Network analysis based on TCGA reveals hub genes in colon cancer

Fenzan Wu[1], Guoping Yuan[2], Junjie Chen[2], Chengzu Wang[2]

[1]Science and Education Division, Cixi Affiliated Hospital of Wenzhou Medical University, Zhejiang, China
[2]Clinical Laboratory, Cixi Affiliated Hospital of Wenzhou Medical University, Zhejiang, China

Colorectal cancer (CRC) is the third most widespread cancer in the world. Although many advances have been made in molecular biology, novel approaches are still required to reveal molecular mechanisms for the diagnosis and therapy of colon cancer. In this study, we aimed to determine and analyse the hub genes of CRC. First, we explored the mRNA and microRNA (miRNA) expression profiles of colon carcinoma, then we screened target genes of differentially expressed miRNAs and obtained the intersection between differently expressed genes and target genes. Gene Ontology (GO) classification and KEGG pathway analysis of differently expressed genes were performed, and gene-miRNA and TF-gene-miRNA networks were constructed to identify hub genes, miRNAs, and TFs. In total, 3436 significant differentially expressed genes (1709 upregulated and 1727 downregulated) and 216 differentially expressed miRNAs (99 upregulated and 117 downregulated) were identified in colon cancer. These differentially expressed genes were significantly enriched in GO terms and KEGG pathways, such as cell proliferation, cell adhesion, and cytokine-cytokine receptor interaction signalling pathways. *GCNT4*, *EDN2*, and so on were located in the central hub of the co-expression network. *MYC, WT1,* mir-34a, and LEF1 were located in the central hub of the network of TF-gene-miRNA. These findings increase our understanding of the molecular mechanisms of colon cancer and will aid in identifying potential targets for diagnostic and therapeutic usage.

Key words: Co-expression network, TF-miRNA-gene network, The Cancer Genome Atlas, colon cancer.

Introduction

Colorectal cancer (CRC) is the third most widespread cancer in the world; over 1 million newly diagnosed cases and over half a million CRC-related deaths are estimated to occur each year [1, 2]. It is generally accepted that CRC mainly develops through two different genetic pathways: the chromosomal instability (CIN) pathway, which is responsible for approximately 70–80% of all CRC cases [3], and the microsatellite instability (MSI) pathway, which is related to inactivation of the mismatch repair (MMR) gene system, leading to inactivation of mutated tumour suppressor genes. Recently, a third pathway has been identified; this pathway is called the serrated pathway due to the serrated appearance of tumours upon histological analysis. This pathway involves hypermethylation of specific DNA regions near the promoter genes [4]. Although much progress has been made in molecular biology, novel approaches are still required to reveal the molecular mechanisms of CRC.

MicroRNAs (miRNAs) are short, 18–27-nucleotide, non-coding RNA sequences that function in post-transcriptional regulation by binding mRNA. Many studies have shown that miRNAs are involved in cell differentiation, proliferation, and apoptosis and act as tumour suppressor genes or proto-oncogenes [5]. Transcription factors (TFs) are modular proteins that regulate gene transcription by binding to the promoter region of target genes via their DNA-binding domain [6]. Both TFs and miRNAs are involved in the regulation of various biological processes, including cell proliferation, differentiation, and apoptosis.

With the rapid development of high-throughput technologies, such as RNA sequencing, gene expression analysis based on high-throughput data has become more important for exploring the biological mechanisms of disease. The Cancer Genome Atlas (TCGA) provides an additional high-throughput sequence dataset from thousands of tumours and patient-matched normal tissues, which may contribute to improvements in diagnostic methods, treatment standards, and, ultimately, cancer prognosis [7]. In this study, the expression profiles of genes and miRNAs were downloaded from TCGA and analysed. First, the differentially expressed genes and miRNAs were identified, and then Gene Ontology (GO) classification and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway analysis were performed. Last, the gene-miRNA co-expression network and TF-miRNA-gene network were established to study the molecular mechanisms and find new strategies for the prevention, diagnosis, and treatment of colon cancer (Fig. 1).

Material and methods

Data resources and preprocessing

The level 3 miRNASeq and RNAseqV2 data of colorectal adenocarcinoma (COAD) were downloaded from the TCGA database (http://cancergenome.nih.gov/) by the TCGA Assembler package in R (version 3.1.1) [8, 9]. The miRNA
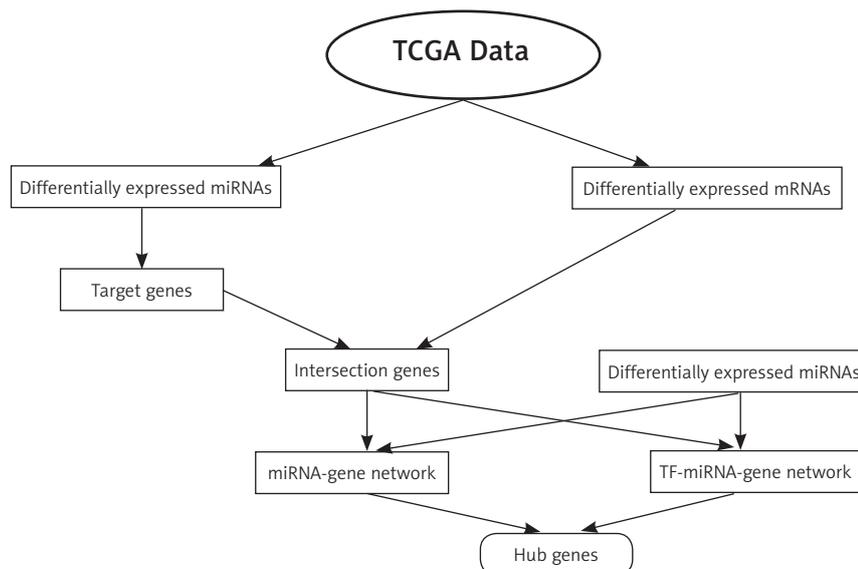
Fig. 1. Flowchart of study

and mRNA expression profiles were generated using Illumina HiSeq 2000 miRNASeq and Illumina HiSeq RNASeq sequencing platforms, respectively, and the RNA-Seq by the Expectation-Maximisation (RSEM) software package [10]. Raw count data and read count data were used to represent the mRNA and miRNA expression levels, respectively. There were 280 samples in total, including 272 colon adenocarcinomas and eight normal tissue samples; these samples were classified into two cohorts: a tumour group and a normal group. Each sample included the corresponding miRNAseq and RNAseq data. Because the data were obtained from TCGA, approval by an Ethics Committee was not needed. This study meets the TCGA publication guidelines.

Identification of differentially expressed genes and miRNAs

Differentially expressed genes and miRNAs were selected based on their fold change and adjusted p-values, which were generated by the DESeq package [11]. The inclusion criteria were set as follows: 1) FDR$Padj$< 0.05 and |log2(fold change)| > 1; and 2) gene and miRNA expression level > 0.

Screening of target genes of differentially expressed miRNAs

To understand the regulatory relationships between the differentially expressed miRNAs and genes, the target genes were predicted with the miRWalk2.0 database [12] (http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/index.html) and those that had been validated by experiments were chosen. The intersections of genes from target genes and differentially expressed mRNAs were then obtained to construct the miRNA-gene co-expression network.

GO classification and KEGG pathway analysis

We used the DAVID tool (Database for Annotation, Visualisation, and Integrated Discovery) [13] for GO classification and KEGG pathway analysis of the differentially expressed genes. The DAVID tool is a web-accessible program that pro-vides a comprehensive set of functional annotation tools for investigators to understand the biological meaning behind large lists of genes (http://david.abcc.ncifcrf.gov/). We selected the top 10 GO terms by their p value for analysis. We used R packages "ggplot2" and "clusterProfiler" to annotate modules with GO terms and to perform KEGG pathway enrichment analysis, respectively.

Construction of the miRNA-gene interaction network

First, we constructed the co-expression network from the identified differentially expressed genes and miRNAs based on their topological properties, such as a distance measure, using the WCGNA (Weighted Gene Co-expression Network Analysis) [14] package in R. Because we applied WGCNA to the gene and miRNA expression data in our analysis, the soft power threshold of β = 16 was selected, which is the smallest value that reaches level 0.9 on the independence scale (Fig. 2). Finally, a network graph was constructed and visualised using Cytoscape v3.3.0 [15] (http://cytoscape.org/). The core region was identified with MCODE (Molecular Complex Detection) [16], a Cytoscape plugin. MCODE can detect densely connected regions in large protein-protein interaction networks.

TF-miRNA-gene network

TFmiR was used (http://service.bioinformatik.uni-saarland.de/tfmir) to identify the TFs that might regulate these differentially expressed genes and miRNAs and to construct a TF-miRNA-gene network. TFmi [17] is a web server used for constructing and analysing disease-specific TF and miRNA co-regulatory networks.

Results

Differentially expressed genes and miRNAs

In total, 3436 differentially expressed genes (1709 up-regulated and 1727 downregulated) and 216 differentially
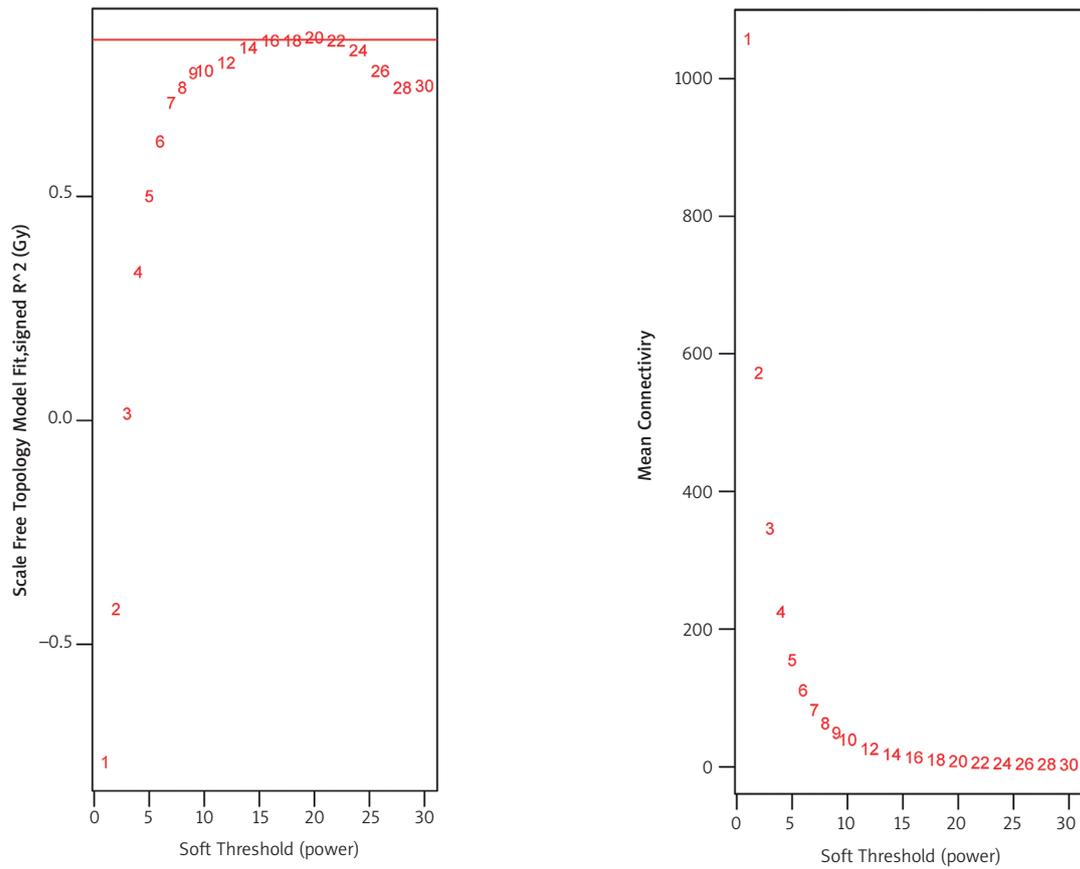
Fig. 2. Analysis of network topology for various soft-thresholding powers
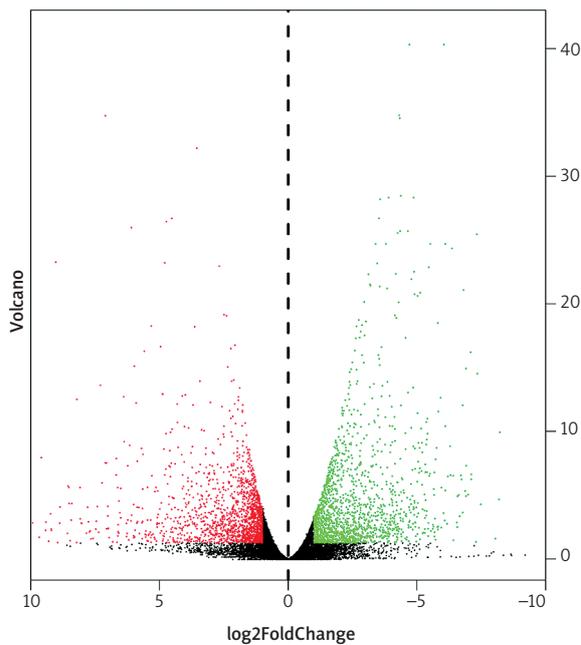

Fig. 3. Volcano plots for expression of differentially expressed genes (Red represents the up-regulated genes, while green represents the downregulated genes)
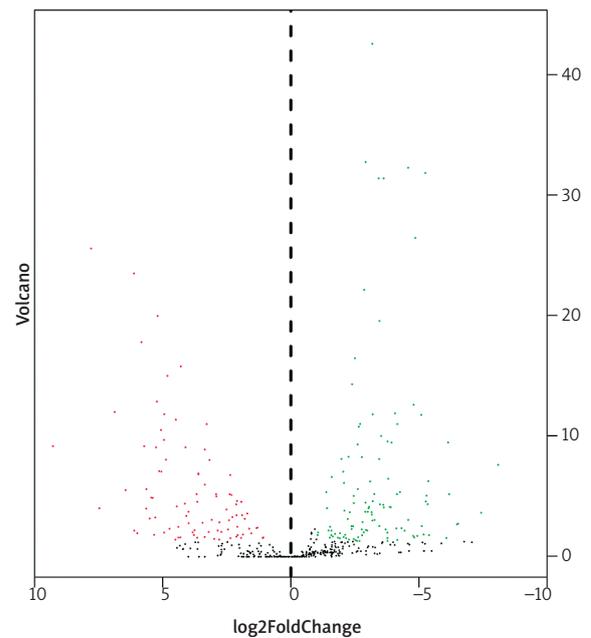

Fig. 4. Volcano plots for expression of differentially expressed microRNAs (Red represents the up-regulated miRNAs, while green represents the downregulated miRNAs)
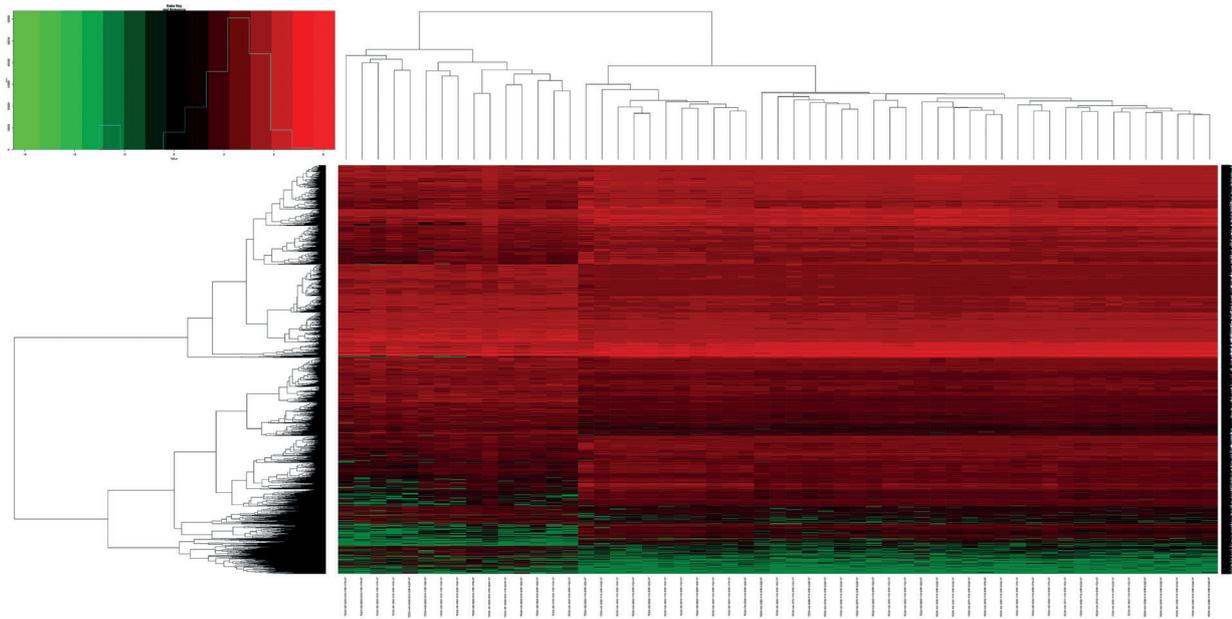
Fig. 5. Heat map showing hierarchical gene clustering analysis of the differentially expressed genes. Values used are normalised with log10. (The right longitudinal axis: the names of genes, the left longitudinal axis: the clustering information of genes, the upper abscissa axis: the clustering information of samples, the under abscissa axis: the sample code of cancer and normal. Red represents the up-regulated genes, while green represents the down-regulated genes)
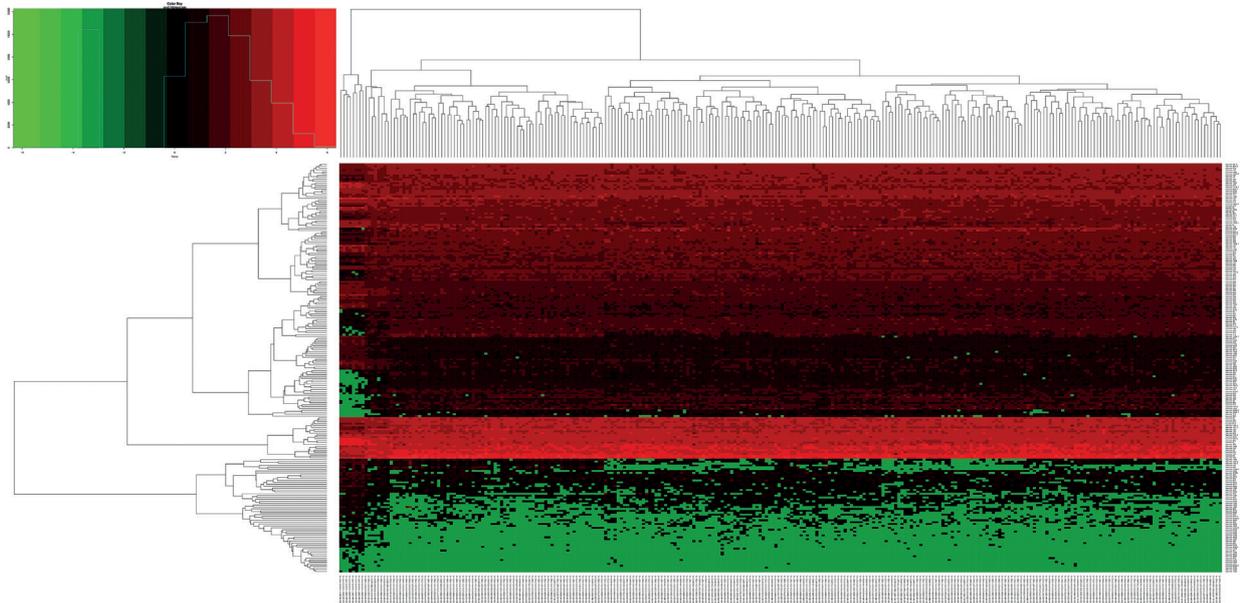

Fig. 6. Heat map showing hierarchical gene clustering analysis of the differentially expressed microRNAs. Values used are normalised with log10. (The right longitudinal axis: the names of miRNAs, the left longitudinal axis: the clustering information of miRNAs, the upper abscissa axis: the clustering information of samples, the under abscissa axis: the sample code of cancer and normal. Red represents the up-regulated miRNAs while green represents the down-regulated miRNAs).

expressed miRNAs (99 upregulated and 117 downregulated) were identified in colon cancer in comparison with normal tissues, with cut-offs of *padj* < 0.05 and |log2(fold change)| > 1. Volcano plots and heat maps were generated to select differentially expressed genes and miRNAs, as shown in Figs. 3–6. The top ten most differentially expressed genes and miRNAs are also shown in Table 1.

Functional enrichment analysis of the differentially expressed genes

We performed GO classification and KEGG pathway enrichment analysis for differentially expressed genes. We found some significant biological processes, including regulation of cell proliferation, cell adhesion, biological ad-

Table 1. The top ten most differential genes and miRNAs

Gene symbol	Log2 (fold change)	Padj	Change
CA7	−6.0	4.37E-41	down
SCARA5	−4.7	4.37E-41	down
CLEC3B	−4.3	1.53E-35	down
KRT80	7.1	1.63E-35	up
SLC4A4	−4.3	2.60E-35	down
TRIB3	3.5	5.79E-33	up
SPIB	−4.4	3.10E-29	down
CADM3	−4.9	4.26E-29	down
ADAMDEC1	−3.9	4.26E-29	down
C2orf88	−3.6	5.76E-29	down
hsa-let-7b	−3.2	2.44E-43	down
hsa-mir-99b	−2.9	1.65E-33	down
hsa-mir-197	−4.6	4.89E-33	down
hsa-mir-328	−5.2	1.33E-32	down
hsa-mir-125a	−3.6	3.69E-32	down
hsa-let-7d	−3.4	3.69E-32	down
hsa-mir-139	−4.9	3.25E-27	down
hsa-mir-374a	7.8	2.49E-26	up
hsa-mir-21	6.1	2.94E-24	up
hsa-mir-181a-1	−2.9	6.82E-23	down

Fig. 7. GO analysis of differently expressed genes. (Y axis: Statistically Enriched GO biological process terms, X axis: the number of differentially expressed genes in each GO Term)

hesion, cell cycle process, response to wounding, and others. The 10 top-ranked GO categories are shown in Fig. 7. Our results revealed that the differentially expressed genes were mainly involved in the cytokine-cytokine receptor interaction signalling pathway, cell adhesion molecule (CAM) signalling pathway, calcium signalling pathway, neuroactive ligand-receptor interaction signalling pathway, complement and coagulation cascade signalling pathway, and others (Fig. 8). In addition, 84 genes were enriched in the cytokine-cytokine receptor interaction signalling pathway (Fig. 9).

Analysis of the miRNA-gene regulatory network

WGCNA is a well-established method designed for constructing co-expression networks. Many network analysis studies have reported that a weighted network retains more information and is more robust than an unweighted network [18, 19]. According to the node connectivity, genes can be further classified into hub genes. Hub genes are very important nodes and represent a small proportion of nodes with maximal information exchange with other nodes in the gene co-expression network. In total, there were 106 nodes and 121 pairs of interaction relationships in the network (Fig. 10). The network included 13 miRNAs and 93 genes. To further explore the hub genes in detail, we identified a highly-connected core region with MCODE. Through MCOMD algorithm analysis, we identified six molecular complexes whose correlation integral values were higher than 2. The complex with the highest score contained seven nodes and 13 edges. The hub genes in-
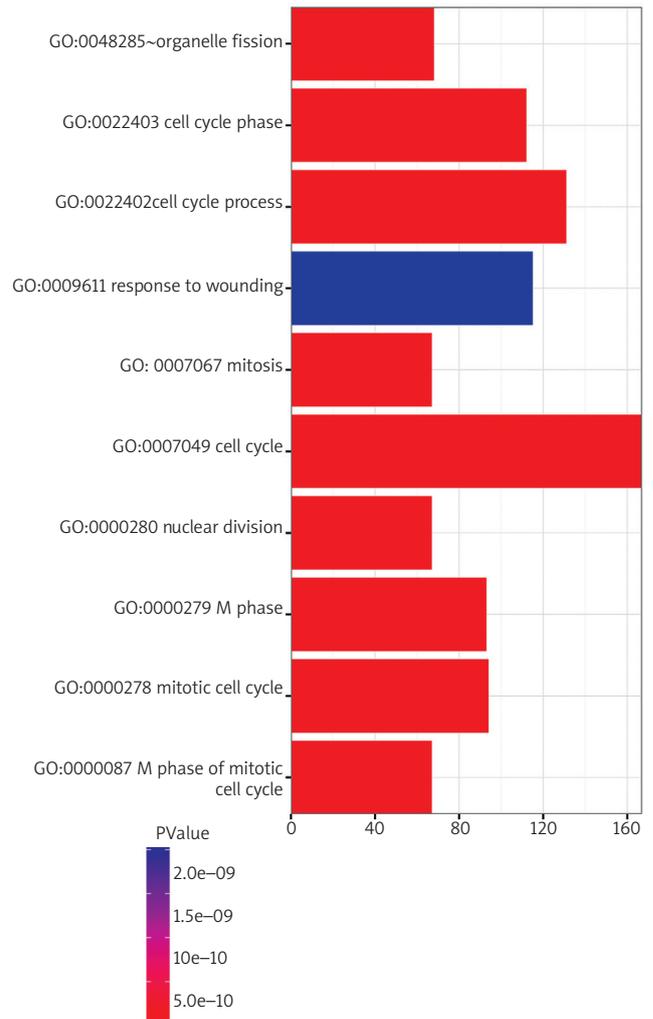
cluded GCNT4 (glucosaminyl transferase 4, core 2), EDN2 (endothelin 2, also called ET-2), miR-1295, miR-615, and others (Fig. 11).

TF-miRNA-gene network

We constructed the TF-miRNA-gene network of the differentially expressed miRNAs and genes using TFmiR, which is a web server, as shown in Fig. 12. LEF1 (lymphoid enhancer-binding factor 1) is the core TF that regulates the most differentially expressed miRNAs and genes. The most central mRNAs were MYC (also called c-myc) and WT1, and the most central miRNA was has-mir-34a.

Discussion

In this study, we analysed the differences in the gene expression profiles of normal and colon cancer samples and investigated the functions and pathways with which differentially expressed genes were mainly associated. We also constructed a miRNA-gene regulatory network and TF-miRNA-gene network based on the closely related

mRNA and miRNA expression data to better clarify the cellular mechanisms of colon cancer at the molecular level.

Here, we found that some hub genes, such as *GCNT4, EDN2,* and miR-1295, are massively altered in colon carcinoma. These genes and miRNAs can potentially be applied not only as novel biomarkers but also as therapeutic targets. For example, *GCNT4* belongs to the glucosaminyl (N-acetyl) transferases family. Altered glycosylation is considered a universal cancer hallmark. Though little is currently known about *GCNT4*, the other genes (*GCNT2* and *GCNT3*) of the glucosaminyl transferases family were identified to be related to colon cancer. Nakamura *et al*. showed that the expression of *GCNT2* was aberrant, which is explained by promoter DNA hypermethylation. Hypomethylation of the *GCNT2* variant 2 can reflect lymph node metastasis of CRC in the tumour and in normal tissues [20]. Additional, low *GCNT3* expression was proven to be a promising prognostic biomarker for colon cancer, which could be used to identify early-stage colon cancer patients at high risk of relapse [21]. Therefore, we speculated that the *GCNT4* may also have significant relationships with colon cancer. *EDN2* encodes a member of the endothelin protein family of secretory vasoconstrictive peptides. The overexpression of *ET-2* and *ET-3* significantly suppressed the migration and invasion of human colon cancer cells [22]. According to the miRNA-gene network, *GCNT4* and *EDN2* were regulated by miR615 and miR-1295. MiR-615 was reported to be a tumour suppressor in many cancers [23–25]. Because of promoter hypermethylation, miR-615 was abnormally downregulated in pancreatic cancer cells, which suppressed pancreatic cancer cell proliferation, migration, and invasion by directly targeting IGF2 and other genes, such as the proto-oncogene JUNB, and by interfering with the insulin signalling pathway [24]. The functions of miR-1295 are unclear.

KEGG pathway analysis showed that these genes were mainly enriched in cytokine-cytokine receptor interaction. Cytokines are soluble extracellular proteins that are crucial intercellular regulators and mobilisers of cells engaged in innate and adaptive inflammatory host defences; cell growth, differentiation, and death; angiogenesis; and development and repair processes aimed at restoring homeostasis. The aberrant expression of some cytokines may be associated with colon cancer; for example, IL-6 is greatly upregulated in colon cancer, and its effect on epithelial cells is multifaceted [26, 27]. Interleukin 6 aggravated tumorigenesis through STAT3 signalling in colon epithelial cells, which enhanced epithelial cell survival and proliferation [28]. Interleukin 6 also downregulated the tumour suppressor p53 and upregulated the oncogene *c-myc* in colon epithelial cells and EMT [29].

Gene expression is usually regulated in one of two ways: regulation of the transcription from DNA to RNA by TFs, or regulation of RNA stability by miRNAs [30]. In our study, LEF1*, MYC, WT1,* and miR-34a were located in the central hub of our TF-miRNA-gene network. As a transcript factor, LEF1 is involved in the Wnt signalling pathway, which has been demonstrated to be involved in CRC. LEF-1 expression is involved in the presence of KRAS mutations and has prognostic value in colorectal carcinoma
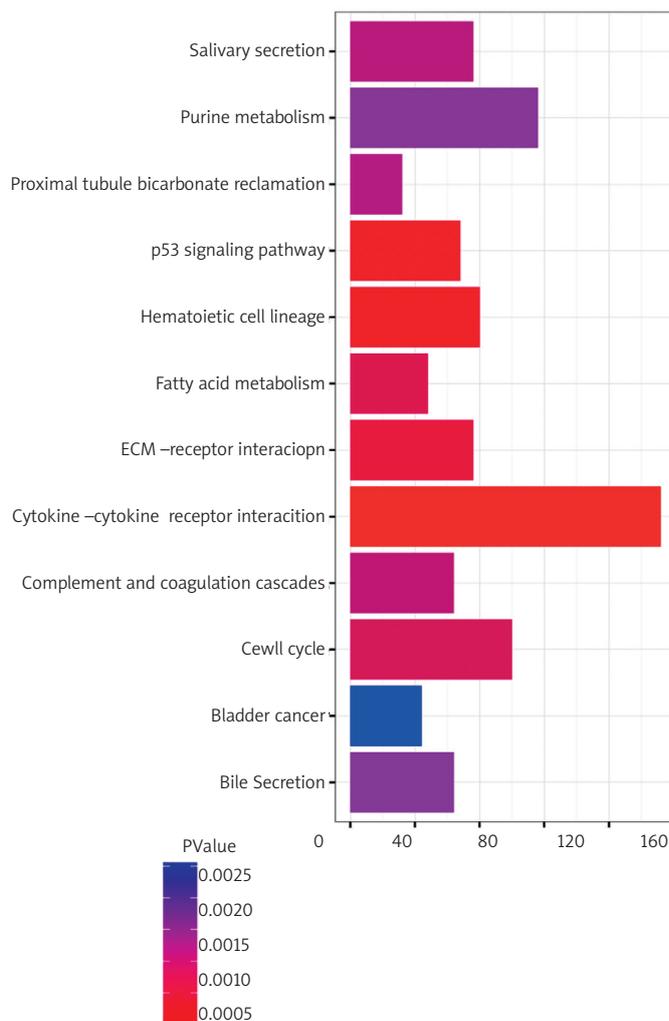

Fig. 8. KEGG pathway enrichment analysis. (Y axis: Statistically Enriched KEGG pathway, X axis: the number of differentially expressed genes in each pathway)

as a trend of worse overall survival [31]. LEF1 can target *CDH1* and *CTNNB1* genes, which closely interact in EMT activation [32]. LEF1 expression levels are associated with infiltration depth, lymph node, distant metastases, and advanced TNM stages of colon cancers as well as poor overall survival rate in patients with colon cancer [33]. As a negative regulator, miR-34a was reported to inhibit colon cancer proliferation and metastasis by targeting platelet-derived growth factor receptor α (PDGFRA) [33]. MiR-34a inhibits cell migration and invasion by regulating the SIRT1/p53 pathway in humans [34, 35]. *MYC* mutations, overexpression, rearrangement, and translocation have been associated with a variety of haematopoietic tumours, leukaemias, and lymphomas, including Burkitt's lymphoma. *MYC* plays an essential role in the regulation of many physiological processes, including cell cycle control, apoptosis, protein synthesis, and cell adhesion [36]. Among its related pathways are the PI3K-Akt signalling pathway and MAPK signalling pathway. Diseases associated with WT1 include Denys-Drash syndrome and Frasier syndrome. Among its related pathways are the integrated
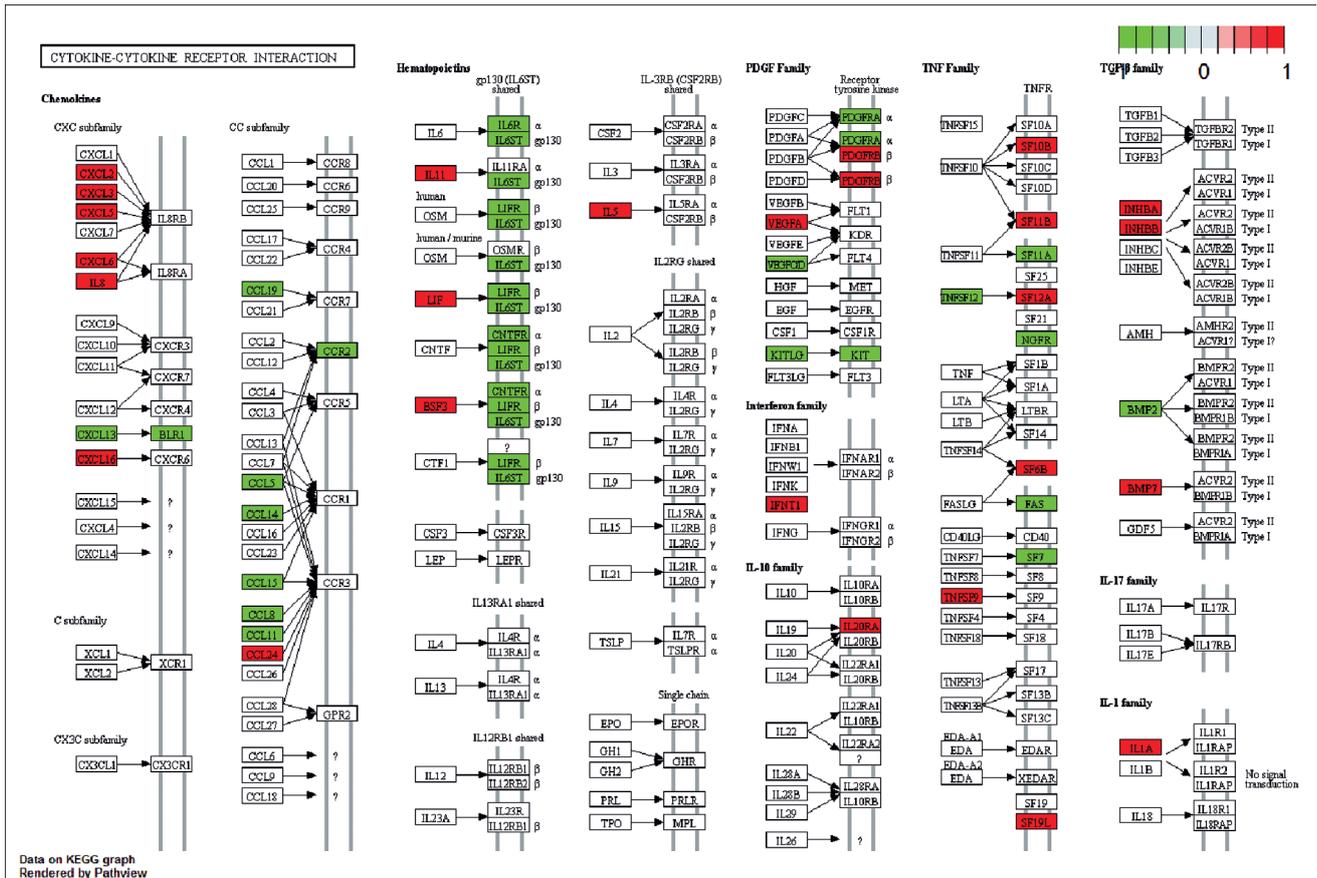
Fig. 9. The signalling pathway of cytokine-cytokine receptor interaction signalling pathway. (Red represents the up-regulated genes, while green represents the downregulated genes)
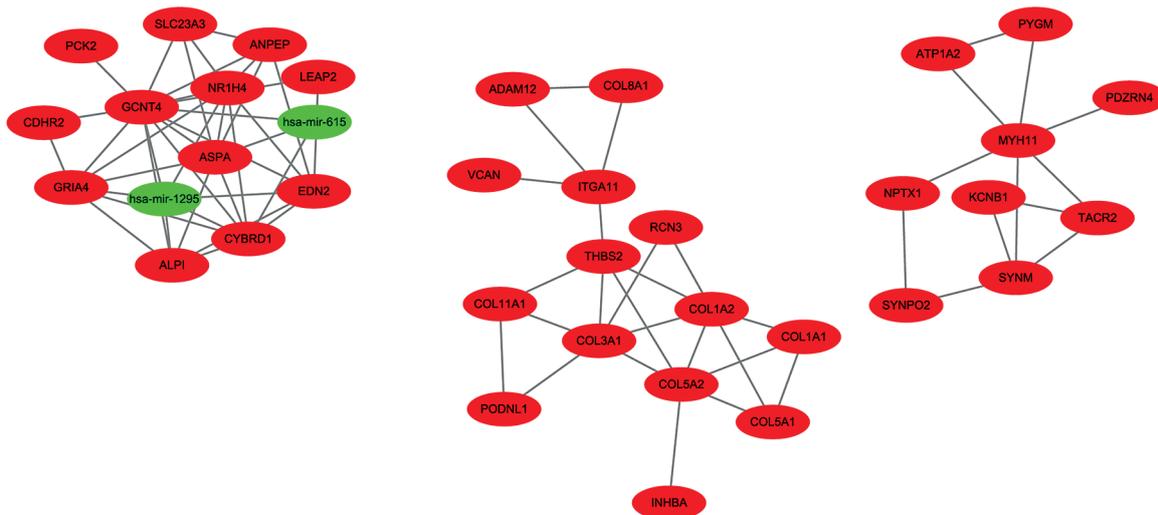

Fig. 10. Co-expression networks between miRNA and genes. (Red represents the genes, while green represents microRNAs)

pancreatic cancer pathway and transcriptional misregulation in cancer. Although aberration expression of *MYC* and *WT1* in colon cancer is frequently observed, future studies are required to explore the mechanism. To conclude, our study may provide useful information for understanding the TF-miRNA-gene network in CRC.
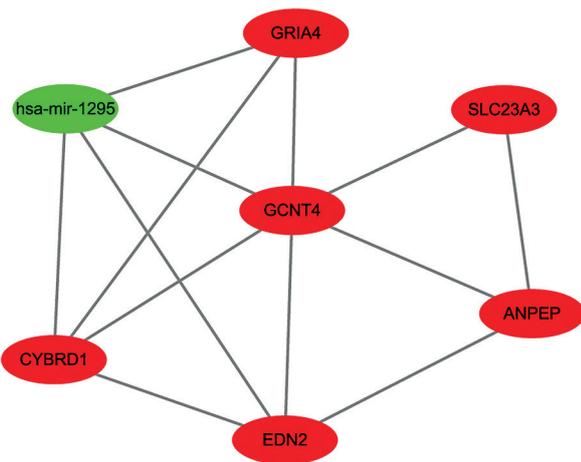
Fig. 11. Molecular complexes obtained by MCOMD algorithm analysis. (Red represents the genes, while green represents microRNAs)

Overall, our findings will improve our understanding of the molecular mechanisms of colon cancer and aid in finding potential targets for diagnostic and therapeutic usage.

Fig. 12. TF-miRNA-gene network in colon cancer. (Rectangles represent differentially expressed genes, hexagons for transcription factors, and ellipses for miRNAs. The lines indicate the interaction relationship. Hub genes are marked in red)

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010; 127: 2893-917.
2. Pawa N, Arulampalam T, Norton JD. Screening for colorectal cancer: established and emerging modalities. Nat Rev Gastroenterol Hepatol 2011; 8: 711-22.
3. Pino MS, Chung DC. The chromosomal instability pathway in colon cancer. Gastroenterology 2010; 138: 2059-72.
4. Leggett B, Whitehall V. Role of the serrated pathway in colorectal cancer pathogenesis. Gastroenterology 2010; 138: 2088-100.
5. Mitchell PS, Parkin RK, Kroh EM, et al. Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A 2008; 105: 10513-8.
6. Lee TI, Young RA. Transcription of eukaryotic protein-coding genes. Annu Rev Genet 2000; 34: 77-137.
7. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) 2015; 19: A68-77.
8. Zhu Y, Qiu P, Ji Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods 2014; 11: 599-600.
9. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing 2014.
10. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011; 12: 323.
11. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010; 11: R106.
12. Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. Nat Methods 2015; 12: 697.
13. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003; 4: P3.
14. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9: 559.
15. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13: 2498-504.
16. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003; 4: 2.
17. Hamed M, Spaniol C, Nazarieh M, Helms V. TFmiR: a web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks. Nucleic Acids Res 2015; 43: W283-8.
18. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005; 4: Article 17.
19. Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S. Weighted gene coexpression network analysis: state of the art. J Biopharm Stat 2010; 20: 281-300.
20. Nakamura K, Yamashita K, Sawaki H, et al. Aberrant methylation of GCNT2 is tightly related to lymph node metastasis of primary CRC. Anticancer Res 2015; 35: 1411-21.
21. Gonzalez-Vallinas M, Vargas T, Moreno-Rubio J, et al. Clinical relevance of the differential expression of the glycosyltransferase gene GCNT3 in colon cancer. Eur J Cancer 2015; 51: 1-8.
22. Wang JL, Hu Y, Kong X, Wang ZH, Chen HY, Xu J, Fang JY. Candidate microRNA biomarkers in human gastric cancer: a systematic review and validation study. PLoS One 2013; 8: e73683.
23. Ellinger J, Schefer A, Holdenrieder S, et al. Circulating microRNAs in serum. A novel biomarker for patients with bladder cancer? Urologe – Ausgabe A 2012; 51: 58.
24. Gao W, Gu Y, Li Z, et al. miR-615-5p is epigenetically inactivated and functions as a tumor suppressor in pancreatic ductal adenocarcinoma. Oncogene 2015; 34: 1629-40.
25. Sun Y, Zhang T, Wang C, Jin X, Jia C, Yu S, Chen J. MiRNA-615-5p functions as a tumor suppressor in pancreatic ductal adenocarcinoma by targeting AKT2. PLoS One 2015; 10: e0119783.
26. Podolsky DK. Inflammatory bowel disease. N Engl J Med 2002; 347: 417-29.
27. Chung YC, Chang YF. Serum interleukin-6 levels reflect the disease status of colorectal cancer. J Surg Oncol 2003; 83: 222-6.
28. Bollrath J, Phesse TJ, von Burstin VA, et al. gp130-mediated Stat3 activation in enterocytes regulates cell survival and cell-cycle
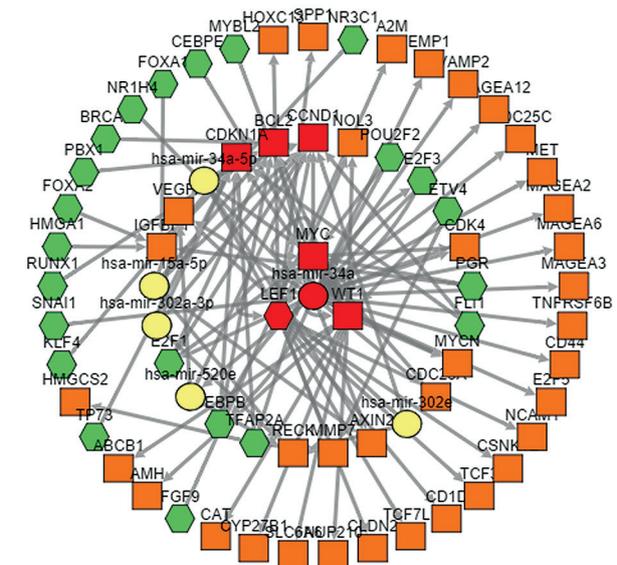
progression during colitis-associated tumorigenesis. Cancer Cell 2009; 15: 91-102.

29. Brighenti E, Calabrese C, Liguori G, Giannone FA, Trerè D, Montanaro L, Derenzini M. Interleukin 6 downregulates p53 expression and activity by stimulating ribosome biogenesis: a new pathway connecting inflammation to cancer. Oncogene 2014; 33: 4396-406.

30. Shi WY, Liu KD, Xu SG, Zhang JT, Yu LL, Xu KQ, Zhang TF. Gene expression analysis of lung cancer. Eur Rev Med Pharmacol Sci 2014; 18: 217-28.

31. Kermanshahi TR, Jayachandran P, Chang DT, Pai R. LEF-1 is frequently expressed in colorectal carcinoma and not in other gastrointestinal tract adenocarcinomas: an immunohistochemical survey of 602 gastrointestinal tract neoplasms. Appl Immunohistochem Mol Morphol 2014; 22: 728-34.

32. Martinelli M, Palmieri A, Rodia MT, et al. CDH1 polymorphisms and low expression of E-cadherin and beta-catenin in colorectal cancer patients. J Biol Regul Homeost Agents 2015; 29: 89-96.

33. Li C, Wang Y, Lu S, Zhang Z, Meng H, Liang L, Zhang Y, Song B. MiR-34a inhibits colon cancer proliferation and metastasis by inhibiting platelet-derived growth factor receptor alpha. Mol Med Rep 2015; 12: 7072-8.

34. Gao J, Li N, Dong Y, et al. miR-34a-5p suppresses colorectal cancer metastasis and predicts recurrence in patients with stage II/III colorectal cancer. Oncogene 2015; 34: 4142-52.

35. Lai M, Du G, Shi R, et al. MiR-34a inhibits migration and invasion by regulating the SIRT1/p53 pathway in human SW480 cells. Mol Med Rep 2015; 11: 3301-7.

36. Dang CV. MYC on the path to cancer. Cell 2012; 149: 22-35.

Address for correspondence

Chengzu Wang
999 South Second Ring Road
CiXi, Zip Code: 315300
e-mail: wczly1988@163.com