




ORIGINAL ARTICLE

A qualitative transcriptional signature for the early diagnosis of colorectal cancer

Qingzhou Guan^{1,2} | Qihong Zeng^{1,2} | Haidan Yan^{1,2} | Jiajing Xie^{1,2} | Jun Cheng^{1,2} |
 Lu Ao^{1,2} | Jun He^{1,2} | Wenyuan Zhao³  | Kui Chen⁴ | You Guo^{1,2}  |
 Guoxian Guan⁵ | Zheng Guo^{1,2} 

¹Department of Bioinformatics, School of Basic Medical Sciences, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Fujian Medical University, Fuzhou, China

²Key Laboratory of Medical Bioinformatics, Fuzhou, China

³Department of Systems Biology, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

⁴Department of General Surgery, Affiliated Fuzhou First Hospital of Fujian Medical University, Fuzhou, China

⁵Department of Colorectal Surgery, The Affiliated Union Hospital of Fujian Medical University, Fuzhou, China

Correspondence

Zheng Guo, Department of Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China.
 Email: guoz@ems.hrbmu.edu.cn

Guoxian Guan, Department of Colorectal Surgery, The Affiliated Union Hospital of Fujian Medical University, Fuzhou, China.
 Email: gy_8637@fjmu.edu.cn

Funding information

The Joint Scientific and Technology Innovation Fund of Fujian Province, Grant/Award Number: 2016Y9044; National Natural Science Foundation of China, Grant/Award Number: 61801118, 81572935 and 81872396; Fujian Provincial Finance Department Special Fund, Grant/Award Number: (2015)1297; Startup Fund for Scientific Research, Fujian Medical University, Grant/Award Number: 2017XQ2002, 2017XQ2006 and 2017XQ2007

Abstract

Currently, using biopsy specimens for the early diagnosis of colorectal cancer (CRC) is not entirely reliable due to insufficient sampling amount and inaccurate sampling location. Thus, it is necessary to develop a signature that can accurately identify patients with CRC under these clinical scenarios. Based on the relative expression orderings of genes within individual samples, we developed a qualitative transcriptional signature to discriminate CRC tissues, including CRC adjacent normal tissues from non-CRC individuals. The signature was validated using multiple microarray and RNA sequencing data from different sources. In the training data, a signature consisting of 7 gene pairs was identified. It was well validated in both biopsy and surgical resection specimens from multiple datasets measured by different platforms. For biopsy specimens, 97.6% of 42 CRC tissues and 94.5% of 163 non-CRC (normal or inflammatory bowel disease) tissues were correctly classified. For surgically resected specimens, 99.5% of 854 CRC tissues and 96.3% of 81 CRC adjacent normal tissues were correctly identified as CRC. Notably, we additionally measured 33 CRC biopsy specimens by the Affymetrix platform and 13 CRC surgical resection specimens, with different proportions of tumor epithelial cells, ranging from 40% to 100%, by the RNA sequencing platform, and all these samples were correctly identified as CRC. The signature can be used for the early diagnosis of CRC, which is also suitable for

Abbreviations: AUC, area under the receiver operating characteristic curve; CRC, colorectal cancer; GEO, Gene Expression Omnibus; IBD, inflammatory bowel diseases; k-TSP, k-Top Scoring Pairs; REO, relative expression ordering; RNA-seq, RNA sequencing; TCGA, The Cancer Genome Atlas; TSP50, testes-specific protease 50.

Guan, Zeng, and Yan contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2019 The Authors. *Cancer Science* published by John Wiley & Sons Australia, Ltd on behalf of Japanese Cancer Association.

minimum biopsy specimens and inaccurately sampled specimens, and thus has potential value for clinical application.

KEYWORDS

biopsy, colorectal cancer, early diagnosis, relative expression orderings, signature

1 | INTRODUCTION

Colorectal cancer (CRC) is the third most commonly diagnosed malignancy and the fourth leading cause of cancer-related deaths in the world.¹ Patients with CRC are easily curable when diagnosed at an early stage,² thus the early diagnosis of CRC is crucial for the fight against this cancer. However, most CRC patients are diagnosed with middle or advanced stage disease.³ Currently, established noninvasive tests, such as the guaiac-based fecal occult blood test, have a low sensitivity⁴ and positive predictive value.^{5,6} Several serum protein biomarkers, including carcinoembryonic antigen, CA19.9, and CA125, can be used for monitoring the prognosis of CRC patients but none of them are recommended for the early diagnosis of CRC.^{7,8} The expression of TSP50 has also been proposed as a diagnostic signature for CRC,⁹ but its sensitivity, specificity, and positive predictive value were 68.4%, 92.5%, and 95.6%, respectively. This signature is based on a risk score summarized from quantitative expression measurements of TSP50 protein, which lacks robustness for clinical applications due to large measurement batch effects.¹⁰

In clinical practice, biopsy sampling with less invasive techniques, such as colonoscopy, are often used for the initial clinical evaluation of CRC.¹¹⁻¹⁵ However, an indeterminate diagnosis often creates a dilemma.¹⁶ It has been reported that the miss rate of CRC after colonoscopy, which is the predominant screening and diagnostic test for CRC,^{12,17} is approximately 15% for patients with IBD.¹² Moreover, the biopsy location can be inaccurate, which might lead to inaccurately sampled adjacent nontumor tissues and degrading the diagnosis performance.¹⁸ However, previously reported diagnostic signatures, such like the transcriptional signatures reported by Zheng et al⁹ and our previous study,¹⁰ all took tumor-adjacent normal tissues as the normal samples to obtain the signature. Thus, these signatures cannot classify inaccurately sampled CRC adjacent normal tissues to CRC. Given that the adjacent nontumor colorectal tissues of CRC patients might have some molecular characteristics of CRC,¹⁹⁻²¹ it is possible to develop a signature to discriminate CRC (including CRC adjacent tissues) from tissues of nontumor (normal or IBD) individuals, which is suitable for minimum biopsy specimens and inaccurately sampled specimens.

Another major limitation of the previously reported transcriptional diagnostic signatures is that their applications are based on risk scores summarized from the quantitative expression measurements of the signature genes,²²⁻²⁴ which are sensitive to batch effects and hardly applicable to individualized diagnoses.^{10,25-27} Notably, several reported quantitative transcriptional disease signatures, including AlloMap,²⁴ have been approved by the US FDA.

However, due to the existence of batch effects, the tissue samples must be sent to specific laboratories for measurement with strict quality control.

In contrast, the REOs of genes within individual samples, which are the qualitative transcriptional characteristics, are robust against experimental batch effects and can be directly applied to samples at the individualized level.²⁸⁻³¹ The robustness property of the REO enables researchers to integrate multiple datasets produced by the same or similar platforms for developing disease signatures or classifiers,^{32,33} which makes it more likely to find robust signatures.^{10,32,34} In addition, the qualitative transcriptional characteristics are highly robust against varied proportions of the tumor epithelial cell in specimens sampled from different tumor locations of the same patients,²⁶ partial RNA degradation during specimen preparation and storage,²⁵ and amplification bias for minimum specimens,²⁷ which are the common factors that lead to the failure of quantitative transcriptional signatures in clinical practice. Therefore, it is worth exploiting the within-sample REOs to identify a robust qualitative signature for the early diagnosis of CRC.

In this study, based on the robust within-sample REOs, we identified a qualitative transcriptional signature consisting of 7 gene pairs for the early diagnosis of CRC. The signature can accurately discriminate CRC tissues, including CRC adjacent normal tissues, from normal or IBD tissues of non-CRC individuals in both biopsy and surgical resection samples.

2 | MATERIALS AND METHODS

2.1 | Samples and data measurement

The gene expression profiles of 33 CRC biopsy specimens were measured by Affymetrix platform in our laboratory³⁵ and this study (NCT02770911) was approved by the Institutional Review Board at Fujian Medical University Union Hospital (No. 2015-23; Fuzhou, China). Written informed consents for all the 33 participants were obtained. The tumor biopsy specimens were obtained by endoscopy. RNA was extracted using the RNeasy Mini Kit (Qiagen), and was measured by Affymetrix GeneChip PrimeView Array. For the raw data (.CEL file) from the array platform, the Robust Multi-Array Average algorithm³⁶ was applied for background adjustment without quantile normalization.

We also measured 13 CRC surgical resection specimens, from 5 CRC patients, with the RNA-seq platform. This study was approved by the institutional review boards of all participating institutions, and written consent forms were obtained from all participants. For

each patient, 3 specimens were sampled from 3 different locations. Of these, 2 specimens were excluded from the subsequent analysis due to poor RNA quality (RNA integrity number less than 6.0). The proportion of tumor epithelial cells for each of the 13 tumor specimens, ranging from 40% to 100% (see Table 1), was measured by pathological section analysis. After surgical resection, the obtained cancer specimens were fresh-frozen for the subsequent RNA extraction. According to the manufacturer's protocol, total RNA was isolated from fresh-frozen CRC tissues using TRIzol reagent (Invitrogen) and the quality of RNA was assessed by Agilent 2200 TapeStation (Agilent Technologies). Then mRNA was captured from 1-2 μg total RNA using NEBNext PolyA mRNA Magnetic Isolation Module and stranded RNA-seq libraries were constructed using a NEBNext Ultra Directional RNA Library Prep Kit. Paired-end sequencing (2×150) was undertaken using an Illumina HiSeqXten and generated raw RNA-seq files (fastq) were preprocessed using Trimmomatic,³⁷ and the reference genome (GRCh37) was used to align reads using hisat2.³⁸ Finally, the fragments per kilobase of transcript per million fragments mapped values of genes were calculated using StringTie.³⁹

2.2 | Public data and preprocessing

Multiple gene expression profiles were downloaded from the GEO repository (<http://www.ncbi.nlm.nih.gov/geo/>), ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>), and TCGA (<http://cancergenome.nih.gov/>), as described in Table 2, including CRC samples, CRC adjacent normal samples, IBD samples, and normal samples. Notably, the cancer samples include CRC samples and CRC adjacent normal samples, and the noncancer samples include IBD samples and normal samples in this study. The normal samples have been proven to have no polyps and no known family history or previous CRC incidence.

For the data measured by the Affymetrix platform, we downloaded the raw mRNA expression data (.CEL files) and used the Robust Multi-array Average algorithm for background adjustment without quantile normalization. For the sequence-based data, the fragments per kilobase of transcript per million fragments mapped or reads per kilobase of transcript per million reads mapped value was downloaded.

For the array-based data, if multiple probes were mapped to a gene, the expression value of the gene was defined as the arithmetic

TABLE 1 Proportions of tumor epithelial cells in colorectal cancer (CRC) tissues

Patient	Proportion 1	Proportion 2	Proportion 3
CRC 1	70%	-	40%
CRC 2	40%	100%	100%
CRC 3	50%	90%	90%
CRC 4	60%	100%	100%
CRC 5	100%	100%	-

-, No sample in the corresponding category due to poor RNA quality.

mean of the values of the multiple probes. If a probe was mapped to zero or multiple genes, then the data of this probe were deleted. For the sequence-based data from ArrayExpress, the gene symbols were mapped to Entrez gene ID with the biological database network.⁴⁰ For the sequence-based data from TCGA, the Ensembl gene IDs corresponding to the unique Entrez gene IDs of protein coding genes were used.

2.3 | Identification of REO-based CRC diagnosis signature

First, within a sample, the REO of two genes, i and j , is denoted as $G_i > G_j$ (or $G_i < G_j$) if the expression level of gene i is higher (or lower) than that of gene j . If the same REO pattern is maintained in a majority of samples, eg 85%, it is called a stable REO and the pair is a stable gene pair. A gene pair with stable REOs in both groups of samples, but the REO patterns are opposite, is called a reversal gene pair. Here, we selected the reversal gene pairs that are stable in noncancer samples and cancer samples, but the REO patterns are reversed in the latter group. They form the candidate REO signature of the cancer.

Then the selected candidate REO signature above were sorted in a descending order according to their reversal degree, where the reversal degree for each reversal gene pair was calculated as follow:

$$\text{avg}R_{ij} = \sqrt{|\text{mean}[R_{ij}(\text{cancer})]| \times |\text{mean}[R_{ij}(\text{non_cancer})]|}$$

where $|\text{mean}[R_{ij}(\text{cancer})]|$ and $|\text{mean}[R_{ij}(\text{non_cancer})]|$ represent the absolute of the means of rank differences of the reversal gene pair (i, j) in cancer samples and noncancer samples, respectively. The rank difference for each reversal gene pair was calculated as follows:

$$R_{ij} = R_i - R_j$$

where R_i and R_j represent the rank of gene i and gene j in a sample, respectively, and R_{ij} is the rank difference between the 2 genes. Obviously, the higher the reversal degree for a gene pair, the higher the cross-platform performance is for this gene pair.

Finally, we used the top- k gene pairs, where k is ranging from 1 to the total number of the reversal gene pairs, to classify the samples based on the majority vote rule. The value of k was chosen when its value reached the highest geometric mean of the sensitivity and specificity in the training data. The top- k gene pairs were selected as the early diagnosis signature of CRC.

2.4 | Performance evaluation

Cancer samples, including cancer and cancer adjacent normal samples, were classified as positive samples; noncancer samples, including normal and IBD samples, were classified as negative samples.

TABLE 2 Description of datasets used in this study

	Platform	Sampling method	Sample size			
			Normal	IBD	Adjacent normal	Cancer
Datasets used for identification of the qualitative signature						
GSE4183	AffymetrixGPL570	Biopsy	8	15	-	15
GSE9348	AffymetrixGPL570	Biopsy	12	-	-	70
GSE35452	AffymetrixGPL570	Biopsy	-	-	-	46
GSE22619	AffymetrixGPL570	Biopsy	10	10	-	-
GSE14580	AffymetrixGPL570	Biopsy	-	24	-	-
GSE13367	AffymetrixGPL570	Biopsy	-	16	-	-
GSE18105	AffymetrixGPL570	Surgery	-	-	17	77
GSE23878	AffymetrixGPL570	Surgery	-	-	24	35
GSE33113	AffymetrixGPL570	Surgery	-	-	6	90
GSE32323	AffymetrixGPL570	Surgery	-	-	17	17
GSE41328	AffymetrixGPL570	Surgery	-	-	10	10
GSE17536	AffymetrixGPL570	Surgery	-	-	-	177
GSE35144	AffymetrixGPL570	Surgery	-	-	-	27
E-GEOD-72819	Illumina GPL11154	Biopsy	-	73	-	-
E-GEOD-50760	Illumina GPL11154	Surgery	-	-	18	36
Datasets used for evaluating the performance of the qualitative signature						
GSE47908	AffymetrixGPL570	Biopsy	15	39	-	-
GSE36807	AffymetrixGPL570	Biopsy	7	28	-	-
GSE16879	AffymetrixGPL570	Biopsy	-	43	-	-
GSE12251	AffymetrixGPL570	Biopsy	-	23	-	-
GSE9452	AffymetrixGPL570	Biopsy	-	8	-	-
GSE45404	AffymetrixGPL570	Biopsy	-	-	-	42
GSE21510	AffymetrixGPL570	Surgery	-	-	25	104
GSE22598	AffymetrixGPL570	Surgery	-	-	17	17
GSE27854	AffymetrixGPL570	Surgery	-	-	-	115
GSE35896	AffymetrixGPL570	Surgery	-	-	-	62
Our_Data1	Affymetrix PrimeView Array	Biopsy	-	-	-	33
Our_Data2	Illumina HiSeqXten	Surgery	-	-	-	13
TCGA	Illumina HiSeq_RNASeqV2	Surgery	-	-	39	556

-, No sample in the corresponding category; IBD, inflammatory bowel disease; TCGA, The Cancer Genome Atlas.

The performance of the signature was evaluated using sensitivity and specificity, which are calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

where TP, TN, FP, and FN represent the number of true-positive, true-negative, false-positive, and false-negative samples, respectively.

The AUCs were calculated with the nonparametric Hanley-McNeil algorithm¹⁸ and 95% confidence intervals for AUCs were determined using an approximate normal distribution.

3 | RESULTS

3.1 | Identification of the qualitative diagnostic signature

The analysis procedure of this study is described in Figure 1. First, using 30 normal samples and 65 IBD samples collected from 5 datasets measured by the Affymetrix platform (see Table 2), 11 558 060 gene pairs with identical REO patterns in at least 85% of both the normal and IBD samples were identified as stable gene pairs of noncancer samples. Similarly, using 564 CRC samples and 74 CRC adjacent normal samples collected from 10 datasets measured by the Affymetrix platform, 106 958 978 gene pairs with identical REO patterns in at least 85% of both the CRC and CRC

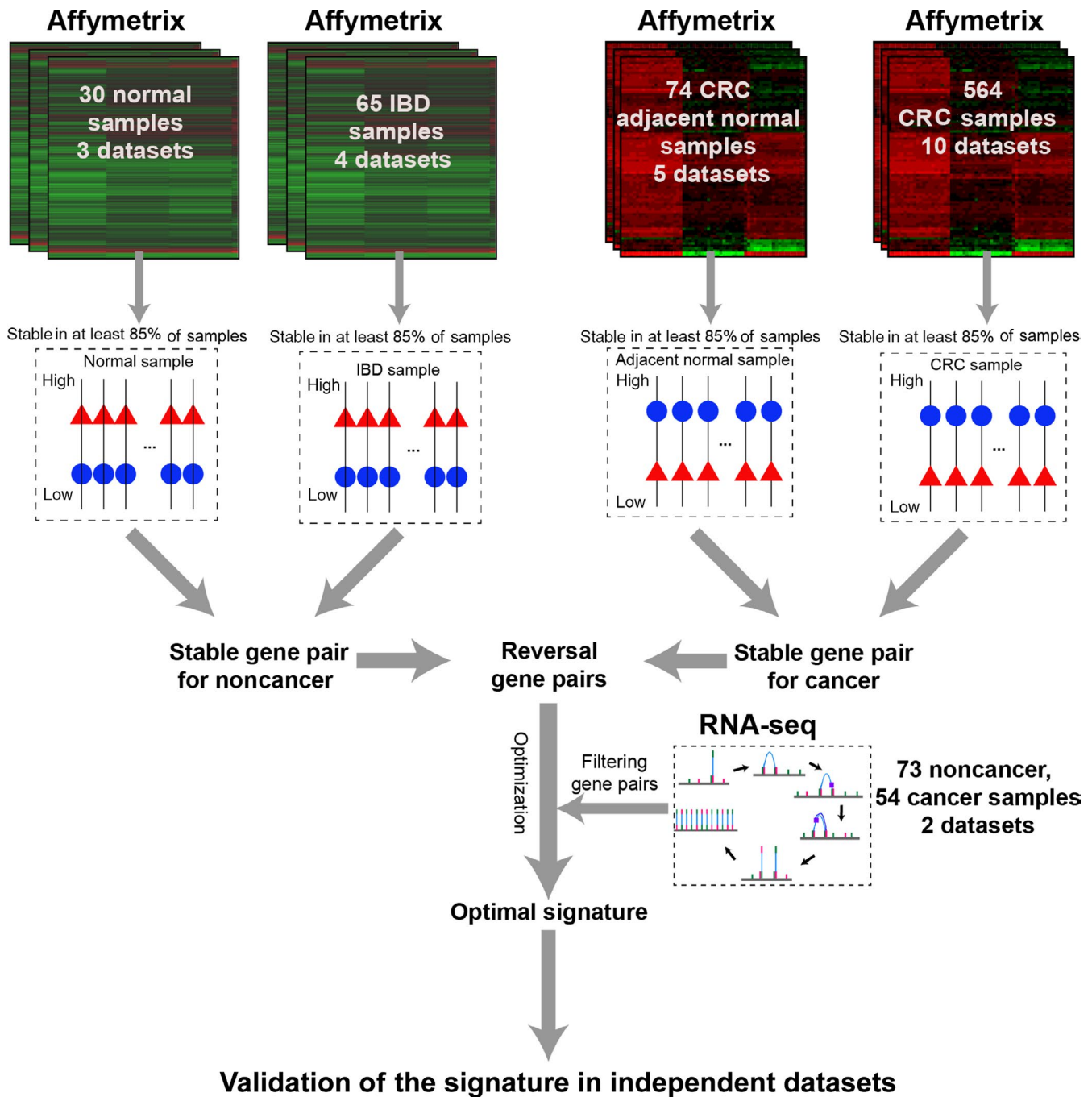


FIGURE 1 Analysis procedure for identifying the colorectal cancer (CRC) diagnosis signature. IBD, inflammatory bowel disease; RNA-seq, RNA sequencing

adjacent normal samples were identified as stable gene pairs of cancer samples. We found 218 reversal gene pairs between the non-CRC and CRC tissues including the adjacent normal tissues from the above 2 lists of gene pairs identified from the data measured by the Affymetrix platform. Among these 218 gene pairs, we further selected 7 gene pairs that had the identical REO pattern in at least 85% of 73 noncancer samples and reversal REO patterns in at least 85% of 54 cancer samples in the combined data from the E-GEOD-50760 and E-GEOD-72819 datasets measured by the RNA-seq platform.

Then, the 7 gene pairs were sorted in a descending order according to their reversal degrees (see Materials and Methods 2.3) between CRC (including CRC and CRC adjacent normal) and non-CRC samples (normal and IBD) in the combined data from the training set, as shown in Table 2. We then used the top-ranked *k* pairs to classify samples according to the majority vote rule. The results showed that, for all possible *k* ranging from 1 to 7, the largest geometric mean of the sensitivity and specificity was 97.08% when *k* = 7 (Figure 2). Thus, these 7 gene pairs, as described in Table 3, were selected as the signature for discriminating CRC samples from noncancer

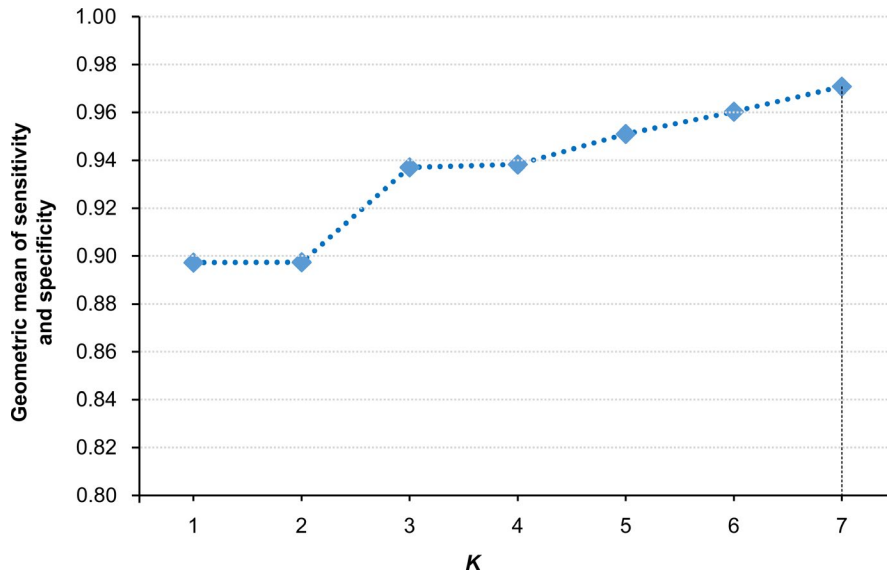


FIGURE 2 Performance of k gene pairs of relative expression ordering-based signatures in the training set of biopsy and surgically resected colorectal cancer and noncancer samples

samples. We additionally showed the expression pattern of the 7 gene pairs (consisting of 13 genes) in the training datasets measured by Affymetrix platform. As shown in Figures S1 and S2, the results showed that, for each gene pair, the REO is stable in both types of samples, but the REO patterns are opposite.

3.2 | Validation of the diagnostic signature in independent datasets

We then validated the performance of the 7 gene pairs in multiple public datasets for biopsy and surgically resected samples. For a total of 977 cancer samples and 163 noncancer samples from these public databases, the geometric mean of the sensitivity and specificity was 96.80% and the AUC was 0.9589 (95% confidence interval, 0.9521-0.9657; Figure 3).

Notably, all the colorectal normal and IBD tissue samples from non-CRC individuals and 42 CRC tissue samples from GSE45404 were obtained by endoscopic biopsy. For these biopsy samples measured by the Affymetrix platform, 90.9% of the 22 normal samples from healthy individuals and 95.0% of the 141 IBD samples of non-CRC patients were correctly identified as non-CRC, while 97.6% of the 42 cancer

samples were correctly identified as CRC. The detailed results of each dataset are shown in Table 4. These results indicated that our signature is suitable for the early diagnosis of CRC based on biopsy specimens.

For surgically resected samples measured by the Affymetrix platform, all of the 298 CRC samples and 42 CRC adjacent normal samples were correctly identified as CRC. For the data measured by the RNA-seq platform, 99.3% of the 556 CRC samples and 92.3% of the 39 CRC adjacent normal samples were correctly identified as CRC. The detailed results of each dataset are shown in Table 5. These results suggested that the 7 gene pairs could identify most of the adjacent nontumor colorectal tissues from CRC patients as CRC, which is suitable for inaccurately sampled specimens.

Among the 556 CRC samples from TCGA, 536 samples included staging information. 99.0% of 96 patients with stage I, 99.0% of 209 patients with stage II, 100.0% of 156 patients with stage III, and 98.7% of 75 patients with stage IV were correctly identified as CRC. The clinical stage status did not affect the validation results using the GEO dataset either. All of the 104 samples from dataset GSE21510, including 13 patients with stage I, 37 patients with stage II, 34 patients with stage III, and 20 patients with stage IV, were correctly identified as CRC. Moreover, all of the 62 CRC samples from the dataset GSE35896 had their gene mutation status information (*KRAS*, *BRAF*, *APC*, *TP53*, *PIK3CA*, and *PTEN*), but all of them were correctly identified as CRC regardless of the mutation status of any gene. Among the dataset GSE35896, 61 of the 62 CRC samples had microsatellite instability information. All of the 56 patients with stable microsatellite status and 5 patients with unstable microsatellite status were correctly identified as CRC, regardless of the microsatellite status. The results further indicated that our signature is robust against clinicopathological variations.

TABLE 3 Seven gene-pair signatures for early diagnosis of colorectal cancer (CRC)

Signature	Gene i	Gene j
Pair 1	<i>AREG</i>	<i>TRIM40</i>
Pair 2	<i>SCARNA2</i>	<i>CHRNE</i>
Pair 3	<i>SCARNA2</i>	<i>CASKIN1</i>
Pair 4	<i>ARHGAP10</i>	<i>KIAA0125</i>
Pair 5	<i>KCNH2</i>	<i>ZNF671</i>
Pair 6	<i>CLCN5</i>	<i>C19orf44</i>
Pair 7	<i>SSBP1</i>	<i>DHRS7</i>

Gene i has a higher expression level than gene j in CRC tissue samples compared with non-CRC tissue samples.

3.3 | Validation of the diagnostic signature in our data

To further validate the signature, using the RNA-seq platform, we additionally measured gene expression profiles of 13 CRC surgical

FIGURE 3 Area under the receiver operating characteristic curve (AUC) of the validation data from public databases of biopsy and surgically resected colorectal cancer and noncancer samples

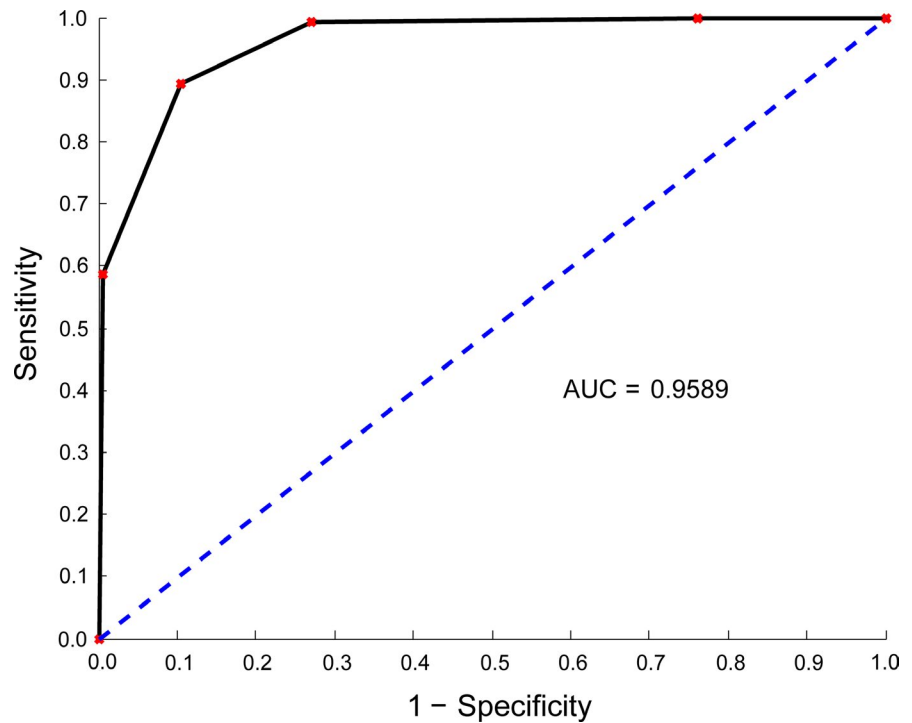


TABLE 4 Performance of the gene signature in the validation datasets for colorectal biopsy samples

	Normal	IBD	Adjacent_normal	Cancer	Specificity	Sensitivity
GSE36807	7	28	-	-	85.71%	-
GSE12251	-	23	-	-	100.00%	-
GSE9452	-	8	-	-	100.00%	-
GSE47908	15	39	-	-	92.59%	-
GSE16879	-	43	-	-	100.00%	-
GSE45404	-	-	-	42	-	97.62%

-, No information in the corresponding category; IBD, inflammatory bowel disease.

TABLE 5 Performance of the gene signature in the validation datasets for surgically resected colorectal samples

	Normal	IBD	Adjacent_normal	Cancer	Specificity	Sensitivity
GSE21510	-	-	25	104	-	100.00%
GSE22598	-	-	17	17	-	100.00%
GSE27854	-	-	-	115	-	100.00%
GSE35896	-	-	-	62	-	100.00%
TCGA	-	-	39	556	-	98.82%

-, No information in the corresponding category; IBD, inflammatory bowel disease; TCGA, The Cancer Genome Atlas.

resection specimens from 5 CRC patients, each with 3 specimens sampled from 3 tumor locations with different proportions of tumor epithelial cells (see Table 1). Two specimens were excluded from the gene expression measurements because of poor RNA quality. All the 13 CRC specimens were correctly identified as CRC by our signature, even when the proportion of tumor epithelial cells was as low as 40%, which further verified that the REO-based signature is robust against varied proportions of tumor epithelial cells for the

same patient with different tumor locations.²⁶ Moreover, for the 33 CRC biopsy specimens measured by the Affymetrix platform in our laboratory,³⁵ all of them were correctly identified as CRC based on our signature.

In summary, the above results together revealed that the signature can accurately discriminate CRC from non-CRC individuals using both surgical resection and biopsy samples measured by different platforms. In particular, the signature is robust against varied

proportions of tumor epithelial cells in specimens sampled from different tumor locations of the same patients.

4 | DISCUSSION

In this study, we identified a robust qualitative signature of 7 gene pairs, consisting of 13 genes, for the early diagnosis of CRC, which can discriminate CRC and CRC adjacent tissues from IBD and normal tissue of non-CRC individuals. It means that, even when the specimens are sampled inaccurately, this signature can still aid the early diagnosis of CRC. The REO-based qualitative transcriptional signature is robust against experimental batch effects and invariant to monotone data transformation, and it can be directly applied to samples at the individualized level.²⁸⁻³¹ For a total of 1023 cancer sample and 163 noncancer samples from the validation datasets, the sensitivity, specificity, and positive predictive value of our signature was 99.22%, 94.48%, and 99.12%, which indicate the robustness of our signature. As shown in Table 4, among the 5 validation datasets with noncancer samples, our signature has 100% specificity in 3 datasets, GSE12251, GSE9452, and GSE16879. For the other 2 datasets, GSE47908 and GSE36807, our signature has 92.59% and 85.71% specificity, respectively. For the dataset GSE47908 with 54 noncancer samples (including 15 normal and 39 IBD samples [including 19 pancolitis and 20 left-sided colitis samples]), all the 15 normal samples and 20 left-sided colitis samples were correctly identified as noncancer, whereas 4 of the 19 pancolitis samples were identified as cancer. Because patients with pancolitis have a higher cancer incidence risk than those with left-sided colitis,⁴¹ we speculate that these 4 pancolitis samples might have some characteristics of cancer. Similarly, for the GSE36807 database with 35 noncancer samples (including 7 normal and 28 IBD samples), 2 normal and 3 IBD (1 Crohn's disease and 2 ulcerative colitis) samples were identified as cancer. Those healthy individuals with normal samples, including that were identified as cancer, were referred for colorectal cancer screening⁴²; we speculated that some of them might also have some characteristics of cancer.

Under many practical situations, with tissue biopsy sampling, it is difficult to obtain sufficient a quantity of RNA molecules for gene expression profiling or other molecular measurements.¹⁸ Fortunately, our recent study showed that the REO-based signatures can be robustly applied to minimum specimens even with approximately 15 cancer cells.²⁷ Therefore, it is highly possible that the 7 gene pairs could be applicable for biopsy samples with minimum sampling amounts. Moreover, the REO-based signature was robust against varied proportions of tumor epithelial cells from the same patient with different tumor locations,²⁶ which is a common factor that could lead to the failure of the quantitative transcriptional signature in clinical practice. This study also showed that 13 specimens from 5 patients with different sampling locations, with different proportions of tumor epithelial cells (see Table 1), were correctly identified as CRC.

As for the other REO-based approaches, such as TSP and k-TSP, we additionally evaluated these approaches using the same

training and validation datasets, as shown in Table 2. Using the *tspair* R package (version 3.3.3) and *ktspair* R package (version 3.3.3), we trained the TSP and k-TSP classifier in the combined sample data from the training datasets measured by the Affymetrix and RNA-seq platforms, respectively. In the validation data, the k-TSP classifier performed better than the TSP classifier but poorer than our signature, as shown in Tables S1 and S2. For example, for the 33 CRC samples measured by our laboratory, our REO signature could identify 100% of the 33 CRC samples correctly, but the k-TSP signature identified only 30.3% CRC samples correctly. One possible reason could be that the difference in the proportion of samples from the Affymetrix and RNA-seq platforms will bias the signature to the platform with larger samples when using the *tspair* R package and *ktspair* R package. In the training process for our REO signature, the gene pairs that were consistently detected in the data produced by the 2 platforms were used for the final signature selection (7 gene pairs in this study). Therefore, our method is intuitive and simple with the ability to identify very robust disease signatures.

Some genes in our signature, including *AREG*, *SSBP1*, *KCNH2*, and *TRIM40*, are well known CRC-related genes that might play a key role in the development of CRC. For example, *AREG* could induce the upregulation of *EGFR*, which is a key mediator of intestinal neoplastic transformation, and high gene expression level of *AREG* is a favorable prognostics biomarker for metastatic CRC.⁴³ Another gene, *SSBP1*, has highly abundant gene expression levels in CRC and is closely related with poor outcomes of CRC patients.⁴⁴ In cisplatin-resistant CRC cells, *KCNH2* inhibitors had a synergistic action with cisplatin in triggering apoptosis and inhibiting proliferation.⁴⁵ Additionally, *TRIM40* might provide therapeutic benefits, not only for inhibition of the growth of gastrointestinal cancers but also for the prevention of chronic IBDs.⁴⁶ In addition, *ARHGAP10*,⁴⁷ *DHR57*,⁴⁸ and *ZNF671*⁴⁹ have also been reported to be closely correlated with other types of cancer, such as lung and prostate cancer. The above results indicated that the genes of the signature might play important roles in the carcinogenesis of CRC and these functions need to be further studied in future work.

In summary, our signature, consisting of 7 gene pairs, could robustly be applied for aiding the early diagnosis of CRC in multiple datasets of both biopsy and surgically resected samples, which is also suitable for minimum biopsy specimens and inaccurately sampled specimens. The clinical value of the 7 gene pairs for early diagnosis of CRC is worthy of further verification. Moreover, as the cost of high-throughput sequencing decreases markedly, for a limited amount of precious tissue sample at the clinical scene, we could measure all the genes or a set of genes of different biomarkers for diagnosis, histological classification, prognosis, and drug resistance evaluation of CRC ("a sequencing for all").

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant nos. 81872396, 61801118, and

81572935), the Startup Fund for Scientific Research, Fujian Medical University (grant nos. 2017XQ2002, 2017XQ2007, and 2017XQ2006), the Doctoral Research Foundation of the First Affiliated Hospital of Gannan Medical University, the Joint Scientific and Technology Innovation Fund of Fujian Province (grant no. 2016Y9044), and the Fujian Provincial Finance Department Special Fund (No. (2015)1297).

CONFLICT OF INTEREST

The authors have no conflict of interest.

ORCID

Wenyuan Zhao  <https://orcid.org/0000-0002-6477-9434>

You Guo  <https://orcid.org/0000-0002-2751-3899>

Zheng Guo  <https://orcid.org/0000-0003-4466-6026>

REFERENCES

- Arnold M, Sierra MS, Laversanne M, et al. Global patterns and trends in colorectal cancer incidence and mortality. *Gut*. 2017;66:683-691.
- Haggard FA, Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin Colon Rectal Surg*. 2009;22:191-197.
- Liu J, Zhou Q, Xu J, et al. Detection of EGFR expression in patients with colorectal cancer and the therapeutic effect of cetuximab. *J BUON*. 2016;21:95-100.
- Adelstein BA, Macaskill P, Chan SF, et al. Most bowel cancer symptoms do not indicate colorectal cancer and polyps: a systematic review. *BMC Gastroenterol*. 2011;11:65.
- Brown JP, Wooldrage K, Wright S, et al. High test positivity and low positive predictive value for colorectal cancer of continued faecal occult blood test screening after negative colonoscopy. *J Med Screen*. 2018;25:70-75.
- Azimaousse Assogba GF, Jezewski-Serra D, Lastier D, et al. Impact of subsequent screening episodes on the positive predictive value for advanced neoplasia and on the distribution of anatomic subsites of colorectal cancer: A population-based study on behalf of the French colorectal cancer screening program. *Cancer Epidemiol*. 2015;39:964-971.
- Soler M, Estevez MC, Villar-Vazquez R, et al. Label-free nanoplasmonic sensing of tumor-associated autoantibodies for early diagnosis of colorectal cancer. *Anal Chim Acta*. 2016;930:31-38.
- Duffy MJ, van Dalen A, Haglund C, et al. Tumour markers in colorectal cancer: European Group on Tumour Markers (EGTM) guidelines for clinical use. *Eur J Cancer*. 2007;43:1348-1360.
- Zheng L, Xie G, Duan G, et al. High expression of testes-specific protease 50 is associated with poor prognosis in colorectal carcinoma. *PLoS ONE*. 2011;6:e22203.
- Guan Q, Yan H, Chen Y, et al. Quantitative or qualitative transcriptional diagnostic signatures? A case study for colorectal cancer. *BMC Genom*. 2018;19:99.
- Wolf AMD, Fontham ETH, Church TR, et al. Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA Cancer J Clin*. 2018;68:250-281.
- Wang YR, Cangemi JR, Loftus EV Jr, et al. Rate of early/missed colorectal cancers after colonoscopy in older patients with or without inflammatory bowel disease in the United States. *Am J Gastroenterol*. 2013;108:444-449.
- Fusco V, Ebert B, Weber-Eibel J, et al. Cancer prevention in ulcerative colitis: long-term outcome following fluorescence-guided colonoscopy. *Inflamm Bowel Dis*. 2012;18:489-495.
- von Karsa L, Patnick J, Segnan N, et al. European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. *Endoscopy*. 2013;45:51-59.
- Kaminski MF, Polkowski M, Kraszewska E, et al. A score to estimate the likelihood of detecting advanced colorectal neoplasia at colonoscopy. *Gut*. 2014;63:1112-1119.
- Ahmed A, VandenBussche CJ, Ali SZ, et al. The dilemma of "indeterminate" interpretations of pancreatic neuroendocrine tumors on fine needle aspiration. *Diagn Cytopathol*. 2016;44:10-13.
- Rex DK, Johnson DA, Anderson JC, et al. American College of Gastroenterology guidelines for colorectal cancer screening 2009. *Am J Gastroenterol*. 2009;104:739-750.
- Ao L, Zhang Z, Guan Q, et al. A qualitative signature for early diagnosis of hepatocellular carcinoma based on relative expression orderings. *Liver Int*. 2018;38:1812-1819.
- Guo H, Zeng W, Feng L, et al. Integrated transcriptomic analysis of distance-related field cancerization in rectal cancer patients. *Oncotarget*. 2017;8:61107-61117.
- Park SK, Song CS, Yang HJ, et al. Field cancerization in sporadic colon cancer. *Gut Liv*. 2016;10:773-780.
- Cherkezyan L, Stypula-Cyrus Y, Subramanian H, et al. Nanoscale changes in chromatin organization represent the initial steps of tumorigenesis: a transmission electron microscopy study. *BMC Cancer*. 2014;14:189.
- Yang Z, Zhuan B, Yan Y, et al. Identification of gene markers in the development of smoking-induced lung cancer. *Gene*. 2016;576:451-457.
- Panebianco F, Mazzanti C, Tomei S, et al. The combination of four molecular markers improves thyroid cancer cytologic diagnosis and patient management. *BMC Cancer*. 2015;15:918.
- Pham MX, Teuteberg JJ, Kfoury AG, et al. Gene-expression profiling for rejection surveillance after cardiac transplantation. *N Engl J Med*. 2010;362:1890-1900.
- Chen R, Guan Q, Cheng J, et al. Robust transcriptional tumor signatures applicable to both formalin-fixed paraffin-embedded and fresh-frozen samples. *Oncotarget*. 2017;8:6652-6662.
- Cheng J, Guo Y, Gao Q, et al. Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites. *Oncotarget*. 2017;8:30265-30275.
- Liu H, Li Y, He J, et al. Robust transcriptional signatures for low-input RNA samples based on relative expression orderings. *BMC Genom*. 2017;18:913.
- Eddy JA, Sung J, Geman D, et al. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technol Cancer Res Treat*. 2010;9:149-159.
- Wang H, Sun Q, Zhao W, et al. Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics*. 2015;31:62-68.
- Guan Q, Chen R, Yan H, et al. Differential expression analysis for individual cancer samples based on robust within-sample relative gene expression orderings across multiple profiling platforms. *Oncotarget*. 2016;7:68909-68920.
- Ao L, Song X, Li X, et al. An individualized prognostic signature and multiomics distinction for early stage hepatocellular carcinoma patients with surgical resection. *Oncotarget*. 2016;7:24097-24110.
- Xu L, Tan AC, Winslow RL, et al. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics*. 2008;9:125.
- Xu L, Tan AC, Naiman DQ, et al. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*. 2005;21:3905-3911.

34. Qi L, Chen L, Li Y, et al. Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer. *Brief Bioinform.* 2016;17:233-242.
35. Guo Y, Jiang W, Ao L, et al. A qualitative signature for predicting pathological response to neoadjuvant chemoradiation in locally advanced rectal cancers. *Radiother Oncol.* 2018;129:149-153.
36. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003;4:249-264.
37. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30:2114-2120.
38. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 2015;12:357-360.
39. Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* 2015;33:290-295.
40. Mudunuri U, Che A, Yi M, et al. bioDBnet: the biological database network. *Bioinformatics.* 2009;25:555-556.
41. Fujita M, Matsubara N, Matsuda I, et al. Genomic landscape of colitis-associated cancer indicates the impact of chronic inflammation and its stratification by mutations in the Wnt signaling. *Oncotarget.* 2018;9:969-981.
42. Montero-Melendez T, Llor X, Garcia-Planella E, et al. Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling. *PLoS ONE.* 2013;8:e76235.
43. Jing C, Jin YH, You Z, et al. Prognostic value of amphiregulin and epiregulin mRNA expression in metastatic colorectal cancer patients. *Oncotarget.* 2016;7:55890-55899.
44. Li Q, Qu F, Li R, et al. A functional polymorphism of SSBP1 gene predicts prognosis and response to chemotherapy in resected gastric cancer patients. *Oncotarget.* 2017;8:110861-110876.
45. Pillozzi S, D'Amico M, Bartoli G, et al. The combined activation of KCa3.1 and inhibition of Kv11.1/hERG1 currents contribute to overcome Cisplatin resistance in colorectal cancer cells. *Br J Cancer.* 2018;118:200-212.
46. Noguchi K, Okumura F, Takahashi N, et al. TRIM40 promotes neddylation of IKKgamma and is downregulated in gastrointestinal cancers. *Carcinogenesis.* 2011;32:995-1004.
47. Teng JP, Yang ZY, Zhu YM, et al. The roles of ARHGAP10 in the proliferation, migration and invasion of lung cancer cells. *Oncol Lett.* 2017;14:4613-4618.
48. Seibert JK, Quagliata L, Quintavalle C, et al. A role for the dehydrogenase DHR57 (SDR34C1) in prostate cancer. *Cancer Med.* 2015;4:1717-1729.
49. Yeh CM, Chen PC, Hsieh HY, et al. Methyloomics analysis identifies ZNF671 as an epigenetically repressed novel tumor suppressor and a potential non-invasive biomarker for the detection of urothelial carcinoma. *Oncotarget.* 2015;6:29555-29572.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Guan Q, Zeng Q, Yan H, et al. A qualitative transcriptional signature for the early diagnosis of colorectal cancer. *Cancer Sci.* 2019;110:3225-3234. <https://doi.org/10.1111/cas.14137>