A Gene Expression Profile of Peripheral Blood in Colorectal Cancer

Chi-Shuan Huang1, Harn-Jing Terng2, Chi-Wen Chang3, Yu-Chin Chou2, Yu-Tien Chang2,4, Chin-Yu Chen2,4, Su-Lung Su3, Woan-Jen Lee2, Chien-An Sun2, Ching-Huang Lai2, Yun-Wen Shih2,3, Lu Pai2,4, Thomas Wetter7, Kang-Hwa Chen2,4, and Chi-Ming Chu2,4*

1 Division of Colorectal Surgery, Cheng Hsin Rehabilitation Medical Center, Taiwan
2 Adypharma, Inc., Taipei, Taiwan
3 School of Public Health, National Defense Medical Center, Taiwan
4 Division of Biomedical Statistics and Informatics, School of Public Health, National Defense Medical Center, Taiwan
5 Department of Public Health, Fu Jen Catholic University, Taiwan
6 Taipei Medical University, Taiwan
7 Institute of Medical Biometry and Informatics, Heidelberg University, Germany and Department of Biomedical Informatics and Medical Education, University of Washington Seattle, USA
8 Department of Nursing, College of Medicine, Chang Gung University, Taiwan

Abstract

Background: Optimal molecular markers for detecting colorectal cancer (CRC) in a blood-based assay were evaluated. Microarray technology has shown a great potential in the colorectal cancer research. Genes significantly associated with cancer in microarray studies, were selected as candidate genes in the study. Pooling Internet public microarray data sets can overcome the limitation by the small number of samples in previous studies.

Objective: Using public microarray data sets verifies gene expression profiles for colorectal cancer.

Methods: Logistic regression analysis was performed, and odds ratios for each gene were determined between CRC and controls. Public microarray datasets of GSE 4107, 4183, 9348, 10961, 13294, 13433, 15960, 17538, and 18105 included 519 cases of adenocarcinoma and 88 controls of normal mucosa, which were used to verify the candidate genes from logistic models and estimated its external generality.

Results: A 7-gene model of CPEB4, EIF2S3, MGC20553, MAS4A1, ANXA3, TNFAIP6 and IL2RB was pairwise selected that showed the best results in logistic regression analysis (H-L p=1.000, R2=0.951, AUC=0.999, accuracy=0.968, specificity=0.966 and sensitivity=0.994).

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

Keywords: Colorectal cancer; Gene expression; Microarray; Internet

Background

Colorectal cancer (CRC) is a common cancer worldwide [1]. An estimated 146,970 new cases of colon and rectal cancer and 49,920 deaths are expected to occur in 2009 in the United States [2]. CRC screening can possibly reduce the incidence of advanced disease and provide better overall, 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% to 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. Individual circulating tumor cells may be the earliest detectable form of metastasis [5]. PCR-based analyses of mRNA from cytokeratins, the carcinoembryonic antigen (CEA), and epidermal growth factor receptor (EGFR) genes in peripheral blood samples from CRC patients have been reported [6]. 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 [7,8].

In the present study, expression levels of 28 cancer-associated candidate genes in the 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 AUC of the discrimination model are reported. Meanwhile, from this study (Model 1: 5 genes), Marshall et al. [7] (Model 2: 7 genes) and Han et al. [8] (Model 3: 5 genes), the 17 candidate genes were validated by pooling12 public microarray data sets as well as the external validation.

Methods

Patients, controls, and blood samples

One hundred eleven patients with histologically confirmed colorectal cancer were enrolled (2006-2009) in a prospective study.
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 BD vacutainer CPT™ tubes containing sodium citrate as an anti-coagulant (Becton Dickinson, NJ, USA) and were stored at 4°C.

The healthy controls were 227 volunteers who had come in for a routine health examination and had no evidence of any clinically detectable cancer disease. 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 BD vacutainer CPT™ tubes (Becton Dickinson, NJ, USA), 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 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 absorbance of each RNA at 260 and 280 nm (A260:A280) was confirmed to be greater than 1.7, which is an indicator of RNA purity [9]. One microgram of total RNA was used for cDNA synthesis with random hexamer primers (Amersham Bioscience, UK) and Superscript II reverse transcriptase (Invitrogen, USA).

Quantitative real-time PCR
Real-time PCR was performed using pre-designed, gene-specific amplification primer sets purchased from Adypharma, Inc. (Taiwan), nucleotide probes from Universal ProbeLibrary™ (Roche, Germany) and TaqMan® Master Mix (Roche) on a Roche LightCycler® I.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 [10,11]. The cycle number for each candidate gene, Ct(test), was normalized against the cycle number of HPRT1, that is, C(HK). The calculation is performed as follows: ∆C(t)=Ct(HK) - Ct(test). The derived (normalized) value, ∆C, for each candidate gene is presented as the relative difference as compared to the mRNA expression level of the reference gene [12].

Preliminary screening of investigating genes
In the preliminary screening for CRC-associated genes, we selected candidate genes from the published microarray study [14], and tested for the relative range of expression levels using real-time PCR. There were totally 28 gene candidates for first run of screening, including 12 genes, which were reported as risk for cancer prognosis [14], 14 genes identified as correlated with the incidence of tumor tissues (unpublished data), and two genes with elevated expression in colon cancer patients, A3 adenosine receptor and CCSP-2 [15,16]. Since the measurement of a higher cycle number (i.e., Ct greater than 30) generally implies lower amplification efficiency [17,18], 15 genes were used for further analysis (Table 2) after eliminating genes with low amplification efficiency.

Statistical analysis
The chi-square 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 in order to study the association of candidate genes with CRC. The power of the study was 100% for each candidate gene [13]. The statistical alpha level was 0.05.

Multivariate logistic regression was used to analyze the relationship of the cases and controls to the ∆C(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.

### Table 1: Characteristics of the training and testing sets*§

<table>
<thead>
<tr>
<th>Training set (n=162)</th>
<th>Testing set (n=176)</th>
<th>P-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, yr (S.E.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRC (n=55)</td>
<td>Non-CRC (n=107)</td>
<td>0.335</td>
<td>CRC (n=56)</td>
</tr>
<tr>
<td>Gender, no. (%)</td>
<td>Male</td>
<td>66 (58.2)</td>
<td>53 (49.5)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>33 (41.8)</td>
<td>47 (50.5)</td>
</tr>
<tr>
<td>Stage, no. (%)</td>
<td>I</td>
<td>21 (38.2)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>10 (18.2)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>III</td>
<td>14 (25.5)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>IV</td>
<td>10 (18.2)</td>
<td>-</td>
</tr>
<tr>
<td>Tumor site, no. (%)</td>
<td>Colon</td>
<td>28 (50.9)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Rectum</td>
<td>22 (40.0)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Cecum + Rectum</td>
<td>4 (7.3)</td>
<td>-</td>
</tr>
</tbody>
</table>

**CRC: Colon+Rectal Cancer; *Data are given as means (SE) or as the number of cases (%); §P values were estimated using the t-test.**

---

**Table 1:** Characteristics of the training and testing sets*§.
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. Although this full model seemed capable of discriminating between the CRC cases and controls, it may have resulted in over fitting (Table 2). The logistic regression analysis further resulted in the selection of five genes of significance (i.e., P-value<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, which statistical powers of the five genes were 1.00 between case and control groups in training and testing sets.

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. A five-gene logistic regression model provided good discriminative performance with 87.0% accuracy, 78.0% sensitivity, 92.0% specificity, 90.7% positive predictive value (PPV), and 80.7% negative predictive value (NPV) in the training set. 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 (receiver operating characteristic) curve of 0.978 (0.912-1) in the external validation. 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 genes 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% C.I.: 0.730-0.879), as compared to other one-gene models (AUC: 0.49-0.69). Distinct increases in the AUC of up to 0.905 (95% C.I.: 0.849-0.960) resulted from the combination of five genes (Table 3). The five-gene model fulfilled the criteria of good performance for diagnostic tests as well as accuracy (87%), sensitivity (78%), and specificity (92%); in addition, the Hosmer-Lemeshow test was non-significant (P-value=0.108).

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% (Table 4). The five-gene model performed stably for the discrimination between CRC cases and controls in the training set, with accuracies ranging from 73.9% to 87.0%, with sensitivity of 95%, or with specificity of 95%. In addition, a well performance in the testing set was obtained using the discrimination model, with 84% accuracy, 66% sensitivity, 92% specificity, 89% PPV and 73% NPV (Table 4b).
Table 3: Discrimination power and ROC analysis of different combinations of CRC-associated genes in training set.

<table>
<thead>
<tr>
<th>Genes used for models</th>
<th>AUC</th>
<th>S.E.</th>
<th>P-value</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUSP6, CPEB4, EIF253, MDM2, MMD</td>
<td>0.905</td>
<td>0.028</td>
<td>&lt;0.001</td>
<td>0.849</td>
<td>0.960</td>
</tr>
<tr>
<td>DUSP6, CPEB4, EIF253, MMD2</td>
<td>0.895</td>
<td>0.030</td>
<td>&lt;0.001</td>
<td>0.838</td>
<td>0.953</td>
</tr>
<tr>
<td>DUSP6, CPEB4, EIF253</td>
<td>0.882</td>
<td>0.032</td>
<td>&lt;0.001</td>
<td>0.820</td>
<td>0.945</td>
</tr>
<tr>
<td>DUSP6, CPEB4</td>
<td>0.855</td>
<td>0.032</td>
<td>&lt;0.001</td>
<td>0.791</td>
<td>0.919</td>
</tr>
<tr>
<td>DUSP6</td>
<td>0.804</td>
<td>0.038</td>
<td>&lt;0.001</td>
<td>0.730</td>
<td>0.879</td>
</tr>
</tbody>
</table>

Table 4: Performance of the statistical model based on the five-gene profile.

<table>
<thead>
<tr>
<th>Training set</th>
<th>Testing set</th>
<th>External validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Cancers</td>
<td>107</td>
<td>120</td>
</tr>
<tr>
<td>True negative</td>
<td>98</td>
<td>72</td>
</tr>
<tr>
<td>False positive</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Colorectal Cancers</td>
<td>55</td>
<td>519</td>
</tr>
<tr>
<td>False negative</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>True positive</td>
<td>43</td>
<td>504</td>
</tr>
<tr>
<td>Total</td>
<td>162</td>
<td>176</td>
</tr>
</tbody>
</table>

Logit(P): Logistic Probabilities; PPV: Positive Predictive Value; NPV: Negative Predictive Value

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 [19]. 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 [6]. One study, performed with a newly identified molecular marker known as ProtM [20], also attained unsatisfactory sensitivity.

Circulating tumor cells from any cancer type are capable of disseminating from solid tumor tissues, penetrating and invading blood vessels and circulating in the peripheral blood [21,22]. The number of circulating tumor cells has been used to predict the clinical outcome of cancer patients [23]. On the basis of the presence of circulating tumor cells, we identified five molecular markers, MDM2, DUSP6, CPEB4, MMD, and EIF253, which were differentially expressed between peripheral blood samples of CRC patients and healthy controls. The application of multivariate logistic regression analysis resulted in

Pooling 12 microarray studies to verify the 17 candidate genes and estimate its external generality

Furthermore, we performed the multivariate logistic regression analysis for pooled 12 public microarray data sets as well as the external validation to verify the CRC-associated genes from 3 studies (the present one of Chu et al., Marshall et al. and Han et al.) [7,8]. As the Table 5, we validated the 17 CRC-associated genes from this study (Model 1: 5 genes), Marshall et al. [7] (Model 2: 7 genes) and Han et al. [8] (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 goodness-of-fit test of Hosmer-Lemeshow (H-L) showed statistical significance (p=0.044) for Model 2 of Marshall et al. [7], which observed event rates did not match expected event rates in subgroups of the model population. Models for which expected and observed event rates in subgroups are similar are called well calibrated (Model 1, 3 and 4). A 7-gene model (Model 4 with genes CPEB4, EIF253, MGC20553, MAS4A1, ANXA3, TNFAIP6 and IL2RB) was pairwise selected from genes of Model 1, 2 and 3 that showed the best results in logistic regression analysis (H-L p=1.000, R2=0.951, AUC=0.999, accuracy=0.968, specificity=0.966 and sensitivity=0.994).
a five-gene discrimination model, which achieved good diagnostic performance and provided stable conditions with accuracies ranging from 73.9% to 87.0%, with sensitivity of 95%, or with specificity of 95%.

Both mRNAs and proteins in the peripheral blood have been tested for diagnostic use to detect circulating tumor cells of different solid tumors or to determine prognoses of various cancers. We confirmed, in our study, 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 with all five genes. More and more clinical studies show improvements in the sensitivity of cancer detection by assaying transcript levels of multiple genes in patient peripheral blood [7,8,24].

A higher sensitivity or specificity of the discriminatory performance of our five-gene model was achieved by adjusting the cut-off value of Logit(P) (Table 4a). 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 specificity of 90%, 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% to 86%) were achieved with Logit(P), ranging from 0.0511 to 0.4747. In the testing set, the five-gene model performed with satisfactory accuracy, sensitivity, and specificity.

Two reports [7,8] with similar screening approaches used different gene sets to detect CRC (Table 5). 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.

The strategy of our study was to choose genes clinically confirmed to be cancer-associated in tumor tissues and to validate in peripheral blood samples. Five genes (MDM2, DUSP6, CPEB4, MMD and EIF2S3) identified here for discrimination between CRC patients and healthy controls showed strong association with CRC. MDM2 (Mouse double minute 2 homolog) gene, also known as HDM2 gene in human, is a negative regulator of the tumor suppressor protein p53 [25]. Overexpression of MDM2 gene was reported in several human tumor types, including osteosarcomas, melanoma, non-small cell lung cancer (NSCLC), esophageal cancer, leukemia, and non-Hodgkin’s lymphoma [26-31]. Inhibition of MDM2 can restore p53 activity in cancers containing wild-type p53 and has recently become a strategy to develop anti-tumor drug [32-35].

**Table 5:** The logistic regression models for pooled 12 microarray data sets as the external validation of CRC-associated genes from 3 studies.

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>S.E.</td>
<td>p</td>
<td>B</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>MDM2</td>
<td>6.069</td>
<td>1.461</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CPEB4</td>
<td>-3.177</td>
<td>0.383</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>EIF2S3</td>
<td>1.462</td>
<td>0.244</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average AUC</td>
<td>0.978</td>
<td>0.985</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.818</td>
<td>0.986</td>
<td>0.971</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.971</td>
<td>0.988</td>
<td>0.987</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-32.758</td>
<td>6.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.460</td>
<td>0.194</td>
<td>1.000</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.853</td>
<td>0.941</td>
<td>0.991</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.818</td>
<td>0.986</td>
<td>0.976</td>
</tr>
</tbody>
</table>

Model 1: 5 candidate genes of this study; Model 2: 7 candidate genes of Marshall et al. ; Model 3: 5 candidate genes of Han et al. ; Model 4: stepwise 7 candidate genes from model 1, 2 and 3; B: logistic regression coefficient Beta; S.E.: 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.
carcinoma both at the mRNA and protein level [39,40] and with KRAS mutant colon cancer [41]. In addition, higher expression level of DUSP6 was found in the tamoxifen-resistant breast tumors group compared with the tamoxifen-sensitive tumor group [42] and tumor growth promotion in glioblastoma [43]. The DUSP6 function might vary in different cancer types. On the contrary, some other reports demonstrated that DUSP6 gene was a candidate tumor suppressor gene, for instance, in pancreatic cancer [44], esophageal squamous cell carcinoma [45], and lung cancer with 17.7% cases of study sample [46].

CPEB4, cytoplasmic polyadenylation element binding protein, targets mRNAs and promote translation by inducing cytoplasmic polyadenylation [47,48]. Overexpressed CPEB4 was identified in pancreatic ductal adenocarcinomas and glioblastomas compared with its corresponding normal tissue [49]. Increased CPEB4 mRNA was considered as a prognostic marker for poor clinical outcome in non-small cell lung cancer (NSCLC) patients (hazard ratio=1.8) [14]. In contrast, reduced or weaker CPEB4 expression was observed in most of hepatocellular carcinoma samples compared with normal tissues using IHC staining [50]. In addition, Xu and Liu [51] proposed that CPEB4 gene might be selectively overexpressed in metastatic cancers, such as metastatic prostate cancer, and potentially as a biomarker for chemotherapy resistance.

MMD is an integral membrane protein with seven putative transmembrane segments [52]. Its biological function is still unclear. EIF2S3 is the largest subunit (gamma) of eukaryotic translation initiation factor 2 (EIF2) [53] and might be indirectly involved in inhibition of prostate cancer metastasis through N-myc downstream regulated gene 1 [54].

Our study has firstly presented that four expressing genes in PBMC-derived fractions, including MDM2, CPEB4, EIF2S3 and MMD, have the direct association with CRC with significance. As many clinical studies have been reported, MDM2, DUSP6 and CPEB4 have been showed their association with other pathologies, especially different cancer types. These observations might provide the evidence that these biomarkers play central roles during carcinogenesis or malignance of tumor, but with different strength depending on cancer type. Indeed, it is important to have multiple biomarkers integrated in developed diagnostic or prognostic methods, while each candidate gene has its independent power and the efficacy to discriminate cancer and normal subject (Table 3) [note: deletion of “could be incredible elevated”].

There are several limitations of current study. Since the small number of different stages in the study CRC sample, we were not able to know whether individual gene expression or the five gene signature is stage-dependent. Secondly, the change of gene expression level in the BPMC fraction of CRC patient before and after treatment was not studied due to the restriction of single blood drawing of IRB-approved clinical protocol. Thirdly, the collection of survival status information is not completed and prognostic value of biomarkers could not be evaluated in this study due to many censored cases (over 50%; 5-year survival rate of CRC patients is around 50%).

Further investigation is warranted on the potential of gene signature for evaluation of clinical staging, metastatic probability and survival in CRC patients, when information for the disease progression and survival is completely collected. In addition, the application of currently identified gene signature for CRC detection is very important and it is also the goal for assay development. Discussion with physicians will be planning for integration the test of our CRC-specific gene signature into the national screening program for CRC. Especially, the diagnostic performance between this CRC-specific gene signature and current screening method, such fecal occult blood test (FOBT) or colonoscopy should be approached. As well as the potential of the individual gene expression or gene signature for evaluation of therapeutic response should be planned as future work.

Furthermore, we verified the CRC-associated genes by pooling 12 public microarray data sets that the four logistic models performed similar AUCs without statistically significant difference. In the future, the 7-gene logistic regression model (Model 4: CPEB4, EIF2S3, MGC20553, MA5A41, ANXA3, TNFAP6 and IL2RB) showed the best results that can be further verified for more samples. Meanwhile, the causal relations are needed to confirm among the selected genes and CRC. The expression signature of these CRC-associated genes should be evaluated for early detection of CRC, with more samples randomly screened from the population; in addition, subjects who eventually receive a diagnosis of CRC should be evaluated as well. 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.

Conclusion

In conclusion, we found that the expression profile of 7 genes, CPEB4, EIF2S3, ANXA3, TNFAIP6, IL2RB, MGC20553 and MA5A41, is highly associated with colorectal cancer. Detection of cancer cell-specific biomarkers in the peripheral blood can be an effective screening strategy for CRC.

Competing Interests

No other potential conflict of interest relevant to this article was reported. The authors have declared that no competing interests exist. Dr. Terng, Woan-Jen Lee, and Chin-Yu Chen report being employees of Advypharma. Authors of Advypharma involved in partial sample collection, laboratory experiment and manuscript.

Financial Disclosure

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Finance was supported by Taiwan’s SBIR promoting program from the Department of Industrial Technology of Ministry of Economic Affairs and by research grants from National Defense Medical Center of Ministry of Defense.

Ethics Statement

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 were obtained.

Authors’ contributions


Acknowledgments

Research Funding: This work was supported by Taiwan’s SBIR promoting program from the Department of Industrial Technology of the Ministry of Economic Affairs, Advypharma, Inc., and the National Defense Medical Center (NDMC),
References