World Journal of Gastroenterology

Submit a Manuscript: http://www.wjgnet.com/esps/
Help Desk: http://www.wjgnet.com/esps/helpdesk.aspx
DOI: 10.3748/wjg.v20.i39.14463

World J Gastroenterol 2014 October 21; 20(39): 14463-14471
ISSN 1007-9327 (print) ISSN 2219-2840 (online)

RETROSPECTIVE STUDY

Gene expression profile of peripheral blood in colorectal cancer

Yu-Tien Chang, Chi-Shuan Huang, Chung-Tay Yao, Sui-Lung Su, Harn-Jing Terng, Hsiu-Ling Chou, Yu-Ching Chou, Kang-Hua Chen, Yun-Wen Shih, Chian-Yu Lu, Ching-Huang Lai, Chen-En Jian, Chiao-Huang Lin, Chien-Ting Chen, Yi-Syuan Wu, Ke-Shin Lin, Thomas Wetter, Chi-Wen Chang, Chi-Ming Chu

Yu-Tien Chang, Yun-Wen Shih, Chen-En Jian, Chiao-Huang Lin, Chien-Ting Chen, Yi-Syuan Wu, Ke-Shin Lin, Chi-Ming Chu, Division of Biomedical Statistics and Informatics, School of Public Health, National Defense Medical Center, Taipei 114, Taiwan
Chi-Shuan Huang, Division of Colorectal Surgery, Cheng Hsin Rehabilitation Medical Center, Taipei 112, Taiwan
Chung-Tay Yao, Department of Surgery, Cathay General Hospital, Taipei 114, Taiwan
Sui-Lung Su, Ching-Huang Lai, Yu-Ching Chou, Department of Epidemiology, School of Public Health, National Defense Medical Center, Taipei 114, Taiwan
Harn-Jing Terng, Advpharma, Inc., New Taipei 114, Taiwan
Hsiu-Ling Chou, Department of Nursing, Far Eastern Memorial Hospital and Oriental Institute of Technology, New Taipei 114, Taiwan
Kang-Hua Chen, Chi-Wen Chang, Department of Nursing, School of Medicine, Chang Gung University, Taoyuan 333, Taiwan
Chian-Yu Lu, Air Force Combatant Command, National Defense Ministry, Taipei 114, Taiwan
Thomas Wetter, Institute of Medical Biometry and Informatics, Heidelberg University, Germany and Department of Biomedical Informatics and Medical Education, University of Washington Seattle, United States
Supported by Taiwan's SBIR promoting program from the Department of Industrial Technology of the Ministry of Economic Affairs, Advpharma, Inc., and the National Defense Medical Center (NDMC), Bureau of Military Medicine, Ministry of Defense, Taiwan
Author contributions: Chu CM and Chang CW designed the research; Shih YW, Chang YT, Terng HJ and Wetter T performed the research; Chu CM, Chang CW, Shih YW, Chang YT, Terng HJ, Wetter T, Chou YC, Su SL, Lai CH, Chen KH, Yao CT, Chou HL, Huang CS, Shih YW, Lu CY, Chen KH, Jian CE, Lin CH, Chen CT, Wu YS and Lin KS analyzed the data; Chu CM, Chang YT, Chang CW, Shih YW, Terng HJ and Wetter T wrote the paper.
Correspondence to: Chi-Ming Chu, PhD, Professor, Division of Biomedical Statistics and Informatics, School of Public Health, National Defense Medical Center, Mingquan East Road 161, Taipei 114, Taiwan. chuchiming@web.de
Telephone: +886-963367484 Fax: +886-2-87923147
Received: December 31, 2013 Revised: April 8, 2014
Accepted: June 12, 2014
Published online: October 21, 2014

Abstract

AIM: Optimal molecular markers for detecting colorectal cancer (CRC) in a blood-based assay were evaluated.

METHODS: A matched (by variables of age and sex) case-control design (111 CRC and 227 non-cancer samples) was applied. Total RNAs isolated from the 338 blood samples were reverse-transcribed, and the relative transcript levels of candidate genes were analyzed. The training set was made of 162 random samples of the total 338 samples. A logistic regression analysis was performed, and odds ratios for each gene were determined between CRC and non-cancer. The samples (n = 176) in the testing set were used to validate the logistic model, and an inferred performance (generality) was verified. By pooling 12 public microarray datasets(GSE 4107, 4183, 8671, 9348, 10961, 13067, 13294, 13471, 14333, 15960, 17538, and 18105), which included 519 cases of adenocarcinoma and 88 controls of normal mucosa, we were able to verify the selected genes from logistic models and estimate their external generality.

RESULTS: The logistic regression analysis resulted in the selection of five significant genes (P < 0.05; *MDM2*, *DUSP6*, *CPEB4*, *MMD*, and *EIF2S3*), with odds ratios of 2.978, 6.029, 3.776, 0.538 and 0.138, respectively. The five-gene model performed stably for the discrimination of CRC cases from controls in the training set, with accuracies ranging from 73.9% to 87.0%, a sensitivity of 95% and a specificity of 95%. In addition, a good performance in the test set was obtained using the discrimination model, providing 83.5% ac-

curacy, 66.0% sensitivity, 92.0% specificity, a positive predictive value of 89.2% and a negative predictive value of 73.0%. Multivariate logistic regressions analyzed 12 pooled public microarray data sets as an external validation. Models that provided similar expected and observed event rates in subgroups were termed well calibrated. A model in which *MDM2*, *DUSP6*, *CPEB4*, *MMD*, and *EIF2S3* were selected showed the result in logistic regression analysis (H-L P = 0.460, R2= 0.853, AUC = 0.978, accuracy = 0.949, specificity = 0.818 and sensitivity = 0.971).

CONCLUSION: A novel gene expression profile was associated with CRC and can potentially be applied to blood-based detection assays.

Key words: Colorectal cancer; Gene expression; Microarray; Internet

Core tip: A novel gene expression profile was associated with colorectal cancer and can potentially be applied to blood-based detection assays. The model that selected *MDM2*, *DUSP6*, *CPEB4*, *MMD*, and *EIF2S3* showed the result in logistic regression analysis (H-L P = 0.460, R2 = 0.853, AUC = 0.978, accuracy = 0.949, specificity = 0.818 and sensitivity = 0.971).

Chang YT, Huang CS, Yao CT, Su SL, Terng HJ, Chou HL, Chou YC, Chen KH, Shih YW, Lu CY, Lai CH, Jian CE, Lin CH, Chen CT, Wu YS, Lin KS, Wetter T, Chang CW, Chu CM. Gene expression profile of peripheral blood in colorectal cancer. *World J Gastroenterol* 2014; 20(39): 14463-14471 Available from: URL: http://www.wjgnet.com/1007-9327/full/v20/i39/14463.htm DOI: http://dx.doi.org/10.3748/wjg.v20.i39.14463

INTRODUCTION

Colorectal cancer (CRC) is a common cancer worldwide[1]. An estimated 146970 new cases of CRC and 49920 deaths were expected to occur in 2009 in the United States[2]. CRC screening can possibly reduce the incidence of advanced disease and provide better overall and progression-free survival. Conventional CRC screening tests include fecal occult blood testing, flexible sigmoidoscopy, double-contrast barium enema X-ray, and colonoscopy[3]. Although they are commonly used, these tests have limitations, including highly variable sensitivity (*i.e.*, 37%-80%) and diet-test interactions[4].

The dissemination of malignant cells from a primary neoplasm is the pivotal event in cancer progression. In many clinical cases, tumor cells metastasize before the primary tumor is diagnosed[5-11]. Individual circulating tumor cells may be the earliest detectable form of metastasis[12]. PCR-based analyses of mRNA from cytokeratins, identified the carcinoembryonic antigen (CEA), and epidermal growth factor receptor (EGFR) genes in peripheral blood samples from CRC patients[13]. However, the low sensitivities and specificities for these well-known genes are not considered acceptable for the detection of colorectal cancer. Recently, multiple biomarkers were reported for the detection of colorectal cancer that delivered a better sensitivity or specificity[14-15].

In the present study, expression levels of 28 cancer-associated candidate genes from the study of Quyun *et al*[16] in peripheral blood samples from 111 colorectal cancer patients and 227 non-cancer controls were analyzed using quantitative real time-PCR. Genes correlated with CRC were selected, and a discrimination model was constructed using multivariate logistic regression. Sensitivity, specificity, accuracy, positive and negative predictive values, and the area under the curve (AUC) of the discrimination model are reported. Meanwhile, models from the present study (Model 1: five genes), Marshall *et al*[14] (Model 2: seven genes) and Han *et al*[15] (Model 3: five genes) were used to validate 17 selected genes by pooling 12 public microarray data sets, in addition to external validation.

MATERIALS AND METHODS

Patients, controls, and blood samples

One hundred eleven patients with histologically confirmed colorectal cancer were enrolled (2006-2009) in a prospective investigational protocol, which was approved by the Institutional Review Board at Cheng Hsin Rehabilitation Medical Center (Taipei, Taiwan). CRC patients at different stages were classified according to the TNM system (Table 1). Peripheral blood samples (6-8 mL) were drawn from patients before any therapeutic treatment, including surgery, but after written informed consent was obtained. All blood samples were collected using a BD vacutainer CPT™ tubes containing sodium citrate as an *anti*-coagulant (Becton Dickinson, NJ, United States) and were stored at 4 ℃.

The healthy controls were 227 volunteers matched by variables of age and sex who had come in for a routine health examination and had no evidence of any clinically detectable cancer. Each participant gave informed consent for the analysis. The same volume of peripheral blood was collected from controls as from patients. Samples were randomly divided into a training set (n = 162) and a testing set (n = 176). There were no significant differences in age, sex, cancer stage or tumor site between the two sets (Table 1).

RNA isolation and reverse transcription

The mononuclear cell (MNC) fraction was isolated within three hours after blood collection, using a BD vacutainer CPT™ tubes (Becton Dickinson), according to the manufacturer's instructions. Total RNA was then extracted from the MNC fraction using the Super RNApure™ kit (Genesis, Taiwan), according to the manufacturer's instructions. The average yield of total RNA per milliliter

Table 1 Characteristics of the training and testing sets[1,2] *n* (%)

	Training set (*n* = 162)			Testing set (*n* = 176)			*P* value	
	CRC (*n* = 55)	Non-CRC (*n* = 107)	*P* value	CRC (*n* = 56)	Non-CRC (*n* = 120)	*P* value	Cases	Controls
Age, yr (S.E.)	66.47 (1.50)	68.31 (1.12)	0.335	67.38 (1.83)	69.99 (1.03)	0.216	0.704	0.270
Gender			0.630			0.176	0.387	0.313
Male	32 (58.2)	58 (54.2)		28 (50.0)	73 (60.8)			
Female	23 (41.8)	49 (45.8)		28 (50.0)	47 (39.2)			
Stage			-			-	0.447	-
Ⅰ	21 (38.2)	-		15 (26.8)	-			
Ⅱ	10 (18.2)	-		9 (16.1)	-			
Ⅲ	14 (25.5)	-		21 (37.5)	-			
Ⅳ	10 (18.2)	-		11 (19.6)	-			
Tumor site		-	-		-	-	0.286	-
Colon	28 (50.9)			30 (53.6)				
Rectum	22 (40.0)			16 (28.6)				
Cecum	4 (7.3)			5 (8.9)				
Colon+Rectum	1 (1.8)			5 (8.9)				

[1]Data are given as means (SE) or as the number of cases (%); [2]*P*-values were estimated using the *t*-test. CRC: Colorectal cancer.

Table 2 Multivariate analysis of colorectal cancer-related molecular markers and the discrimination model based on age, sex, and 15 genes, using the logistic regression model on the training set

	B	OR	95%CI of OR		*P* value
			Upper	Lower	
Sex	0.577	1.780	7.582	0.418	0.435
Age	0.028	1.028	1.083	0.976	0.293
MCM4	0.142	1.152	4.504	0.295	0.838
ZNF264	1.450	4.265	18.208	0.999	0.050
RNF4	-0.550	0.577	5.146	0.065	0.622
GRB2	2.009	7.456	37.131	1.497	0.014
MDM2	1.359	3.892	15.166	0.999	0.050
STAT2	-1.178	0.308	1.466	0.065	0.139
WEE1	1.264	3.540	14.784	0.848	0.083
DUSP6	2.465	11.769	40.330	3.435	1.33E-11
CPEB4	2.045	7.725	27.695	2.155	0.002
MMD	-1.067	0.344	0.865	0.137	0.023
NF1	-1.417	0.243	1.517	0.039	0.130
IRF4	0.057	1.059	3.350	0.335	0.923
EIF2S3	-2.105	0.122	0.718	0.021	0.020
EXT2	-1.933	0.145	1.235	0.017	0.077
POLDIP2	-1.294	0.274	1.515	0.050	0.138

B: Coefficient of logistic regression; OR: Odds ratio; CI: Confidence interval.

of peripheral blood was 1.6 μg. The mRNA quality was assessed by the electrophoresis of total RNA, followed by staining with ethidium bromide, which showed two clear rRNA bands of 28S and 18S. Using a spectrophotometer, the ratio of the absorbances of each RNA at 260 and 280 nm (A_{260}:A_{280}) was confirmed to be greater than 1.7, which is an indicator of RNA purity[17]. One microgram of total RNA was used for cDNA synthesis with random hexamer primers (Amersham Bioscience, United Kingdom) and Superscript™ Ⅱ reverse transcriptase (Invitrogen, United States).

Quantitative real-time polymerase chain reaction

Real-time polymerase chain reaction (PCR) was performed using pre-designed, gene-specific amplification primer sets purchased from Advpharma, Inc. (Taiwan), nucleotide probes from Universal ProbeLibrary™ (Roche, Germany) and TaqMan® Master Mix (Roche) on a Roche LightCycler® 1.5 instrument. The hypoxanthine phosphoribosyltransferase 1 (*HPRT1*) gene was used as the internal control because its expression accurately reflects the mean expression of multiple commonly used normalization genes[18-19]. The cycle number for each candidate gene, Ct(test), was normalized against the cycle number of *HPRT1*,Ct(HK). The calculation was performed as follows: ΔCt(test) = Ct(HK)-Ct(test). The derived (normalized) value, ΔCt(test), for each candidate gene was presented as the relative difference compared with the mRNA expression level of the reference gene[20]. The transcripts of 14 genes were identified as being correlated with the incidence of tumor tissues and were associated with clinical outcomes in a microarray study[21]. Two genes with elevated expression in colon cancer patients[22-23], encoding the A3 adenosine receptor and CCSP-2, were also assayed at the beginning of our study. Since the measurement of a higher cycle number (*i.e.*, Ct greater than 30) generally implies lower amplification efficiency[24], 15 genes were used for further analysis (Table 2) after eliminating genes with low amplification efficiencies.

Statistical analysis

The χ^2 test and *t*-test were performed to characterize sex and age distributions between cases and controls. The transcript levels of candidate genes were tested statistically for differences between the case and control samples using the *t*-test. A logistic regression was performed, and odds ratios were determined to study the association of candidate genes with CRC. The power of the study was 100% for each candidate gene. The statistical alpha level was 0.05. The Bonferroni adjustment for multiple testing was performed using SISA[25] to control for a family-wise error rate of 0.05, for which a significance level was considered as 0.05/42 = 0.00114. The *P*-values in the tables are reported in scientific notation if too many digits were needed for the evaluation and to address the issue of

Table 3 Discrimination power and receiver operating characteristic analysis of different combinations of colorectal cancer-associated genes in the training set

Genes used for models	AUC	SE	*P* value	95%CI	
				Lower	Upper
DUSP6	0.804	0.038	< 0.001	0.73	0.879
DUSP6, CPEB4	0.855	0.032	< 0.001	0.791	0.919
DUSP6, CPEB4, EIF2S3	0.882	0.032	< 0.001	0.820	0.945
DUSP6, CPEB4, EIF2S3, MDM2	0.895	0.030	< 0.001	0.838	0.953
DUSP6, CPEB4, EIF2S3, MDM2, MMD	0.905	0.028	< 0.001	0.849	0.960

P-values for AUC were estimated using the *Z* test. ROC: Receiver operating characteristic; AUC: Area under the ROC curve; SE: Standard Error; CI: Confidence interval.

multiple testing.

Multivariate logistic regression was used to analyze the relationship of the cases and controls to the ΔCt(test) values of candidate genes. The logistic probabilities were calculated using the modeling equations from logistic regression analysis. Diagnostic performances were further used to evaluate multivariate logistic models, including sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). We used the Hosmer-Lemeshow test to check goodness-of-fit. A receiver operating characteristic (ROC) curve analysis was performed to determine the cut-off logistic probabilities and the areas under the ROC curves (AUC), to identify the performance of each candidate gene and combinations of multiple genes. A sensitivity analysis demonstrated the influence on performance of different cut-off logistic probabilities [Logit(P)] in the logistic model.

Internet public microarray data sets

The microarray gene expression data were obtained searches using "colon cancer" AND "human [organism]" AND "expression profiling by array [dataset type]" as the key words in the GEO database of the National Center for Biotechnology Information (NCBI). The eligible criteria were (1) the examined samples were frozen tissue sections of normal human colorectal mucosa, primary colorectal cancer or hepatic metastases from colorectal cancer; (2) the microarray platform used was limited to single-color, whole genome gene chips from Affymetrix; and (3) the data were presented as gene expression levels. The exclusion criteria were (1) data from cultured cell lines or other in vitro assays; (2) datasets without the original gene expression level data files; and (3) those with redundant sub-datasets. A total of 175 GEO series (GSE) datasets were excluded, leaving 12 public microarray dataset: GSE 4107, 4183, 8671, 9348, 10961, 13067, 13294, 13471, 14333, 15960, 17538, and 18105. These data included 519 cases of adenocarcinoma and 88 controls of normal mucosa.

Furthermore, we validated the 17 CRC-associated genes from the studies (Model 1: 5 genes), Marshall *et al*[14] (Model 2: 7 genes) and Han *et al*[15] (Model 3: 5 genes) and performed the multivariate logistic regression analysis using the pooled 12 public microarray data sets, in addition to external validation.

RESULTS

Genes correlated with colorectal cancer

A multivariate analysis based on age, sex and 15 genes was used in a logistic regression model in the training set because the peripheral blood samples were drawn from patients before any therapeutic treatment (Table 2). However this full model seemed capable of discriminating between the CRC cases and controls, it may have resulted in overfitting.

Discrimination of colorectal cancer and non-cancer controls using five genes

Five genes, *i.e.*, *MDM2, DUSP6, CPEB4, MMD, and EIF2S3*, were significantly associated with CRC. Discrimination models can be constructed with one of the five genes selected, based on forward multivariate logistic regression analysis using the training set. AUCs were used to compare the performance of discrimination models for single gene and combinations of two, three, four, or five marker genes. The *DUSP6* model (Table 3) displayed the best discrimination ability, with an AUC of 0.804 (95%CI: 0.730-0.879) compared with the other one-gene models (AUC: 0.49-0.69). Distinct increases in the AUC of up to 0.905 (95%CI: 0.849-0.960) resulted from the combination of the five genes. The logistic regression analysis (Table 3) resulted in the selection of five significant genes (*i.e.*, $P < 0.05$), *MDM2*, *DUSP6*, *CPEB4, MMD,* and *EIF2S3*, with odds ratios of 2.978, 6.029, 3.776, 0.538 and 0.138, respectively. This model was reduced to a panel of five genes in a forward stepwise regression, in which the statistical powers of the five genes were 1.00 between case and control groups in training and testing sets (Table 4).

The cut-off value of Logit(P) for the five-gene model could also be adjusted to achieve high sensitivity or specificity, *i.e.*, 99%, 95% or 90%. The five-gene model performed stably to discriminate between CRC cases and controls in the training set (Table 5), with accuracies ranging from 73.9% to 87.0%, a sensitivity of 95%, and a specificity of 95%. The five-gene model fulfilled the criteria of good performance for diagnostic tests, as well as accuracy (87.0%), sensitivity (78%), and specificity (92%); in addition, the Hosmer-Lemeshow test was not significant ($P = 0.108$). In addition, a good performance in the testing set (Table 6) was obtained using the discrimination model, with 84% accuracy, 66% sensitivity, 92% specificity, 79% PPV and 85% NPV. In external validation (Tables 6 and 7), the five-gene model performed with 94.9% accuracy, 97.1% sensitivity, 81.8% specificity, 96.9% PPV, 82.8% NPV, and an area under the ROC curve of 0.978 (0.912-1).

Table 4 Mean expression levels, standard error and statistical power of selected genes between case and control groups in the training and testing sets

Selected genes	Training set			Testing set		
	Case (n = 55)	Control (n = 107)	Power	Case (n = 56)	Control (n = 120)	Power
MDM2	-0.4225 (0.08945)	-0.8913 (0.04572)	1	-0.3270 (0.09063)	-0.9209 (0.03618)	1
DUSP6	2.5483 (0.13248)	1.5458 (0.06415)	1	2.0335 (0.12041)	1.7462 (0.06135)	1
CPEB4	1.3413 (0.11016)	0.3932 (0.09799)	1	1.4595 (0.11851)	0.4014 (0.06980)	1
MMD	2.0567 (0.15441)	1.3178 (0.09799)	1	1.7029 (0.15958)	1.4320 (0.07806)	1
EIF2S3	3.4489 (0.07883)	3.6158 (0.05331)	1	3.4311 (0.05937)	3.5620 (0.03815)	1

Values in cells: Mean expression levels (standard error); α-level is 0.05.

Table 5 Performance of the statistical model based on the five-gene profile logistic probabilities for the training set

Logit(P)	Sensitivity	Specificity	PPV	NPV	Accuracy
0.020	99%	16%	2.3%	99.9%	44.2%
0.051	95%	63%	12.1%	99.6%	73.9%
0.178	90%	72%	41.1%	97.1%	78.1%
0.500	78%	92%	82.7%	89.1%	87.0%
0.475	80%	90%	87.8%	83.3%	86.6%
0.685	61%	95%	96.4%	52.9%	83.5%
0.901	25%	99%	99.6%	12.6%	73.9%

Logit(P): Logistic probabilities; PPV: Positive predictive value; NPV: Negative predictive value.

Table 6 Performance of the statistical model on the training, testing sets and external validation dataset from 12 public microarray studies with Logit(P) = 0.5

	Training set	Testing set	External validation
Non-Cancers	107	120	88
True negative	98	110	72
False positive	9	10	16
Colorectal Cancers	55	56	519
False negative	12	19	15
True positive	43	37	504
Total	162	176	607
Sensitivity	78.2%	66.1%	97.1%
Specificity	91.5%	91.7%	81.8%
PPV	82.7%	78.7%	96.9%
NPV	89.1%	85.3%	82.8%
Accuracy	87.0%	83.5%	94.9%

Logit(P): Logistic probabilities; PPV: Positive predictive value; NPV: Negative predictive value.

Pooling 12 microarray studies to verify the 17 selected genes and estimate their external generality.

Furthermore, we performed multivariate logistic regression analysis for the 12 pooled public microarray data sets, as well as the external validation (Tables 6 and 7), to verify the CRC-associated genes from three studies (the present one, Marshall *et al*[14] and Han *et al*[15]). As shown in Table 7, we validated the 17 CRC-associated genes from this study (Model 1: 5 genes), Marshall *et al*[14] (Model 2: 7 genes) and Han *et al*[15] (Model 3: 5 genes) by pooling 12 public microarray dataset of GSE 4107, 4183, 8671, 9348, 10961, 13067, 13294, 13471, 14333, 15960, 17538, and 18105, which included 519 cases of adenocarcinoma and 88 controls of normal mucosa. The Hosmer-Lemeshow (H-L) goodness-of-fit test showed statistical significance (P = 0.044) for Model 2 of Marshall *et al*[14], in which the observed event rates did not match the expected event rates in the subgroups of the model population. Models showing similar expected and observed event rates in subgroups were called well calibrated (Model 1 and 3).

DISCUSSION

Common serum tumor markers used in primary care practice have not demonstrated a survival benefit in randomized controlled trials for screening in the general population. Most of them showed elevated levels only in some early-stage or late-stage cancer patients[26]. A recent review of real-time PCR-based assays with single molecular markers, such as CEA, CK19, and CK20, demonstrated low sensitivity, ranging from 4% to 35.9%, 25.9% to 41.9%, and 5.1% to 28.3%, respectively[13]. One study, performed with a newly identified molecular marker known as ProtM[27], also attained unsatisfactory sensitivity.

Circulating cancer cells from any cancer type are capable of disseminating from solid tumor tissues, penetrating and invading blood vessels, and circulating in the peripheral blood[28-29]. The number of circulating tumor cells has been used to predict the clinical outcome of cancer patients[30-31]. On the basis of the presence of circulating tumor cells, we identified five molecular markers, *MDM2*, *DUSP6*, *CPEB4*, *MMD* and *EIF2S3*, which were differentially expressed between peripheral blood samples of CRC patients and healthy controls. The application of multivariate logistic regression analysis resulted in a five-gene discrimination model, which achieved good diagnostic performance and provided stable conditions, with accuracies ranging from 73.9% to 87.0%, a sensitivity of 95% and a specificity of 95%.

Both mRNAs and proteins in the peripheral blood have been tested for their diagnostic utility to detect circulating tumor cells of different solid tumors or to determine prognoses of various cancers. In the present study, we confirmed that the AUCs of the discrimination models greatly improved from 0.80 for the model based on a single gene (DUSP6) to 0.91 for the combined model

Table 7 Logistic regression models for 12 pooled microarray data sets as the external validation of colorectal cancer -associated genes from three studies

	Model 1			Model 2			Model 3		
	B	S.E.	*P* value	B	S.E.	*P* value	B	S.E.	*P* value
Five selected genes of this study:									
MDM2	6.069	1.461	< 0.001						
DUSP6	1.360	0.235	< 0.001						
CPEB4	-3.177	0.383	< 0.001						
MMD	0.335	0.442	0.448						
EIF2S3	1.462	0.244	< 0.001						
Seven selected genes of Marshall *et al*[14]									
ANXA3				0.559	0.212	0.008			
CLEC4D				46.259	9.918	< 0.001			
LMNB1				1.883	0.330	< 0.001			
PRRG4				-1.284	0.371	0.001			
TNFAIP6				1.787	0.377	< 0.001			
VNN1				0.207	0.159	0.194			
IL2RB				0.269	0.216	0.213			
Five selected genes of Han *et al*[15]									
CDA							-0.496	0.090	< 0.001
MGC20553							-1.386	0.197	< 0.001
BANK1							0.565	0.373	0.129
BCNP1							-0.944	1.148	0.411
MS4A1							-1.483	0.457	0.001
Constant	-32.758	6.001	< 0.001	-124.678	25.437	< 0.001	16.601	2.995	< 0.001
H-L		0.460			0.044			0.194	
R^2		0.853			0.841			0.693	
AUC		0.978			0.985			0.957	
Accuracy		0.949			0.974			0.939	
Specificity		0.818			0.886			0.716	
Sensitivity		0.971			0.988			0.977	

Model 1: Five selected genes of this study; Model 2: Seven selected genes of Marshall *et al*[14]; Model 3: Five selected genes of Han *et al*[15]; B: Logistic regression coefficient beta; SE: Standard error of B; *P*: *P* value with statistical significance; H-L: Hosmer and Lemeshow test *P* value R2: Nagelkerke R Square; AUC: Area under ROC.

with all five genes. An increasing number of clinical studies have shown improvements in the sensitivity of cancer detection by assaying transcript levels of multiple genes in patient peripheral blood[14-15,32].

A higher sensitivity or specificity of the discriminatory performance of our five-gene model (Table 5) was achieved by adjusting the cut-off value of Logit(P). This five-gene discrimination model with Logit(P) = 0.0511 had a sensitivity of 95%, a specificity of 63% and an accuracy of 74%, which is ideal for screening colorectal cancer. However, setting Logit(P) to 0.4747 resulted in a specificity of 90%, a sensitivity of 80% and an accuracy of 86%, which indicates that our five-gene model is robust and highly accurate for discriminating CRC from healthy or benign conditions. Similar accuracy rates (*i.e.*, 80%-86%) were achieved with Logit(P) values ranging from 0.0511 to 0.4747. In the testing set, the five-gene model performed with satisfactory accuracy, sensitivity and specificity.

Two reports[14-15] with similar screening approaches used different gene sets to detect CRC (Table 7). The two gene sets were obtained by direct selection from differentially expressed genes in peripheral blood samples using microarray techniques, followed by real-time PCR. The biomarkers they selected may more or less reflect the static and dynamic changes of the immune system in response to cancer. In our study, genes clinically confirmed to be cancer-associated in tumor tissues were chosen for selection and validation in peripheral blood samples.

The five genes identified here for discrimination between CRC patients and healthy controls might be useful to evaluate the therapeutic responses and prognoses of colorectal cancer patients. They could also be selected as targets for the development of therapies because of their strong association with CRC. MDM2 is a negative regulator of the tumor suppressor protein p53[33]. Higher MDM2 expression has been reported in a variety of human stromal and epithelial malignancies, including colorectal cancers[33-38]. DUSP6, also known as MAPK phosphatase 3 (MKP3), inactivates MAPK1/ERK2[39-42]. Elevated DUSP6 transcript levels have been reported as a risk factor for poor prognosis in non-small cell lung cancer patients[21] and tamoxifen resistance in breast cancer patients[43]. In contrast, DUSP6 is a candidate tumor suppressor gene in pancreatic cancer[42] and primary human ovarian cancer cells. CPEB4 binds to the cytoplasmic polyadenylation element (CPE) of target mRNAs and

controls cytoplasmic polyadenylation and translational activation during development[44-46]. MMD is an integral membrane protein with seven putative transmembrane segments[47]. Its biological function is still unclear. EIF2S3 is the largest subunit (gamma) of eukaryotic translation initiation factor 2[48], and might be indirectly involved in the inhibition of prostate cancer metastasis through N-myc downstream regulated gene 1[49]. This is the first study to show an association of *MDM2*, *DUSP6*, *CPEB4*, *MMD* and *EIF2S3* with CRC.

Meanwhile, we verified the CRC-associated genes by pooling 12 public microarray data sets such that the three logistic models performed similar AUCs without statistically significant difference. In the future, the causal relations should be confirmed among the selected genes and CRC. In future works, the expression signature of these CRC-associated genes should be evaluated for early detection, with more samples randomly screened from the population. In addition, subjects who eventually receive a diagnosis of CRC should be evaluated. Early CRC detection could provide inherent benefits to the patient and could also enable screening for post-operative residual tumor cells and occult metastases, an early indicator of tumor recurrence. Early detection could thus improve survival in patients before symptoms are detectable, during treatment, or during remission.

In conclusion, we found the gene expression profile of peripheral blood that five genes (*MDM2, DUSP6, CPEB4*, *MMD,* and *EIF2S3*) are highly associated with colorectal cancer. Detection of cancer cell-specific biomarkers in the peripheral blood can be an effective screening strategy for CRC.

COMMENTS

Background

The five genes (*MDM2*, *DUSP6*, *CPEB4*, *MMD*, and *EIF2S3*) identified here for discrimination between colorectal cancer (CRC) patients and healthy controls might be useful in evaluating the therapeutic responses and prognoses of colorectal cancer patients. They could also be selected as targets for the development of therapies because of their strong association with CRC.

Research frontiers

The present study is the first to translate a cancer tissue microarray into clinical practice for peripheral blood samples of case-control study, and to use pools of 12 datasets of public microarray studies as the external validation for of the expression profiles of the five selected genes. The authors were able to verify the selected genes from logistic models and estimate their external generality and inferred performance.

Innovations and breakthroughs

The gene expression profiles in peripheral blood of five genes (*MDM2*, *DUSP6*, *CPEB4*, *MMD*, and *EIF2S3*) are highly associated with CRC.

Applications

Detection of cancer cell-specific biomarkers in peripheral blood can be an effective screening strategy for CRC.

Peer review

This paper is very well written and examines the possibility of using a panel of genes as a potential biomarker of CRC. Peripheral leucocyte gene expression was quantified using PCR. The authors used a pooled multivariate analysis to select genes of interest from a list of CRC candidate genes. The authors then compared their own panel of genes from peripheral blood, to microarray data sets from colonic tissue (CRC and control).

REFERENCES

1 **Parkin DM**, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin* 2005; **55**: 74-108 [PMID: 15761078]

2 **Labianca R**, Beretta GD, Kildani B, Milesi L, Merlin F, Mosconi S, Pessi MA, Prochilo T, Quadri A, Gatta G, de Braud F, Wils J. Colon cancer. *Crit Rev Oncol Hematol* 2010; **74**: 106-133 [PMID: 20138539 DOI: 10.1016/j.critrevonc.2010.01.010]

3 **Jemal A**, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ. Cancer statistics, 2008. *CA Cancer J Clin* 2008; **58**: 71-96 [PMID: 18287387]

4 **Nannini M**, Pantaleo MA, Maleddu A, Astolfi A, Formica S, Biasco G. Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives. *Cancer Treat Rev* 2009; **35**: 201-209 [PMID: 19081199 DOI: 10.1016/j.ctrv.2008.10.006]

5 **Lancashire LJ**, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies. *Brief Bioinform* 2009; **10**: 315-329 [PMID: 19307287 DOI: 10.1093/bib/bbp012]

6 **Cardoso J**, Boer J, Morreau H, Fodde R. Expression and genomic profiling of colorectal cancer. *Biochim Biophys Acta* 2007; **1775**: 103-137 [PMID: 17010523 DOI: 10.1016/j.bbcan.2006.08.004]

7 **Sagynaliev E**, Steinert R, Nestler G, Lippert H, Knoch M, Reymond MA. Web-based data warehouse on gene expression in human colorectal cancer. *Proteomics* 2005; **5**: 3066-3078 [PMID: 16041676 DOI: 10.1002/pmic.200402107]

8 **Shih W**, Chetty R, Tsao MS. Expression profiling by microarrays in colorectal cancer (Review). *Oncol Rep* 2005; **13**: 517-524 [PMID: 15706427]

9 **Chan SK**, Griffith OL, Tai IT, Jones SJ. Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. *Cancer Epidemiol Biomarkers Prev* 2008; **17**: 543-552 [PMID: 18349271]

10 **Smith RA**, Cokkinides V, Eyre HJ. American Cancer Society guidelines for the early detection of cancer, 2006. *CA Cancer J Clin* 2006; **56**: 11-25; quiz 49-50 [PMID: 16449183]

11 **Levin B**, Lieberman DA, McFarland B, Smith RA, Brooks D, Andrews KS, Dash C, Giardiello FM, Glick S, Levin TR, Pickhardt P, Rex DK, Thorson A, Winawer SJ. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA Cancer J Clin* 2008; **58**: 130-160 [PMID: 18322143]

12 **Fidler IJ**. Critical factors in the biology of human cancer metastasis: twenty-eighth G.H.A. Clowes memorial award lecture. *Cancer Res* 1990; **50**: 6130-6138 [PMID: 1698118]

13 **Sergeant G**, Penninckx F, Topal B. Quantitative RT-PCR detection of colorectal tumor cells in peripheral blood--a systematic review. *J Surg Res* 2008; **150**: 144-152 [PMID: 18621394 DOI: 10.1016/j.jss.2008.02.012]

14 **Marshall KW**, Mohr S, Khettabi FE, Nossova N, Chao S, Bao W, Ma J, Li XJ, Liew CC. A blood-based biomarker panel for stratifying current risk for colorectal cancer. *Int J Cancer* 2010; **126**: 1177-1186 [PMID: 19795455 DOI: 10.1002/ijc.24910]

15 **Han M**, Liew CT, Zhang HW, Chao S, Zheng R, Yip KT, Song ZY, Li HM, Geng XP, Zhu LX, Lin JJ, Marshall KW, Liew CC. Novel blood-based, five-gene biomarker set for the detection of colorectal cancer. *Clin Cancer Res* 2008; **14**: 455-460 [PMID: 18203981 DOI: 10.1158/1078-0432.CCR-07-1801]

16 **Quyun C**, Ye Z, Lin SC, Lin B. Recent patents and advances in genomic biomarker discovery for colorectal cancers. *Recent Pat DNA Gene Seq* 2010; **4**: 86-93 [PMID: 20426761]

17 **Sambrook J**, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual. 2nd ed. NY: Cold Spring Harbor Labo-

ratory Press, 1989
18 **Vandesompele J**, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 2002; **3**: RESEARCH0034 [PMID: 12184808]
19 **de Kok JB**, Roelofs RW, Giesendorf BA, Pennings JL, Waas ET, Feuth T, Swinkels DW, Span PN. Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Lab Invest* 2005; **85**: 154-159 [PMID: 15543203 DOI: 10.1038/labinvest.3700208]
20 **Livak KJ**, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 2001; **25**: 402-408 [PMID: 11846609 DOI: 10.1006/meth.2001.1262]
21 **Chen HY**, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, Singh S, Chen WJ, Chen JJ, Yang PC. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007; **356**: 11-20 [PMID: 17202451 DOI: 10.1056/NEJMoa060096]
22 **Xin B**, Platzer P, Fink SP, Reese L, Nosrati A, Willson JK, Wilson K, Markowitz S. Colon cancer secreted protein-2 (CCSP-2), a novel candidate serological marker of colon neoplasia. *Oncogene* 2005; **24**: 724-731 [PMID: 15580307 DOI: 10.1038/sj.onc.1208134]
23 **Gessi S**, Cattabriga E, Avitabile A, Gafa' R, Lanza G, Cavazzini L, Bianchi N, Gambari R, Feo C, Liboni A, Gullini S, Leung E, Mac-Lennan S, Borea PA. Elevated expression of A3 adenosine receptors in human colorectal cancer is reflected in peripheral blood cells. *Clin Cancer Res* 2004; **10**: 5895-5901 [PMID: 15355922 DOI: 10.1158/1078-0432.CCR-1134-03]
24 **Pfaffl MW**. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 2001; **29**: e45 [PMID: 11328886]
25 **Møller P**, Clark N, Mæhle L. A SImplified method for Segregation Analysis (SISA) to determine penetrance and expression of a genetic variant in a family. *Hum Mutat* 2011; **32**: 568-571 [PMID: 21309035 DOI: 10.1002/humu.21441]
26 **Perkins GL**, Slater ED, Sanders GK, Prichard JG. Serum tumor markers. *Am Fam Physician* 2003; **68**: 1075-1082 [PMID: 14524394]
27 **Schuster R**, Max N, Mann B, Heufelder K, Thilo F, Gröne J, Rokos F, Buhr HJ, Thiel E, Keilholz U. Quantitative real-time RT-PCR for detection of disseminated tumor cells in peripheral blood of patients with colorectal cancer using different mRNA markers. *Int J Cancer* 2004; **108**: 219-227 [PMID: 14639606 DOI: 10.1002/ijc.11547]
28 **Bogenrieder T**, Herlyn M. Axis of evil: molecular mechanisms of cancer metastasis. *Oncogene* 2003; **22**: 6524-6536 [PMID: 14528277 DOI: 10.1038/sj.onc.1206757]
29 **Carmeliet P**, Jain RK. Angiogenesis in cancer and other diseases. *Nature* 2000; **407**: 249-257 [PMID: 11001068 DOI: 10.1038/35025220]
30 **Cristofanilli M**, Budd GT, Ellis MJ, Stopeck A, Matera J, Miller MC, Reuben JM, Doyle GV, Allard WJ, Terstappen LW, Hayes DF. Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *N Engl J Med* 2004; **351**: 781-791 [PMID: 15317891 DOI: 10.1056/NEJMoa040766]
31 **Cristofanilli M**, Hayes DF, Budd GT, Ellis MJ, Stopeck A, Reuben JM, Doyle GV, Matera J, Allard WJ, Miller MC, Fritsche HA, Hortobagyi GN, Terstappen LW. Circulating tumor cells: a novel prognostic factor for newly diagnosed metastatic breast cancer. *J Clin Oncol* 2005; **23**: 1420-1430 [PMID: 15735118 DOI: 10.1200/JCO.2005.08.140]
32 **Shen C**, Hu L, Xia L, Li Y. Quantitative real-time RT-PCR detection for survivin, CK20 and CEA in peripheral blood of colorectal cancer patients. *Jpn J Clin Oncol* 2008; **38**: 770-776 [PMID: 18845519 DOI: 10.1093/jjco/hyn105]
33 **Reifenberger G**, Liu L, Ichimura K, Schmidt EE, Collins VP. Amplification and overexpression of the MDM2 gene in a subset of human malignant gliomas without p53 mutations. *Cancer Res* 1993; **53**: 2736-2739 [PMID: 8504413]
34 **Bueso-Ramos CE**, Manshouri T, Haidar MA, Huh YO, Keating MJ, Albitar M. Multiple patterns of MDM-2 deregulation in human leukemias: implications in leukemogenesis and prognosis. *Leuk Lymphoma* 1995; **17**: 13-18 [PMID: 7773150 DOI: 10.3109/10428199509051698]
35 **Bueso-Ramos CE**, Manshouri T, Haidar MA, Yang Y, McCown P, Ordonez N, Glassman A, Sneige N, Albitar M. Abnormal expression of MDM-2 in breast carcinomas. *Breast Cancer Res Treat* 1996; **37**: 179-188 [PMID: 8750585]
36 **Bueso-Ramos CE**, Yang Y, deLeon E, McCown P, Stass SA, Albitar M. The human MDM-2 oncogene is overexpressed in leukemias. *Blood* 1993; **82**: 2617-2623 [PMID: 8219216]
37 **Marchetti A**, Buttitta F, Girlando S, Dalla Palma P, Pellegrini S, Fina P, Doglioni C, Bevilacqua G, Barbareschi M. mdm2 gene alterations and mdm2 protein expression in breast carcinomas. *J Pathol* 1995; **175**: 31-38 [PMID: 7891224 DOI: 10.1002/path.1711750106]
38 **Marchetti A**, Buttitta F, Pellegrini S, Merlo G, Chella A, Angeletti CA, Bevilacqua G. mdm2 gene amplification and overexpression in non-small cell lung carcinomas with accumulation of the p53 protein in the absence of p53 gene mutations. *Diagn Mol Pathol* 1995; **4**: 93-97 [PMID: 7551299]
39 **Zhou B**, Wu L, Shen K, Zhang J, Lawrence DS, Zhang ZY. Multiple regions of MAP kinase phosphatase 3 are involved in its recognition and activation by ERK2. *J Biol Chem* 2001; **276**: 6506-6515 [PMID: 11104775 DOI: 10.1074/jbc.M009753200]
40 **Zhou G**, Wang H, Liu SH, Shahi KM, Lin X, Wu J, Feng XH, Qin J, Tan TH, Brunicardi FC. p38 MAP kinase interacts with and stabilizes pancreatic and duodenal homeobox-1. *Curr Mol Med* 2013; **13**: 377-386 [PMID: 23331010]
41 **Zhou Q**, Li G, Deng XY, He XB, Chen LJ, Wu C, Shi Y, Wu KP, Mei LJ, Lu JX, Zhou NM. Activated human hydroxy-carboxylic acid receptor-3 signals to MAP kinase cascades via the PLC-dependent PKC and MMP-mediated EGFR pathways. *Br J Pharmacol* 2012; **166**: 1756-1773 [PMID: 22289163 DOI: 10.1111/j.1476-5381.2012.01875.x]
42 **Furukawa T**, Yatsuoka T, Youssef EM, Abe T, Yokoyama T, Fukushige S, Soeda E, Hoshi M, Hayashi Y, Sunamura M, Kobari M, Horii A. Genomic analysis of DUSP6, a dual specificity MAP kinase phosphatase, in pancreatic cancer. *Cytogenet Cell Genet* 1998; **82**: 156-159 [PMID: 9858808]
43 **Cui Y**, Parra I, Zhang M, Hilsenbeck SG, Tsimelzon A, Furukawa T, Horii A, Zhang ZY, Nicholson RI, Fuqua SA. Elevated expression of mitogen-activated protein kinase phosphatase 3 in breast tumors: a mechanism of tamoxifen resistance. *Cancer Res* 2006; **66**: 5950-5959 [PMID: 16740736 DOI: 10.1158/0008-5472.CAN-05-3243]
44 **Huang YS**, Kan MC, Lin CL, Richter JD. CPEB3 and CPEB4 in neurons: analysis of RNA-binding specificity and translational control of AMPA receptor GluR2 mRNA. *EMBO J* 2006; **25**: 4865-4876 [PMID: 17024188 DOI: 10.1038/sj.emboj.7601322]
45 **Hake LE**, Mendez R, Richter JD. Specificity of RNA binding by CPEB: requirement for RNA recognition motifs and a novel zinc finger. *Mol Cell Biol* 1998; **18**: 685-693 [PMID: 9447964]
46 **Hake LE**, Richter JD. CPEB is a specificity factor that mediates cytoplasmic polyadenylation during Xenopus oocyte maturation. *Cell* 1994; **79**: 617-627 [PMID: 7954828]
47 **Rehli M**, Krause SW, Schwarzfischer L, Kreutz M, Andreesen R. Molecular cloning of a novel macrophage maturation-associated transcript encoding a protein with several potential transmembrane domains. *Biochem Biophys Res Commun* 1995; **217**: 661-667 [PMID: 7503749 DOI: 10.1006/

bbrc.1995.2825]
48 **Gaspar NJ**, Kinzy TG, Scherer BJ, Hümbelin M, Hershey JW, Merrick WC. Translation initiation factor eIF-2. Cloning and expression of the human cDNA encoding the gamma-subunit. *J Biol Chem* 1994; **269**: 3415-3422 [PMID: 8106381]
49 **Tu LC**, Yan X, Hood L, Lin B. Proteomics analysis of the interactome of N-myc downstream regulated gene 1 and its interactions with the androgen response program in prostate cancer cells. *Mol Cell Proteomics* 2007; **6**: 575-588 [PMID: 17220478]

P- Reviewer: Ventham NT **S- Editor**: Qi Y
L- Editor: Stewart G **E- Editor**: Wang CH

Published by **Baishideng Publishing Group Inc**
8226 Regency Drive, Pleasanton, CA 94588, USA
Telephone: +1-925-223-8242
Fax: +1-925-223-8243
E-mail: bpgoffice@wjgnet.com
Help Desk: http://www.wjgnet.com/esps/helpdesk.aspx
http://www.wjgnet.com

ISSN 1007-9327