Received: 13 August 2020 | Revised: 19 February 2021 | Accepted: 1 March 2021

DOI: 10.1111/jcmm.16467

ORIGINAL ARTICLE

WILEY

Two novel qualitative transcriptional signatures robustly applicable to non-research-oriented colorectal cancer samples with low-quality RNA

Jun Cheng[1,2,3] | Yating Guo[2,3] | Guoxian Guan[4] | Haiyan Huang[2,3] | Fengle Jiang[2,3] | Jun He[2,3] | Junling Wu[2,3] | Zheng Guo[2,3] | Xing Liu[4] | Lu Ao[2,3]

[1]Affiliated Foshan Maternity and Child Healthcare Hospital, Southern Medical University (Foshan Maternity & Child Healthcare Hospital), Foshan, China

[2]Department of Bioinformatics, Fujian Key Laboratory of Medical Bioinformatics, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China

[3]Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Fujian Medical University, Fuzhou, China

[4]Department of Colorectal Surgery, The Affiliated Union Hospital of Fujian Medical University, Fuzhou, China

Correspondence

Lu Ao, Department of Bioinformatics, Fujian Key Laboratory of Medical Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, 350122, China.
Email: lukey@fjmu.edu.cn

Xing Liu, Department of Colorectal Surgery, The Affiliated Union Hospital of Fujian Medical University, Fuzhou 350001, China.
Email: liu9503@126.com

Funding information

Joint Scientific and Technology Innovation Fund of Fujian Province, Grant/Award Number: 2018Y9065; Joint research program of health and education in Fujian Province, Grant/Award Number: 2019-WJ-32; National Natural Science Foundation of China, Grant/Award Number: 61801118 and 81602738; Natural Science Foundation of Fujian Province, Grant/Award Number: 2020J01600

Abstract

Currently, due to the low quality of RNA caused by degradation or low abundance, the accuracy of gene expression measurements by transcriptome sequencing (RNA-seq) is very challenging for non-research-oriented clinical samples, majority of which are preserved in hospitals or tissue banks worldwide with complete pathological information and follow-up data. Molecular signatures consisting of several genes are rarely applied to such samples. To utilize these resources effectively, 45 stage II non-research-oriented samples which were formalin-fixed paraffin-embedded (FFPE) colorectal carcinoma samples (CRC) using RNA-seq have been analysed. Our results showed that although gene expression measurements were significantly affected, most cancer features, based on the relative expression orderings (REOs) of gene pairs, were well preserved. We then developed two REO-based signatures, which consisted of 136 gene pairs for early diagnosis of CRC, and 4500 gene pairs for predicting post-surgery relapse risk of stage II and III CRC. The performance of our signatures, which included hundreds or thousands of gene pairs, was more robust for non-research-oriented clinical samples, compared to that of two published concise REO-based signatures. In conclusion, REO-based signatures with relatively more gene pairs could be robustly applied to non-research-oriented CRC samples.

KEYWORDS

colorectal cancer, low-quality RNA, non-research-oriented clinical samples, relative expression orderings, transcriptional signature

Jun Cheng and Yating Guo are contributed equally to this work.

1 | INTRODUCTION

With technological advancement and reduced cost, transcriptome sequencing (RNA-seq) has become the primary technology for gene expression measurements.[1,2] This platform generally requires input of pre-selected high-quality RNA samples, designated as research-oriented clinical samples, to obtain reliable research results. For example, adequate amounts of RNA extracted from fresh-frozen (FF) samples containing at least 60% or 70% of the tumour nuclei,[3] and high RNA integrity (RIN) scores, such as RIN > 6 or RIN > 7,[4] are required. However, realistically, millions of samples obtained in hospitals and tissue banks worldwide are considered to be non-research-oriented clinical samples with low-quality RNA,[5] characterized by RNA degradation or fragmentation,[6] low amounts of RNA,[7] low tumour purities,[8] or above multiple features simultaneously. Although amplification technology can be used for these samples containing low amounts of RNA, it may introduce amplification bias.[9] Thus, for these non-research-oriented clinical samples, the accuracy of gene expression measurements would be significantly challenged. As a result, the transcriptional signatures based on risk scores that are summarized from the expression levels of signature genes are rarely applied to these non-research-oriented samples. Rather, these samples are primarily limited to pathological or immunohistochemical analysis[10]; however, these samples contain valuable pathological information and follow-up data,[11] which are precious resources in disease-related research. Therefore, it is imperative that transcriptional signatures should be developed that are applicable to non-research-oriented clinical samples with low-quality RNA.

Colorectal carcinoma (CRC) is one of the most common malignant tumours with high morbidity and mortality,[12] which is mainly transformed from acquired pre-cancerous lesions. Inflammatory bowel disease (IBD), including ulcerative colitis (UC) and Crohn's disease (CD), is a main type of pre-cancerous colorectal lesions, which could result in dysplasia, eventually develops and progresses to CRC.[13] Some studies have been reported that long-term exposure to chronic inflammation is the primary risk factor for CRC pathogenesis.[14] Meanwhile, some patients with stage II and III CRC after surgery treatment commonly have relapse risk. Currently, it is essential to timely discriminate early CRC patients from patients with inflammation and accurately predict the recurrence risk for stages II and III CRC patients after surgery.[15,16] However, established non-invasive tests, such as the guaiac-based faecal occult blood test and faecal haemoglobin, usually lack proper sensitivity and specificity for early diagnosis. Carcinoembryonic antigens, CA125 and CA19.9, which have already been applied into clinical practice, are not highly promising diagnostic or prognostic targets for personalized medicine.[17] Therefore, there is a critical need to develop highly robust and reliable biomarkers for diagnosis and prognosis of CRC patients.

Specific methods, such as TSP (top scoring pairs),[18] *k*-TSP[19] and other adjusted methods,[20] that take advantage of the qualitative transcriptional features of genes, which are based on the relative expression orderings (REOs) of gene pairs within sample, have been proposed to develop disease-related signatures. Our previous work has demonstrated that most of the REO patterns of gene pairs were insensitive to samples with degraded RNA, low amounts of RNA or varying tumour purities.[7,10,21] Hence, to facilitate clinical translational application, some concise REO-based signatures with several or dozens of gene pairs have been developed in our previous studies, including seven gene pairs for early diagnosis of CRC,[16] 44 gene pairs for predicting post-surgery relapse risk of stage II and III CRC[22] and so on.[23] Nevertheless, gene expression measurements could be widely and significantly affected by the low-quality non-research-oriented clinical samples. If the expression measurements of one or several signature genes are severely influenced, and even become zero, the performance of these concise REO-based signatures with several gene pairs may be seriously weakened or even rendered unfeasible. In consideration of the rapid development and decreasing cost of high-throughput sequencing technology, we proposed that the REO-based signatures should include relatively more gene pairs, potentially even hundreds or thousands of gene pairs, to obtain robust performance for the non-research-oriented clinical samples with low-quality RNA.

To this end, herein we analysed 45 stage II CRC non-research-oriented samples that were formalin-fixed paraffin-embedded (FFPE) samples without location pre-selection or pre-purification of tumour cells, measured using RNA-seq, and evaluated the influences of low-quality samples on their gene expression measurements. For these widely preserved non-research-oriented clinical samples, two REO-based signatures with relatively more gene pairs were developed and robustly applied to diagnosis and recurrence prediction of individuation. It had great value for clinical translational applications.

2 | MATERIALS AND METHODS

2.1 | Samples and data measurement

A total of 45 stage II CRC FFPE samples, denoted as CRC45, including 24 non-relapse and 21 relapse samples, were collected from FFPE tissue blocks which have been preserved at room temperature for about 6 years. The RIN scores, overall alignment rate of sequencing reads and clinical information of the 45 samples are shown in Table S1. Each FFPE tissue block without location pre-selection or pre-purification of tumour cells was cut into 6-10 slides of approximately 5 μm thickness. Then, the slides with frozen preservation were directly sent to the sequencing company. The whole process was about 72 hours. Next, according to the manufacture's protocol, total RNA was isolated from each sample using the RNAprep pure FFPE kit (Tiangen Biotech) and the quality of the RNA was assessed by electrophoresis on an Agilent 2100 Bioanalyzer system (Agilent Technologies). Ribosomal RNA was removed using the Globin-Zero Gold rRNA Removal kit & directional library, and the stranded RNA-seq library was constructed using the NEB Next® Ultra™ RNA

Library Prep kit. Paired-end sequencing (2 × 150) was performed on the Illumina HiSeq X Ten system (Illumina). Subsequently, the generated raw RNA-seq (FASTQ) files were pre-processed using the Trimmomatic,[24] and the reads were aligned to the reference genome (GRCh38) using HISAT2.[25] Finally, number of the Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) values of the genes were calculated. This research has been approved by the Institutional Review Board at Fujian Medical University Union Hospital, and written consent forms were obtained from all participants.

2.2 | Public data and pre-processing

All public gene expression profiles measured by RNA-seq were downloaded from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) and The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) database, as described in detail in Table 1. In total, 662 CRC samples, 149 normal samples and 353 IBD samples, including those from patients with UC and CD, were assessed. The FPKM or number of Reads Per Kilobase of transcript per Million reads mapped (RPKM) values were directly downloaded. Next, the Ensembl gene IDs were mapped to the Entrez gene IDs. The Ensembl gene ID which was mapped to zero or multiple Entrez gene IDs were deleted. If multiple Ensembl gene IDs were mapped to a Entrez gene ID, the expression value of the gene was defined as the arithmetic mean of the values of the multiple Ensembl gene IDs. The FPKM values of 13 CRC samples (CRC13) from our previous study were used directly.[16]

TABLE 1 Description of datasets used in this study

	Sample size		
Data	Normal	IBD	Cancer
TCGA			
FF			
-	51	–	–
I	–	–	106
II	–	–	231
III	–	–	176
IV	–	–	90
NA	–	–	22
FFPE			
I-IV	–	–	19
GSE72819	–	73	–
GSE109142	20	206	–
GSE83687	60	74	–
GSE50760	18	–	18
CRC13	–	–	13
CRC45	–	–	45

Note: NA, CRC samples without the information of stage.

2.3 | Evaluation of REO-based cancer features in non-research-oriented clinical samples

For a gene pair with two genes, for example gene *i* and gene *j*, the REO was denoted as $G_i > G_j$ (or $G_i < G_j$), where the expression measurement of gene *i* was higher (or lower) than that of gene *j* within a sample. A gene pair exhibiting the same REO pattern in most samples from one group, for example 95% or 99%, was defined as a stable gene pair. A stable gene pair with an opposite REO pattern between two groups was defined as a stable opposite gene pair. Using a hypergeometric cumulative distribution model, we evaluated whether a specific REO pattern, for example $G_i > G_j$ in one group of samples, was significantly opposited into the pattern $G_i < G_j$ in the other group of samples. A gene pair of which the REO pattern was significantly opposited between two groups was defined as a significant opposite gene pair. A stable or significant opposite gene pair, $G_i > G_j$, representing the REO pattern in CRC, was defined as an REO-based cancer feature. The stable and significant opposite gene pairs between stage II FF CRC samples and FF normal samples from TCGA were selected to evaluate the maintenance of the REO-based cancer features in each non-research-oriented clinical sample, respectively. The retention rate was calculated as follows:

$$\mathrm{ratio} = k/m \tag{1}$$

where *m* was the number of stable or significant opposite gene pairs, and *k* was the number of gene pairs that maintained the cancer features in a non-research-oriented sample. Notably, the gene pair was removed if two expression measurements of the gene pair both carried a value of zero. Additionally, the gene pairs containing a gene that was not measured were also removed.

2.4 | Identification of an REO-based signature for the early diagnosis of CRC

The stable opposite gene pairs between stage I FF CRC samples and FF normal samples were selected as candidate gene pairs for the REO-based signature for the early diagnosis of CRC. The gene expression profiles were then transformed into gene expression rank profiles according to their expression levels. All genes were sorted in ascending order, and the rank difference (*RD*) for a gene pair (*i*, *j*) in sample *t* was calculated as:

$$RD_{tij} = R_{ti} - R_{tj} \tag{2}$$

where R_{ti} and R_{tj} were the expression ranks of gene *i* and gene *j* in sample *t*, respectively. Next, the *RD* for each gene pair was calculated in each stage I sample or normal sample, respectively.

$$\mathrm{avgRD}_{ij} = \sqrt{\left|\mathrm{mean}\left[\mathrm{RD}_{ij}(\mathrm{cancer})\right]\right| \times \left|\mathrm{mean}\left[\mathrm{RD}_{ij}(\mathrm{normal})\right]\right|} \tag{3}$$

where |mean[RD_{ij}(cancer)]| and |mean[RD_{ij}(normal)]| represented the absolute mean RD value of the opposite gene pair (i, j) in all the samples of stage I CRC and normal groups, respectively. Subsequently, the geometric mean of the absolute mean[RD_{ij}(cancer)] and the absolute mean [RD_{ij}(normal)] was calculated to evaluate the opposite degree of the gene pair between two types of samples. The larger the opposite degree of the REO for the gene pair between stage I samples and normal samples, the larger the geometric mean (avgRD). If a gene appeared in multiple gene pairs, a redundancy removal process was performed that only the gene pair with the largest avgRD was retained. Based on the assumption that, for non-research-oriented clinical samples, the REO-based signature should include relatively more gene pairs, such as hundreds of gene pairs, all candidate gene pairs were pooled together as the REO-based signature for the early diagnosis of CRC. For the REO pattern of CRC, a sample was assigned to the CRC group when the retention rate of gene pairs in the signature was more than a certain cut-off, for example 60%; otherwise, it was assigned to the non-cancer group. Similarly, the gene pairs including two genes both with expression values of zero, or the gene pairs containing a non-measured gene, were excluded.

2.5 | Development of an REO-based signature for predicting post-surgery relapse risk of stages II and III CRC

Compared with a stringent cut-off of 99% for stable opposite gene pairs, a false discovery rate (FDR) control of 1% for significant opposite gene pairs was relatively loosen. Because the transcriptional differences between stage I and stage IV samples were less than that between normal and CRC, it is difficult to select hundreds or thousands of stable opposite gene pairs to construct the REO-based signature for predicting post-surgery relapse risk of CRCs. Thus, in order to obtain sufficient gene pairs with more advantageous discriminating ability for CRC relapse, we selected the significant opposite gene pairs from the differentially expressed genes (DEGs) between stage I and stage IV samples identified by the RankCompV2 algorithm.[26] The stage I CRC samples without new tumour events and metastasis were analysed. Briefly, the RankCompV2 algorithm firstly identified the significantly stable gene pairs in two distinct groups by using the binomial test with FDR control (<20%), respectively. Next, based on the overlap between the two lists of stable gene pairs, the concordant and opposite REOs were identified between two distinct groups. Finally, DEGs that may disrupt the REOs of genes were identified using the Fisher's exact test with FDR control (<5%). Using a hypergeometric cumulative distribution model, the significant opposite gene pairs with at least one DEG were identified between stage I and stage IV samples and further filtered by 13 paired FFPE and FF samples from TCGA. For a significant opposite gene pair, the same REO pattern should be kept in at least ten paired FF and FFPE samples.

The candidate gene pairs were then sorted in descending order according to their relative coverage difference between two groups, which was calculated as follows:

$$C_{IVI} = C_{IV} - C_{I} \tag{4}$$

where C_{IV} and C_I represent the coverage of a gene pair with the REO pattern ($G_i > G_j$) in stage IV and stage I samples, respectively. For example, for a gene pair with the REO pattern ($G_i > G_j$) in m of n stage IV samples, its coverage will be m/n. The samples that had expression values of zero for both genes of a gene pair were not counted. C_{IVI} represents the relative frequency difference of the gene pair (i, j) between stage IV and stage I samples.

All candidate gene pairs were classified into several groups representing the candidate signatures. The classification performance of each candidate signature was evaluated by the voting rule that states that a sample was to be classified as high-risk relapse when the proportion of the same REO pattern ($G_i > G_j$) was more than the threshold. Considering that the accuracy for stage IV should be as high as possible, the classification threshold was evaluated separately as 50% ± 1%, and the candidate signature with better classification performance and relatively higher robustness was selected as the signature. Finally, a Cox proportional-hazards regression model was used to evaluate the association between the predictive signature and the disease-free interval time (DFI) of patients with stage II and III CRC [27] and the Schoenfeld residuals test was used to test the proportional hazard assumption in the Cox model. The Kaplan-Meier method and log-rank test were used to estimate the survival curves. All statistical analyses were performed by R 3.6.0. The R scripts were provided in Supplementary scripts.

2.6 | Protein-protein interactions and functional enrichment analysis

A regulatory protein-protein interaction (PPI) network for the genes of interest was constructed based on the integrated data from the HSNet (Human Signaling Network, version 6)[28] and the SIGNOR databases.[29] Particularly, the Ensembl gene IDs corresponding to the unique Entrez gene IDs of protein-coding genes were analysed. Functional enrichment analysis for the genes of interest was subsequently performed based on KEGG (the Kyoto Encyclopedia of Genes and Genomes).[30] The hypergeometric distribution model was used to calculate the enrichment significance of biological pathways, whereas the Benjamini-Hochberg method was adopted to estimate the FDR.

3 | RESULTS

3.1 | Quality evaluation of non-research-oriented clinical samples

Here, we firstly evaluated the RNA qualities of the 45 non-research-oriented FFPE samples of stage II CRC measured in our laboratory. Compared to the requirement of high-quality FF samples with RIN scores of 6.0 for RNA-seq, the RIN scores of total RNA in these non-research-oriented clinical samples ranged from 2.1 to 2.7. These results suggest that the RNA of these non-research-oriented samples was seriously degraded and fragmented. Meanwhile, in more than half of the 45 non-research-oriented FFPE samples, more than 28% of genes had an expression value of zero, and the highest percentage of genes with an expression value of zero in FFPE samples reached 53.90%. In contrast, approximately 10% of the 231 stage II FF samples from TCGA had slightly more than 28% of genes with expression values of zero. These results indicated that the gene expression measurements of these non-research-oriented samples were seriously affected.

Next, the REO-based cancer features in these non-research-oriented samples were evaluated through comparison with the stable opposite REO patterns and significant opposite REO patterns between the FF stage II CRC samples and normal samples from TCGA, respectively. To weaken the biased influences of sample size for two groups, we selected significant or stable opposite gene pairs between the first third (77) of the 231 stage II FF samples and 51 normal samples. With a 95% cut-off for stable opposite gene pairs, 177 122 stable opposite gene pairs were identified between 77 stage II FF CRCs ($G_i > G_j$) and 51 FF normal samples ($G_i < G_j$). Taking these REO patterns of stable opposite gene pairs in CRC samples as cancer features, approximately 95% of the stable opposite gene pairs were also retained in the remaining 154 stage II FF CRCs (as shown in Figure S1), and the retention rate was approximately 80% in the 45 non-research-oriented clinical samples (as shown in Figure 1A). Similarly, with the control of FDR < 0.05, 43 439 977 significant opposite gene pairs were identified between 77 stage II FF CRCs and 51 FF normal samples. Taking these REO patterns of the significant opposite gene pairs in CRC samples as cancer features, we found that compared to approximately 60% of the retention rate for the 77 and the remaining 154 FF CRC samples, approximately 55% of the significant opposite gene pairs were retained in the non-research-oriented clinical samples (as shown in Figures 1B and S1). On the contrary, the retention rate of stable and significant opposite gene pairs ($G_i > G_j$) in the normal samples were only kept about 5% and 20%, respectively. These results indicated that most of the REO-based cancer features were well preserved in the non-research-oriented clinical samples.

3.2 | Development of an REO-based signature for the early diagnosis of CRC

Recently, we have reported a concise REO-based signature consisting of seven gene pairs for discriminating early CRC from IBD samples, including UC and CD samples.[16] However, the seven gene pairs were able to only distinguish 31.11% of the 45 non-research-oriented CRCs as cancer because the expression measurements of two signature genes were zero in all non-research-oriented samples. The results showed that the seven gene pairs were not robust enough against the non-research-oriented samples with low-quality RNA. It is, therefore, necessary to develop a more robust signature for the early diagnosis of CRC.

The development of the early diagnosis signature for CRC was summarized in Figure 2A. First, with a cut-off of 99%, 29 135 stable opposite gene pairs were identified between the 106 stage I FF

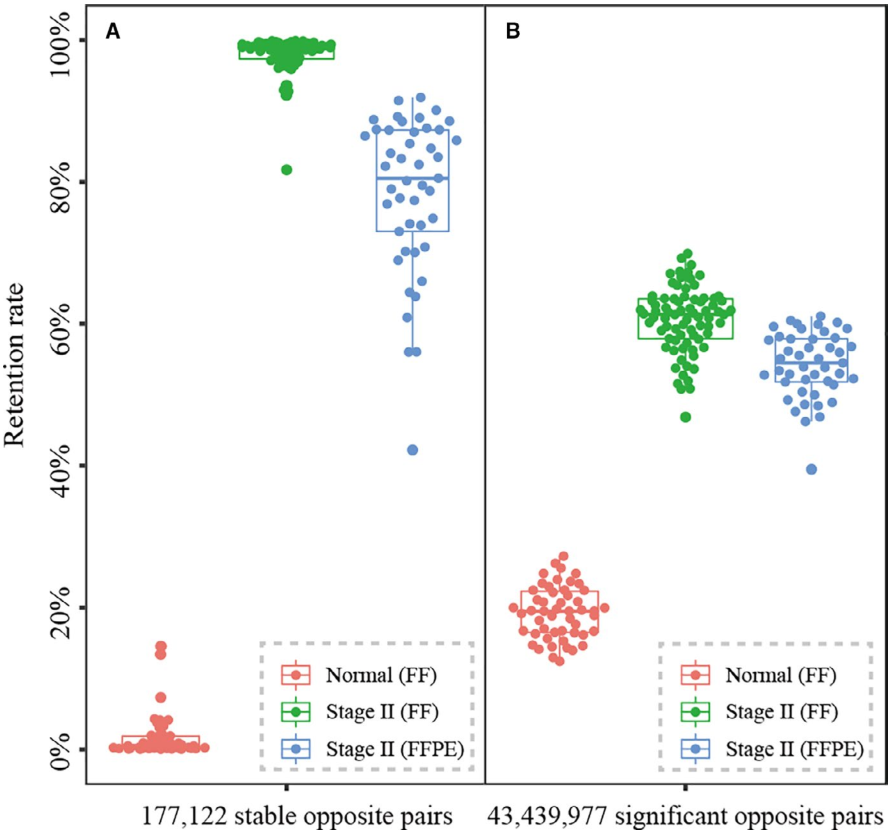

FIGURE 1 Evaluation of the REO-based cancer features in non-research-oriented clinical samples using the 177 122 stable opposite gene pairs (A) or the 43 439 977 significant opposite gene pairs (B) which were selected between 77 FF tumour and 51 FF normal samples. The retention rates of gene pairs with the specific cancer pattern ($G_i > G_j$) in the public FF CRCs (green), in-house FFPE CRCs (blue) and normal FF samples (red) were shown

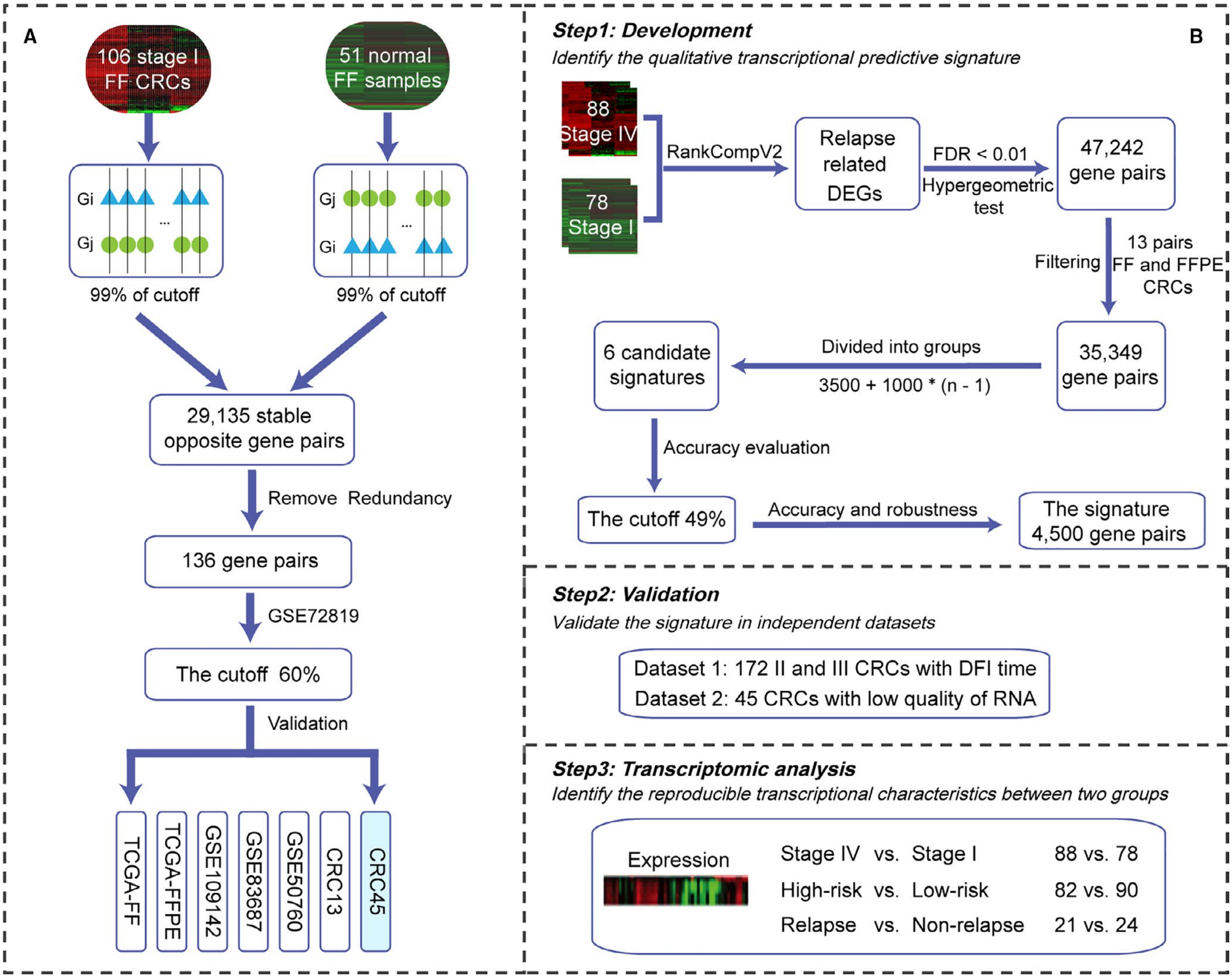
A
106 stage I FF CRCs
51 normal FF samples
Gi
Gj
...
Gj
Gi
...
99% of cutoff
99% of cutoff
29,135 stable opposite gene pairs
Remove Redundancy
136 gene pairs
GSE72819
The cutoff 60%
Validation
TCGA-FF
TCGA-FFPE
GSE109142
GSE83687
GSE50760
CRC13
CRC45
B
Step1: Development
Identify the qualitative transcriptional predictive signature
88 Stage IV
78 Stage I
RankCompV2
Relapse related DEGs
FDR < 0.01
Hypergeometric test
47,242 gene pairs
Filtering
13 pairs FF and FFPE CRCs
35,349 gene pairs
Divided into groups
3500 + 1000 * (n - 1)
6 candidate signatures
Accuracy evaluation
The cutoff 49%
Accuracy and robustness
The signature 4,500 gene pairs
Step2: Validation
Validate the signature in independent datasets
Dataset 1: 172 II and III CRCs with DFI time
Dataset 2: 45 CRCs with low quality of RNA
Step3: Transcriptomic analysis
Identify the reproducible transcriptional characteristics between two groups
Expression
Stage IV vs. Stage I 88 vs. 78
High-risk vs. Low-risk 82 vs. 90
Relapse vs. Non-relapse 21 vs. 24

FIGURE 2 The flow chart for the development of the REO-based signatures for the early diagnosis of CRC (A) or predicting post-surgery relapse risk of stage II and III CRC (B), respectively

CRCs and 51 FF normal samples from TCGA. Second, we narrowed down the number of gene pairs via a redundancy removal process and 136 stable opposite gene pairs were identified as the early diagnosis signature for CRC (see Section 2, Table S2). The cut-off of the voting rule was set to 60% as the 73 colitis samples in the GSE72819 dataset were all correctly assigned to the non-cancer group. For a sample, if the retention ratio of the specific cancer pattern ($G_i > G_j$) was ≥60%, the sample was labelled as CRC; otherwise, it was labelled as non-cancer (see Section 2). In the training datasets, 106 CRC and 51 normal samples were all correctly assigned to the CRC group and the non-cancer group, respectively.

The 136 gene pair signature was further validated in multiple public datasets. For the remaining CRC (untrained FF samples) surgical samples from TCGA, 99.42% of the 519 FF CRCs and 89.47% of the 19 FFPE CRCs were correctly classified as cancer. The 13 CRC samples with various tumour purities from the CRC13 dataset were all correctly assigned to the CRC group. For the 206 colitis and 20 normal biopsy samples from the GSE109142 dataset, all were correctly assigned to the non-cancer group. For the GSE83687 dataset, 97.30% of the 74 IBD surgical samples and 100% of the 60 normal surgical samples were correctly classified into the non-cancer group. For the GSE50760 dataset, two thirds of the 18 primary cancer surgical samples were classified as cancer samples and 100% of the normal surgical samples were correctly designated as non-cancer samples. These results demonstrate that our signature can effectively facilitate the early diagnosis of CRC, regardless of whether the samples are obtained via surgery or biopsy, or whether the samples have varying tumour purities.

Next, the 136 gene pair signature was further verified in the 45 non-research-oriented samples. Similar results were observed that 95.56% of the 45 non-research-oriented samples were correctly classified as cancer. As shown in Table 2, 99.60% of the total 502 non-cancer samples, and 97.92% of the total 720 cancer samples, were correctly identified. The retention rates of the early diagnosis signature for all analysed samples were shown in Figure S2. Taken together, these results reveal that the REO-based signature with 136 gene pairs can be robustly applied to non-research-oriented clinical samples with low-quality RNA.

TABLE 2 The performance of the early diagnosis signature in multiple datasets

Dataset	Accuracy/sample size Normal	IBD	Cancer
The performance of the signature in the training datasets			
TCGA-1	100% (51/51)	–	100% (106/106)
GSE72819	–	100% (73/73)	–
The performance of the signature in the validation datasets			
TCGA-FF	–	–	99.42% (516/519)
TCGA-FFPE	–	–	89.47% (17/19)
GSE109142	100% (20/20)	100% (206/206)	–
GSE83687	100% (60/60)	97.30% (72/74)	–
GSE50760	100% (18/18)	–	66.67% (12/18)
CRC13	–	–	100% (13/13)
CRC45	–	–	95.56% (43/45)
Total	99.60% (500/502)		98.19% (707/720)

Note: IBD represents inflammatory bowel diseases samples. TCGA-1: stage I CRC and normal FF samples; TCGA-FF: the non-stage I FF CRC samples; TCGA-FFPE: the FFPE CRC samples.

3.3 | Identification of an REO-based signature for predicting post-surgery relapse risk of stage II and III CRC

The process for developing an REO-based signature for predicting the post-surgery relapse risk of stage II and III CRC was summarized in Figure 2B. First, 619 CRC samples with survival information were selected from TCGA.[27] Based on the hypothesis that the relapse of stage II and III CRC could be attributed to micro-metastasis,[31,32] 644 DEGs were identified between the 88 stage IV samples (the metastatic samples) and 78 stage I samples (the non-metastatic samples) using the RankcompV2 algorithm. With the control of FDR < 0.01, 47 242 significant opposite gene pairs, including at least one DEG, were identified between stage IV and stage I samples, of which 35 349 gene pairs were kept after further filtering using 13 paired FF and FFPE CRC samples (as shown in Table S3). Next, the sorted significant opposite gene pairs were categorized into six groups for six candidate signatures based on the rule: the *n*th candidate signature consisted of 3500 + 1000 * (*n* − 1) gene pairs. Specifically, the six candidate signatures included 3500, 4500, 5500, 6500, 7500 and 7849 gene pairs, respectively. For example, the first candidate signature was assigned the top 1 to 3500 gene pairs, and the top 3501 to 8000 gene pairs were assigned to the second candidate signature and so on. The accuracy of classification for each candidate signature was evaluated in the training dataset. The cut-off was set to 49% in comparison with 50% and 51%, as the higher accuracy for stage IV samples. For a given sample, if more than 49% of the gene pairs in the signature contained the specific REO pattern for relapse, the sample was labelled as high-risk relapse, and vice versa (see Section 2).

Using the cut-off of 49%, the accuracies of two candidate signatures were found to both be above 80% in stage I and stage IV. The first candidate signature including 3500 gene pairs could correctly classify 82.05% of the stage I samples and 81.82% of the stage IV samples in the training dataset, whereas the second candidate signature including 4500 gene pairs correctly classified 80.77% of the stage I and 81.82% of the stage IV samples. For the sake of robustness, the second candidate signature that included relatively more gene pairs was ultimately chosen as the signature (see Table S4).

For 172 CRCs with DFI time obtained from TCGA, 80 samples and 92 samples were classified by the signature as high relapse risk and low relapse risk, respectively, of which the former had significantly worse DFI survival than the latter (univariate Cox, HR = 2.78, 95% CI = 1.15-6.72, log-rank test $P = 0.018$, Figure 3A). And there was no separate residual for the high relapse risk and the low relapse risk groups in the 172 CRCs from TCGA (the Schoenfeld residuals test, $P = 0.25$). The retention rates for 172 CRC samples were show n in Figure S3. Meanwhile, similar results were also observed within the 5-year (univariate Cox, HR = 3.42, 95% CI = 1.34-8.74, log-rank test $P = 0.0064$, Figure 3B) and 3-year DFI time (univariate Cox, HR = 3.65, 95% CI = 1.29-10.29, log-rank test $P = 0.0090$, Figure 3C). Moreover, our signature was able to robustly predict the post-surgery relapse risk in stage II CRC (Figure 3D-F). Unfortunately, for stage III patients, the significant difference of DFI time between high and low relapse risk groups was not observed, as shown in Figure 3G-I. Actually, one patient with 85 years who had cancer relapse in about four months was classified into the low relapse risk group. After excluding the sample, the DFI time between the low and high relapse risk groups in stage III CRC patients was moderately different (univariate Cox, HR = 5.2, 95% CI = 0.65-41.71, log-rank test $P = 0.083$, Figure S4A). Moreover, the 5-year and 3-year DFI time was significantly different (Figure S4B,C). Meanwhile, we also performed multivariable cox proportional-hazards regression analyses for the CRC relapse signature with 4500 gene pairs. Because the MSI information of 37 patients were lost, 135 II-III colorectal cancer patients, including 27 patients with MSI-high, 21 patients with MSI-low and 87 patients with MSI-stable, were evaluated. After correction for stage, gender, age and MSI, there was a modest difference

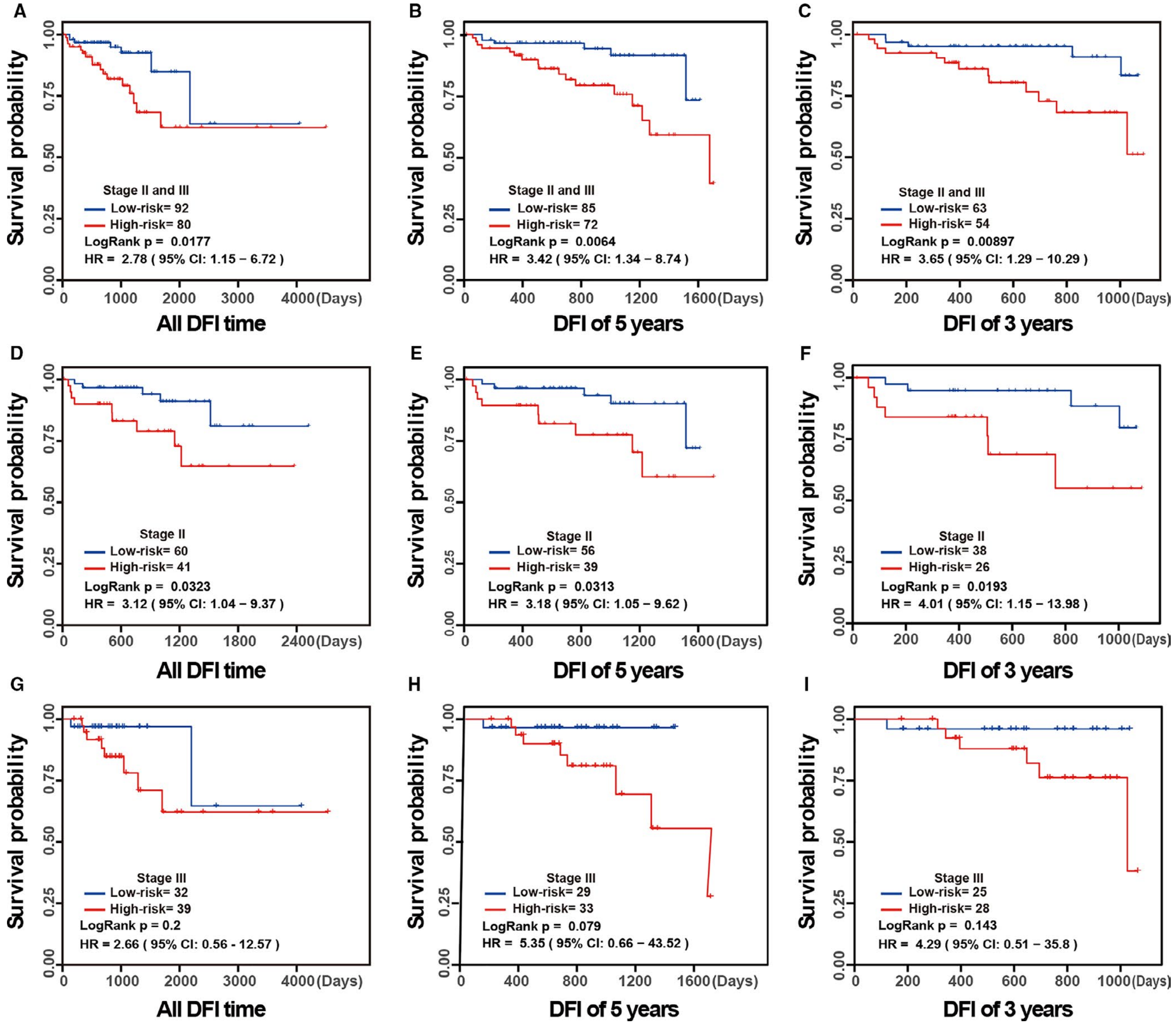

FIGURE 3 The predictive performance of the 4500 gene pair signature. A-C, Kaplan-Meier curves of DFI for patients with stage II and III CRC. D-F, Kaplan-Meier curves of DFI for patients with stage II CRC. G-I, Kaplan-Meier curves of DFI for patients with stage III CRC

on DFI time (HR = 2.65, 95% CI = 0.91-7.69, log-rank *P* = 0.074, as shown in Table S5). To reduce the impact of sample reduction due to the missing MSI information, multivariable cox proportional-hazards regression analyses were additionally performed in 172 II-III colorectal cancer patients. Significant difference on DFI time was observed after correction for stage, gender and age (HR = 2.65, 95% CI = 1.06-6.62, log-rank test *P* = 0.037).

For the 45 stage II non-research-oriented CRC samples measured in our laboratory, the 23 patients predicted as low relapse risk had significantly better DFI survival than the 22 patients predicted as high relapse risk (univariate Cox, HR = 3.87, 95% CI = 1.49-10.05, log-rank test *P* = 0.0028, as shown in Figure 4A). There was no separate residual for the high relapse risk and the low relapse risk groups in the 45 CRCs (the Schoenfeld residuals test, *P* = 0.67). Moreover, there were 70.83% of the 24 non-relapse samples, and 71.42% of the 21 relapse samples, correctly identified. Compared to the predictive performance of our previous 44 gene pairs, which predicted six patients (25% of non-relapse samples) as low relapse risk and 39 as high relapse risk (univariate Cox, log-rank test *P* = 0.033, Figure 4B), the results further demonstrated that our signature with 4500 gene pairs had higher prognosis capacity, and more robust predictive performance in non-research-oriented samples.

3.4 | The potential relapse mechanism of CRC

Using the RankCompV2 algorithm with FDR < 0.05, 3109 DEGs were identified between the 80 high and 92 low relapse risk stage II and III CRCs, whereas 784 DEGs were identified between 88 stage IV and 78 stage I samples. A total of 503 DEGs were overlapped

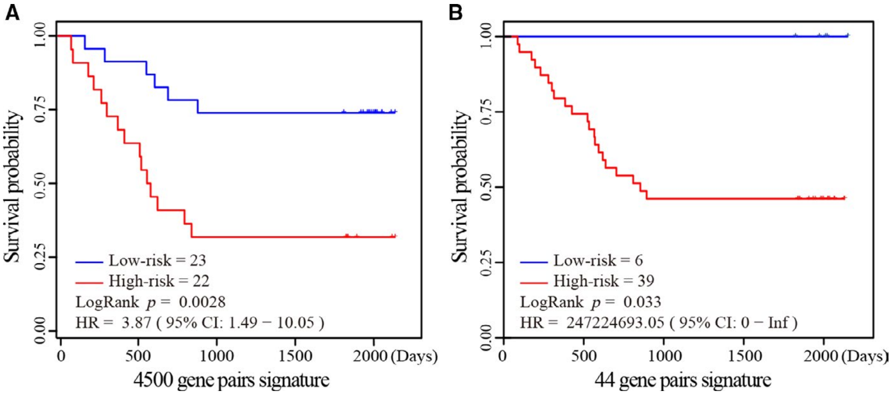

FIGURE 4 The predictive performances of the 4500 gene pairs signature (A) and 44 gene pairs signature (B) in the non-research-oriented clinical samples

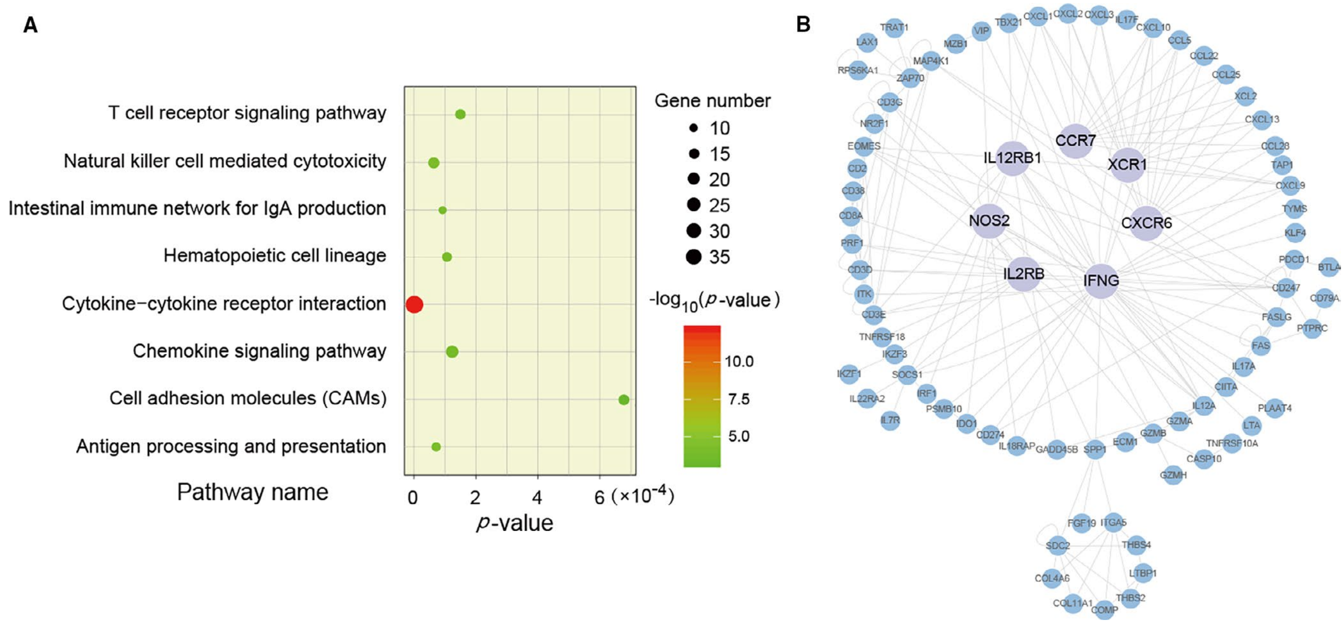

FIGURE 5 The function analysis of the common DEGs. A, KEGG function enrichment analysis with the 499 DEGs. B, The largest sub-network in PPI analysis

between the two DEG lists, of which 499 DEGs had consistent dysfunction direction. The consistency was 99.20%, which was significantly higher than what was expected by chance (binomial test, $P < 2.2 \times 10^{-16}$). The functional enrichment analysis of 499 DEGs showed that they were significantly enriched in the immune-related pathways, including ‘natural killer cell-mediated cytotoxicity’ and ‘T cell receptor signalling pathway’. (FDR < 0.05, as shown in Figure 5A). Notably, based on the 1811 immune-related genes downloaded from the ImmPort database, 117 of the 499 DEGs were immune-related genes. By mapping the 499 DEGs into the integrated data from both the HSNet and SIGNOR databases, a regulatory PPI network including 102 DEGs with 196 edges was constructed, 61.78% of which were immune-related genes. The largest sub-network was shown in Figure 5B. We also observed that seven hub DEGs (*IFNG*, *IL2RB*, *IL12RB1*, *CCR7*, *XCR1*, *CXCR6* and *NOS2*) with more than ten PPI interactions were all immune-related genes.

Furthermore, using the RankCompV2 algorithm (FDR < 0.05), 956 DEGs were detected between 24 non-relapse and 21 relapse samples measured in our laboratory, of which 42 DEGs overlapped with the above-mentioned 499 DEGs. The concordance score between the two DEG lists was as high as 85.71% (36 DEGs), which was unlikely to occur by chance (binomial test, $P < 2.2 \times 10^{-07}$). Meanwhile, the PPI analysis showed that the 18 DEGs directly interacted with 301 other genes (as shown in Figure S5) and these 319 genes were also significantly enriched in immune-related functional pathways, including ‘T cell receptor signalling pathway’ and ‘B cell receptor signalling pathway’ (Table S6). These results further suggest that the potential relapse mechanism of CRC might be closely related to immune dysfunction.

4 | DISCUSSION

In this study, we first demonstrated that most of the REO-based cancer features were preserved in non-research-oriented clinical samples, although their gene expression measurements were seriously affected. Second, we developed an REO-based signature with 136 gene pairs for early diagnosis of CRC and an REO-based signature with 4500 gene pairs for predicting post-surgery relapse risk of stage II and III CRC. The results demonstrate that the REO-based signatures with relatively more gene pairs, rather than several or dozens of gene pairs, could be robustly applied to non-research-oriented clinical samples containing low-quality RNA.

Additionally, we found that CRC relapse could be closely related to immune dysfunction. Previous studies have shown that the immune-related DEGs, *NOS2*, *CCR7* and *IFNG*, which were hub nodes in Figure 5B, could affect the prognosis of CRC patients. For example, Thomas et al[33] have shown that *NOS2* is highly expressed in different cancers and may be a powerful prognostic biomarker, and *NOS2* polymorphisms could be used to predict whether metastatic CRC patients may benefit from first-line chemotherapy.[34] The expression of *CCR7* in tumour infiltrating CD8+ T cells may lead to a tumour-specific immune response with potential antitumour activity, leading to a favourable prognosis for metastatic CRC patients.[35] Moreover, *CCR7* has been suggested as a potential target in cancer therapy as it plays an important role in the metastasis of several cancers.[35,36] Ganapathi et al[37] have indicated that low expression of *IFNG* could be the reason for the progression of stage IV CRC.[37] *IL12RB1* has been reported that its mutation played a causal role in non-polyposis CRC pre-disposition.[38] Furthermore, the other three hub genes, *IL2RB*, *XCR1* and *CXCR6*, have been reported to be related with the prognosis of other cancers, such as early breast cancer,[39] salivary adenoid cystic carcinoma,[40] bladder and hepatocellular cancers.[41,42]

Obviously, the subtle quantitative information associated with gene expression would be missed in REO patterns. However, this genetic information is quite generally error-prone and sensitive to batch effects[43,44] and data normalization.[23] Nevertheless, the REO patterns that take advantage of the qualitative features of genes within samples could be readily applied for individualized clinical applications.[45] With the rapid development and sharp decrease in costs associated with sequencing technology, signatures with hundreds or thousands of genes could now be feasible in clinical settings. Moreover, including additional genes in the REO-based signatures could increase robustness against the dysfunction of some signature genes, while weakening the influence of non-research-oriented clinical samples with low-quality RNA, particularly in widely preserved FFPE samples. However, it is certainly necessary to also develop new technologies to extract high-quality RNA from non-research-oriented samples and optimize the protocols or workflow, including the extraction, amplification and labelling methods.[6,46,47]

In contrast with RNA biomarkers in FFPE samples, the DNA biomarkers have a minor risk of degradation. Two DNA-related biomarkers, Cologuard[48] and Epi proColon® 2.0 CE,[49] have been approved by the FDA for colorectal cancer screening. However, they remain too expensive and technically complex.[50] Some DNA methylation biomarkers that could be directly detected by blood and stool have also been reported.[51-54] Additionally, other DNA biomarkers, such as mutations of *KRAS*, *TP53* and *APC*, and hypermethylation of tumour suppressor genes at the promoter regions, have been developed.[55,56] But until now, there are no recognized prognostic biomarkers in clinical practice for CRC patients.[15] Therefore, development of robust RNA or DNA biomarkers for clinical application is worthy to further study.

Colorectal adenoma is a major type of pre-cancerous colorectal lesions. The risk of developing colorectal cancer for patients with adenomas is two to four times higher than those patients without adenomas. However, a lack of adenomas measured by RNA-seq in the public database caused that no adenoma samples datasets were included to develop the early diagnosis signature. We also noticed that the sample size of non-research-oriented clinical samples was small in this study. In our CRC45 dataset, only approximately 70% of the non-relapse and relapse samples were correctly identified using the 4500 gene pairs signature. Thus, additional non-research-oriented samples are needed to further validate and optimize our REO-based signatures. Another limitation was that only the sequencing data were analysed. The predictive power of our signatures is not high in all datasets across platforms. For instance, in the GSE50760 dataset, one third of the 18 cancer patients were incorrectly identified by the 136 gene pairs signature, which may result from platform differences. In next work, we will pool microarray and sequencing data together to develop REO-based signatures for the non-research-oriented clinical samples, to allow them to readily be applied across different platforms.

In summary, the REO-based signature with relatively more gene pairs could be robustly applied to the non-research-oriented clinical samples containing low-quality RNA, and it holds significant value for clinical translational applications.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos 81602738 and 61801118) and the Joint Scientific and Technology Innovation Fund of Fujian Province (Grant No: 2018Y9065), the Joint Research Program of Health and Education in Fujian Province (2019-WJ-32) and the Natural Science Foundation of Fujian Province (Grant No. 2020J01600).

CONFLICT OF INTEREST

The authors confirm that there are no conflicts of interest.

AUTHOR CONTRIBUTION

Jun Cheng: Conceptualization (equal); Writing-original draft (lead); Writing-review & editing (equal). **Yating Guo:** Data curation (equal); Formal analysis (equal); Visualization (equal); Writing-review & editing (equal). **Guoxian Guan:** Data curation (equal). **Haiyan Huang:** Visualization (equal). **Fengle Jiang:** Formal analysis (equal). **Jun He:** Writing-original draft (supporting). **Junling Wu:** Visualization (equal). **Zheng Guo:** Conceptualization (equal). **Xing Liu:** Data curation (equal). **Lu Ao:** Conceptualization (equal); Writing-review & editing (equal).

DATA AVAILABILITY STATEMENT

The in-house data used and analysed during the current study are available from the corresponding authors upon reasonable request.

ORCID

Lu Ao https://orcid.org/0000-0001-7378-4967

REFERENCES

1. 't Hoen PAC, Ariyurek Y, Thygesen HH, et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res*. 2008;36(21):e141-e141.
2. Fumagalli D, Blanchet-Cohen A, Brown D, et al. Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology. *BMC Genom*. 2014;15:1008.
3. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun*. 2015;6:8971.
4. Bossel Ben-Moshe N, Gilad S, Perry G, et al. mRNA-seq whole transcriptome profiling of fresh frozen versus archived fixed tissues. *BMC Genom*. 2018;19(1):419.
5. Blow N. Tissue preparation: tissue issues. *Nature*. 2007;448(7156):959-963.
6. Li J, Fu C, Speed TP, Wang W, Symmans WF. Accurate RNA sequencing from formalin-fixed cancer tissue to represent high-quality transcriptome from Frozen tissue. *JCO Precis Oncol*. 2018;2018(2):1-9.
7. 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(1):913.
8. Clarke J, Seo P, Clarke B. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*. 2010;26(8):1043-1049.
9. Bhargava V, Head SR, Ordoukhanian P, Mercola M, Subramaniam S. Technical variations in low-input RNA-seq methodologies. *Sci Rep*. 2014;4:3678.
10. 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(4):6652-6662.
11. Tang W, David FB, Wilson MM, Barwick BG, Leyland-Jones BR, Bouzyk MM. DNA extraction from formalin-fixed, paraffin-embedded tissue. *Cold Spring Harbor Protocols*. 2009;2009(2):pdb.prot5138.
12. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Research Support, Non-U.S. Gov't. *Gut*. 2017;66(4):683-691.
13. Bjerrum JT, Nielsen OH, Riis LB, et al. Transcriptional analysis of left-sided colitis, pancolitis, and ulcerative colitis-associated dysplasia. Comparative Study Research Support, Non-U.S. Gov't. *Inflamm Bowel Dis*. 2014;20(12):2340-2352.
14. Axelrad JE, Lichtiger S, Yajnik V. Inflammatory bowel disease and cancer: the role of inflammation, immunosuppression, and cancer treatment. *Review. World J Gastroenterol*. 2016;22(20):4794-4801.
15. Das V, Kalita J, Pal M. Predictive and prognostic biomarkers in colorectal cancer: a systematic review of recent advances and challenges. *Rev Syst Rev Biomed Pharmacother*. 2017;87:8-19.
16. Guan Q, Zeng Q, Yan H, et al. A qualitative transcriptional signature for the early diagnosis of colorectal cancer. *Cancer Sci*. 2019;110(10):3225-3234.
17. Gao Y, Wang J, Zhou Y, Sheng S, Qian SY, Huo X. Evaluation of Serum CEA, CA19-9, CA72-4, CA125 and Ferritin as Diagnostic Markers and Factors of Clinical Parameters for Colorectal Cancer. Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't. *Sci Rep*. 2018;8(1):2732.
18. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol*. 2004;3:Article19.
19. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*. 2005;21(20):3896-3904.
20. Winslow RL, Trayanova N, Geman D, Miller MI. Computational medicine: translating models to clinical care. *Sci Transl Med*. 2012;4(158):158rv11.
21. 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(18):30265-30275.
22. Song K, Guo Y, Wang X, et al. Transcriptional signatures for coupled predictions of stage II and III colorectal cancer metastasis and fluorouracil-based adjuvant chemotherapy benefit. *FASEB J*. 2019;33(1):151-162.
23. 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(2):233-242.
24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120.
25. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12(4):357-360.
26. Cai H, Li X, Li J, et al. Identifying differentially expressed genes from cross-site integrated data based on relative expression orderings. *Int J Biol Sci*. 2018;14(8):892-900.
27. Liu J, Lichtenberg T, Hoadley KA, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*. 2018;173(2):400-416 e11.
28. Wang E. Human Signaling Network (Version 6). 2014.
29. Perfetto L, Briganti L, Calderone A, et al. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res*. 2016;44(D1):D548-D554.
30. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids Res*. 2010;38(Database issue):D355-D360.
31. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*. 2017;67(2):93-99.
32. Sloothaak DA, Sahami S, van der Zaag-Loonen HJ, et al. The prognostic value of micrometastases and isolated tumour cells in histologically negative lymph nodes of patients with colorectal cancer: a systematic review and meta-analysis. *Eur J Surg Oncol*. 2014;40(3):263-269.
33. Thomas DD, Wink DA. NOS2 as an emergent player in progression of cancer. Editorial Introductory. *Antioxid Redox Signal*. 2017;26(17):963-965.
34. Schirripa M, Zhang WU, Yang D, et al. NOS2 polymorphisms in prediction of benefit from first-line chemotherapy in metastatic colorectal cancer patients. Research Support, N.I.H., Extramural. *PLoS One*. 2018;13(3):e0193640.
35. Correale P, Rotundo MS, Botta C, et al. Tumor infiltration by T lymphocytes expressing chemokine receptor 7 (CCR7) is predictive of favorable outcome in patients with advanced colorectal carcinoma. Clinical Trial, Phase III Randomized Controlled Trial Research Support, Non-U.S. Gov't. *Clin Cancer Res*. 2012;18(3):850-857.
36. Mishan MA, Ahmadiankia N, Bahrami AR. CXCR4 and CCR7: Two eligible targets in targeted cancer therapy. Review. *Cell Biol Int*. 2016;40(9):955-967.
37. Ganapathi SK, Beggs AD, Hodgson SV, Kumar D. Expression and DNA methylation of TNF, IFNG and FOXP3 in colorectal cancer and

their prognostic significance. Research Support, Non-U.S. Gov't. *Br J Cancer*. 2014;111(8):1581-1589.

38. Belhadj S, Terradas M, Munoz-Torres PM, et al. Candidate genes for hereditary colorectal cancer: mutational screening and systematic review. *Hum Mutat*. 2020;41(9):1563-1576.
39. Lee H, Kwon MJ, Koo BM, Park HG, Han J, Shin YK. A novel immune prognostic index for stratification of high-risk patients with early breast cancer. *Sci Rep*. 2021;11(1):128.
40. Mays AC, Feng X, Browne JD, Sullivan CA. Chemokine and chemokine receptor profiles in metastatic salivary adenoid cystic carcinoma. *Anticancer Res*. 2016;36(8):4013-4018.
41. Lee JT, Lee SD, Lee JZ, Chung MK, Ha HK. Expression analysis and clinical significance of CXCL16/CXCR6 in patients with bladder cancer. *Oncol Lett*. 2013;5(1):229-235.
42. Gao Q, Zhao YJ, Wang XY, et al. CXCR6 upregulation contributes to a proinflammatory tumor microenvironment that drives metastasis and poor patient outcomes in hepatocellular carcinoma. *Cancer Res*. 2012;72(14):3546-3556.
43. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11(10):733-739.
44. Nygaard V, Rodland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*. 2016;17(1):29-39.
45. Wang H, Sun Q, Zhao W, et al. Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics*. 2015;31(1):62-68.
46. Amini P, Ettlin J, Opitz L, Clementi E, Malbon A, Markkanen E. An optimised protocol for isolation of RNA from small sections of laser-capture microdissected FFPE tissue amenable for next-generation sequencing. *BMC Mol Biol*. 2017;18(1):22.
47. Landolt L, Marti HP, Beisland C, Flatberg A, Eikrem OS. RNA extraction for RNA sequencing of archival renal tissues. *Scand J Clin Lab Invest*. 2016;76(5):426-434.
48. Imperiale TF, Ransohoff DF, Itzkowitz SH, et al. Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J Med*. 2014;370(14):1287-1297.
49. Church TR, Wandell M, Lofton-Day C, et al. Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. *Gut*. 2014;63(2):317-325.
50. Loktionov A. Biomarkers for detecting colorectal cancer non-invasively: DNA, RNA or proteins? *World J Gastrointest Oncol*. 2020;12(2):124-148.
51. Vega-Benedetti AF, Loi E, Moi L, et al. Colorectal cancer early detection in stool samples tracing CpG Islands methylation alterations affecting gene expression. *Int J Mol Sci*. 2020;21(12):4494.
52. Jensen SO, Ogaard N, Orntoft MW, et al. Novel DNA methylation biomarkers show high sensitivity and specificity for blood-based detection of colorectal cancer-a clinical biomarker discovery and validation study. *Clin Epigenetics*. 2019;11(1):158.
53. Pickhardt PJ. Emerging stool-based and blood-based non-invasive DNA tests for colorectal cancer screening: the importance of cancer prevention in addition to cancer detection. *Abdom Radiol*. 2016;41(8):1441-1444.
54. Barault L, Amatu A, Siravegna G, et al. Discovery of methylated circulating DNA biomarkers for comprehensive non-invasive monitoring of treatment response in metastatic colorectal cancer. *Gut*. 2018;67(11):1995-2005.
55. Moon S, Balch C, Park S, Lee J, Sung J, Nam S. Systematic inspection of the clinical relevance of TP53 missense mutations in gastric cancer. Research Support, Non-U.S. Gov't. *IEEE/ACM Trans Comput Biol Bioinform*. 2019;16(5):1693-1701.
56. Yang Q, Huo S, Sui Y, et al. Mutation status and immunohistochemical correlation of KRAS, NRAS, and BRAF in 260 Chinese Colorectal and Gastric Cancers. *Front Oncol*. 2018;8:487.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Cheng J, Guo Y, Guan G, et al. Two novel qualitative transcriptional signatures robustly applicable to non-research-oriented colorectal cancer samples with low-quality RNA. *J Cell Mol Med*. 2021;25:3622–3633. https://doi.org/10.1111/jcmm.16467