Microarray analysis of differentially-expressed genes and linker genes associated with the molecular mechanism of colorectal cancer

XINGJIE SHEN1*, MENG YUE1*, FANSHENG MENG2, JINGYU ZHU1, XIAOYAN ZHU1 and YAKUN JIANG1

Departments of 1Gastroenterology and 2Pharmacy, Jinan Central Hospital Affiliated to Shandong University, Jinan, Shandong 250013, P.R. China

Received April 8, 2015; Accepted June 16, 2016

DOI: 10.3892/ol.2016.5122

Abstract. Colorectal cancer (CRC) is one of the most prevalent malignancies worldwide and remains the third leading cause of cancer-associated mortality. The present study aimed to fully elucidate the pathogenesis of CRC and identify associated genes in tumor development. Microarray GSE44076, GSE41328 and GSE44861 datasets were downloaded from the Gene Expression Omnibus database and integrated with meta-analysis. Differentially-expressed genes (DEGs) were identified from CRC samples compared with adjacent non-cancerous controls using the Limma package in R, followed by functional analysis using the Database for Annotation, Visualization, and Integrated Discovery online tool. A protein-protein interaction (PPI) network of DEGs and linker genes was constructed using NetBox software and modules were also mined. Functional annotation was performed for modules with a maximum number of nodes. Subsequent to meta-analysis to pool the data, one dataset that included 327 samples involved in 11,081 genes was obtained. A total of 697 DEGs were identified between CRC samples and adjacent non-cancerous controls. In the PPI network, modules 1 and 5 contained the maximum number of nodes. Collagen, type I, α1 (COL1A1), COL1A2 and matrix metallopeptidase 9 (MMP9) in module 1 and UDP-glucose 6-dehydrogenase (UGDH), aldehyde dehydrogenase 1 family, member A1 (ALDH1A1), fatty acid binding protein 4 (FABP4) and monoglyceride lipase (MGLL) in module 5 exhibited a high degree of connectivity. Functional analysis indicated that the genes in module 1 were involved in extracellular matrix (ECM)-associated functions and that the genes in module 5 were involved in metabolism-related functions. Overall, significant DEGs and linker genes, namely COLIA1, COLIA2, MMP9, UGDH, ALDH1A1, FABP4 and MGLL, play a crucial role in the development of CRC via regulating the ECM and cell metabolism.

Introduction

Colorectal cancer (CRC), also known as colon cancer, is one of the most prevalent malignancies worldwide and remains the third leading cause of cancer-associated mortality (1). In 2014, ~65,000 women and 71,830 men were estimated to be diagnosed with CRC (2).

Similar to the majority of other complex tumors, CRC has been the subject of multiple studies with regards to its pathogenesis, diagnosis and therapy. Much has been elucidated about the molecular mechanism of CRC in recent years. It is widely recognized that chromosomal instability is the most common genetic abnormality to occur in CRC and has been found in almost 85% of all CRC cases (3). Key genes involved in this pathway include Kirsten rat sarcoma viral oncogene homolog (KRAS), deleted in colorectal carcinoma (DCC), SMAD family member 2 (SMAD2) and SMAD4. KRAS is a proto-oncogene that plays a critical role in the transduction of intracellular signals. The activation of KRAS by binding to guanosine triphosphate could regulate downstream mediator mitogen-activated protein kinase, which is involved in cell division (4). Additionally, SMAD2 and SMAD4 play a vital role in the transforming growth factor-β signaling pathway, which in involved in the regulation of cell proliferation, differentiation and apoptosis (5). DCC has been shown to correlate with metastasis and a poor prognosis in CRC (6). Moreover, chronic inflammation has been shown to increase the incidence of bowel cancer (7). In the process of inflammation, cyclooxygenase-2 (COX2) is a key molecule highlighted by experiments (8). Previous studies into the downstream effects of COX have illustrated that basic fibroblast growth factor and vascular endothelial growth factor are activated by COX2 via prostaglandin E2, all of which are involved in the regulation of cell proliferation and angiogenesis contributing to tumor development (9-11). However, the pathogenesis of CRC is
complex and multifactorial. The complete elucidation of its etiology remains to be defined.

The present study analyzed three microarray datasets, comparing between colon tumor samples and adjacent normal mucosa tissue samples. Differentially-expressed genes (DEGs) were identified and functional annotation was performed for significant genes, followed by protein-protein interaction (PPI) network construction. The study aimed to detect the molecular mechanisms and associated genes in the development of CRC.

Materials and methods

Microarray data. A total of 3 microarray datasets were downloaded from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/), including GSE44076 (12), GSE41328 (13) and GSE44861 (14). Expression data from GSE44076, which included colon tumor samples from 98 patients and adjacent paired normal mucosa tissues from 50 healthy donors, were obtained using platform GPL13667 (Affymetrix Human Genome U219 Arrays). Microarray data from GSE41328, which included 5 colorectal adenocarcinomas samples and 5 matched normal colon tissue, were generated with the [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array platform. Expression data from GSE44861, which included 56 tumor samples and 55 adjacent non-cancerous tissue samples, were obtained through the [HT_HG-U133A] Affymetrix HT Human Genome U133A Array platform.

Data preprocessing. Prior to analysis, probe identifications in each dataset were converted into standard gene symbols. For genes with more than one probe set in the array, the average value for the probes was obtained as the expression value of the gene. By contrast, the probe set was deleted when mapped to more than one gene. As the genes were different in the 3 datasets, meta-analysis was performed of these studies, pooling the microarray data across different platforms. In the combined process, batch effects are inevitable. To adjust the data for these batch effects, the surrogate variable analysis (SVA) method was applied, and normalization was performed using the preprocessCore package (16) in R.

Identification of DEGs in CRC. To identify significant DEGs in colon tumor samples compared with adjacent non-cancerous controls, preprocessed data were exported to Limma package in R language (17). An adjusted P-value was estimated using the Benjamini & Hochberg (BH) method (18). Significant DEGs were identified as those with log2 FC (fold-change)>1 and an adjusted P-value of <0.05.

Functional annotation of DEGs in CRC. Testing for functional enrichment of DEGs in CRC was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) online tool (19). Categories analyzed included Gene Ontology (GO) terms (20) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (21). Data from the GO annotations was used to construct a functional enrichment network, which was visualized by the enrichment map plugin in Cytoscape (22). The BH correction for multiple testing was performed with a cutoff for an adjusted P-value of <0.05.

PPI network construction. NetBox software, which is written in the Java language, is used to store and establish the Human Interaction Network based on public databases consisting of Reactome (23,24), the Human Protein Reference Database (25), Memorial Sloan-Kettering Cancer Center Cancer Cell Map (26) and the National Cancer Institute-Nature Pathway Interaction Database (27). Linker genes with statistical significance, which are not differentially-expressed in colon tumors, but interact with DEGs, were obtained through mapping DEGs onto the network. Cytoscape software (28) was used to visualize the molecular interaction. Besides the PPI network under the criteria, NetBox also divided the network into modules. The modules with the maximum number of nodes in the PPI network were subjected to GO terms and Swiss-Prot and Protein Information Resource Keywords enrichment analysis with the DAVID online tool.

Results

Preprocessed results and DEGs in CRC. Following meta-analysis of these 3 studies to pool microarray data across the different platforms, one dataset was obtained that included 327 samples and 11,081 genes. The dataset was preprocessed and normalized, followed by further analysis. The normalized results are shown in Fig. 1. A total of 697 genes were selected as DEGs, including 286 upregulated and 411 downregulated genes, between CRC samples and adjacent non-cancerous control.

Significant functions and pathways of DEGs. To annotate these DEGs in the tumor samples, DAVID was used for GO function and KEGG pathway analysis, with the threshold of the adjusted P-value at <0.05. Functional enrichment networks of upregulated and downregulated DEGs are shown in Figs. 2 and 3. The results showed that upregulated DEGs were significantly enriched in cell cycle-related functions, including the cell cycle process, the regulation of the mitotic cell cycle and the regulation of cell proliferation. Downregulated DEGs were mainly enriched in homeostasis-related functions, including chemical homeostasis and cellular ion homeostasis.

PPI network analysis. Significant DEGs and linker genes were used to construct the PPI network (Figs. 4 and 5). In the PPI network for upregulated DEGs (Fig. 4), there were 2,508 edges and 296 genes, including 140 DEGs and 156 linker genes. The network was divided into 9 modules by NetBox, in which module 1 contained the maximum number of nodes. Additionally, in the PPI network for downregulated DEGs, there were 301 edges and 165 genes, including 113 DEGs and 42 linker genes. The network was divided into 18 modules, in which module 5 contained the maximum number of nodes. In the PPI network, the hub genes were mined with the top-five degrees of connectivity in the different modules (Table I). The upregulated minichromosome maintenance complex component 7 gene in module 0, linker genes collagen type I, α1 (COL1A1) and COL1A2, and differentially-expressed matrix metallopeptidase 9 (MMP9) in module 1, and linker genes polo-like kinase 1 and exportin 1 in module 2 exhibited a connectivity degree of >20. Downregulated genes
UDP-glucose 6-dehydrogenase (*UGDH*), aldehyde dehydrogenase 1 family, member A1 (*ALDH1A1*), fatty acid binding protein 4, adipocyte (*FABP4*) and monoglyceride lipase (*MGLL*) in module 5 exhibited a connectivity degree of >20.

The functional annotation results showed that the DEGs in module 1 were mainly enriched in extracellular region-related functions and extracellular matrix (ECM)-associated functions (Table II). Downregulated DEGs in module 5 were significantly enriched in metabolic process and biosynthetic process-related functions (Table III).

### Discussion

Using a meta-analysis approach to group 3 microarray datasets, including GSE44076, GSE41328 and GSE44861, DEGs were identified in CRC mucosa compared with adjacent normal mucosa samples. The results suggested that there were 697 DEGs, including 286 upregulated genes. Functional annotation results showed that the upregulated DEGs were involved in cell cycle-related functions, in comparison with the downregulated DEGs, which were enriched in homeostasis-associated functions. In the PPI network, the linker genes *COL1A1* and *COL1A2*, and the DEGs MMP9, *UGDH*, *ALDH1A1*, *FABP4* and *MGLL*, which exhibited a connectivity degree of >20, participated in the development of CRC.

After the upregulated and downregulated networks were divided into multiple modules, modules 1 and 5 with the maximum number of nodes were subjected to functional annotation. In module 1, the linker genes *COL1A1* and *COL1A2*, and the DEG MMP9 exhibited the highest degree of connectivity. *COL1A1* and *COL1A2*, two type I collagen members, are major components of the ECM. Growing evidence has shown that the ECM plays a critical role in promoting epithelial-to-mesenchymal transition (EMT), which is associated with tumor invasion and metastasis (29).

Additionally, EMT is indicated to confer tumor cell resistance to apoptosis and to promote the escape of tumor cells from the senescence process (30,31). Moreover, a pioneer study uncovered the fact that EMT has the capacity of endowing tumor cells with cancer stem cell-like characteristics, which could promote tumor development and chemoresistance (32).

MMP9 (also known as gelatinase B), a member of the MMP family, has been proven to degrade various components of the ECM, including type I collagen (33). Notably, an elevated level of *MMP9* has been found in CRC (34), which is consistent with the present analysis. Numerous studies have shown that MMP9 plays crucial roles in invasion, metastasis, cell proliferation and angiogenesis (35,36). Angiogenesis and cell proliferation are critically important for tumor development and metastatic spreading (37). From the results of the functional annotation in the present study, the three important genes in module 1 were mainly enriched in ECM-related functions and the ECM-receptor interaction pathway. Accordingly, *COL1A1*, *COL1A2* and *MMP9* are involved in CRC tumorigenesis and metastasis via regulation of ECM-associated functions.

In module 5, *UGDH*, *ALDH1A1*, *FABP4* and *MGLL* were downregulated in colorectal tumor samples and were significantly involved in metabolism-related functions. *UGDH*...
is the four-electron transfer enzyme and is associated with the biosynthesis of hyaluronan (HA), which participates in tissue organization, development and cell proliferation (38). A previous study showed that elevated levels of HA are directly involved in the progression of various cancers, and UGDH has been proposed as a biomarker for prostate cancer (39).
In parallel, ALDH1A1, which belongs to a superfamily of enzymes, has been identified as a crystalline in the lens and cornea (40). Notably, ALDH1A1 also plays a critical role in regulating lipid metabolism and gluconeogenesis (41). In addition, FABP4, known as a new adipokine, is involved in fatty acid trafficking from the cytoplasm to the nucleus and in lipid metabolism (42). Furthermore, FABP4 is considered as a candidate biomarker of lipodystrophy and metabolic syndrome (43). It is well known that MGLL is a member of the serine hydrolase superfamily, which hydrolyze intracellular triglyceride and cholesteryl ester into free fatty acid as an important fuel in mammals (44). More recently, MGLL was found to be abnormally expressed in aggressive human cancer, and to promote cell proliferation and tumor growth (45). It is now clear that the conversion of cells from a normal to cancerous state requires metabolic alterations,
including changes in lipid metabolism and gluconeogenesis, in order to support tumor growth and survival. As a result, 
**UGDH, ALDH1A1, FABP4, and MGLL** play a key role in metabolism-related functions and regulate the tumorigenesis of CRC.

Taken together, the present results suggest that **COL1A1, COL1A2, and MMP9** in module 1, and **UGDH, ALDH1A1, FABP4, and MGLL** in module 5 serve as key hub genes in CRC development, where the genes regulate ECM and cell metabolism-associated functions that are important for tumor
Figure 4. Protein-protein interaction network of up-regulated genes and linker genes. The coloration of the nodes is representative of the genes in different modules. Circular nodes represent differentially-expressed genes and rhombic nodes represent linker genes.

Figure 5. Protein-protein interaction network of downregulated genes and linker genes. The coloration of the nodes is representative of the genes in different modules. Circular nodes represent differentially-expressed genes and rhombic nodes represent linker genes.
growth. However, additional experiments will be required to confirm the bioinformatic results.

References

1. Kemp Z, Thrillwel C, Sieber O, Silver A and Tomlison I: An update on the genetics of colorectal cancer. Hum Mol Genet 13: 2667-77, 2004.
2. Siegel R, DeSantis C and Jemal A: Colorectal cancer statistics, 2014. CA Cancer J Clin 64: 104-117, 2014.
3. Grady WM and Carethers JM: Genomic and epigenetic instability in colorectal cancer pathogenesis. Gastroenterology 135: 1079-1091, 2008.
4. Hatzivassiliou G, Song K, Yen I, Brandhuber BJ, Anderson DJ, Alvarado R, Ludmij M, Stokoe D, Grool SL, Vigers G, et al: RAF inhibitors prime wild-type RAF to activate the MAPK pathway and enhance growth. Nature 464: 431-435, 2010.
5. Bellam N and Pasche B: Tgf-beta signaling alterations and colon cancer. Cancer Treat Res 155: 85-103, 2010.
6. Chang SC, Lin JK, Yang SH, Wang HS, Li AF and Chi CW: Relationship between genetic alterations and prognosis in sporadic colorectal cancer. Int J Cancer 118: 1721-1727, 2006.
7. Harraz M N and Abou-Serour A D: Colon cancer dysplasia in chronic inflammatory bowel disease: Pathology, clinical implications and pathogenesis. Arch Pathol Lab Med 134: 876-895, 2010.
8. Zisman TL and Rubin DT: Colorectal cancer and dysplasia in inflammatory bowel disease. World J Gastroenterol 16: 2660-2669, 2010.
9. Wang D and Dubois RN: Prostaglandins and cancer. Gut 55: 115-122, 2006.
10. Eisinger AL, Prescott SM, Jones DA and Stafforini DM: The role of cyclooxygenase-2 and prostaglandins in colon cancer. Prostaglandins Other Lipid Mediat 82: 147-154, 2007.
11. Doherty GA, Byrne SM, Molloy ES, Malhotra V, Austin SC, Kay EW, Murray FE and Fitzgerald DJ: Proenoneplastic effects of PGE2 mediated by EP4 receptor in colorectal cancer. BMC Cancer 9: 207, 2009.
12. Sanz-Pamplos R, Berenguer A, Cordero D, Mollevi DG, Crous-Bou M, Sole X, Paré-Brunet L, Guino E, Salazar R, Santos C, et al: Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. Mol Cancer 13: 46, 2014.
13. Lin G, He X, Ji H, Shi L, Davis RW and Zhong S: Reproducibility probability score-incorporating measurement variability across laboratories for gene selection. Nat Biotechnol 24: 1476-1477, 2006.
14. Ryan BM, Canetti KA, Robles AJ, Schetter AJ, Goodman J, Hayes RB, Huang WY, Gunter MJ, Yeager M, Burdette L, et al: Germline variation in NCF4, an innate immunity gene, is associated with an increased risk of colorectal cancer. Int J Cancer 134: 1399-1407, 2014.
15. Reiss D, Williams E, Parker HS, Fertig EJ, Jaffe AE and Storey JD: Package ‘SVA’: Surrogate Variable Analysis. R package version 3, 2013. https://www.biocductor.org/packages/devel/bioc/manuals/sva/man/sva.pdf.
16. BolstadBW: Package ‘preprocessCore’: A collection of pre-processing functions. R package version 1, 2013. https://www.bioconductor.org/packages/devel/bioc/manuals/preprocessCore/man/preprocessCore.pdf.
17. Smyth GK: Limma: Linear models for microarray data. In: Bioinformatics and computational biology solutions using R and Bioconductor. Springer, NY, pp 397-420, 2005.
18. Ferreira JA: The Benjamini-Hochberg method in the case of discrete test statistics. Int J Biostat 3: 1, 2007.
19. Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA: DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol 4: P3, 2003.
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29, 2000.
21. Kanehisa M and Goto S: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-43, 2000.
22. Merico D, Isserlin R and Bader GD: Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map. Methods Mol Biol 781: 257-277, 2011.
23. Yoshi-Tope G, Gillespie M, Vastrik L, D'Eustachio P, Schmidt E, de Bonito L, Gopinath GR, Wu GR, Matthews L, et al: Reactome: A knowledgebase of biological pathways. Nucleic Acid Res 33: D428-D432, 2005.