top three significantly enriched pathways were extracellular matrix (ECM)–receptor interaction (P=1.4×10^-16), focal adhesion (P=3.4×10^-4) and protein digestion and absorption (P=9.0×10^-8). For the Ggreen, the top three significantly enriched pathways were vascular smooth muscle contraction (P=1.0×10^-8), calcium signaling pathway (P=1.0×10^-5) and dilated cardiomyopathy (P=1.2×10^-5). The miRNAs in the Mblack were also subjected to pathway analysis, and the top three significantly enriched pathways were proteoglycans in cancer (P=2.3×10^-10), axon guidance (P=6.8×10^-7) and mucin type O-glycan biosynthesis (P=1.2×10^-6). More
importantly, seven pathways, including proteoglycans in cancer, focal adhesion, ECM–receptor interaction, PI3K–Akt signaling pathway, adrenergic signaling in cardiomyocytes, arrhythmogenic right ventricular cardiomyopathy (ARVC) and platelet activation, were common pathways of miRNAs and genes. Table 1 lists the top three significant pathways of Gmagenta, Ggreen and Mblack as well as the common pathways shared by genes and miRNAs. Tables S6 and S7 list all the significant pathways of genes and miRNAs.

Identification of hub genes and miRNAs
Hub genes or miRNAs in a module have relative high MM, which may be functionally significant. Since the pathological stage was a key parameter to predict prognosis and the three modules (Gmagenta, Ggreen and Mblack) were associated with the pathological stage, we calculated the GS based on the correlation between gene or miRNA and the pathological stage. The top ten genes or miRNAs with the highest MM were identified as hub genes or miRNAs. The hub genes and hub miRNAs as well as the corresponding MM and GS are listed in Table 2. In the Gmagenta, the SPARC was the most central gene with an MM of 0.95 and a GS of 0.12. In the Ggreen, the TNS1 had the highest MM value (MM=0.96). As for the Mblack, the hsa-miR-125b-5p was the top hub miRNA with the highest MM of 0.90 and a GS of 0.11.
Validation of hub genes

The association between hub genes and the pathological stage was validated using the dataset GSE29621. Three hub genes in the Ggreen module were successfully validated (TAGLN: \( P=0.048 \), DACT3: \( P=0.011 \) and GNAO1: \( P=0.049 \)). Furthermore, seven of the ten hub genes in the Gmagenta were also found to be significantly differentially expressed across different stages, which indicated that they were consistently associated with the pathological stage even in an independent dataset (SPARC: \( P=0.015 \), COL1A2: \( P=0.034 \), COL5A2: \( P=0.014 \), COL1A1: \( P=0.047 \), COL5A1: \( P=0.039 \), BGN: \( P=0.008 \) and THBS2: \( P=0.016 \) (Figure 5)). Considering the Mblack was not successfully validated, we only analyzed the association between hub miRNAs’ expression and stage based on TCGA data. As shown in Figure S7, the hsa-miR-199a-5p, hsa-miR-218-5p, hsa-miR-125b-2-3p and hsa-miR-1-3p were significantly related with the pathological stage (\( P=0.026 \), \( P=0.017 \), \( P=0.045 \) and \( P=0.011 \), respectively). The hsa-miR-125b-5p, hsa-miR-100-5p, hsa-miR-199a-5p, hsa-miR-218-5p and hsa-miR-125b-2-3p were significantly associated with CRC. Increased serum CEA levels are correlated with an advanced stage. In addition, there are numerous reports on the correlation between serum cancer antigen 19-9 (CA19-9) levels and the stage of CRC. We analyzed the correlation between hub miRNA expression and serum CEA and CA19-9 levels using GSE28364. We found that all the hub miRNAs were positively related with CEA and CA19-9 levels, but did not reach the significant level (all \( P \geq 0.05 \)).

Correlation between hub miRNA expression and serum carcinoembryonic antigen (CEA) and CA19-9 levels

The CEA is a widely recognized prognostic biomarker in CRC. Increased serum CEA levels are correlated with an advanced stage. In addition, there are numerous reports on the correlation between serum cancer antigen 19-9 (CA19-9) levels and the stage of CRC. We analyzed the correlation between hub miRNA expression and serum CEA and CA19-9 levels using GSE28364. We found that all the hub miRNAs were positively related with CEA and CA19-9 levels, but did not reach the significant level (all \( r>0 \), all \( P>0.05 \); Figure S8). It should be pointed out that we found that the
CEA and CA19-9 levels were not significantly correlated with the TNM stage in the GSE28364 dataset (CEA: \( r = 0.292, P = 0.84 \); CA19-9: \( r = 0.32, P = 0.056 \)). We also observed that none of the ten hub miRNAs showed a significant correlation (all \( P < 0.05 \)), which validated the co-expression relationships of hub miRNAs.

### Hub genes and hub miRNAs are predominantly expressed in tumor stroma

We observed that four collagen coding genes (COL1A2, COL1A1, COL5A2 and COL5A1), which are known to be expressed in stroma, were in the list of hub genes. In addition, the proteoglycans in the cancer pathway was the most significant pathway of miRNAs, and proteoglycans are a major component of stroma. We wondered whether these two gene modules and one miRNA module reflected the expression pattern of cancer cells or stromal (Str) cells since the TCGA data were derived from macrodissected cancer samples, which contained both cancer cells and Str cells. To address this issue, we used the microarray dataset, GSE35602, which contained miRNA and gene expression profiles of epithelium and stroma of CRC tissues dissected by a laser microdissection technology. We compared the expression levels of hub genes and hub miRNAs in cancer epithelium and stroma. As demonstrated in Figure 6, we found that 18 of the 20 hub genes, except DACT3 and GNAO1, were predominantly expressed in tumor stroma (all \( P \)-values <0.001). Furthermore, seven of the ten hub miRNAs, except hsa-let-7c-5p, hsa-miR-133a-3p and hsa-miR-125b-2-3p, were significantly upregulated in tumor stroma (\( P < 0.05 \)). To further validate the results, we queried The Human Protein Atlas (http://www.proteinatlas.org/) to evaluate the gene expression on the protein level. Since it is widely recognized that collagen genes are mainly expressed in stroma, we selected the SPARC and BGN in the Gmagenta as well as the TNS1 and MYLK in the Ggreen to assess their expression. The representative images of immunohistochemical staining are presented in Figure S9. The staining...
of the four genes was stronger in cancer stroma compared with cancer cells, which was consistent with the results of the RNA level. More importantly, the staining of SPARC and BGN in the normal stroma was weaker than that in the tumor stroma, which indicated that stroma was implicated in tumorigenesis and progression of CAC. These findings suggested that the expression levels of these hub genes and hub miRNAs mainly depended on tumor stroma. The tumor stroma includes a large quantity of cells, such as CAF, glial, Ep, vascular and immune cells. Next, we employed a microarray dataset, GSE39396, to explore the cell-type specificity of the hub genes. The GSE39396 contained the transcriptome data of four cell populations (Epcam+: Ep cancer cells, CD45+: leukocytes, CD31+: endothelial cells and FAP+: CAFs) isolated by fluorescence-activated cell sorting (FACS) from six CRC samples. As shown in Figure S10, most of the hub genes were predominantly expressed in CAFs, although some genes showed a moderate level of expression in the endothelial cells. This finding indicated that the CAFs were a key contributor to the expression levels of hub genes.

Survival analysis on hub genes and hub miRNAs

The tumor stroma provides a favorable microenvironment for cancer proliferation and infiltration.40 The interaction between stroma and cancer cells may influence the malignant behaviors of cancer and, ultimately, lead to a poor prognosis. Since the hub genes and hub miRNAs were selected from

### Table 2 Hub genes and hub miRNAs

<table>
<thead>
<tr>
<th>Module</th>
<th>Symbol</th>
<th>MM</th>
<th>GS</th>
<th>Log$_{FC}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gmagenta</td>
<td>SPARC</td>
<td>0.95</td>
<td>0.12</td>
<td>1.46</td>
</tr>
<tr>
<td>Gmagenta</td>
<td>COL1A2</td>
<td>0.94</td>
<td>0.10</td>
<td>1.61</td>
</tr>
<tr>
<td>Gmagenta</td>
<td>COL1A1</td>
<td>0.93</td>
<td>0.08</td>
<td>2.67</td>
</tr>
<tr>
<td>Gmagenta</td>
<td>COLSA2</td>
<td>0.93</td>
<td>0.08</td>
<td>1.80</td>
</tr>
<tr>
<td>Gmagenta</td>
<td>COLSA1</td>
<td>0.93</td>
<td>0.10</td>
<td>1.60</td>
</tr>
<tr>
<td>Gmagenta</td>
<td>BGN</td>
<td>0.91</td>
<td>0.13</td>
<td>2.56</td>
</tr>
<tr>
<td>Gmagenta</td>
<td>THBS2</td>
<td>0.90</td>
<td>0.09</td>
<td>2.35</td>
</tr>
<tr>
<td>Gmagenta</td>
<td>SULF1</td>
<td>0.90</td>
<td>0.07</td>
<td>1.87</td>
</tr>
<tr>
<td>Gmagenta</td>
<td>WISP1</td>
<td>0.90</td>
<td>0.03</td>
<td>2.67</td>
</tr>
<tr>
<td>Gmagenta</td>
<td>THY1</td>
<td>0.90</td>
<td>0.10</td>
<td>1.38</td>
</tr>
<tr>
<td>Ggreen</td>
<td>TNS1</td>
<td>0.96</td>
<td>0.11</td>
<td>-2.68</td>
</tr>
<tr>
<td>Ggreen</td>
<td>LMOD1</td>
<td>0.96</td>
<td>0.10</td>
<td>-3.45</td>
</tr>
<tr>
<td>Ggreen</td>
<td>MYLK</td>
<td>0.96</td>
<td>0.09</td>
<td>-3.29</td>
</tr>
<tr>
<td>Ggreen</td>
<td>HSP8B</td>
<td>0.95</td>
<td>0.10</td>
<td>-3.34</td>
</tr>
<tr>
<td>Ggreen</td>
<td>PDZRN4</td>
<td>0.95</td>
<td>0.09</td>
<td>-4.00</td>
</tr>
<tr>
<td>Ggreen</td>
<td>TAGLN</td>
<td>0.95</td>
<td>0.12</td>
<td>-2.83</td>
</tr>
<tr>
<td>Ggreen</td>
<td>DACT3</td>
<td>0.95</td>
<td>0.13</td>
<td>-2.41</td>
</tr>
<tr>
<td>Ggreen</td>
<td>GNAO1</td>
<td>0.94</td>
<td>0.11</td>
<td>-3.35</td>
</tr>
<tr>
<td>Ggreen</td>
<td>TMEM200B</td>
<td>0.94</td>
<td>0.11</td>
<td>-1.64</td>
</tr>
<tr>
<td>Ggreen</td>
<td>HAND2</td>
<td>0.94</td>
<td>0.10</td>
<td>-3.82</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-miR-125b-5p</td>
<td>0.90</td>
<td>0.11</td>
<td>-1.64</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-miR-145-5p</td>
<td>0.84</td>
<td>0.14</td>
<td>-2.72</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-miR-100-5p</td>
<td>0.81</td>
<td>0.10</td>
<td>1.06</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-miR-145-3p</td>
<td>0.76</td>
<td>0.12</td>
<td>2.15</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-miR-199a-5p</td>
<td>0.75</td>
<td>0.15</td>
<td>1.48</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-let-7c-5p</td>
<td>0.74</td>
<td>0.10</td>
<td>-1.40</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-miR-218-5p</td>
<td>0.73</td>
<td>0.13</td>
<td>4.52</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-miR-133a-3p</td>
<td>0.71</td>
<td>0.15</td>
<td>-2.75</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-miR-125b-2-3p</td>
<td>0.68</td>
<td>0.09</td>
<td>-1.03</td>
</tr>
<tr>
<td>Mblack</td>
<td>hsa-miR-1-3p</td>
<td>0.67</td>
<td>0.15</td>
<td>6.04</td>
</tr>
</tbody>
</table>

Abbreviations: miRNA, microRNA; MM, module membership; GS, gene significance; FC, fold change.

Figure 5 Pathological stage-associated hub genes validated by GSE29621.

Notes: (A) TAGLN, (B) DACT3, (C) GNAO1, (D) SPARC, (E) COL1A2, (F) COL5A2, (G) COL1A1, (H) COL5A1, (I) BGN, (J) THBS2. I–IV: stages I–IV.
Figure 6 The expression levels of hub genes (A) and hub miRNAs (B) in Ep and Str areas from CRC tissues (GSE35602).

Note: ***P<0.001 and **P<0.01.

Abbreviations: miRNA, microRNA; Ep, epithelial; Str, stromal; CRC, colorectal cancer.
the modules related to patients’ pathological stages, it was of great significance to evaluate the potentiality of them to serve as prognostic biomarkers. Among the ten hub genes of Gmagenta, five genes were found to be significantly associated with patient prognosis (COL1A2: $P=0.017$, COL1A1: $P=0.036$, BGN: $P=0.0058$, THBS2: $P=0.017$ and SULF1: $P=0.011$), of which upregulation indicated a poor prognosis. In addition, six of the ten hub genes in the Ggreen were significantly associated with prognosis (TNS1: $P=0.03$; HSPB8: $P=0.022$; TAGLN: $P=0.014$; DACT3: $P=0.044$; TMEM200B: $P=0.035$; HAND2: $P=0.033$), and a high expression level indicated a poor prognosis. Five of the ten hub miRNAs were also identified to be related to patient prognosis (hsa-miR-125b-5p: $P=0.014$, hsa-miR-145-5p: $P=0.045$, hsa-let-7c-5p: $P=0.034$, hsa-miR-218-5p: $P=0.047$ and hsa-miR-125b-2-3p: $P=0.034$; Figure S11). The prognosis-related hub genes and miRNAs have the potentiality to be prognostic biomarkers of CAC in the future.

Analysis of the miRNA–gene interaction network

In the present study, we noticed seven pathways were shared by genes and miRNAs, including proteoglycans in cancer, focal adhesion, ECM–receptor interaction, PI3K–Akt signaling pathway, adrenergic signaling in cardiomyocytes, ARVC and platelet activation. Given that miRNAs–genes comprise networks to participate in many cellular pathways, we constructed a miRNA–gene interaction network based on the miRNAs and genes involved in the same pathway to achieve a better understanding of the regulatory relationships. In total,
there were 51 nodes (26 miRNAs and 25 genes) and 144 pairs of interaction relationships in the network (Figure 7). The miRNA hsa-let-7c-5p was the top hub node in the network with a degree of 19, which suggested its central role in the network. A highly connected core region containing nine nodes (hsa-let-7c-5p, COL1A1, COL4A1, COL5A2, COL5A1, COL1A2, COL5A3 and THBS2) and 46 edges with a score of 8.5 was detected by MCODE.

**Discussion**

In the present study, we applied a systems biology approach, WGCNA, to identify CAC pathological stage-associated gene and miRNA modules for the first time. We found that three modules, including two gene modules (Gmagenta and Ggreen) and one miRNA module (Mblack), were significantly related to the pathological stage, which indicated the implication of these modules in the progress of CAC. A total of 20 hub genes and ten hub miRNAs were identified. The stability and module–clinical trait relationships of the two gene modules were successfully validated. Most of the hub genes and hub miRNAs were significantly upregulated in the tumor stroma. Survival analysis indicated that 11 hub genes and five hub miRNAs were related with the prognosis of CAC patients, suggesting their prognostic values. In addition, the miRNA–gene interaction network analysis revealed a highly connected miRNA–gene interaction region. The findings could contribute to a better understanding of the implication of tumor stroma in the progress of CAC and provide prognostic biomarkers for clinical decisions.

A tumor is composed of tumor cells and tumor stroma. The stroma consists of Str cells, such as CAF, glial, Ep, vascular and immune cells, as well as ECM components, such as collagen, fibronectins, proteoglycans and glycosaminoglycans. As the most basic components of the ECM, collagen is closely related with the function of ECM. The implication of collagen in tumorigenesis and progression of CRC have been widely reported. The COL11A1 and COL5A2 are not expressed in normal colon tissues but expressed in the stroma of CRC tissues. The COL1A2 is also overexpressed in CRC stroma.

The functions of type I collagen in CRC have been intensively studied by cell culture experiments. CRC cells grown on the type I collagen medium show an epithelial–mesenchymal-like transition (EMT-like) morphology. Type I collagen inhibits cell differentiation, while promotes the expression of stem cell markers of CRC cells. Collagen type I induces a transient transcriptional downregulation of Cdx2, which is associated with a loss of differentiation. In addition, CRC cells grown on type I collagen show enhanced growth, survival, chemoresistance and invasion.

In the present study, we also found that COL1A2 and COL5A2 were upregulated in the tumor stroma, which was in line with the previous studies. More importantly, we identified that genes shared a similar expression pattern with type I collagen genes, which provided new insight into the implication of ECM in the progress of CAC.
angiogenesis and inflammation. Furthermore, CAF-specific proteins have the potential to be prognostic markers and therapeutic targets.\(^\text{54}\) The myofibroblast is a principal member of CAFs. The electron microscope indicates that there are peripheral myofilaments in the myofibroblast. The contraction of myofilaments causes the tumor stroma to become stiffer, which, as a result, enhances cell growth, survival and promotes migration.\(^\text{55}\) In consideration of the five pathways related with myofilament movement, we speculated that this module could be related to the functions of myofibroblast. For example, CAFs in the gastric carcinomas enhance the expression of TAGLN to promote cell migration and invasion.\(^\text{56}\) In addition, the TNS1 plays an important role in the myofibroblast differentiation and ECM formation.\(^\text{57}\) The functions of these hub genes in the CAC remain largely unknown and molecular biology experiments are required in the future.

As for the Mblack, we found that many cancer-related pathways, such as proteoglycans in cancer, Hippo signaling pathway, ErbB signaling pathway and CRC were enriched in the module, which indicated the complex regulatory mechanisms of the miRNA module in the progression of CAC. Interestingly, we found seven overlaps between gene pathways and miRNA pathways. The proteoglycans in cancer, focal adhesion, ECM–receptor interaction, PI3K–Akt signaling pathway, adrenergic signaling in cardiomyocytes, ARVC and platelet activation were overlapped pathways. Since the three modules (Gmagenta, Ggreen and Mblack) were related to the pathological stage of CAC patients, we speculated that the genes and miRNAs involved in the same pathway could have regulatory relationships, which may promote the deterioration of CAC. Thus, we constructed a miRNA–gene interaction network aiming to find out the core regulatory relationships. A highly connected core region containing nine nodes (hsa-let-7c-5p, COL1A1, COL4A1, COL5A2, COL5A1, COL1A2, COL5A3 and THBS2) was found. Among the core region, hsa-let-7c-5p was regarded as the central regulator since it targeted the other eight genes. Interestingly, seven of the eight (COL11A1, COL4A1, COL5A2, COL5A1, COL1A1, COL1A2 and COL5A3) genes are collagen-coding genes, and the remaining one (THBS2) modulates collagen fibrillogenesis and plays a crucial role in the ECM assembly.\(^\text{44,56}\) These findings indicate that hsa-let-7c-5p may regulate ECM-related genes in the progression of CAC. Hsa-let-7c-5p belongs to the conservative let-7 family, which is one of the well-studied miRNA that involves in many cancer types.\(^\text{59-61}\) The expression of let-7c is negatively correlated with COL1A2 expression in chronic hepatitis tissues,\(^\text{62}\) which indicates that let-7c may target COL1A2. However, there was no significant difference in the hsa-let-7c-5p levels between tumor cells and tumor stroma. The high conservation of let-7 family may partly explain this phenomenon.\(^\text{63}\) The increased hsa-let-7c-5p level was related with a poor survival, which suggested its potential prognostic value.

Some limitations in our study should be mentioned. First, the datasets used to identify co-expressed modules were derived from macrodissected cancer samples, which included tumor cells and tumor stroma. Since great heterogeneity exists in cancer cells, utilizing comprehensive gene expression profiles from both tumor and Str cells may result in failing to identify tumor cell-specific gene modules. Second, the preservation of the Mblack was not strong enough and the module–clinical trait relationship of Mblack was not very well validated. Readers should be careful when utilizing the information from Mblack. Third, the miRNA–gene relationships in CAC were only based on target predication. Further molecular biology experiments are required to prove the interaction.

**Conclusion**

By applying WGCNA and other methods, the present study identifies two gene modules and one miRNA module associated with the pathological stage and provides some potential prognostic biomarkers of CAC. These findings provide insights about the mechanisms underlying CAC progression, which is of great significance to the treatment of CAC.

**Acknowledgments**

This work was supported by Guangxi Natural Science Fund for Innovation Research Team (2013GXNSFFA019002, 2016GXNSFGA38006), Guangxi Collaborative Innovation Center for Genomic and Personalized medicine (201319) and Student’s Platform for Innovation and Entrepreneurship Training Program (201710598063). We thank TCGA and the GEO for providing all data in this research.

**Author contributions**

Conceived and designed the experiments: Q-YW, Z-NM, X-GZ and X-LH. Performed the data statistics and analysis: X-GZ, X-LH, S-YL, S-MT, S-KW and T-TH. Contributed materials/analysis tools: S-YL, S-MT, S-KW and T-TH. Performed the data statistics and analysis: X-GZ and X-LH. Contributed to the writing of the manuscript: X-GZ, X-LH, S-YL, S-MT, S-KW and T-TH. Contributed to the writing of the manuscript: X-GZ, X-LH, S-YL, S-MT, S-KW and T-TH. Revised the manuscript: Q-YW and Z-NM. All authors contributed toward data analysis, drafting and revising the paper and agree to be accountable for all aspects of the work.
Disclosure
The authors report no conflicts of interest in this work.

References


