

RESEARCH ARTICLE

Open Access



# Quantitative or qualitative transcriptional diagnostic signatures? A case study for colorectal cancer

Qingzhou Guan<sup>1†</sup>, Haidan Yan<sup>1†</sup>, Yanhua Chen<sup>1†</sup>, Baotong Zheng<sup>1</sup>, Hao Cai<sup>1</sup>, Jun He<sup>1</sup>, Kai Song<sup>3</sup>, You Guo<sup>1,4</sup>, Lu Ao<sup>1</sup>, Huaping Liu<sup>1</sup>, Wenyuan Zhao<sup>3</sup>, Xianlong Wang<sup>1\*</sup> and Zheng Guo<sup>1,2,3\*</sup>

## Abstract

**Background:** Due to experimental batch effects, the application of a quantitative transcriptional signature for disease diagnoses commonly requires inter-sample data normalization, which would be hardly applicable under common clinical settings. Many cancers might have qualitative differences with the non-cancer states in the gene expression pattern. Therefore, it is reasonable to explore the power of qualitative diagnostic signatures which are robust against experimental batch effects and other random factors.

**Results:** Firstly, using data of technical replicate samples from the MicroArray Quality Control (MAQC) project, we demonstrated that the low-throughput PCR-based technologies also exist large measurement variations for gene expression even when the samples were measured in the same test site. Then, we demonstrated the critical limitation of low stability for classifiers based on quantitative transcriptional signatures in applications to individual samples through a case study using a support vector machine and a naïve Bayesian classifier to discriminate colorectal cancer tissues from normal tissues. To address this problem, we identified a signature consisting of three gene pairs for discriminating colorectal cancer tissues from non-cancer (normal and inflammatory bowel disease) tissues based on within-sample relative expression orderings (REOs) of these gene pairs. The signature was well verified using 22 independent datasets measured by different microarray and RNA\_seq platforms, obviating the need of inter-sample data normalization.

**Conclusions:** Subtle quantitative information of gene expression measurements tends to be unstable under current technical conditions, which will introduce uncertainty to clinical applications of the quantitative transcriptional diagnostic signatures. For diagnosis of disease states with qualitative transcriptional characteristics, the qualitative REO-based signatures could be robustly applied to individual samples measured by different platforms.

**Keywords:** Classifiers, Diagnostic signature, Relative expression orderings, Platform, Batch effects

## Background

In clinical, biopsy sampling with less-invasive techniques such as colonoscopy and endoscopic ultrasound-guided fine needle aspiration is often used for the initial clinical evaluation of cancer [1–6]. However, an indeterminate diagnosis often creates a dilemma [7]. Taking colorectal cancer as an example, it has been reported that the miss

rate of colorectal cancer after colonoscopy, which is the predominant screening and diagnostic test for colorectal cancer [2, 8, 9], is about 15% for patients with inflammatory bowel diseases (IBD) [2]. Thus, it is necessary to find a molecular biomarker as an auxiliary diagnostic method for colonoscopy.

With the wide application of high throughput gene expression profiling techniques, many classifiers based on quantitative transcriptional signatures for cancer subtyping [10–12] or early detection [13–17] have been developed. However, clinical applications of these transcriptional signatures are scarce due to technological,

\* Correspondence: wang.xianlong@139.com; guoz@ems.hrbmu.edu.cn

†Equal contributors

<sup>1</sup>Fujian Key Laboratory of Medical Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Fujian Medical University, Fuzhou 350122, China

Full list of author information is available at the end of the article



mathematical and translational barriers [18]. Besides factors such like tissue sampling [19] and sample preparation quality [20], a well-known factor is that gene expression data are often “noisy” and subject to lab and batch effects introduced by the differences in laboratory conditions and personnel [21–23]. As reported by the MicroArray Quality Control (MAQC) project [24], for the high-throughput microarray platforms, the median values of coefficient of variation (CV) of gene expression levels in replicate samples measured by the same platforms ranged from 5 to 15% within the same test sites and became 10 to 20% for replicate samples measured across different test sites. Similarly, as demonstrated in this study, the quantitative measurements of gene expression in replicate samples measured by the low-throughput PCR-based technologies, such as Standardized (Sta) RT-PCR™ Assays and TaqMan® Gene Expression Assays, also exist large variations even in the same test sites. The large variation of quantitative measurements will introduce uncertainty of such signatures in applications. Due to this problem, the application of classical classifiers based on quantitative transcriptional signatures requires data normalization. This means that the analysis of a single sample requires this sample to be normalized along with a set of samples measured together. This constraint makes the classifiers hardly applicable under common clinical settings. Especially for prognostic signatures, the risk score of a patient is dependent on the risk composition of the other samples adopted for normalization together, introducing substantial uncertainty for risk predication [25–27].

Notably, among the vast number of reported quantitative disease signatures, several signatures have been approved by the Food and Drug Administration (FDA). One of the FDA approved signatures is MammaPrint® for predicting the recurrence risk of early stage (I and II) breast cancer patients with lymph node negative and tumor size < 5.0 cm treated with surgical resection [28–30]. However, currently the tissue samples must be sent to one of the two Agendia laboratories (one in Amsterdam, The Netherlands, and the other in Irvine, CA) for measurement with strict quality control and data normalization, which greatly limits the wide application of the signature. Another FDA approved signature is AlloMap® [31] for identifying the probability of transplant rejection for heart transplant recipients, which also requires patients’ samples to be sent to a central laboratory (XDx reference laboratory, based in Brisbane, California) [31, 32]. The same problem exists in other transcriptional signatures incorporated into clinical recommendations and guidelines, such like the Oncotype DX genomic assay (Genomic Health, Inc. Redwood City, CA, USA) used for predicting recurrence risk of early stage breast cancer and in decision making with respect to systemic therapy [33]. Therefore, obviation of the impact

of the batch effects and the need of inter-sample normalization is an urgent issue.

In contrast, it has been found that the within-sample relative expression orderings (REOs) of gene pairs, which is also called Relative Expression Analysis (RXA) [34], are robust against experimental batch effects and invariant to monotone data transformation [34, 35]. Besides, the within-sample REOs of gene pairs are robust against variations of the tumor epithelial cell proportions in tissues sampled from different sites of a tumor [19, 36], partial RNA degradation in the sample preparation process and during the storage stage [20] and amplification bias for minimum specimens even with about 15–25 cancer cells [37], which are also important factors leading to the failure of validation and clinical application of the quantitative transcriptional signatures. The robustness property of the within-sample REOs enables researchers to integrate multiple datasets produced by the same or similar platforms for selecting disease signatures and training classifiers [20, 38, 39], which makes it more likely to find robust signatures [25, 38, 40]. Based on this unique advantage, some REO-type classifiers, such as TSP [41], K-TSP [42] and other adjusted methods [26, 43] were proposed to identify signatures for discriminating cancer subtypes [18, 38, 39, 44–46]. Recently, we have reported several REO-based prognostic signatures for specific medical issues for various cancers such as non-small cell lung cancer [25, 47], colorectal cancer [48] and other cancers [49–51], which have been well verified in multiple data sources produced by different laboratories, obviating the need of inter-sample data normalization. These results provide strong evidences of the clinical applicability of the type of signatures based on the robust qualitative REO information extracted from the quantitative measurements of gene expression, rather than the “exact” quantitative measurements themselves [52]. As revealed recently, although different platforms (e.g., Affymetrix and Illumina platforms) have different measurement principles, it would be highly likely that a REO-based signature consistently detected by two or more platforms could be robustly applied to samples measured by other platforms [53].

In this article, in addition to the previous results for the high-throughput platforms reported by the MicroArray Quality Control (MAQC) project [24], we firstly demonstrated that the quantitative values of gene expression in replicate samples measured by two low-throughput PCR-based technologies (StaRT-PCR™ Assays and TaqMan® Gene Expression Assays) in the same test site also exist large variations. Then, through a case study of building a support vector machine (SVM) and a naïve Bayesian classifier for discriminating colorectal cancer samples from normal samples, we demonstrated that

the classical classifiers based on quantitative transcriptional signatures cannot be robustly applied to independent samples measured by the same platform used for the training data, let alone the samples measured by different platforms, which makes this type of signatures being hardly applicable under clinical settings. Then, we developed a within-sample REO-based signature that could discriminate colorectal cancer from non-cancer samples (IBD and normal samples) without the need of inter-sample data normalization or experimental batch adjustment. The signature was validated using data from multiple sources measured by different laboratories with different platforms.

## Results

### Technical variations of quantitative measurement

Firstly, we evaluated the CV of gene expression measurements in replicates for sample A and sample B measured in the same test site by two PCR-based technologies, StaRT-PCR™ Assays and TaqMan® Assays, respectively.

For a total of 199 genes with non-zero measurements assayed by StaRT-PCR™ for 3 replicates of sample A, about 32.7% genes showed at least 10% CV and 15.1% genes showed at least 15% CV. Similarly, for a total of 195 genes with non-zero measurements assayed by StaRT-PCR™ for 3 replicates of sample B, about 34.4% genes showed at least 10% CV and 17.4% genes showed at least 15% CV, the results were also shown in Fig. 1.

For a total of 964 genes with non-zero measurements assayed by TaqMan® for sample A, about 13.1% genes showed at least 10% CV and 7.8% genes showed at least

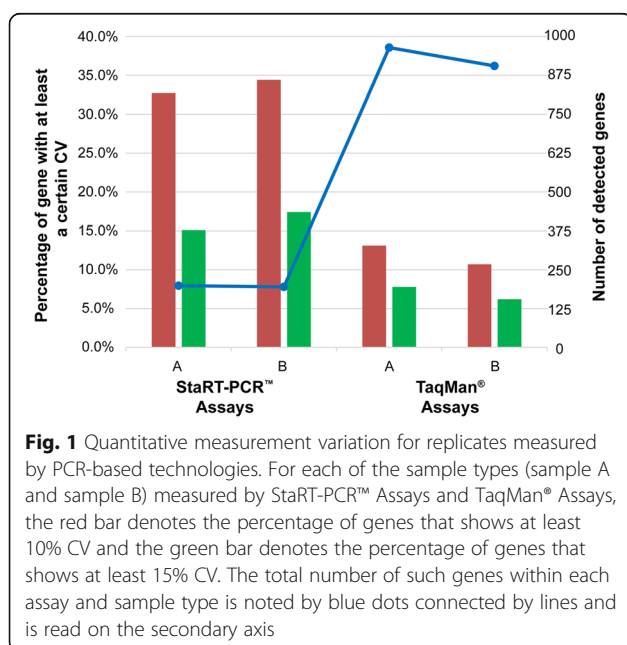
15% CV. Similarly, for a total of 905 genes with non-zero measurements assayed by TaqMan® for sample B, about 10.7% genes showed at least 10% CV and 6.2% genes showed at least 15% CV, as shown in Fig. 1. Although TaqMan® Assays showed smaller variations than StaRT-PCR™ Assays, the variations were still not negligible even in samples measured in the same test site, and it could expect that the variations would increase for measurements from different test sites.

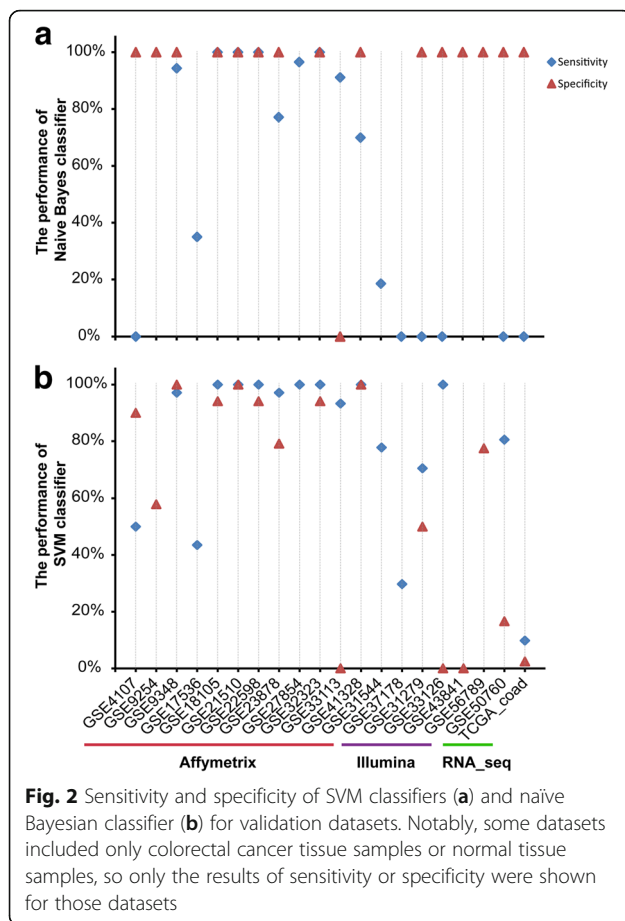
### Limitation of classifiers based on quantitative transcriptional signatures

Due to large experimental batch effects, quantitative transcriptional measurement data from different experiments or profiled with different platforms could not be directly put together to train traditional SVM and naïve Bayesian classifiers. Because we could not find a single dataset with sufficient samples for colorectal cancer, normal and IBD tissues simultaneously, we were unable to train SVM and naïve Bayesian classifiers based on quantitative measurements for discriminating colorectal cancer and non-cancer (normal or IBD) tissue samples. Thus, we constructed the SVM and naïve Bayesian classifiers for a simpler problem, discriminating colorectal cancer and normal tissue samples, to demonstrate the limitations of quantitative transcriptional signatures.

Between the 32 cancer samples and 32 normal samples from dataset GSE8671, 7028 differentially expressed genes were detected using Student's *t*-test with 1% FDR control. Using these 7028 genes as feature genes, a SVM classifier with radial basis function (RBF) kernel was trained with tenfold cross-validation [54, 55] using the training dataset GSE20916 with 91 cancer and 44 normal tissue samples. The sensitivity and specificity of the SVM classifier were 98.9% and 100.0% in the training dataset, respectively. However, when tested by validation datasets without applying inter-sample normalization, the classifier failed badly in many cases as described in Fig. 2a and Additional file 1: Table S1. For example, only 35.0% of the 177 cancer samples from the dataset GSE17536 were correctly classified and none of the 12 cancer samples from the dataset GSE4107 were correctly classified. Both the datasets were measured by the same Affymetrix platform with the training dataset. When the SVM classifier was applied to the datasets measured by other platforms, none of the 365 cancer samples from three datasets (GSE31279 measured by the Illumina platform; GSE50760 and TCGA measured by the RNA\_seq platform) were correctly classified. Similar results were also observed for the naïve Bayesian classifier, as shown in Fig. 2b and Additional file 1: Table S1.

More comprehensive evaluation results were shown in Supplementary Result and Additional file 1: Table S2-S6. These results clearly show that the classical classifiers





based on the quantitative transcriptional signatures cannot be robustly applied to independent samples even measured by the same platform as the training datasets, let alone the samples measured by different platforms. This problem limits the applicability of these classifiers to clinical applications.

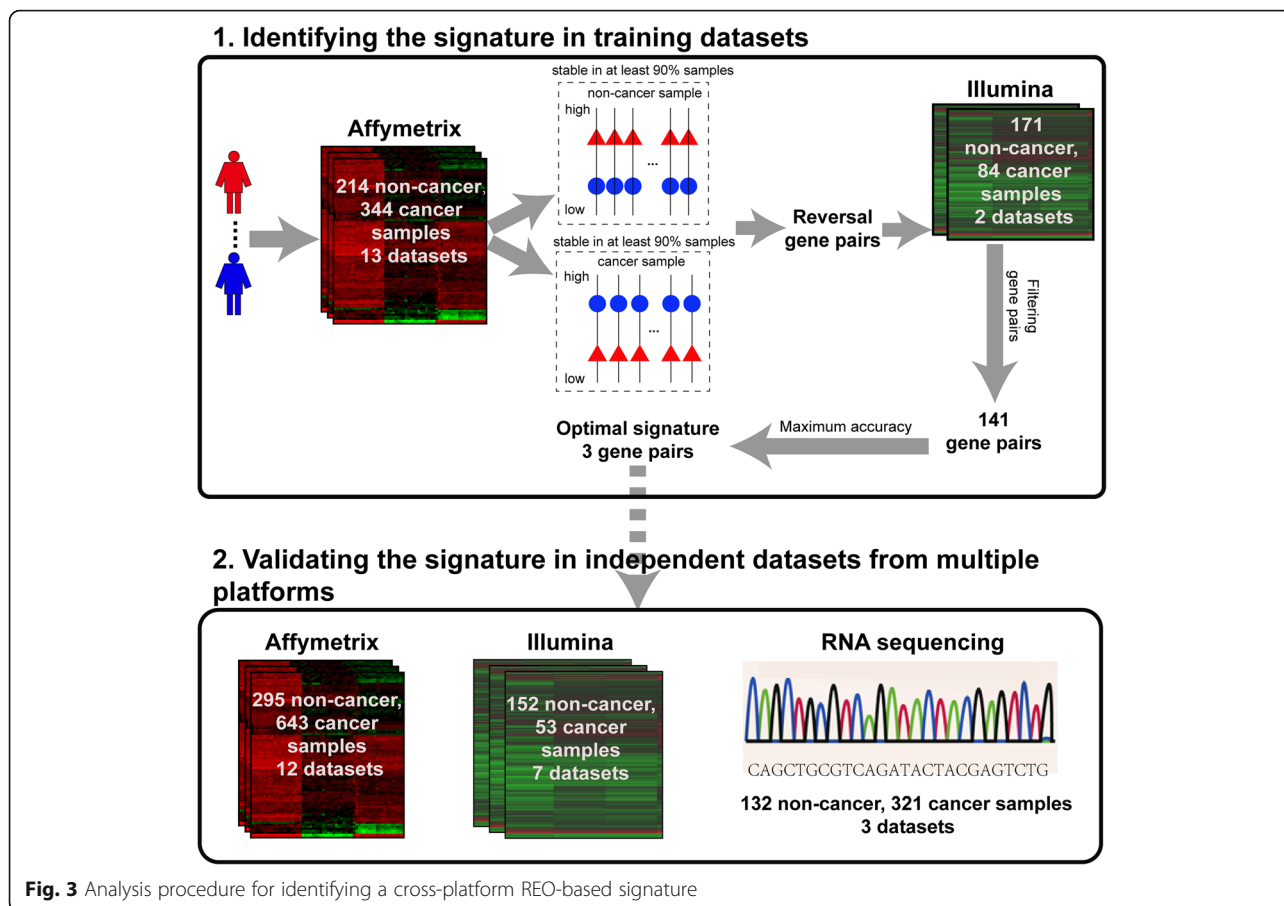
#### Identification and application of REO-based signature

The analysis procedure is described in Fig. 3. Firstly, using 91 normal samples and 123 IBD samples measured by the Affymetrix platform collected from 11 datasets (see Table 2), we identified 144,090,213 gene pairs with identical REOs in at least 90% of both the normal samples and the IBD samples. Similarly, using 344 colorectal cancer tissue samples from 9 datasets measured by the Affymetrix platform (see Table 2), we identified 149,446,895 gene pairs with identical REOs in at least 90% of the cancer tissues. We found 843 gene pairs that have reversal REOs from the above two lists of gene pairs. Among these 843 gene pairs, we further selected 141 gene pairs that had the identical REOs in at least 90% of 171 non-cancer samples and reversed REOs in at least 90% of 84 cancer samples in the combined

GSE48634 and GSE37178 datasets measured by the Illumina platform, the list of the 141 gene pairs were shown in Additional file 1: Table S7. These 141 gene pairs were sorted in a descending order according to their reversal coverage rates (see Methods) between all the cancer samples and all the non-cancer samples in the training data collected from 13 datasets measured by Affymetrix platform (see Table 2). We then used the top-ranked  $k$  pairs, where  $k$  is an odd integer, to classify samples according to the majority vote rule. The results showed that for all possible  $k$  values ranging from 1 to 141, the largest geometric mean of sensitivity and specificity was 94.8% when  $k = 3$  (Fig. 4). Thus, these three gene pairs, as described in Table 1, were selected as the signature for discriminating colorectal cancer samples from non-cancer samples.

The performance of the signature was evaluated using independent test datasets measured by multiple different platforms. As shown in Fig. 5 and Additional file 1: Table S8, the performance of the signature in each of the 12 datasets measured by the Affymetrix GPL570 platform is excellent. In total, 98.3% of the 643 colorectal cancer samples and 96.6% of the 295 non-cancer samples were identified correctly. Similar results were observed for the independent test datasets measured by the Illumina platforms, as shown in Fig. 5 and Additional file 1: Table S8. Especially, the signature was also verified in the datasets measured by the RNA sequencing platforms which have no data used in obtaining the signature. For the TCGA dataset, 97.9% of the 285 colorectal cancer samples and 97.6% of the 41 normal colorectal samples were identified correctly. For the GSE72819 dataset which did not include colorectal cancer samples, 94.5% of the 73 non-cancer tissue samples were correctly identified. The above results indicate that the classifier based on the within-sample REOs of gene pairs can be applied to the analysis of individual samples measured by different platforms, obviating the need of inter-sample data normalization.

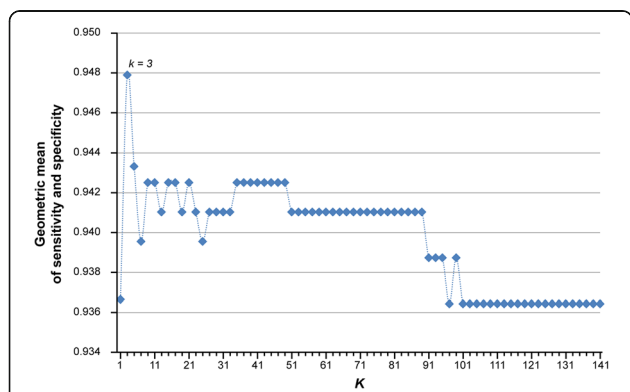
Moreover, to explore the generalization of the signature, we used all the possible top-ranked  $k$  (where  $k$  is an odd integer) pairs from the 141 gene pairs to classify samples according to the majority vote rule. With different top-rank  $k$ , similar performances were achieved in the validation datasets, as shown in Additional file 2: Figure S1. However, for the dataset GSE68570, the classification performance decreased slightly when  $k$  increased to 77 or larger. The possible reason of the decreased performance for GSE68570 should be that the gene pairs with relatively low reversal rates in the training data might be unstable in data measured by other platforms [53]. In general, the generalization of the signature with three gene pairs is good enough.



**Discussion**

We demonstrated that, besides high-throughput gene expression profiling platforms, the low-throughput PCR-based quantitative measurements also exist large variation in replicate samples measured in the same or different test sites. Thus, the classifiers based on quantitative transcriptional signatures could not be robustly applied to individual samples measured by the

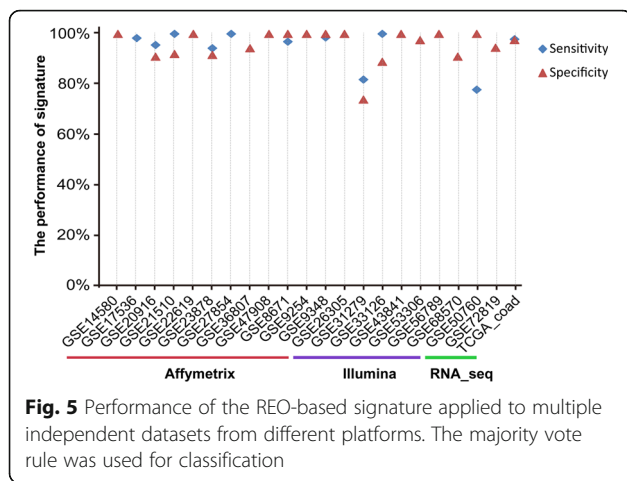
same platform as the training samples, let alone those individual samples measured by different platforms. This could explain the problem mentioned in Introduction that some quantitative transcriptional signatures approved by FDA or incorporated into clinical guidelines must be sent to a central laboratory for measurement with strict quality control and data normalization. Besides the batch effects, the quantitative measurements of gene expression are commonly affected by partial RNA degradation [20] and different sampling sites of tumor for the same patient [19], which will increase the uncertainty for clinical applications of quantitative transcriptional diagnostic signatures.



**Table 1** The REO-based signature

Gene pair	REO ( $G_i > G_j$ ) <sup>a</sup>
1	<i>GPAT3</i> > <i>TRIP13</i>
2	<i>PYY</i> > <i>CKAP2</i>
3	<i>SDCBP2</i> > <i>DAP3</i>

Note:  
<sup>a</sup>Relative expression ordering (REO) of a gene pair,  $G_i > G_j$  denotes that the expression value of gene  $i$  is larger than the expression value of gene  $j$  in 90% of non-cancer samples but is less than the expression value of gene  $j$  in 90% of colorectal cancer samples



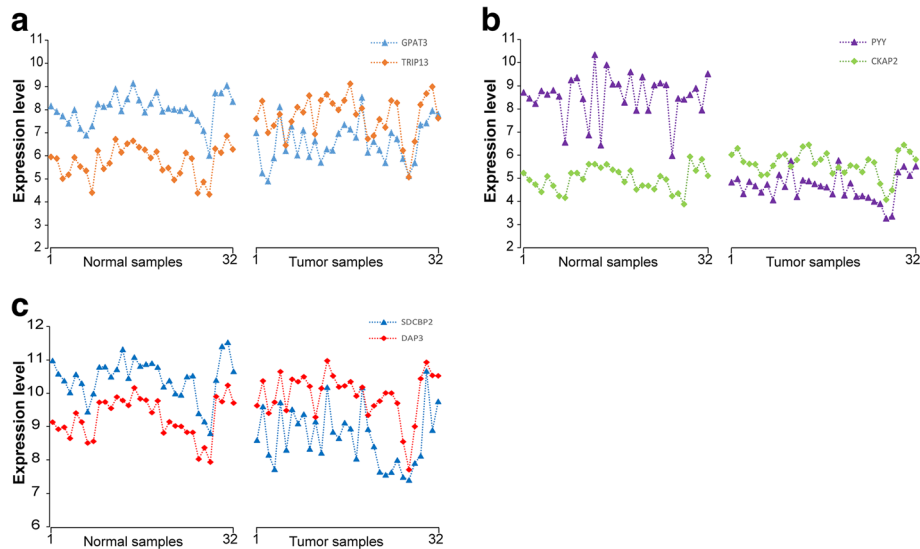
Fortunately, as demonstrated in our previous studies [19, 20, 36] and in this study, the REO-based transcriptional signatures could circumvent the above-mentioned problems. As a case study, we identified a signature consisting of three gene pairs for discriminating colorectal cancer from non-cancer (normal and IBD) tissue samples based on the within-sample REOs of the gene pairs. The result showed that the REO-based signature obtained from samples measured by two different platforms could be robustly applied to classify individual samples measured by multiple different platforms, including the RNA\_seq platform that did not participate in the training process. However, in the GSE31279 dataset measured by the Illumina GPL6104 platform which did not participate in the training process, the signature performed relatively poor: only 81.8% of the 44 cancer samples and 73.8% of the 42 normal samples from were correctly identified. Although the within-sample REOs tend to be rather robust to data measured by different platforms, a certain degree of uncertainty still exists due to different measurement principles of the platforms [53]. Ideally, a REO-based signature should be applied to data measured by the platforms participating the train and validation of the signature.

Even with sufficient high-quality data, it is difficult to interpret the signature used in complex classifiers to gain biological insights about the biomarkers [18]. In contrast, we can readily gain biological insights for a signature consisting of only a few genes. The three gene pairs of the signature for colorectal cancer diagnosis consist of *GPAT3* and *TRIP13*, *PYY* and *CKAP2*, *SDCBP2* and *DAP3*. These genes were found in the differentially expressed genes (Student’s *t*-test, *FDR* < 0.01) detected between the 32 cancer samples and 32 normal samples in the GSE8671 dataset. For *GPAT3-TRIP13* gene pair, both up-regulation of *TRIP13* and down-regulation of *GPAT3* contribute to the reversal REO in

colorectal cancer samples. Similarly, for *PYY-CKAP2* and *SDCBP2-DAP3* gene pairs, up-regulation of *CKAP2*, *DAP3* and down-regulation of *PYY*, *SDCBP2* contribute to the reversal REO in colorectal cancer samples. Some of these genes, such as *TRIP13* [56], *PYY* [57], are known to be cancer-associated. *TRIP13* is a novel mitotic checkpoint-silencing protein, whose overexpression is associated with poor prognosis in breast cancer patients [56, 58, 59]. The decreased expression of *PYY* may be relevant to the development and progression of colon adenocarcinoma [57]. We additionally showed the distribution of the expression level of the 6 genes in dataset GSE8671. As shown in Fig. 6, the fold changes of each signature gene pair across samples for the two phenotypes were quite different. For the *GPAT3 - TRIP13* gene pair, as shown in Fig. 6a, the fold change of the expression levels between *GPAT3* and *TRIP13* took values ranging from 1.26 to 1.72 with the median of 1.38 in the normal samples, while in the tumor samples the fold change took values ranging from 0.63 to 1.06 with the median of 0.87. Similar results for the other two gene pairs, *PYY - CKAP2* and *SDCBP2 - DAP3*, were shown in Fig. 6b and c, respectively. The above results showed that the fold changes of each signature gene pair are quite different across different samples for each of the two phenotypes but the relative expression levels of the gene pair are stably.

The REO-based method is based on a single binary “switch” that compares the ordering of expression between two genes. The simplicity does not necessarily limit its prediction performance and the method is not prone to the overfitting issue. Arguably, REO-based signatures may lose some subtle quantitative information on gene expression. However, considering that subtle quantitative information of gene expression measurements tends to be unreliable and even the ratios of expression values of gene pairs are affected by the batch effects [25, 60], the apparent disadvantage of REOs analysis is in fact a unique advantage in terms of robustness [20]. The REO-based signature identified for colorectal cancer obviates the need of data normalization, which makes it feasible to clinical settings for colorectal cancer diagnosis and surveillance of patients with long-term IBD using biopsies obtained by colonoscopy or other improved techniques [61–65]. Notably, we have applied both the tissue samples and biopsy samples for training and validation. Thus, the signature based on the REOs is suitable for tissue samples and biopsy samples [10].

The main purpose of this study is to systematically demonstrate the critical limitations of the traditional classifiers based on the quantitative transcriptional measurements, which are sensitive to batch effects and detection platforms and could not be applied directly to the data measured by different laboratories. As for the



**Fig. 6** The distribution of the expression levels of the 3 gene-pairs in GSE8671. The gene expression levels of *GPAT3* and *TRIP13* (a), *PYY* and *CKAP2* (b) and *SDCBP2* and *DAP3* (c)

REO-based method, other approaches, such as TSP and k-TSP, could be applied to the data measured by different laboratories or platforms. Here, we additionally evaluated other rank based approaches using the same training and validation datasets. Using the *tspair* R package (version 3.3.3), we trained the TSP classifier in the training samples directly combined from data measured by the Affymetrix and Illumina platforms. In the training set, 97.0% of the 428 cancer samples and 94.3% of the 385 non-cancer sample were correctly identified. However, the classifier failed badly in many validation datasets as described in Additional file 1: Table S9. Using the *ktspair* R package (version 3.3.3), we also trained the k-TSP classifier. In the training set, with the default five-fold cross-validation, 5 gene pairs were selected as the classification signature which correctly identified 96.7% of the 428 cancer samples and 98.0% of the 385 non-cancer sample. In the validation data, the k-TSP classifier performed better than the TSP classifier but poorer than our signature, as shown in Additional file 1: Table S10. For example, for the dataset GSE23878, our REO signature could identify 91.7% of the 24 non-cancer sample correctly, but the k-TSP signature identified only 41.7% non-cancer samples correctly. One possible reason should be that the difference in the proportion of samples from Affymetrix and Illumina platform will make the signature to be unable to characterize the common features of the two platforms but biased to the platform with larger samples. Other approaches such as CART [66] should have the same problem. In the training process for our REO signature, the gene pairs (141 gene pairs) that were consistently detected in the data produced by the two

platforms were used for the final signature selection (3 gene pairs in this study). Therefore, our method is intuitive and simple with the ability to identify very robust disease signatures.

In conclusion, REO-based signatures circumvent the critical limitation of quantitative transcriptional signatures and the REO-based classifying method should be also applicable for classifying other tissue samples. Moreover, because the data normalization problem also exists in miRNA [67] and DNA methylation profile analyses, the REO-based analysis of these multi-omic data should be taken into account in the further study.

## Conclusions

Because the subtle quantitative information of gene expression measurements currently tends to be greatly affected by many random factors, the disease diagnostic signatures based on the quantitative measurements lack robustness for clinical applications. Thus, we should make more efforts to capture the qualitative differences of gene expression patterns between cancer and non-cancer and between cancer subtypes to exploit robust qualitative signatures for disease diagnosis.

## Methods

### Data and preprocessing

The gene expression profiles analyzed in this study are described in Table 2. The array-based data measured by the Affymetrix and Illumina platforms were downloaded from Gene Expression Omnibus [68] (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) and the mRNA-seq data measured by the Illumina platform were downloaded from

**Table 2** Data used in this study

GEO Acc	Platform	Sample size <sup>a</sup>		
		Normal	IBD	Tumor
Training				
GSE32323	Affymetrix GPL570	17		17
GSE22598	Affymetrix GPL570	17		17
GSE41328	Affymetrix GPL570	10		10
GSE4107	Affymetrix GPL570	10		12
GSE4183	Affymetrix GPL570	8	15	15
GSE18105	Affymetrix GPL570	17		94
GSE12251	Affymetrix GPL570		23	
GSE13367	Affymetrix GPL570		16	
GSE9452	Affymetrix GPL570		8	
GSE16879	Affymetrix GPL570	6	61	
GSE35144	Affymetrix GPL570			27
GSE35896	Affymetrix GPL570			62
GSE33113	Affymetrix GPL570	6		90
GSE37178	Illumina GPL6947			84
GSE48634	Illumina GPL10558	69	102	
Validation				
GSE9348	Affymetrix GPL570	12		70
GSE23878	Affymetrix GPL570	24		35
GSE47908	Affymetrix GPL570	15	39	
GSE36807	Affymetrix GPL570	7	28	
GSE27854	Affymetrix GPL570			115
GSE22619	Affymetrix GPL570	10	10	
GSE21510	Affymetrix GPL570	25		123
GSE17536	Affymetrix GPL570			177
GSE14580	Affymetrix GPL570	6	24	
GSE8671	Affymetrix GPL570	32		32
GSE9254	Affymetrix GPL570	19		
GSE20916	Affymetrix GPL570	44		91
GSE53306	Illumina GPL10558	12	28	
GSE31279	Illumina GPL6104	42		44
GSE33126	Illumina GPL6947	9		9
GSE68570	Illumina GPL10558	5	6	
GSE26305	Illumina GPL6884	2	2	
GSE56789	Illumina GPL10558	40		
GSE43841	Illumina GPL14951	6		
GSE50760 <sup>b</sup>	Illumina GPL11154	18		36
GSE72819 <sup>b</sup>	Illumina GPL11154		73	
TCGA_coad <sup>b,c</sup>	IlluminaHiSeq_RNASeqV2	41		285

**Notes:**<sup>a</sup>Empty cells indicate that there is no sample in the corresponding category<sup>b</sup>These samples are measured by the RNA-sequencing platform<sup>c</sup>Denotes the colorectal adenocarcinoma sample from TCGA

ArrayExpress [69] (<http://www.ebi.ac.uk/arrayexpress/>) and The Cancer Genome Atlas [70] (TCGA, <http://cancergenome.nih.gov/>).

For the data measured by the Affymetrix platform, we downloaded the raw mRNA expression data (.CEL files) and used the Robust Multi-array Average (RMA) algorithm for background adjustment without quantile normalization [71]. For the data measured by the Illumina platform, we directly downloaded the processed data. For the sequence-based data from TCGA, we directly downloaded the level 3 data measured by the UNC IlluminaHiSeq\_RNASeqV2 platform.

For the array-based data, each probe ID was mapped to Entrez gene ID with the corresponding platform file. If a probe was mapped to multiple or zero genes, then the data of this probe were deleted. If multiple probes were mapped to the same gene, the expression value of the gene was defined as the arithmetic mean of the values of multiple probes. For the sequence-based data from ArrayExpress, the gene symbols were mapped to Entrez gene ID with the biological database network [72] (bioDBnet, <https://biobdbnet-abcc.ncifcrf.gov/db/db2db.php>).

**Variation analysis of quantitative measurement**

In the MicroArray Quality Control (MAQC) project, two commercially available Reference RNA samples (sample A and sample B) with multiple replicates were measured by multiple microarray platforms and PCR-based technologies [24]. The MAQC project has reported the large measurement variations of the high-throughput microarray platforms [24]. Here, we additionally analyzed the variations of quantitative gene expression levels measured by two PCR-based technologies, Standardized (Sta) StaRT-PCR™ and TaqMan® Gene Expression Assays.

The MAQC PCR-based data, as described in Table 3, were directly downloaded from GSE5350. Notably, for the 3 replicates of sample A and sample B measured by StaRT-PCR™ Assays. If the measurement of a gene was 0 or “nan” in at least one replicate of a sample, then this gene was not included for further analysis. Thus, the total number of genes was not identical for sample A and sample B. For 4 replicates of sample A or sample B measured by TaqMan® Assay, a gene was considered absent in a sample when the average cycle threshold (CT) exceeds 35 [24]. For sample A or sample B, if the measurement of a gene was absent in at least one replicate, this gene was not included for the further analysis. Thus,

**Table 3** MAQC PCR-based data used in this study

GEO Acc	Protocol	Platform	Sample A	Sample B
GSE5350	StaRT-PCR™ Assays	GPL4198	3	3
GSE5350	TaqMan® Assays	GPL4097	4	4



the total number of genes for sample A and sample B obtained from TaqMan® Assays was also not identical.

For the gene expression levels of a certain gene in the replicates for sample A or sample B measured by each platform, the coefficient of variation (CV), calculated as the ratio of the standard deviation and arithmetic mean for the expression levels of this gene in the replicates, is used to measure the degree of variation of quantitative measurements. For the sample A and sample B measured by each platform, we calculated the percentage of genes that shows at least 10% and 15% CV, respectively, to reveal the degree of variation or uncertainty of quantitative measurements.

### SVM and naïve Bayesian classifiers

The SVM classifier using radial basis function (RBF) kernel [55] and the naïve Bayesian classifier, implanted in the WEKA software (version 3–6-13) with the default settings [54], were used for the case study. Each of the classifiers was trained with tenfold cross-validation in the training data. The performance of a trained classifier was evaluated in multiple independent data with or without normalization.

We called cancer samples as positive samples, non-cancer samples, either normal or IBD, as negative samples, and evaluated the performance of the classification signature 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 denote the number of true positives, true negatives, false positives and false negatives, respectively.

### Identification of the REO-based diagnosis signature

First, in the training dataset, each gene measurement is converted to its rank within each sample (the smallest measurement corresponding to the minimum rank, and the largest measurement corresponding to the maximum rank). Then, pairwise comparisons are performed for all genes to identify gene pairs with stable ordering in samples for a particular tissue type. For a pair of genes ( $i, j$ ), the relationship of their relative ranks,  $G_i$  and  $G_j$ , within one sample, has only two possibilities,  $G_i > G_j$  or  $G_i < G_j$ . The relationship is called the relative expression ordering (REO). If the same REO pattern is maintained in a majority of samples, e.g. 90%, it is called a highly stable REO and the pair is a highly stable gene pair. Furthermore, if a gene pair ( $i, j$ ) is highly stable in both a group of non-cancer samples and a group of cancer samples,

respectively, but with reversal REO patterns ( $G_i < G_j$  in one group but  $G_i > G_j$  in the other group), the pair is called a reversal gene pair. Here, we selected the reversal gene pairs which are highly stable in non-cancer samples and cancer samples, respectively, but the REO patterns are reversed in the latter group. They form the candidate REO signature of the cancer.

Then, the candidate REO signatures selected above were sorted in a descending order according to their reversal coverage rates, where the reversal coverage rate of a reversal gene pair is defined as the geometric mean of the percentage of the highly stable REO pattern in the non-cancer samples and the percentage of the reversed REO pattern in the cancer samples. Obviously, the higher the reversal coverage rate is for a gene pair, the higher the classification ability is for this gene pair.

Thirdly we used the top  $k$  gene pairs, where  $k$  is an odd integer 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 as the smallest number of gene pairs that reached the highest geometric mean of the sensitivity and specificity in the training data.

Finally, the signature was tested in independent samples.

### Additional files

**Additional file 1:** Supplementary results for this manuscript. (DOCX 57 kb)

**Additional file 2: Figure S1.** Performances of all possible top-ranked  $k$  (from 1 to 141,  $k$  is an odd integer) gene pairs in the independent datasets measured by different platforms. The majority vote rule was used for classification. (TIFF 1848 kb)

### Abbreviations

CT: Cycle threshold; CV: Coefficient of variation; GEO: Gene Expression Omnibus; IBD: Inflammatory bowel diseases; MAQC: MicroArray Quality Control; RBF: Radial basis function; REO: Relative expression ordering; RMA: Robust Multi-array Average; RXA: Relative Expression Analysis; SVM: Support vector machine; TCGA: The Cancer Genome Atlas

### Acknowledgements

Not applicable.

### Funding

This work was supported by the National Natural Science Foundation of China [grant numbers. 81372213, 81572935, 21534008, 81602738, 61601151] and the Joint Scientific and Technology Innovation Fund of Fujian Province [grant numbers. 2016Y9044].

### Availability of data and materials

All data analyzed in this study were downloaded from the public database: GEO, ArrayExpress and TCGA.

### Authors' contributions

QZG, HDY and YHC conceived the study, analysed the data, made figures, performed the statistical analysis, and drafted the manuscript. BTZ, HC, JH, KS, YG and HPL searched the data and participated in the statistical analysis. LA and WYZ participated in discussing and revising the manuscript. ZG and XLW conceived of the study, and participated in its design and coordination, helped to draft the manuscript and supervised the work. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare no conflict of interest.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Author details**

<sup>1</sup>Fujian Key Laboratory of Medical Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Fujian Medical University, Fuzhou 350122, China. <sup>2</sup>Fujian Key Laboratory of Tumor Microbiology, Fujian Medical University, Fuzhou 350122, China. <sup>3</sup>College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China. <sup>4</sup>Department of Preventive Medicine, School of Basic Medicine Sciences, Gannan Medical University, Ganzhou 341000, China.

Received: 26 September 2017 Accepted: 11 January 2018

Published online: 29 January 2018

**References**

- Brawley OW, Flenaugh EL. Low-dose spiral CT screening and evaluation of the solitary pulmonary nodule. *Oncology (Williston Park)*. 2014;28(5):441–6.
- Wang YR, Cangemi JR, Loftus EV, Jr., Picco MF: 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(3):444–449.
- Fusco V, Ebert B, Weber-Eibel J, Jost C, Fleige B, Stolte M, Oberhuber G, Rinneberg H, Lochs H, Ortner M. Cancer prevention in ulcerative colitis: long-term outcome following fluorescence-guided colonoscopy. *Inflamm Bowel Dis*. 2012;18(3):489–95.
- European Colorectal Cancer Screening Guidelines Working Group, von Karsa L, Patnick J, Segnan N, Atkin W, Halloran S, Lansdorp-Vogelaar I, Malila N, Minozzi S, Moss S, et al. European guidelines for quality assurance in colorectal cancer screening and diagnosis: overview and introduction to the full supplement publication. *Endoscopy*. 2013;45(1):51–9.
- Kaminski MF, Polkowski M, Kraszewska E, Rupinski M, Butruk E, Regula J. A score to estimate the likelihood of detecting advanced colorectal neoplasia at colonoscopy. *Gut*. 2014;63(7):1112–9.
- Lastra RR, Pramick MR, Crammer CJ, LiVolsi VA, Baloch ZW. Implications of a suspicious afirma test result in thyroid fine-needle aspiration cytology: an institutional experience. *Cancer Cytopathol*. 2014;122(10):737–44.
- Ahmed A, VandenBussche CJ, Ali SZ, Olson MT. The dilemma of "indeterminate" interpretations of pancreatic neuroendocrine tumors on fine needle aspiration. *Diagn Cytopathol*. 2016;44(1):10–3.
- Gross CP, Andersen MS, Krumholz HM, McAvay GJ, Proctor D, Tinetti ME. Relation between Medicare screening reimbursement and stage at diagnosis for older patients with colon cancer. *JAMA*. 2006;296(23):2815–22.
- Rex DK, Johnson DA, Anderson JC, Schoenfeld PS, Burke CA, Inadomi JM. American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. *Am J Gastroenterol*. 2009;104(3):739–50.
- Price ND, Trent J, El-Naggar AK, Cogdell D, Taylor E, Hunt KK, Pollock RE, Hood L, Shmulevich I, Zhang W. Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. *Proc Natl Acad Sci U S A*. 2007;104(9):3414–9.
- Lockwood WW, Wilson IM, Coe BP, Chari R, Pikor LA, Thu KL, Solis LM, Nunez MI, Behrens C, Yee J, et al. Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development. *PLoS One*. 2012;7(5):e37775.
- Zhang A, Wang C, Wang S, Li L, Liu Z, Tian S. Visualization-aided classification ensembles discriminate lung adenocarcinoma and squamous cell carcinoma samples using their gene expression profiles. *PLoS One*. 2014;9(10):e110052.
- Yang Z, Zhuan B, Yan Y, Jiang S, Wang T. Identification of gene markers in the development of smoking-induced lung cancer. *Gene*. 2016;576(1 Pt 3):451–7.
- Gesthalter YB, Vick J, Steiling K, Spira A. Translating the transcriptome into tools for the early detection and prevention of lung cancer. *Thorax*. 2015;70(5):476–81.
- Rossi ED, Larocca LM, Fadda G. Can a gene-expression classifier with high negative predictive value solve the indeterminate thyroid fine-needle aspiration dilemma? *Cancer Cytopathol*. 2013;121(7):403.
- Tomei S, Marchetti I, Zavaglia K, Lessi F, Apollo A, Aretini P, Di Coscio G, Bevilacqua G, Mazzanti C. A molecular computational model improves the preoperative diagnosis of thyroid nodules. *BMC Cancer*. 2012;12:396.
- Panebianco F, Mazzanti C, Tomei S, Aretini P, Franceschi S, Lessi F, Di Coscio G, Bevilacqua G, Marchetti I. The combination of four molecular markers improves thyroid cancer cytologic diagnosis and patient management. *BMC Cancer*. 2015;15:918.
- Winslow RL, Trayanova N, Geman D, Miller ML: Computational medicine: translating models to clinical care. *Sci Transl Med* 2012, 4(158):158rv111.
- Cheng J, Guo Y, Gao Q, Li H, Yan H, Li M, Cai H, Zheng W, Li X, Jiang W, et al. Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites. *Oncotarget*. 2017;8(18):30265–75.
- Chen R, Guan Q, Cheng J, He J, Liu H, Cai H, Hong G, Zhang J, Li N, Ao L, et al. Robust transcriptional tumor signatures applicable to both formalin-fixed paraffin-embedded and fresh-frozen samples. *Oncotarget*. 2016;
- Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11(10):733–9.
- Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27.
- 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.
- MAQC Consortium, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES et al: The MicroArray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006, 24(9):1151–1161.
- Qi L, Chen L, Li Y, Qin Y, Pan R, Zhao W, Gu Y, Wang H, Wang R, Chen X, 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):233–42.
- Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst*. 2015;107(1):357.
- Patil P, Bachant-Winner PO, Haibe-Kains B, Leek JT. Test set bias affects reproducibility of gene signatures. *Bioinformatics*. 2015;31(14):2318–23.
- Sapino A, Roepman P, Linn SC, Snel MH, Delahaye LJ, van den Akker J, Glas AM, Simon IM, Barth N, de Snoo FA, et al. MammaPrint molecular diagnostics on formalin-fixed, paraffin-embedded tissue. *J Mol Diagn*. 2014;16(2):190–7.
- Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, Pierga JY, Brain E, Causeret S, DeLorenzi M, et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N Engl J Med*. 2016;375(8):717–29.
- Bueno-de-Mesquita JM, van Harten WH, Retel VP, van't Veer LJ, van Dam FS, Karsenberg K, Douma KF, van Tinteren H, Peterse JL, Wesseling J, et al. Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). *Lancet Oncol*. 2007;8(12):1079–87.
- Pham MX, Teuteberg JJ, Kfoury AG, Starling RC, Deng MC, Cappola TP, Kao A, Anderson AS, Cotts WG, Ewald GA, et al. Gene-expression profiling for rejection surveillance after cardiac transplantation. *N Engl J Med*. 2010;362(20):1890–900.
- Pham MX, Deng MC, Kfoury AG, Teuteberg JJ, Starling RC, Valentine H. Molecular testing for long-term rejection surveillance in heart transplant recipients: design of the invasive monitoring attenuation through gene expression (IMAGE) trial. *J Heart Lung Transplant*. 2007;26(8):808–14.
- McVeigh TP, Kerin MJ. Clinical use of the Oncotype DX genomic test to guide treatment decisions for patients with invasive breast cancer. *Breast Cancer (Dove Med Press)*. 2017;9:393–400.

34. Eddy JA, Sung J, Geman D, Price ND. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technol Cancer Res Treat*. 2010;9(2):149–59.
35. Wang H, Zhang H, Dai Z, Chen MS, Yuan Z. TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection. *BMC Med Genet*. 2013;6(Suppl 1):S3.
36. Cheng J, Guo Y, Gao Q, Li H, Yan H, Li M, Cai H, Zheng W, Li X, Jiang W, et al. Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites. 2017.
37. Liu H, Li Y, He J, Guan Q, Chen R, Yan H, Zheng W, Song K, Cai H, Guo Y, et al. Robust transcriptional signatures for low-input RNA samples based on relative expression orderings. *BMC Genomics*. 2017;18(1):913.
38. Xu L, Tan AC, Winslow RL, Geman D. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics*. 2008;9:125.
39. Xu L, Tan AC, Naiman DQ, Geman D, Winslow RL. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*. 2005;21(20):3905–11.
40. Yasrebi H, Sperisen P, Praz V, Bucher P. Can survival prediction be improved by merging gene expression data sets? *PLoS One*. 2009;4(10):e7431.
41. 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
42. 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):3896–904.
43. Li H, Hong G, Xu H, Guo Z. Application of the rank-based method to DNA methylation for cancer diagnosis. *Gene*. 2015;555(2):203–7.
44. Xu L, Geman D, Winslow RL. Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics*. 2007;8:275.
45. Zhao H, Logothetis CJ, Gorlov IP. Usefulness of the top-scoring pairs of genes for prediction of prostate cancer progression. *Prostate Cancer Prostatic Dis*. 2010;13(3):252–9.
46. Patnaik SK, Kannisto E, Knudsen S, Yendamuri S. Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non-small cell lung cancer after surgical resection. *Cancer Res*. 2010;70(1):36–45.
47. Qi L, Li Y, Qin Y, Shi G, Li T, Wang J, Chen L, Gu Y, Zhao W, Guo Z. An individualised signature for predicting response with concordant survival benefit for lung adenocarcinoma patients receiving platinum-based chemotherapy. *Br J Cancer*. 2016;115(12):1513–9.
48. Zhao W, Chen B, Guo X, Wang R, Chang Z, Dong Y, Song K, Wang W, Qi L, Gu Y, et al. A rank-based transcriptional signature for predicting relapse risk of stage II colorectal cancer identified with proper data sources. *Oncotarget*. 2016;7(14):19060–71.
49. Cai H, Li X, Li J, Ao L, Yan H, Tong M, Guan Q, Li M, Guo Z. Tamoxifen therapy benefit predictive signature coupled with prognostic signature of post-operative recurrent risk for early stage ER+ breast cancer. *Oncotarget*. 2015;6(42):44593–608.
50. Li X, Cai H, Zheng W, Tong M, Li H, Ao L, Li J, Hong G, Li M, Guan Q, et al. An individualized prognostic signature for gastric cancer patients treated with 5-fluorouracil-based chemotherapy and distinct multi-omics characteristics of prognostic groups. *Oncotarget*. 2016;7(8):8743–55.
51. Ao L, Song X, Li X, Tong M, Guo Y, Li J, Li H, Cai H, Li M, Guan Q, et al. An individualized prognostic signature and multiomics distinction for early stage hepatocellular carcinoma patients with surgical resection. *Oncotarget*. 2016;7(17):24097–110.
52. SM-1 C. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat Biotechnol*. 2014;32(9):903–14.
53. Guan Q, Chen R, Yan H, Cai H, Guo Y, Li M, Li X, Tong M, Ao L, Li 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(42):68909–20.
54. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *Acm Sigkdd Explorations Newsletter*. 2008;11(1):10–8.
55. Lin CJ. A practical guide to support vector classification. In. 2003;2003:012004.
56. Wang K, Sturt-Gillespie B, Hittle JC, Macdonald D, Chan GK, Yen TJ, Liu ST. Thyroid hormone receptor interacting protein 13 (TRIP13) AAA-ATPase is a novel mitotic checkpoint-silencing protein. *J Biol Chem*. 2014;289(34):23928–37.
57. Tseng WW, Liu CD. Peptide YY and cancer: current findings and potential clinical applications. *Peptides*. 2002;23(2):389–95.
58. Martin KJ, Patrick DR, Bissell MJ, Fournier MV. Prognostic breast cancer signature identified from 3D culture model accurately predicts clinical outcome across independent datasets. *PLoS One*. 2008;3(8):e2994.
59. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*. 2004;101(25):9309–14.
60. Loven J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. Revisiting global gene expression analysis. *Cell*. 2012;151(3):476–82.
61. Choi CH, Rutter MD, Askari A, Lee GH, Warusavitarne J, Moorghen M, Thomas-Gibson S, Saunders BP, Graham TA, Hart AL. Forty-year analysis of Colonoscopic surveillance program for neoplasia in ulcerative colitis: an updated overview. *Am J Gastroenterol*. 2015;110(7):1022–34.
62. Fornaro R, Caratto M, Caratto E, Caristo G, Fornaro F, Giovinozzio D, Sticchi C, Casaccia M, Andorno E. Colorectal cancer in patients with inflammatory bowel disease: the need for a real surveillance program. *Clin Colorectal Cancer*. 2016;15(3):204–12.
63. Kaltenbach T, Leite G, Soetikno R. Colonoscopy surveillance and Management of Dysplasia in inflammatory bowel disease. *Curr Treat Options Gastroenterol*. 2016;14(1):103–14.
64. Sengupta N, Yee E, Feuerstein JD. Colorectal cancer screening in inflammatory bowel disease. *Dig Dis Sci*. 2016;61(4):980–9.
65. Mooiweer E, van der Meulen-de Jong AE, Ponsioen CY, van der Woude CJ, van Bodegraven AA, Jansen JM, Mahmmod N, Kremer W, Siersema PD, Oldenburg B, et al. Incidence of interval colorectal cancer among inflammatory bowel disease patients undergoing regular colonoscopic surveillance. *Clin Gastroenterol Hepatol*. 2015;13(9):1656–61.
66. Breiman L, Friedman JH, Olshen R, Stone CJ. Classification and regression trees. *Encyclopedia of Ecology*. 2008;40(3):582–8.
67. Peng F, Zhang Y, Wang R, Zhou W, Zhao Z, Liang H, Qi L, Zhao W, Wang H, Wang C, et al. Identification of differentially expressed miRNAs in individual breast cancer patient and application in personalized medicine. *Oncogene*. 2016;5:e194.
68. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;41(Database issue):D991–5.
69. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, et al. ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*. 2011;39(Database):D1002–4.
70. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, et al. International network of cancer genome projects. *Nature*. 2010;464(7291):993–8.
71. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.
72. Mudunuri U, Che A, Yi M, Stephens RM. bioDBnet: the biological database network. *Bioinformatics*. 2009;25(4):555–6.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

