RESEARCH ARTICLE

Analysis of Key Genes and Pathways Associated with Colorectal Cancer with Microarray Technology

Yan-Jun Liu, Shu Zhang*, Kang Hou,Yun-Tao Li, Zhan Liu, Hai-Liang Ren, Dan Luo, Shi-Hong Li

Abstract

Objective: Microarray data were analyzed to explore key genes and their functions in progression of colorectal cancer (CRC). Methods: Two microarray data sets were downloaded from Gene Expression Omnibus (GEO) database and differentially expressed genes (DEGs) were identified using corresponding packages of R. Functional enrichment analysis was performed with DAVID tools to uncover their biological functions. Results: 631 and 590 DEGs were obtained from the two data sets, respectively. A total of 32 common DEGs were then screened out with the rank product method. The significantly enriched GO terms included inflammatory response, response to wounding and response to drugs. Two interleukin-related domains were revealed in the domain analysis. KEGG pathway enrichment analysis showed that the PPAR signaling pathway and the renin-angiotensin system were enriched in the DEGs. Conclusions: Our study to systemically characterize gene expression changes in CRC with microarray technology revealed changes in a range of key genes, pathways and function modules. Their utility in diagnosis and treatment now require exploration.

Keywords: Colorectal cancer - differentially expressed genes - microarray - pathway - functional enrichment analysis

Asian Pacific J Cancer Prev, **14 (3)**, 1819-1823

Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer in the world, but it is more common in developed countries. Around 60% of cases were diagnosed in the developed world. It is estimated that worldwide, in 2008, 1.23 million new cases of colorectal cancer were clinically diagnosed, and that it killed 608,000 people (Ferlay et al., 2010).

Bad lifestyle and aging are the main risk factors while genetic disorders also contribute to the incidence of CRC. It typically starts in the lining of the bowel and if left untreated, can grow into the muscle layers underneath, and then through the bowel wall. Therefore, early diagnosis is critical to decrease the mortality.

Gene mutations (Johnson et al., 2005; Talseth-Palmer et al., 2010) and SNPs (Menin et al., 2006; Aizat et al., 2011) have been found to be linked with CRC and some of them are suggested to be biomarkers. Discovering and validating protein biomarkers are the research hotspots (Zhai et al., 2012). Liu et al suggest that CD73 is a novel prognostic biomarker for human colorectal cancer (Liu et al., 2012). Of course circulating proteins exhibit more usefulness in clinical applications and some have been validated, like cytokeratin 18 (Greystoke et al., 2012) and M2-pyruvate kinase (Meng et al., 2012). Besides, Takahashi et al report that MiR-148a can act as a predictive biomarker in patients with advanced colorectal cancer (Takahashi et al., 2012). Various technologies are utilized to carry out the researches, among which proteomic tool (de Wit et al., 2012) and microarray technology (Amid et al., 2012) are two powerful methods enabling global identification of potential biomarkers.

Though many achievements have been obtained, our understandings about the disease are far from enough to control it clinically. In present study, two microarray data sets were analyzed to identify differentially expressed genes (DEGs). These DEGs can be potential biomarkers. Functional enrichment analysis was then carried out for the DEGs to elucidate their biological functions and thus help to uncover the pathogenesis of CRC.

Materials and Methods

Microarray data and DEGs

Microarray data (GSE4107 (Hong et al., 2007) and GSE8671 (Sabates-Bellver et al., 2007)) were downloaded from GEO database (Edgar et al., 2002). Both raw data were collected in the following platform: GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. Package limma (Smyth, 2004) of R was used to identify DEGs from each data set. The original data

*Department of General Surgery, The Third People's Hospital of Chengdu, The Second Clinical College Affiliated to Chongqing Medical University, Chengdu, Sichuan, China. *For correspondence: zhangshucd@hotmail.com*

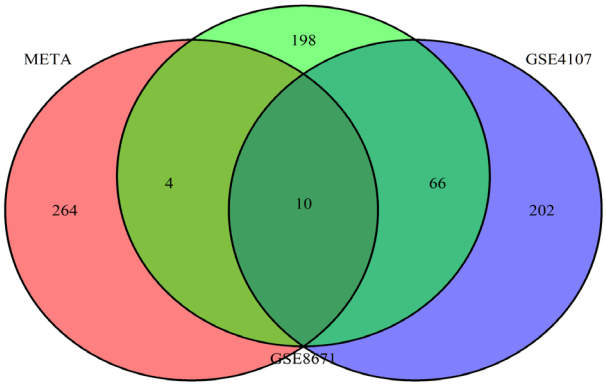

Figure 1. Venn Diagram for the GO Terms from GSE4107, GSE8671 and Meta-analysis. The gray area is the intersection of the three groups

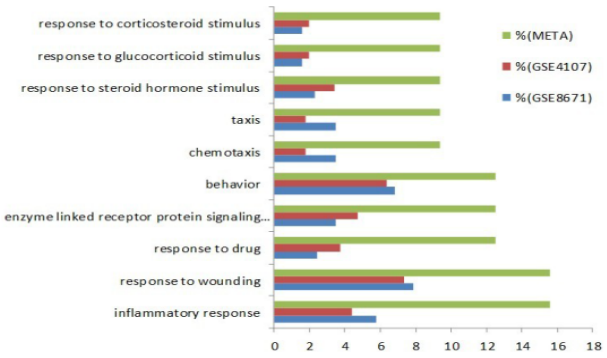

Figure 2. Percentage of DEGs for Each Enriched GO Terms in GSE4107 (red), GSE8671 (blue) and Meta-analysis (green). Only significantly enriched GO terms are shown (hypergeometric test, *p*-value < 0.05)

were processed by Bioconductor with RMA method and default settings, and then linear model was adopted. Fold change of >2 and *p*-value of < 0.05 were set as the cut-offs to screen DEGs.

Screening of common DEGs

The rank product package (Hong et al., 2006) was used to identify the common DEGs between control group and disease group. Briefly, genes were ranked based on up- or down-regulation by the disease group in each experiment. Then a combined probability was calculated for each gene as a rank product (RP). The RP values were used to rank the genes based on how likely it was to observe them by chance at that particular position on the list of DEGs. The RP can be interpreted as a p-value. To determine significance levels, the RP method uses a permutation-based estimation procedure to transform the p-value into an e-value that addresses the multiple testing problems derived from testing many genes simultaneously. Genes with a percentage of false-positives (PFP) ≤ 0.05 were considered differentially expressed between treatments and control in each experiment.

GO enrichment and IntroPro domain analysis

Gene Ontology (GO) Biological Process (BP) data and functional domain data were extracted using the DAVID (Huang da et al., 2009). GO terms and domains with less than 2 genes were discarded. Over-represented groups of GO BP terms and IntroPro functional domains (Hunter et al., 2009) were identified using a hypergeometric test, with a threshold of *p*-value <0.05.

Pathway analysis

We adopted an impact analysis that includes the statistical significance of the set of pathway genes but also considers other crucial factors such as the magnitude of each gene's expression change, the topology of the signaling pathway, their interactions, etc (Draghici et al., 2007).

In this model, the Impact Factor (IF) of a pathway Pi is calculated as the sum of two terms:

$$IF(Pi) = \log(\frac{1}{pi}) + \frac{\sum_{g \in Pi} |PF(g)|}{|\Delta E| \cdot N_{de}(Pi)}$$

The first term is a probabilistic term that captures the significance of the given pathway Pi from the perspective of the set of genes. This term captures the information provided by the currently used classical statistical approaches and can be calculated using either an ORA (Doniger et al., 2003) or contingency tables (Pan et al., 2003). The Pi value corresponds to the probability of obtaining a value of the statistic used at least as extreme as the one observed, when the null hypothesis is true. The results presented here were obtained using the hypergeometric model in which pi is the probability of obtaining at least the observed number of differentially expressed gene, N_{de}, just by chance.

The second term is a functional term that depends on the identity of the specific genes that are differentially expressed as well as on the interactions described by the pathway.

Results

Differentially expressed genes

A total of 631 genes from GSE4107 and 590 genes from GSE8671 were selected out as their fold change larger than 2 and p-value less than 0.05. The rank product package was used to further determine the differential expression and finally 32 genes with a percentage of false-positives (PFP) < 0.05 were obtained as the common DEGs.

Functional enrichment analysis results

Biological processes enrichment analysis was performed for the DEGs obtained from GSE4107, GSE8671 and meta-analysis with DAVID tool to gain insights into their functions. *P* value <0.05 was set as the threshold and 278, 155 and 13 GO terms were obtained, respectively. Package VennDiagram of R was chosen to generate Venn diagram (Figure 1). A total of 10 terms were shared by the three groups. They were related with inflammatory response, response to wounding, response to drug, etc. (Figure 2).

DOMAIN analysis results

To add meaningful information to the results from the GO term enrichment, we extended our investigation to IntrPro protein domains. Common and significant

Table 1. DEGs Contained in Each Enriched Domain in CRC

Term	META		GSE4107		GSE8671	
	Count	*P* value	Count	*P* value	Count	*P* value
IPR001811:Small chemokine, interleukin-8-like	3	2.30E-03	6	9.75E-03	16	6.23E-13
IPR002473:Small chemokine, C-X-C/Interleukin 8	2	2.41E-02	3	4.98E-02	8	7.77E-08

Table 2. Significant Pathways Involved in CRC

Pathway Name	META		GSE4107		GSE8671	
	Impact Factor	*p*-value	Impact Factor	*p*-value	Impact Factor	*p*-value
PPAR signaling pathway	5.228	5.36E-03	5.154	5.78E-03	7.013	9.00E-04
Renin-angiotensin system	3.595	2.75E-02	4.118	1.63E-02	4.297	0.013605

functional domains (p-value<0.05, hypergeometric test) were shown. Most of significantly overrepresented groups included domains were related with interleukin (Table1).

KEGG pathway enrichment analysis results

We carried out an impact analysis integrating many factors including the statistical significance of differentially expressed genes in the pathway, the expression level change, the topology of the signaling pathway, their interactions and so on (p-value<0.05, hypergeometric test). The impact analysis method revealed many significant pathways contained PPAR signaling pathway (Takahashi et al., 2005) and renin-angiotensin system (Burrell et al., 2004) (Table 2).

Discussion

Microarray technology is an effective tool to globally uncover changes in gene expression and thus elucidate the molecular mechanisms of complex diseases like CRC. In present study, two microarray data sets were obtained to identify DEGs. A total of 32 DEGs were observed in both data sets, suggesting a high confidence.

5 out of the 32 DEGs were associated with inflammatory response and they were CR2 (CD21), IL8, chemokine (C-C motif) ligand 21 (CCL21), chemokine (C-X-C motif) ligand 13 (CXCL13) and EPH receptor A3 (EPHA3). Since inflammation is closely related with cancer (Marx,2004), it is not strange that inflammation-related DEGs account for a high percentage. CCL21 is a member of chemokines, which are involved in immunoregulatory and inflammatory processes. They stimulate chemotaxis for different types of immunocytes. Shields et al. suggest that CCL21 is involved in altering the microenvironment and thus facilitates tumor progression (Shields et al., 2010). Koizumi et al report that CCL21 promote the metastasis of human non-small cell lung cancer (Koizumi et al., 2007). The study by Li and others further indicate that CCL21 plays a key role in colon cancer metastasis through regulation of matrix metalloproteinase-9 (Dong et al., 2011). CXCL13 is a member of CXC chemokines that promotes the migration of B lymphocytes (Ansel et al., 2002). It has been found to be related with breast cancer (Panse et al., 2008) and prostate cancer (Singh et al., 2009). El-Haibi et al further point out that CXCL13 mediates prostate cancer cell proliferation through JNK signaling and invasion through ERK activation (El-Haibi et al., 2011). EPHA3 belongs to the ephrin receptor subfamily of the protein-tyrosine kinase family and its implication in cancer has been indicated (Surawska et al., 2004; Pasquale, 2010). While several studies investigate the somatic mutations in cancer (Davies et al., 2005; Wood et al., 2006), some look into the underlying mechanisms (Lisabeth et al., 2012) and potential clinical applications as targets (Garber, 2010). The study by Xi el al. confirms the clinicopathological significance and prognostic value of EphA3 expression in colorectal carcinoma (Xi and Zhao, 2011). IL8 is also closely related with cancer (Lokshin et al., 2006). The study by Rubie et al suggest an association between IL-8 expression, induction and progression of colorectal carcinoma and the development of colorectal liver metastases (Rubie et al., 2007). In accordance with previous findings, IL8-related domain was significantly enriched in present study.

Another important group of DEGs are associated with response to drugs and they are UDP glucuronosyltransferase 1 family, polypeptide A6 (UGT1A6), butyrylcholinesterase (BCHE), fatty acid binding protein 4, adipocyte (FABP4) and ATP-binding cassette, sub-family G (WHITE), member 2 (ABCG2). UGT1A6 is a UDP-glucuronosyltransferase participating in transforming small lipophilic molecules like drugs and hormones. It has been found that genetic variants of UGT1A6 influence risk of colorectal adenoma recurrence (Hubner et al., 2006). Bigler et al report that UGT1A6 genotypes influence the protective effect of aspirin on colon adenoma risk (Bigler et al., 2001), implying its regulatory role in the effectiveness of chemopreventive drugs (Samowitz et al., 2006). BCHE also encodes an enzyme related with drug metabolism, like suxamethonium. Brass et al first report the amplification of the genes BCHE in 40% of squamous cell carcinoma of the lung (Brass et al., 1997). Bernardi et al find similar phenomenon in breast cancer (Bernardi et al., 2010). Montenegro et al find that butyrylcholinesterase activity decreases in human colon adenocarcinoma (Montenegro et al., 2006).

Several DEGs implicated in enzyme linked receptor protein signaling pathway were also uncovered: chordin-like 1 (CHRDL1), angiopoietin-like 1 (ANGPTL1), gremlin 2 (GREM2) and EPHA3. CHRDL1 and GREM2 are involved in regulating the intestinal stem cell niche (Ricci-Vitiani et al., 2009; Todaro et al., 2010), and further studies may reveal the whole regulatory mechanisms and thus support the targeted therapy.

Besides, we found that PPAR signaling pathway was significantly enriched, which was in accordance with

previous study (Yangand Frucht,2001). The relationship between the rennin-angiotensin system and malignancy is also determined (Ager et al., 2008; George et al., 2010). These results further confirm the usefulness of our findings.

Overall, our study provides a range of DEGs, some of which have been confirmed to be related with CRC. Subsequent functional enrichment analysis indicates their biological roles and the results are beneficial to promote relevant studies. Like EphA3, which have been validated to be a good biomarker, more targets would be identified if further studies were carried out, which will improve the clinical outcomes for patients with CRC.

References

Ager EI, Neo J, Christophi C (2008). The renin-angiotensin system and malignancy. *Carcinogenesis*, **29**, 1675-84.

Aizat AA, Shahpudin SN, Mustapha MA, et al (2011). Association of Arg72Pro of P53 polymorphism with colorectal cancer susceptibility risk in Malaysian population. *Asian Pac J Cancer Prev*, **12**, 2909-13.

Amid A, Wan Chik WD, Jamal P, Hashim YZ (2012). Microarray and quantitative PCR analysis of gene expression profiles in response to treatment with tomato leaf extract in mcf-7 breast cancer cells. *Asian Pac J Cancer Prev*, **13**, 6319-25.

Ansel KM, Harris RB, Cyster JG (2002). CXCL13 is required for B1 cell homing, natural antibody production, and body cavity immunity. *Immunity*, **16**, 67-76.

Bernardi CC, Ribeiro Ede S, Cavalli IJ, et al (2010). Amplification and deletion of the ACHE and BCHE cholinesterase genes in sporadic breast cancer. *Cancer Genet Cytogenet*, **197**, 158-65.

Bigler J, Whitton J, Lampe JW, et al (2001). CYP2C9 and UGT1A6 genotypes modulate the protective effect of aspirin on colon adenoma risk. *Cancer Res*, **61**, 3566-9.

Brass N, Racz A, Heckel D, et al (1997). Amplification of the genes BCHE and SLC2A2 in 40% of squamous cell carcinoma of the lung. *Cancer Res*, **57**, 2290-4.

Burrell LM, Johnston CI, Tikellis C, Cooper ME (2004). ACE2, a new regulator of the renin-angiotensin system. *Trends Endocrinol Metab*, **15**, 166-9.

Davies H, Hunter C, Smith R, et al (2005). Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res*, **65**, 7591-5.

de Wit M, Fijneman RJ, Verheul HM, et al (2012). Proteomics in colorectal cancer translational research: Biomarker discovery for clinical applications. *Clin Biochem*, **46**, 466-79.

Dong P, He XW, Gu J, et al (2011). Vimentin significantly promoted gallbladder carcinoma metastasis. *Chin Med J Beijing*, **124**, 4236.

Doniger SW, Salomonis N, Dahlquist KD, et al (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, **4**, R7.

Draghici S, Khatri P, Tarca AL, et al (2007). A systems biology approach for pathway level analysis. *Genome Res*, **17**, 1537-45.

Edgar R, Domrachev M, Lash AE (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**, 207-10.

El-Haibi CP, Singh R, Sharma PK, et al (2011). CXCL13 mediates prostate cancer cell proliferation through JNK signalling and invasion through ERK activation. *Cell Prolif*, **44**, 311-9.

Ferlay J, Shin HR, Bray F, et al (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer,* **127**, 2893-917.

Garber K (2010). Of Ephs and ephrins: companies target guidance molecules in cancer. *J Natl Cancer Inst*, **102**, 1692-4.

George AJ, Thomas WG, Hannan RD (2010). The renin-angiotensin system and cancer: old dog, new tricks. *Nat Rev Cancer*, **10**, 745-59.

Greystoke A, Dean E, Saunders MP, et al (2012). Multi-level evidence that circulating CK18 is a biomarker of tumour burden in colorectal cancer. *Br J Cancer*, **107**, 1518-24.

Hong F, Breitling R, McEntee CW, et al (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825-7.

Hong Y, Ho KS, Eu KW, Cheah PY (2007). A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res*, **13**, 1107-14.

Huang da W, Sherman BT, Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44-57.

Hubner RA, Muir KR, Liu JF, et al (2006). Genetic variants of UGT1A6 influence risk of colorectal adenoma recurrence. *Clin Cancer Res*, **12**, 6585-9.

Hunter S, Apweiler R, Attwood TK, et al (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res*, **37**, D211-5.

Johnson V, Lipton LR, Cummings C, et al (2005). Analysis of somatic molecular changes, clinicopathological features, family history, and germline mutations in colorectal cancer families: evidence for efficient diagnosis of HNPCC and for the existence of distinct groups of non-HNPCC families. *J Med Genet*, **42**, 756-62.

Koizumi K, Kozawa Y, Ohashi Y, et al (2007). CCL21 promotes the migration and adhesion of highly lymph node metastatic human non-small cell lung cancer Lu-99 in vitro. *Oncol Rep*, **17**, 1511-6.

Lisabeth EM, Fernandez C, Pasquale EB (2012). Cancer somatic mutations disrupt functions of the EphA3 receptor tyrosine kinase through multiple mechanisms. *Biochemistry*, **51**, 1464-75.

Liu N, Fang XD, Vadis Q (2012). CD73 as a novel prognostic biomarker for human colorectal cancer. *J Surg Oncol*, **106**, 918-9; author reply 920.

Lokshin AE, Winans M, Landsittel D, et al (2006). Circulating IL-8 and anti-IL-8 autoantibody in patients with ovarian cancer. *Gynecol Oncol*, **102**, 244-51.

Marx J (2004). Cancer research. Inflammation and cancer: the link grows stronger. *Science*, **306**, 966-8.

Meng W, Zhu HH, Xu ZF, et al (2012). Serum M2-pyruvate kinase: A promising non-invasive biomarker for colorectal cancer mass screening. *World J Gastrointest Oncol*, **4**, 145-51.

Menin C, Scaini MC, De Salvo GL, et al (2006). Association between MDM2-SNP309 and age at colorectal cancer diagnosis according to p53 mutation status. *J Natl Cancer Inst*, **98**, 285-8.

Montenegro MF, Ruiz-Espejo F, Campoy FJ, et al (2006). Acetyl- and butyrylcholinesterase activities decrease in human colon adenocarcinoma. *J Mol Neurosci*, **30**, 51-4.

Pan D, Sun N, Cheung KH, et al (2003). PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. *BMC bioinformatics*, **4**, 56.

Panse J, Friedrichs K, Marx A, et al (2008). Chemokine CXCL13 is overexpressed in the tumour tissue and in the peripheral

blood of breast cancer patients. *Br J Cancer*, **99**, 930-8.
Pasquale EB (2010). Eph receptors and ephrins in cancer: bidirectional signalling and beyond. *Nat Rev Cancer*, **10**, 165-80.
Ricci-Vitiani L, Fabrizi E, Palio E, De Maria R (2009). Colon cancer stem cells. *J Mol Med (Berl)*, **87**, 1097-104.
Rubie C, Frick VO, Pfeil S, et al (2007). Correlation of IL-8 with induction, progression and metastatic potential of colorectal cancer. *World J Gastroenterol*, **13**, 4996-5002.
Sabates-Bellver J, Van der Flier LG, de Palo M, et al (2007). Transcriptome profile of human colorectal adenomas. *Mol Cancer Res*, **5**, 1263-75.
Samowitz WS, Wolff RK, Curtin K, et al (2006). Interactions between CYP2C9 and UGT1A6 polymorphisms and nonsteroidal anti-inflammatory drugs in colorectal cancer prevention. *Clin Gastroenterol Hepatol*, **4**, 894-901.
Shields JD, Kourtis IC, Tomei AA, et al (2010). Induction of lymphoidlike stroma and immune escape by tumors that express the chemokine CCL21. *Science*, **328**, 749-52.
Singh S, Singh R, Sharma PK, et al (2009). Serum CXCL13 positively correlates with prostatic disease, prostate-specific antigen and mediates prostate cancer cell invasion, integrin clustering and cell adhesion. *Cancer Lett*, **283**, 29-35.
Smyth GK (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol,* **3**, Article3.
Surawska H, Ma PC, Salgia R (2004). The role of ephrins and Eph receptors in cancer. *Cytokine Growth Factor Rev*, **15**, 419-33.
Takahashi M, Cuatrecasas M, Balaguer F, et al (2012). The clinical significance of MiR-148a as a predictive biomarker in patients with advanced colorectal cancer. *PLoS One*, **7**, e46684.
Takahashi N, Goto T, Kusudo T, et al (2005). [The structures and functions of peroxisome proliferator-activated receptors (PPARs)]. *Nihon Rinsho*, **63**, 557-64.
Talseth-Palmer BA, McPhillips M, Groombridge C, et al (2010). MSH6 and PMS2 mutation positive Australian Lynch syndrome families: novel mutations, cancer risk and age of diagnosis of colorectal cancer. *Hered Cancer Clin Pract*, **8**, 5.
Todaro M, Francipane MG, Medema JP, Stassi G (2010). Colon cancer stem cells: promise of targeted therapy. *Gastroenterology*, **138**, 2151-62.
Wood LD, Calhoun ES, Silliman N, et al (2006). Somatic mutations of GUCY2F, EPHA3, and NTRK3 in human cancers. *Hum Mutat,* **27**, 1060-1.
Xi HQ, Zhao P (2011). Clinicopathological significance and prognostic value of EphA3 and CD133 expression in colorectal carcinoma. *J Clin Pathol*, **64**, 498-503.
Yang WL, Frucht H (2001). Activation of the PPAR pathway induces apoptosis and COX-2 inhibition in HT-29 human colon cancer cells. *Carcinogenesis*, **22**, 1379-83.
Zhai XH, Yu JK, Yang FQ, Zheng S (2012). Identification of a new protein biomarker for colorectal cancer diagnosis. *Mol Med Report*, **6**, 444-8.