Title

Identification and Validation of a Blood-Based 18-Gene Expression Signature in Colorectal Cancer

Authors and Affiliations

Ye Xu1,2*, Qinghua Xu1,3,4*, Li Yang1,2, Xun Ye3,4, Fang Liu3,4, Fei Wu3,4, Shujuan Ni1,5,6,7, Cong Tan1,5,6,7, Guoxiang Cai1,2, Xia Meng3,4, Sanjun Cai1,2, Xiang Du1,5,6,7

1Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China;
2Department of Colorectal Surgery, Fudan University Shanghai Cancer Center, Shanghai, China;
3Fudan University Shanghai Cancer Center – Institut Mérieux Laboratory, Fudan University Shanghai Cancer Center, Shanghai, China;
4bioMérieux (Shanghai) Co., Ltd., Shanghai, China;
5Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai, China;
6Institutes of Biomedical Sciences, Fudan University, Shanghai, China;
7Institute of Pathology, Fudan University, Shanghai, China;

*Ye Xu and Qinghua Xu contributed equally to this work.

Corresponding author: Xiang Du, Fudan University Shanghai Cancer Center, 270 Dong’an Road, Shanghai 200032, China. Phone: 86-21-64175590-8911; Fax: 86-21-64174774; E-mail: dx2008cn@yahoo.com.cn and Sanjun Cai, Phone: 86-21-6417590-1108; Fax: 86-21-64035387; E-mail: caisanjun@csc.org.cn
Running title

Blood Biomarkers in Colorectal Cancer

Keywords

Blood; biomarker; colorectal cancer; screening; gene expression signature

Financial support

This study was supported by grants from bioMérieux. This study was also supported by the Chinese National Clinical Key Discipline (2011–2012) and the Shanghai Science and Technology Commission of Shanghai Municipality (No.10DJ1400500).

Conflicts of interest

XM, XY, QHX, FL and FW are employees of bioMérieux (Shanghai) Co., Ltd. The other authors declare no conflicts of interest. A patent application of the 18-gene panel is in progress.

Text word count: 4597
Abstract word count: 222
Number of figures: 4
Number of tables: 2
Number of references: 50
Translational Relevance

Colorectal cancer is the leading cause of cancer-related deaths worldwide. The early detection of colorectal cancer is crucial for successful treatment and patient survival. However, compliance with current screening methods remains poor and there is a clear need for an accurate in vitro blood test to increase participation in colorectal cancer screening. This study shows that the gene expression profiles of peripheral blood samples can be used to distinguish between colorectal cancer patients and controls. An 18-gene signature was identified and validated as highly sensitive and specific for detecting colorectal cancer in blood samples. These results open an avenue for the further development of blood-based gene expression biomarkers for the diagnosis and early detection of colorectal cancer.
Abstract

Purpose: The early detection of colorectal cancer (CRC) is crucial for successful treatment and patient survival. However, compliance with current screening methods remains poor. This study aimed to identify an accurate blood-based gene expression signature for CRC detection.

Experimental Design: Gene expression in peripheral blood samples from 216 patients with CRC tumors and 187 controls was investigated in the study. We first performed a microarray analysis to select candidate genes that were significantly differentially expressed between cancer patients and controls. A quantitative reverse transcription polymerase chain reaction assay was then used to evaluate the expression of selected genes. A gene expression signature was identified using a training set (n = 200) and then validated using an independent test set (n = 160).

Results: We identified an 18-gene signature that discriminated the CRC patients from controls with 92% accuracy, 91% sensitivity and 92% specificity. The signature performance was further validated in the independent test set with 86% accuracy, 84% sensitivity and 88% specificity. The area under the receiver operating characteristics curve was 0.94. The signature was shown to be enriched in genes related to immune functions.

Conclusions: This study identified an 18-gene signature that accurately discriminated CRC patients from controls in peripheral blood samples. Our results prompt the further development of blood-based gene expression biomarkers for the diagnosis and early detection of CRC.
Introduction

Colorectal cancer (CRC) is the third most common malignancy and the fourth most common cause of cancer mortality worldwide (1). In 2008, more than one million cases were newly diagnosed, and more than 600,000 people died from the disease (2). Given its slow development from removable precancerous lesions and curable early stages, screening for CRC has the potential to reduce both the incidence and mortality of the disease (3). The available screening tools include fecal occult blood testing (FOBT), stool DNA tests, flexible sigmoidoscopy, computed tomographic (CT) colonography, and colonoscopy. Different screening strategies are preferred in various countries. However, compliance with current CRC screening recommendations remains poor. The most reliable screening tool, colonoscopy, is invasive, costly, and performed infrequently. In contrast, the currently most widely used noninvasive screening option, FOBT, has important limitations, including inconvenience and low sensitivity.

The discovery of novel biomarkers based on the analysis of blood samples has become a focus of current research. A novel blood biomarker may offer several practical advantages compared with the currently used screening approaches. First, in vitro blood tests are safe and minimally invasive. Second, no dietary restriction, colon cleansing, or sedation is required. Third, the sample collection and processing procedures may be easier and more convenient. Furthermore, there is no microflora that could degrade the biomarker or hamper the analysis. Thus, a sensitive and specific in vitro blood test that detects CRC in the currently noncompliant patient population will improve participation in screening, increase the accuracy of detection,
and therefore save lives.

Several studies have shown that molecular biomarkers in the peripheral blood can be used to develop *in vitro* tests for the detection of CRC. The DNA methylation biomarker methylated Septin 9 (mSEPT9) previously showed a sensitivity of 70% and a specificity of 90% for discriminating CRC patients from controls (4, 5). Recently, the performance of mSEPT9 was determined in a prospective study of 7940 average-risk individuals undergoing colonoscopy for CRC screening. The sensitivity in the 45 CRC patients was 67%, with a specificity of 88% (6). DNA microarray technology can quantify the expression of several thousand genes simultaneously and may be able to capture the complex biology that underlies colorectal tumorigenesis and progression better than single gene markers. A number of studies have been published, using DNA microarray technology to identify blood-based gene expression signatures for CRC detection. Han et al reported a five-gene signature with 88% sensitivity and 64% specificity (7). Marshall et al recently published a seven-gene signature with a sensitivity of 72% and a specificity of 70% (8). Rosenthal et al also reported a panel of 202 genes with 90% sensitivity and 88% specificity (9).

In this study, we tested the hypothesis that gene expression profiling of peripheral blood cells could yield diagnostic information in a cohort of Chinese patients. An 18-gene signature was identified and validated as highly sensitive and specific for detecting CRC in blood samples.
Materials and Methods

Study design and patients

Study participants were recruited from the Fudan University Shanghai Cancer Center and Shanghai Qibao Community Hospital from 2006 through 2010 in Shanghai, China. All the participants were Chinese. For the CRC group, all patients had blood collected before surgery. None of the cancer patients had received preoperative radiotherapy or chemotherapy before blood collection. The tumors were staged according to the Tumor-Node-Metastasis (TNM) system. Patients suffering from hereditary CRC were excluded. For the control group, FOBT positive participants without any symptoms of inflammatory bowel diseases, polyps or CRC, which had been confirmed by colonoscopy, were enrolled from a population-based screening program. Figure 1 depicts three different phases of the study design and Table 1 summarizes the clinical characteristics of the samples in the study.

In the discovery set, we analyzed the whole blood gene expression profiles of 100 CRC patients and 100 controls using GeneChip® U133plus2 microarrays (Affymetrix, Santa Clara, USA). These was no significant difference in the distribution of age or gender between the CRC and control groups. The CRC group included 41 colon cancer and 59 rectal cancer patients. Sixteen of the cancer patients were stage I, 36 were stage II, 24 were stage III, and 24 were stage IV. The significance analysis of microarray (SAM) method was used to identify genes that were differentially expressed between the CRC and control groups (10). The list of genes was further refined by expression signal intensity, fold change, biological annotation, and probeset grade. Finally, 52 unique genes were selected for further testing by
quantitative reverse transcription PCR (qRT-PCR).

The training set consisted of 100 CRC patients and 100 controls, including 71 CRC patients and 86 controls that were also used in the discovery set to assess the correspondence between the microarray and qRT-PCR measurements. The remaining 43 samples from the discovery set were not applicable for qRT-PCR experiments due to low RNA concentrations and, thus, were replaced these with new samples. The distributions of age and gender were balanced between the CRC and control groups. The CRC group included 51 colon cancer and 49 rectal cancer patients. Eleven of the cancer patients were stage I, 40 were stage II, 22 were stage III, and 27 were stage IV. These samples were used as the training set to identify the gene expression signature for differentiation between the CRC group and the control group.

In the test set, we used an independent cohort of 87 CRC patients and 73 controls. The CRC group included 27 colon cancer and 60 rectal cancer patients. Eleven of the cancer patients were stage I, 31 were stage II, 31 were stage III, and 13 were stage IV. The TNM variables of one patient were not available. The fully specified gene expression signature from the training set was applied to the test set for validating the signature performance in the independent samples.

The study was approved by the Institutional Review Board of Fudan University Shanghai Cancer Center, and written informed consent was obtained from all participants.
**Blood collection and RNA extraction**

For each participant, 2.5 mL of peripheral blood was collected into PAXgene™ Blood RNA tubes, and the total RNA was extracted with the PAXgene™ Blood RNA System (PreAnalytiX, Hilden, Germany). The quantity of total RNA was measured with a spectrophotometer at an optical density of 260 nm, and the quality was assessed using the RNA 6000 Nano LabChip kit on an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, USA). All samples met the quality criterion: RNA Integrity Number > 7.0.

**Microarray hybridization**

For each sample, 50 nanograms of total RNA were reversely transcribed and linearly amplified as single-stranded cDNA using Ribo-SPIA™ technology with the WT-Ovation™ RNA Amplification System (NuGEN Technologies, San Carlos, USA), and the products were purified using the QIAquick™ PCR Purification Kit (QIAGEN, Hilden, Germany). A total of 2 micrograms of amplified and purified cDNA were subsequently fragmented with RQ1 RNase-Free DNase (Promega, Fitchburg, USA) and labeled with biotinylated deoxynucleoside triphosphates using Terminal Transferase (Roche Diagnostics, Indianapolis, USA) and GeneChip® DNA Labeling Reagent. The labeled cDNA was hybridized onto the GeneChip® U133plus2 microarray in a Hybridization Oven 640 (Agilent Technologies) at 60 rotations per minute at 50 °C for 18 hours. After hybridization, the arrays were washed and stained according to the Affymetrix protocol EukGE-WS2v4 using a GeneChip® Fluidics Station 450. The arrays were scanned with a GeneChip® Scanner 3000.
QuantiTect® Reverse Transcription Kit (QIAGEN, Hilden, Germany). The cDNA was amplified using the SYBR® Premix DimerEraser® (Perfect Real Time) Kit (Takara biotechnology, Dalian, China). The amplification was detected in real time using the Applied Biosystems 7900HT Fast Real-Time PCR System (Life Technologies, Carlsbad, USA). The primers were designed primarily within the target sequences of the selected Affymetrix probesets with Beacon Designer™ software (Premier Biosoft, Palo Alto, USA). The primer sequences are provided in Table S1. The primer pairs were experimentally validated with the following criteria: (1) a single gene-specific product was produced; (2) the amplification efficiency ranged between 90% and 110%; and (3) the cycle threshold (Ct) value of the no-template control was over 35.

Six candidate reference genes (ACTB, CRY2, CSNK1G2, DECR1, FARP1 and TRAP1) that have been reported to be consistently expressed in human whole blood samples were selected for investigation (11). The expression stability of the six reference genes in the training set were estimated using four commonly used algorithms: geNorm (12), NormFinder (13), BestKeeper (14) and the comparative cycle threshold (ΔCt) method (15). Each algorithm ranked the reference genes from most stable (Rank #1) to least stable (Rank #6). The overall ranking of candidate reference genes was calculated according to the RefFinder method described by Chen et al (16).
Briefly, the geometric means of the four ranking numbers of each gene were calculated, and then candidate reference genes were ranked according to the geometric mean, the gene with the smaller geometric mean being the most stable reference gene. As a result, **CSNK1G2**, **DECR1** and **FARP1** were shown to be the most stable genes among the candidate list. Therefore, these three genes were selected as reference genes and their geometric mean was used as a normalization factor for qRT-PCR data normalization.

**Statistical analysis**

Microarray data were analyzed using R software and packages from the Bioconductor project (17, 18). Raw data had been deposited in the ArrayExpress public repository and were accessible through the accession number E-MTAB-1532. Raw data were normalized using the robust multichip average (RMA) method (19). The probeset-level data were log2 transformed. In addition, we applied a bioinformatics-based filtering approach using information in the Entrez Gene Database (20). Probe sets without Entrez Gene ID annotation were removed. For multiple probe sets mapping to the same Entrez Gene ID, only probe sets showing the largest inter quantile range were kept, and the rest were excluded. The analysis of differentially expressed genes was conducted using the SAM method implemented in the “samr” package (10).

The qRT-PCR-based gene expression levels were estimated using the comparative ΔCt method of relative quantification (21), normalizing the Ct values relative to the normalization factor. The relative fold change was represented as $2^{-\Delta\Delta Ct}$, where $\Delta\Delta Ct$
\[ \Delta Ct_{CRC} = \text{mean } \Delta Ct_{CRC} - \text{mean } \Delta Ct_{\text{Control}}. \]

We used the minimum redundancy maximum relevance (mRMR) algorithm for gene selection (22), the support vector machine (SVM) algorithm for classification (23), and the leave-one-out-cross-validation (LOOCV) procedure for sampling. Gene Ontology and KEGG pathway analysis were conducted using the GeneCodis bioinformatics tool (24-26).
Results

Microarray analysis and candidate gene selection

The recent release of the U133plus2 array contains over 54,000 probe sets, which represent approximately 38,500 human genes. Confronted with such an overwhelming amount of information, it was necessary to reduce the total number of genes analyzed to a manageable number of genes with well-characterized biological information and use visualization schemes to facilitate the recognition of patterns in the data (27). We thus performed a bioinformatics-based filtering procedure to summarize the probesets at the gene level and exclude those probesets with low-grade biological annotations. After filtering, the expression profiles of 8,662 unique genes in 100 CRC patients and 100 controls were retained for downstream analysis.

Using the SAM method, we identified 263 genes that were differentially expressed between the CRC group and the control group, among which 179 genes were up-regulated and 84 genes were down-regulated in the cancer patients. Using the 263 differentially expressed genes, a hierarchical clustering analysis showed that 78 of 100 controls and 76 of 100 CRC patients were correctly classified (Supplementary Figure S1). The gene list was further refined by multiple criteria: (1) an average expression intensity of greater than 64; (2) a fold change in mean expression intensity of greater than 1.2; (3) genes with known biological function; and (4) a high-grade probe design. This analysis resulted in 52 candidate biomarkers for further qRT-PCR study.
Identification of an 18-gene signature in the training set

The training set included 100 CRC patients and 100 controls, among which 71 CRC patients and 86 controls were also included in the discovery set. First, we compared the expression profiles of the 52 selected genes measured by both microarray and qRT-PCR. For each gene, the fold change between the 71 CRC patients and 86 controls was calculated. Spearman’s correlation analysis showed that the fold change of candidate genes measured by microarray and qRT-PCR was highly comparable ($r = 0.94; 95\% CI, 0.90–0.98; P < 0.001$), although a few exceptions were observed. Three genes were considered outliers because their expression profiles were not consistent between the microarray and qRT-PCR analyses.

Signature identification was subsequently conducted using the SVM classification model, with the selection of significant genes based on the mRMR method, through repetitions of the LOOCV process. As shown in Figure 2, our process could conceptually be broken into six steps:

1. The samples were divided into an inner training set and an inner test set. The inner test set consisted of only a single sample; the remaining 199 samples were placed in the inner training set. The sample in the test set was placed aside and not utilized in the development of the class prediction model.

2. Using only the inner training set, the mRMR algorithm was used to search for a subset of $n$ genes that had maximum relevance with the clinical status and minimum redundancy within the gene set.

3. The $n$-gene based SVM model was built using the linear kernel and was fitted to the inner training set.
4. The developed model was used to predict the class of the test sample. The prediction was based on the expression profile of the test sample, without using knowledge of the true class of the sample. The predicted class was compared to the true class label of the sample. If they disagreed, the prediction was in error.

5. Then, a new training set–test set partition was created. This time another sample was placed in the inner test set, and all of the other samples were placed in the inner training set. A new classification model was constructed using the samples in the new training set. Although the same algorithm for gene selection and parameter estimation was used, because the new model was constructed based on the new training set, it would in general not select exactly the same gene set as the previous model. Again, the model was applied to the expression profile of the test sample. If the predicted class did not agree with the true class label of the test sample, then the prediction was in error. The process was repeated leaving each of the 200 biologically independent samples out of the training set, one at a time. During the steps, 200 different models were created and each model was used to predict the class of the test sample. Eventually, the number of prediction errors was totaled and reported as the LOOCV-based error rate.

6. The number of genes (n) to be selected by mRMR algorithm was set as a predefined variable. Steps 1-5 were repeated 52 times, starting with $n = 1$, adding one gene at time until the $n = 52$. The LOOCV-based performance for each of $N$-gene signature was estimated and used to determine the optimal size of the signature.

At the end of the process, 10,000 different classification models were constructed to
assess signature performance in association with the number of genes included in
the signature. As shown in Figure 3A, the classification accuracy was 53.5% when
only one gene was included in the signature. The performance accuracy increased to
90.0% when nine genes were included in the signature. However, the classification
accuracy decreased to 85.5% when another gene was added, implying that the
observed performance might not be truly stable at a signature size of approximately
10. The performance increased continuously and reached 90.5% and 91.5% accuracy
at signature sizes of 15 and 18, respectively. Subsequently, the signature
performance increased to 92.5% and 93.0% accuracy with signature sizes of 38 and
48, respectively.

Not surprisingly, the classification performance increased continuously as more
genes were added to the signature. However, signature size determination also
needs to consider signature complexity; when more genes are included in the
signature, the signature becomes more complex and less generalizable. Our strategy
was to identify a signature that showed satisfactory performance but maintained a
compact signature size and reasonable complexity for future development. Although
the prediction accuracy could be improved from 91.5% to 93% by adding 30
additional genes (signature size from 18 to 48), this input-output ratio was not truly
effective.

Ultimately, we considered a signature size of 18 to be ideal. As shown in Figure 3A, an
estimated performance of 91.5% accuracy, 91% sensitivity, and 92% specificity was
achievable with a signature composed of the top 18 genes. We then used a
consensus method to assess the stability of candidate genes across the LOOCV process. Through the LOOCV process specifying n = 18, we identified 200 different top-18 gene lists. The appearance rate of each gene among all of the top-18 gene lists was recorded. The maximum appearance rate was 100% (ie, some genes were included in all of the top 18 gene lists during the cross validation process). In contrast, the minimum occurrence rate was zero, indicating that those genes were never included in the top 18 gene lists during the cross validation process. The genes were ranked according to their appearance rates. The 18 genes with the highest appearance rates were selected for the final gene expression signature. The entire training set was used to specify the parameters of the final 18-gene signature.

Table 2 shows the descriptions of the 18 selected genes and their statistical significance across control and CRC samples in the training set. Eight and 10 genes were up- and downregulated, respectively, in CRC patients compared with controls (Figure 3B); NEAT1 (nuclear paraspeckle assembly transcript 1) was the most significantly upregulated gene; and DUSP2 (dual specificity phosphatase 2) was the most significantly downregulated gene. Gene Ontology and KEGG pathway analyses showed that several of the 18 genes are involved in apoptosis (eg, GZMB and IL1B), cell adhesion (eg, CD36 and ITGAM), the MAPK signaling pathway (eg, DUSP2 and IL1B), signal transduction (eg, SH2D2A, PDE4D, and IL1B), and some are hematopoietic cell markers (eg, CD36, ITGAM, and IL1B; Table S2).

Validation of the 18-gene signature in an independent test set

The 18-gene signature was then applied to an independent test set of 87 CRC
patients and 73 controls. These samples had been excluded from the training set and were not used in the development of the 18-gene signature. For each sample, a probability score was calculated, and a threshold of 50% was used to classify samples. Samples with a probability score below 50% were classified as controls, whereas samples with a probability score above 50% were classified as CRC (Figure 4). For the controls, 64 samples were correctly classified, and nine samples were misclassified. For the CRC patients, 73 samples were correctly classified, and 14 samples were misclassified. In the test set, the 18-gene signature had 85.6% accuracy (95% CI, 0.79-0.90), 83.9% sensitivity (95% CI, 0.74-0.90), and 87.7% specificity (95% CI, 0.77-0.94). When the sensitivity was plotted against the specificity in a receiver operating characteristic curve, the area under the curve was 0.94 (95% CI, 0.91-0.98).

Of the nine misclassified controls, four were female, and five were male; three were over 60 years old. Of the 14 misclassified cancer patients, five were female, and nine were male; four were over 60 years old. Additionally, four of the 14 misclassified cancer patients were stage I, four were stage II, five were stage III, and one was stage IV; two had colon cancer, and 12 had rectal cancer. The results obtained from Fisher’s exact test suggested that clinical and pathological variables, such as age, gender, tumor site, and stage, did not obviously affect the prediction outcomes (Table S3).
Discussion

The early detection of CRC is crucial for successful treatment and patient survival. However, the lack of compliance remains the greatest challenge currently limiting CRC screening effectiveness. In this study, we aimed to identify and validate a blood-based gene signature that could distinguish CRC patients from controls with high accuracy. Our work followed Biomarker Development for Early Detection of Cancer guidelines (28). We first performed a microarray study to select genes that were significantly differentially expressed between the controls and CRC patients. The selected genes were then transferred to a qRT-PCR platform. The qRT-PCR study was performed for signature identification and validation using two independent cohorts. We presented a blood-based 18-gene signature that can be used as a biomarker to discriminate between CRC patients and controls with a sensitivity and specificity of 84% and 88%, respectively. Fisher’s exact test showed no association between clinical variables (age, gender, tumor sites and stages) and the prediction outcomes assigned to each sample, suggesting that the 18-gene signature performed equally well for different categories of samples.

An understanding of the function of the genes comprising the signature could provide mechanistic insight into the diagnostic effect of this gene panel. Functional annotations revealed that many of the selected genes were related to immune function. GZMB, which was significantly down-regulated in CRC patients, is crucial for the rapid induction of target cell apoptosis by cytotoxic T lymphocytes and natural killer cells in the cell-mediated immune response (29). IL1B, an important mediator of the inflammatory response, is involved in different mechanisms leading to
tumorigenesis via tumor-associated inflammation and neovascularization (30). IL1B has been shown to upregulate COX2 expression in human CRC cells (31), which may contribute to the growth and metastatic potential of CRC by increasing the expression of the antiapoptotic factor BCL-2 and upregulating specific angiogenic factors (32, 33). Furthermore, polymorphisms in IL1B have been associated with tumor recurrence in stage II colon cancer (34). The expression of SH2D2A is limited in immune system tissues, particularly activated T-cells. SH2D2A has been reported as a positive regulator of proximal T-cell receptor signal transduction (35). CD36 is expressed by various types of cells that are associated with the blood and the immune system. High CD36 expression is related to decreased stromal vascularization and is a predictor of good prognosis in colon cancer (36). A polymorphism in CD36 (A52C) has been associated with an increased risk for CRC (37).

In addition, several of these genes have been associated with other human carcinomas. In lymphoma, MYBL1 was shown to activate the BCL2 P2 promoter through a Cdx-binding site, promoting resistance to apoptosis (38). PDE4D was reported to be overexpressed in human prostate cancers and associated with increased tumor growth and cell migration (39). PDE4D is also expressed in lung cancer, interacting with HIF signaling and promoting lung cancer progression (40). PDZK1IP1 was shown to be overexpressed in a variety of human cancers in the kidney, colon, lung, and breast (41). The overexpression of the PDZK1IP1 protein was correlated with tumor progression in prostate and ovarian carcinomas (42).
Interestingly, a long noncoding RNA, NEAT1, was the most significantly upregulated gene in our analysis. Recent studies have shown that NEAT1 plays an essential role in the assembly and architecture of nuclear paraspeckles and colocalizes with the paraspeckle-associated proteins p54\textsuperscript{nrb}, PSP1, and PSF (43-45). NEAT1 may therefore play an important role in regulating gene expression by governing the nuclear export of mRNAs. Our study is the first to report the deregulation of DHRS13, FAM198B, GLT25D2, ITPRIPL2, NUDT16, P2RY10, and VSIG10 mRNA expression in the peripheral blood in association with colorectal carcinoma.

Recently, several studies have suggested that the gene expression signatures identified in peripheral blood are likely not conventional tumor-derived cancer biomarkers but rather reflect subtle alterations in blood gene expression serving as a systemic immune response to tumorigenesis(8, 46-49). Our results are consistent with these findings as evidenced by the fact that the signature was enriched in genes related to immune functions. Recent work has elucidated the role of distinct immune cells, cytokines, and other immune mediators in virtually all steps of colorectal tumorigenesis, including initiation, promotion, progression and metastasis (50). Although we cannot entirely exclude the fact that metastatic tumor cells or circulating cell-free tumor nucleic acids can affect the gene expression profiles of peripheral blood, we postulate that the distinct gene expression observed in controls and CRC patients is most likely attributed to the interactions between the immune system and tumors. Based on these results, future research is needed to understand the mechanistic relationship and the biological meaning of this complex blood gene expression signature in CRC.
Although the 18-gene signature could be developed as an *in vitro* blood test for general population screening, the initial clinical use of our biomarkers is more likely to serve as a complementary test to existing screening methods. In clinical practice, only 30% of high-risk individuals with FOBT-positive results eventually undergo colonoscopy. Our biomarker may therefore provide an additional risk assessment for noncompliant populations. Because a higher probability score by the 18-gene signature increases the likelihood that a patient has cancer, the probability score may provide clinically actionable risk information. For patients with a probability score below 50%, patients may decide by themselves whether to undergo colonoscopy. For those patients who resist undergoing colonoscopy, repeat blood tests would be recommended at intervals consistent with practice guidelines, eg, annually. For patients with a probability score above 50%, they should be strongly recommended to undergo colonoscopy for confirmation. The combination of blood biomarkers with existing screening methods can provide several major advantages. It addresses the greatest challenge currently limiting CRC screening effectiveness, namely, the lack of compliance. Blood biomarkers may have valuable utility for reaching the segment of the screening population that is resistant to the currently recommended methods as well as providing screening to underserved patients with limited access to endoscopy centers. Additionally, blood biomarkers may help to minimize false-positive FOBTs and thus reduce the number and cost of colonoscopies.

Considerable research effort continues for the development of an accurate, reliable
and minimally invasive blood test for the detection of CRC. Han et al and Marshall et al have identified a five- and seven-gene signatures, respectively, by analyzing gene expression profiles in whole blood samples from CRC patients and controls using qRT-PCR assays (7, 8). Although similar approaches were used for biomarker identification, the reported two signatures share no genes in common with our 18-gene signature. The absence of concordant genes could be related to many different issues, including differences in the study populations, the gene quantification technologies, and the statistical approaches used to generate the gene signatures, highlighting the need for extensive validation before the clinical implementation of these promising biomarkers. Our results represent an encouraging primary step, but several issues remain to be addressed. Additional external validation studies are being conducted to establish a standard testing protocol and to confirm the signature accuracy. The utility of the 18-gene signature for detecting precancerous lesions, such as polyps and adenomas, also needs to be evaluated. Finally, a large prospective study in the target screening population is required to fully determine the clinical performance of the blood-based gene signature in comparison to colonoscopy and to assess the practical feasibility of implementing the blood test in a screening program.

In conclusion, our study describes the development and validation of a blood-based 18-gene signature that differentiates CRC patients from controls with a high degree of accuracy in a large number of participants. Our results open an avenue for the further development of blood-based gene expression biomarkers for the diagnosis and early detection of CRC.
Acknowledgments

We wish to thank HUANG Wencui at the Fudan University Shanghai Cancer Center and JIN Ying, ZHANG Ling, and YE Aijun at the Shanghai Qibao Community Clinic for their excellent work in blood sample collection.

Grant Support

This study was supported by Grants from bioMérieux. This study was also supported by the Chinese National Clinical Key Discipline (2011–2012) and the Shanghai Science and Technology Commission of Shanghai Municipality (No.10DJ1400500).
References:


normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 2002;3: H34.


23. Chang C, Lin C. LIBSVM: A library for support vector machines. ACM Transactions on...


#### Table 1. Characteristics of the CRC and Control populations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Discovery Set</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CRC (n = 100)</td>
<td>Control (n = 100)</td>
<td>CRC (n = 100)</td>
</tr>
<tr>
<td><strong>Age, year</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>57·6</td>
<td>56·5</td>
<td>56·3</td>
</tr>
<tr>
<td>Range</td>
<td>27-78</td>
<td>38-74</td>
<td>27-78</td>
</tr>
<tr>
<td><strong>Sex, no. (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>50 (50·0)</td>
<td>50 (50·0)</td>
<td>49 (49·0)</td>
</tr>
<tr>
<td>Female</td>
<td>50 (50·0)</td>
<td>50 (50·0)</td>
<td>51 (51·0)</td>
</tr>
<tr>
<td><strong>Tumor site, no. (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon</td>
<td>41 (41·0)</td>
<td></td>
<td>51 (51·0)</td>
</tr>
<tr>
<td>Rectum</td>
<td>59 (59·0)</td>
<td>49 (49·0)</td>
<td></td>
</tr>
<tr>
<td><strong>Tumor stage, no. (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>16 (16·0)</td>
<td>11 (11·0)</td>
<td></td>
</tr>
<tr>
<td>Stage II</td>
<td>36 (36·0)</td>
<td>40 (40·0)</td>
<td></td>
</tr>
<tr>
<td>Stage III</td>
<td>24 (24·0)</td>
<td>22 (22·0)</td>
<td></td>
</tr>
<tr>
<td>Stage IV</td>
<td>24 (24·0)</td>
<td>27 (27·0)</td>
<td></td>
</tr>
</tbody>
</table>

*TNM stage was not available for one patient in the Test Set.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Cytoband</th>
<th>UniGene</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD36</td>
<td>CD36 molecule (thrombospondin receptor)</td>
<td>7q11.2</td>
<td>Hs.120949</td>
<td>8.76E-07</td>
</tr>
<tr>
<td>DHRS13</td>
<td>dehydrogenase/reductase (SDR family) member 13</td>
<td>17q11.2</td>
<td>Hs.631760</td>
<td>2.96E-04</td>
</tr>
<tr>
<td>DUSP2</td>
<td>dual specificity phosphatase 2</td>
<td>2q11</td>
<td>Hs.1183</td>
<td>1.69E-10</td>
</tr>
<tr>
<td>FAM198B</td>
<td>family with sequence similarity 198, member B</td>
<td>4q32.1</td>
<td>Hs.567498</td>
<td>7.57E-08</td>
</tr>
<tr>
<td>FKBPS</td>
<td>FK506 binding protein 5</td>
<td>6p21.31</td>
<td>Hs.407190</td>
<td>4.70E-05</td>
</tr>
<tr>
<td>GLT25D2</td>
<td>glycosyltransferase 25 domain containing 2</td>
<td>1q25</td>
<td>Hs.387995</td>
<td>2.39E-07</td>
</tr>
<tr>
<td>GZMB</td>
<td>granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)</td>
<td>14q11.2</td>
<td>Hs.1051</td>
<td>9.68E-09</td>
</tr>
<tr>
<td>IL1B</td>
<td>interleukin 1, beta</td>
<td>2q14</td>
<td>Hs.126256</td>
<td>2.09E-04</td>
</tr>
<tr>
<td>ITGAM</td>
<td>integrin, alpha M (complement component 3 receptor 3 subunit)</td>
<td>16p11.2</td>
<td>Hs.172631</td>
<td>6.27E-09</td>
</tr>
<tr>
<td>ITPRIPL2</td>
<td>inositol 1,4,5-trisphosphate receptor interacting protein-like 2</td>
<td>16p12.3</td>
<td>Hs.530899</td>
<td>7.72E-05</td>
</tr>
<tr>
<td>MYBL1</td>
<td>v-myb myeloblastosis viral oncogene homolog (avian)-like 1</td>
<td>8q22</td>
<td>Hs.445898</td>
<td>2.42E-09</td>
</tr>
<tr>
<td>NEAT1</td>
<td>nuclear paraspeckle assembly transcript 1 (non-protein coding)</td>
<td>11q13.1</td>
<td>Hs.523789</td>
<td>3.48E-10</td>
</tr>
<tr>
<td>NUDT16</td>
<td>nudix (nucleoside diphosphate linked moiety X)-type motif 16</td>
<td>3q22.1</td>
<td>Hs.282050</td>
<td>1.08E-05</td>
</tr>
<tr>
<td>P2RY10</td>
<td>purinergic receptor P2Y, G-protein coupled, 10</td>
<td>Xq21.1</td>
<td>Hs.296433</td>
<td>8.54E-09</td>
</tr>
<tr>
<td>PDE4D</td>
<td>phosphodiesterase 4D, cAMP-specific</td>
<td>5q12</td>
<td>Hs.117545</td>
<td>4.32E-08</td>
</tr>
<tr>
<td>PDZK1IP1</td>
<td>PDZK1 interacting protein 1</td>
<td>1p33</td>
<td>Hs.431099</td>
<td>4.02E-05</td>
</tr>
<tr>
<td>SH2D2A</td>
<td>SH2 domain containing 2A</td>
<td>1q21</td>
<td>Hs.103527</td>
<td>3.01E-08</td>
</tr>
<tr>
<td>VSIG10</td>
<td>V-set and immunoglobulin domain containing 10</td>
<td>12q24.23</td>
<td>Hs.187624</td>
<td>1.60E-06</td>
</tr>
</tbody>
</table>
**Figure legends**

**Figure 1. Study design.** The blood gene expression profiles of 216 CRC patients and 187 controls were investigated in three different phases. We first performed a microarray analysis to select candidate genes that were significantly differentially expressed between cancer patients and controls. Quantitative reverse transcription polymerase chain reaction assays were then applied to evaluate the expression of selected genes. A gene expression signature was identified using a training set (n = 200) and then validated using an independent test set (n = 160). *TNM stage was not available for one patient in the Test Set. CRC, colorectal cancer; FUSCC, Fudan University Shanghai Cancer Center; RT-PCR, reverse transcription polymerase chain reaction.

**Figure 2. Gene signature identification process.** Our process could be conceptually broken down into six steps: (1) The samples were divided into an inner training set and an inner test set. (2) Using only the inner training set, the mRMR algorithm returned a subset of N genes that have maximum relevance with the clinical status and minimum redundancy within the gene set. (3) The N-gene-based SVM classification model was built using the linear kernel and fitted to the inner training set. (4) The developed model was used to predict the class of the test sample. The predicted class was compared to the true class label of the sample. If they disagreed, the prediction was in error. (5) The process was repeated leaving each of the 200 samples out of the training set, one at a time. A total of 200 different models were created, and each model was used to predict the class of the test sample. Eventually, the number of prediction errors was summed up and reported as the LOOCV.
estimate of the prediction error. 6) Steps 1-5 were repeated for 52 times starting with $n = 1$, increasing one at a time until $n = 52$. The LOOCV performance for each of the N-gene signatures was estimated and compared to determine the optimal size of the signature.

**Figure 3. Identification of the 18-gene signature.** (A) The classification accuracy during the LOOCV process for each N-gene signature was estimated, and was used to determine the optimal size of the signature. The size determination took into account the classification accuracy and signature complexity. Ultimately, we considered a signature size of 18 to be ideal. The results of LOOCV process showed that an estimated performance of 91.5% accuracy was achievable with a signature composed of the top 18 genes. (B) Univariate fold changes of 18 genes across 100 CRC patients and 100 controls in the training set. Positive fold changes indicate genes that are upregulated in CRC, whereas negative fold changes indicate genes that are downregulated in CRC.

**Figure 4. Validation of the 18-gene signature in the test set.** (A) For each sample, the probability score was calculated. The dashed line at 50% indicates the CRC vs. control decision threshold. A sample was classified as CRC if the probability score was over 50%. For controls, 64 samples were correctly classified, and nine samples were misclassified. For cancer patients, 73 of 87 samples were correctly classified. Of those misclassified cancer patients, four were stage I, four were stage II, five were stage III, and one was stage IV. (B) The area under the receiver operating characteristics curve reached 0.94, and the 95% confidence interval was 0.91-0.98.
Discovery Set (n = 200)

FUSCC and Shanghai Community Hospital

Control (n = 100)
CRC (n = 100)
Stage I 16
Stage II 36
Stage III 24
Stage IV 24

Microarray screening (HG-U133Plus2 Array)

Candidate genes (n = 52)

Training Set (n = 200)

FUSCC and Shanghai Community Hospital

Control (n = 100)
CRC (n = 100)
Stage I 11
Stage II 40
Stage III 22
Stage IV 27

Quantitative RT-PCR

Signature Identification (18-gene signature)

Performance Estimation

Test Set (n = 160)

FUSCC and Shanghai Community Hospital

Control (n = 73)
CRC* (n = 87)
Stage I 11
Stage II 31
Stage III 31
Stage IV 13

Quantitative RT-PCR

Signature Application

Performance Evaluation
Training Set (100 Control_100 CRC)

1. Leave-One-Out Cross-Validation

2. Inner Training Set (n=199)

3. mRMR Feature Selection (Top-N genes selection)

4. SVM classification model

5. Inner Test Set (n=1)

6. Cross-Validation Performance

Signature Size Determination Top-N genes (n=1,2,...,31,32)
Clinical Cancer Research

Identification and Validation of a Blood-Based 18-Gene Expression Signature in Colorectal Cancer

Ye Xu, Qinghua Xu, Li Yang, et al.

Clin Cancer Res Published OnlineFirst March 27, 2013.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-12-3851

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2013/03/28/1078-0432.CCR-12-3851.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.