Fu *et al. BMC Genomics* (2019) 20:769
https://doi.org/10.1186/s12864-019-6129-8

BMC Genomics

RESEARCH ARTICLE Open Access

A qualitative transcriptional signature for predicting microsatellite instability status of right-sided Colon Cancer

Yelin Fu[1†], Lishuang Qi[1†], Wenbing Guo[1], Liangliang Jin[1], Kai Song[1], Tianyi You[1], Shuobo Zhang[1], Yunyan Gu[1], Wenyuan Zhao[1*] and Zheng Guo[1,2,3*]

Abstract

Background: Microsatellite instability (MSI) accounts for about 15% of colorectal cancer and is associated with prognosis. Today, MSI is usually detected by polymerase chain reaction amplification of specific microsatellite markers. However, the instability is identified by comparing the length of microsatellite repeats in tumor and normal samples. In this work, we developed a qualitative transcriptional signature to individually predict MSI status for right-sided colon cancer (RCC) based on tumor samples.

Results: Using RCC samples, based on the relative expression orderings (REOs) of gene pairs, we extracted a signature consisting of 10 gene pairs (10-GPS) to predict MSI status for RCC through a feature selection process. A sample is predicted as MSI when the gene expression orderings of at least 7 gene pairs vote for MSI; otherwise the microsatellite stability (MSS). The classification performance reached the largest F-score in the training dataset. This signature was verified in four independent datasets of RCCs with the F-scores of 1, 0.9630, 0.9412 and 0.8798, respectively. Additionally, the hierarchical clustering analyses and molecular features also supported the correctness of the reclassifications of the MSI status by 10-GPS.

Conclusions: The qualitative transcriptional signature can be used to classify MSI status of RCC samples at the individualized level.

Keywords: Right-sided colon cancer, Microsatellite instability status, Gene expression profiles, Relative gene expression orderings, Qualitative transcriptional signature

Background

Microsatellite instability (MSI), the insertion or deletion mutations in microsatellites [1], is a molecular hallmark of a deficient mismatch repair (dMMR) system and accounts for about 15% of colorectal cancer (CRC) [2]. Results from some studies seem that the MSI feature is associated with good prognosis, and the stage II and III CRC patients with MSI cannot benefit from 5-fluorouracil (5-Fu)-based adjutant chemotherapy (ACT) [3–5], which is regarded as the standard treatment for stage II and III CRC patients after surgery. So, a precise classification is needed to aid appropriate decisions on 5-Fu-based ACT treatment of patients.

Today, the most common method to test MSI status is polymerase chain reaction (PCR) amplification analysis of specific microsatellite repeats, which is considered as the 'golden standard' method [1, 6]. However, the PCR technology exists high measurement variations between different laboratories [7, 8], which are mainly due to the effects of the tumor cell percentage and DNA degradation during sample storage and preparation [8–11]. Additionally, the presence of instability is defined by comparing the length of microsatellite repeats in the tumor sample and the normal sample [1, 6, 12–14]. MSI can also be detected by immunohistochemistry (IHC). But IHC method only provides a semi-quantitative evaluation of the expression levels of the four MMR

---

* Correspondence: zhaowenyuan@ems.hrbmu.edu.cn; guoz@ems.hrbmu.edu.cn
[†]Yelin Fu and Lishuang Qi contributed equally to this work.
[1]Department of Systems Biology, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China
Full list of author information is available at the end of the article

proteins (MLH1, MSH2, MSH6 and PMS2) [2], and also exists high measurement variations between different laboratories [15–17], which are primarily due to the effects of sample preprocessing, such as fixation of tissues, detection reagents and selection of antibody [16, 17]. Moreover, since the results of IHC can be greatly affected by the interpretation of the specificity of staining, when the levels of the MMR proteins analyzed are low, good performance of IHC requires highly skilled personnel and pathologist's experience in interpretation [14, 15]. Therefore, the traditional PCR and IHC methods both have some limitations in determining the MSI status of CRC.

Recently, many alternative methods based on tumor genomic data via next-generation sequencing (NGS) panels have been developed to determine MSI status of patients [1, 18, 19]. For instance, Vanderwalde et al. used a NGS panel comprised of 592 genes to determine MSI status [1]. However, NGS currently has been limited to some highly specialized laboratories [20]. Besides, there has no consensus about NGS gene panels to determine MSI status and each laboratory determine the appropriate mutation load threshold based on its NGS gene panel and technique [1, 18–20]. What's more, the use of the NGS-based methods that require DNA extraction often leads to false-negative or uncertain results in challenging tumor samples due to the tumor DNA dilution and the percentage of tumor cells within a sample [20, 21]. Another method based on gene expression measurement also has been developed to identify MSI status of CRCs [6], which, however, is sensitive to the systematic inter-laboratory biases especially batch effects of microarray and RNA-sequencing experiments [22]. In general, quantitative transcriptional signatures based on absolute expression values is sensitive to the batch effects and thus lack robustness for clinical applications [23]. In contrast, the type of qualitative transcriptional signatures based on the within-sample relative expression orderings (REOs) of genes have strong robustness against the experimental batch effects and can be applied to individual samples directly [24]. Besides, we have demonstrated that they are rather robust against the proportions of tumor epithelial cell variations in tumor tissues sampled from different tumor locations [25], amplification bias for minimum samples [26], and partial RNA degradation during sample preparation [27]. For example, we have reported that more than 90% of the REOs of gene pairs in the fresh-frozen samples are maintained in their paired formalin-fixed paraffin-embedded samples and largely unaffected by the storage time [27], indicating that the vast majority of the REOs of gene pairs are rather robust. Thus, the gene pairs of signature were less vulnerable to degradation. Therefore, it is worthwhile to apply the within-sample REOs to find robust qualitative transcriptional signatures.

Colorectal cancers deriving from proximal or distal of splenic flexure are classified as right-sided or left-sided colon cancer (RCC or LCC), respectively [28]. Consistent with the differences in anatomy location, RCC and LCC have unique gene expression characteristics, different molecular pathways of carcinogenesis [28, 29] and different clinical features [30]. Therefore, it would be necessary to develop signatures to predict MSI status for RCC and LCC, respectively. Because of the high incidence of MSI in the RCC, we developed a REOs-based qualitative signature for predicting MSI status of RCC patients in this work, which was validated in independent datasets.

Results

Identification and validation of the signature for MSI status of RCC

The GSE39582 with the largest samples of RCC, including 57 MSI and 154 MSS was used as the training data for extracting a REOs-based signature. Firstly, we identified 4769 MSI-related differentially expressed (DE) genes (Student's *t*-test, $FDR < 0.01$, Additional file 1: Table S1) between the 57 MSI RCC samples and the 154 MSS RCC samples. From all the gene pairs formed by these DE genes, we identified 1,654,739 gene pairs, whose specific REO pattern occurred more frequently in the MSI samples than in the MSS samples (Fisher's exact test, $FDR < 0.01$). The larger FD of a REO pattern, the stronger classified ability of the REO pattern can classify the status of MSI. We further narrowed down the number of gene pairs to 1898 through a redundancy removal process by keeping only one with the largest FD value of those gene pairs sharing a common gene (see Methods, Fig. 1a). From these gene pairs, we extracted 10 gene pairs with the FD at least 0.8. These 10 gene pairs were used as the signature for predicting MSI status of RCC, denoted as 10-GPS (Table 1). A RCC sample was predicted as MSI if the REOs of at least seven gene pairs in the 10-GPS vote for MSI; otherwise the MSS. According to the classification rule, the F-score of the signature in the training data was 0.9727, with a sensitivity of 0.9649 and a specificity of 0.9805. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve was 0.9838 (Fig. 2a).

We tested the 10-GPS in four independent cohorts of RCC samples (Fig. 1b), the F-scores of the classification by 10-GPS were 1, 0.9630, 0.9412 and 0.8798, respectively, as shown in the Table 2, and the AUCs were 1, 0.9923, 1 and 0.9244, respectively (Fig. 2b, c, d and e).

Transcriptome assessment of the signature-disconfirmed RCC samples

In the training data, there were a total of five signature-disconfirmed samples. We compared the gene expression

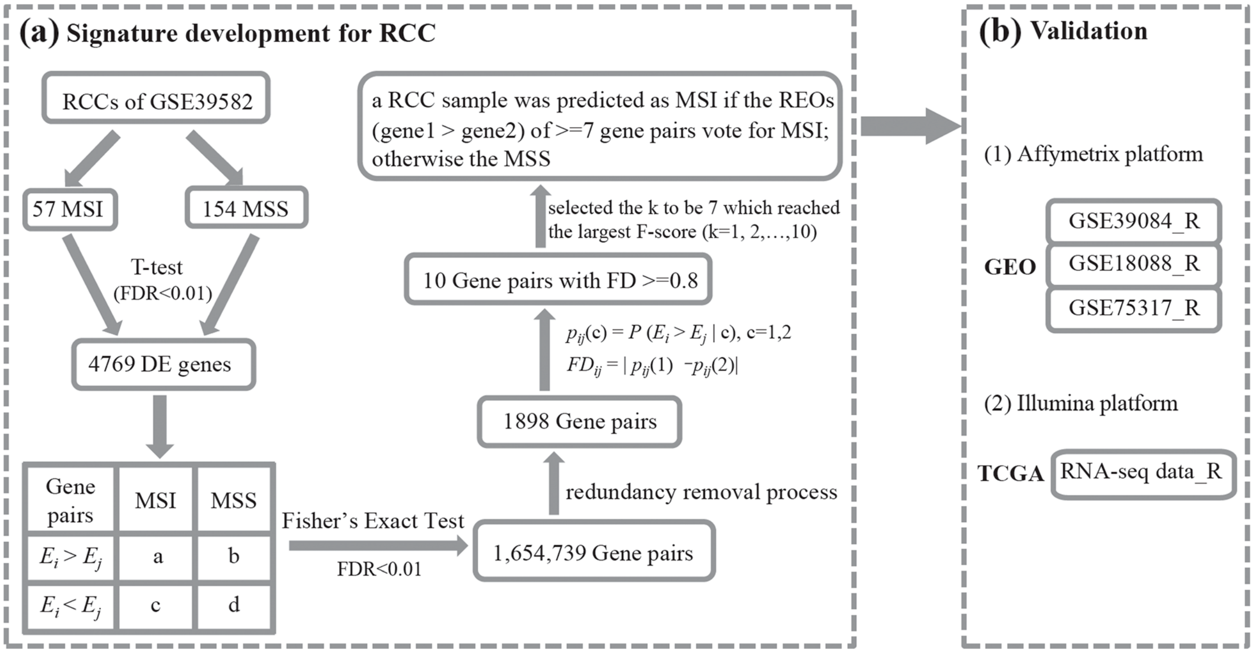
(a) Signature development for RCC
RCCs of GSE39582
57 MSI
154 MSS
T-test
(FDR<0.01)
4769 DE genes
Gene pairs
MSI
MSS
$E_i > E_j$
a
b
$E_i < E_j$
c
d
Fisher's Exact Test
FDR<0.01
1,654,739 Gene pairs
redundancy removal process
1898 Gene pairs
$p_{ij}(c) = P(E_i > E_j \mid c)$, c=1,2
$FD_{ij} = \mid p_{ij}(1) - p_{ij}(2) \mid$
10 Gene pairs with FD >=0.8
selected the k to be 7 which reached the largest F-score (k=1, 2,…,10)
a RCC sample was predicted as MSI if the REOs (gene1 > gene2) of >=7 gene pairs vote for MSI; otherwise the MSS
(b) Validation
(1) Affymetrix platform
GEO
GSE39084_R
GSE18088_R
GSE75317_R
(2) Illumina platform
TCGA
RNA-seq data_R

Fig. 1 The flowchart of this study, as exemplified by the development and validation of predicting MSI status signature for patients with RCC (see Methods)

patterns of the five signature-disconfirmed samples with the 206 signature-confirmed samples through clustering analysis. Firstly, we identified 5664 DE genes (Student's *t*-test, FDR < 0.01) between the 55 signature-confirmed MSI and the 151 signature-confirmed MSS samples. Secondly, using the expression levels of the top 100 significant DE genes, the samples were divided into two subgroups using the complete linkage hierarchical clustering based on the Euclidean distance (Fig. 3a). The results showed that all of the two MSI samples reclassified as MSS by the 10-GPS were clustered with the signature-confirmed MSS samples, and all of the three MSS samples reclassified as MSI were clustered with the signature-confirmed MSI samples.

Similarly in the two of the four validation datasets of RCC (GSE18088 and GSE75317), all of these two MSI samples reclassified as MSS by our signature were clustered with the corresponding signature-confirmed MSS samples (Fig. 3b and c), respectively. These results provided transcriptional evidence of the correctness of the prediction of 10-GPS.

Table 1 The Composition of 10-GPS

signature	gene1	gene2	signature	gene1	gene2
pair1	*HNRNPL*	*CDC16*	pair6	*STRN3*	*TMEM192*
pair2	*MTA2*	*VGF*	pair7	*HPSE*	*BCAS3*
pair3	*CALR*	*SEC22B*	pair8	*PRPF39*	*ATF6*
pair4	*RASL11A*	*CAB39L*	pair9	*CCRN4L*	*GRM8*
pair5	*LYG1*	*DHRS12*	pair10	*AMFR*	*DUSP18*

Notes: A RCC sample was classified as MSI if the REOs (gene1 > gene2) of at least 7 of the gene pairs in the 10-GPS vote for MSI; otherwise the MSS

Genome assessment of the signature-disconfirmed RCC samples

It is known that $\mathrm{BRAF}^{\mathrm{V600E}}$ mutations and CpG island methylator phenotype (CIMP)-positive frequently occur in MSI CRCs, whereas KRAS mutations (in codons 12 or 13) frequently occur in MSS CRCs [2, 31]. In the training data, for the three MSS samples which were reclassified as MSI by 10-GPS, two patients were KRAS wild-type, BRAF mutant and CIMP-positive and one patient was KRAS wild-type. For the two MSI samples which were reclassified as MSS, one was BRAF wild-type and CIMP-negative, as shown in Table 3.

In the TCGA validation dataset of RCC, there were seven signature-disconfirmed samples. Because mutation of MMR genes can result in MSI [32], we observed the mutation status of the MMR genes in the signature-disconfirmed samples. There were only five samples with mutation data. And two of the four MSS samples which were reclassified as MSI by 10-GPS were MSH6 mutant (Additional file 2: Table S1). These results supported that MSI status of these samples reclassified by 10-GPS might be reliable.

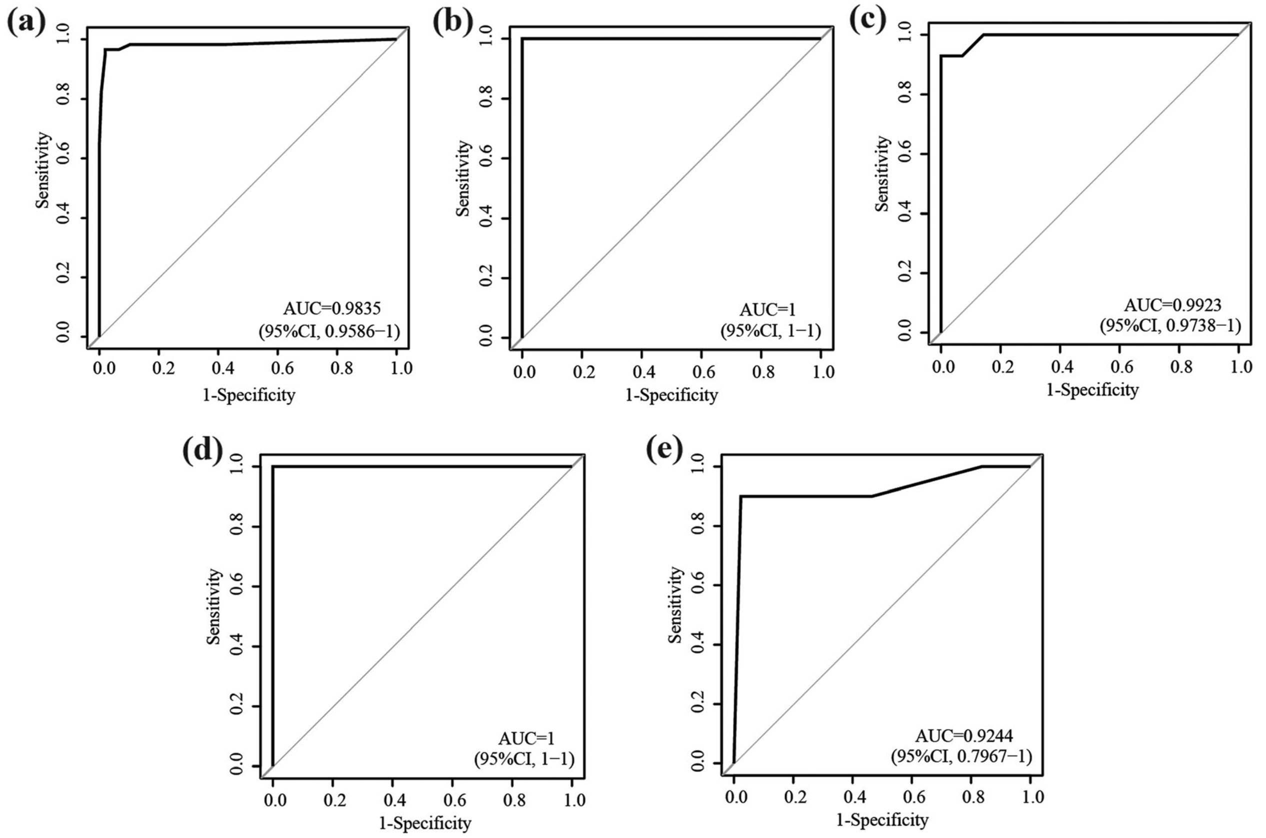

Fig. 2 The ROC curves for 10-GPS in four datasets. **a** the training dataset. **b** the RCCs of GSE39084. **c** the RCCs of GSE18088. **d** the RCCs of GSE75317. **e** the RCCs of TCGA

Prognosis assessment of the signature-disconfirmed RCC samples

Then, we also evaluated the reliability of the reclassifications by 10-GPS through survival analyses based on the knowledge that stage III MSI CRCs treated with surgery only have better prognoses than MSS CRCs [2] and that stage III MSS CRC patients treated with 5-Fu-based ACT after surgery have improved outcomes than patients treated with surgery only [3–5]. And the survival benefit of ACT was only observed in stage III patients [3, 4]. In the 32 stage III RCC samples of the training data for patients treated with surgery only, one of the 19 MSS sample was reclassified as MSI and one of 13 MSI sample was reclassified as MSS by the 10-GPS. In the 54 (46 MSS and 8 MSI) stage III RCC samples for patients receiving 5-Fu-based ACT, the original MSI status of all samples were confirmed by the 10-GPS. By comparison, the MSS patient reclassified as MSI had longer RFS (130 months) than the MSI patient reclassified as MSS (31 months) by the signature. The survival difference between patients with predicted MSI status by 10-GPS were more significant than the difference between patients with the original MSI status due to the two reclassified samples (Additional file 3).

Table 2 The performances of the 10-GPS in RCCs of the independent datasets

	pre-MSI[a] (MSI:MSS)[b]	pre-MSS[a] (MSI:MSS)[b]	sensitivity	specificity	F-score
GSE39084_R	13 (13:0)	18 (0:18)	1	1	1
GSE18088_R	13 (13:0)	15 (1:14)	0.9286	1	0.9630
GSE75317_R	8 (8:0)	18 (1:17)	0.8889	1	0.9412
TCGA_R	15 (9:6)	38 (1:37)	0.900	0.8605	0.8798
Total_RCCs	49 (43:6)	89 (3:86)	0.9348	0.9348	0.9348

Notes: [a] represents the predicted MSI status by 10-GPS; [b] represents the original MSI status; GSE_R represents the RCC samples; Total_RCCs represents all the samples of RCC

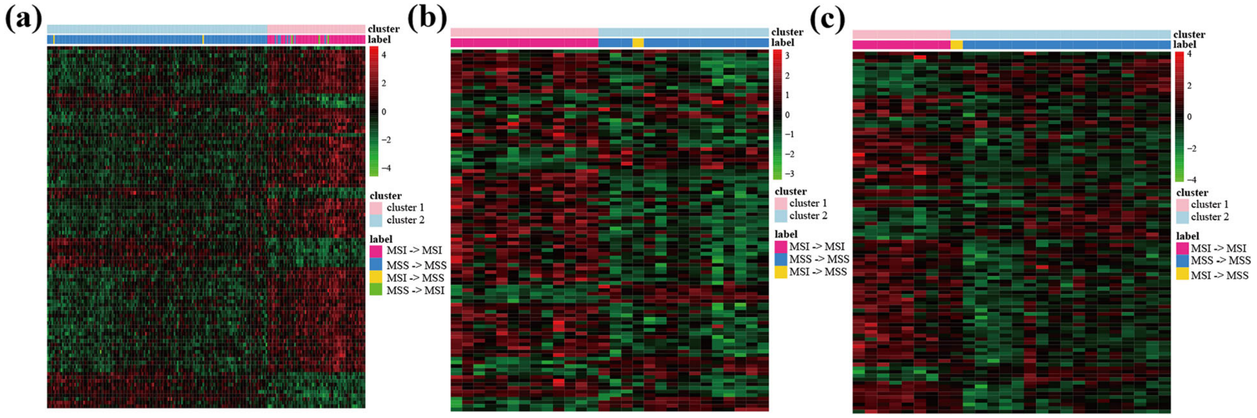

Fig. 3 The complete linkage hierarchical clustering of the RCC samples in the (**a**), training dataset, (**b**) GSE18088 and (**c**) GSE75317 based on the differentially expressed genes between the signature-confirmed MSI and MSS samples. X- > Y, X represents the original MSI status and Y represents the reclassified MSI status by 10-GPS

Identification and validation of the signature for MSI status of LCC

We applied the 10-GPS to LCC samples and CRC samples without clear location information (Additional file 2: Table S2). The results showed that the performance was reduced when applying the 10-GPS to predict MSI status of LCC. Therefore, we also tried to develop a signature to identify MSI status of the LCC patients in the same way as in RCC. Eventually, these six gene pairs were used as the signature for predicting MSI status of LCC, denoted as 6-GPS (Additional file 2: Table S3). A LCC sample was predicted as MSI if the REOs of at least four gene pairs in the 6-GPS vote for MSI; otherwise the MSS. According to the classification rule, the F-score of the signature in the LCC training data was 0.9983, with a sensitivity of 1 and a specificity of 0.9966. And also, the 6-GPS was well validated in four independent cohorts of LCC samples (Additional file 2: Table S4).

Discussion

We developed qualitative transcriptional signatures consisting of 10 and 6 gene pairs to robustly predict MSI status of RCC and LCC at individualized level, which were validated in four independent datasets. Notably, the hierarchical clustering analyses and molecular characteristics supported the correctness of the reclassifications of the MSI status by our signature for some samples whose MSI statuses were determined by the PCR testing. Besides, using gene pairs with large FD (see Methods) of a REO pattern between MSI and MSS samples, we can exclude gene pairs affected by various factors such as RNA degradation and tumor cell percentage and obtain a classifier with high predictive performance of MSI status. Thus, it is possible for our signature to identify the MSI status of CRC samples, which could not be determined by traditional standard methods.

Table 3 The molecular characteristics of the five signature-disconfirmed RCC samples in the training dataset

original_MSI. status	predicted_MSI. status	KRAS. status	BRAF. status	CIMP. status
MSS	MSI	wild type	NA	NA
MSS	MSI	wild type	mutation	+
MSS	MSI	wild type	mutation	+
MSI	MSS	wild type	mutation	+
MSI	MSS	NA	wild type	–

In this study, we selected MSI-related gene pairs formed by the DE genes identified between MSI and MSS RCC samples. Some DE genes are known to be associated with microsatellite instability, prognosis and metastasis of CRC. For example, *CAB39L* have mononucleotide repeats in the coding regions that could be targets for frameshift mutation in CRC with microsatellite instability [33]. Another gene, *MTA2*, is one of metastasis-associated tumor gene family members and was an important prognosis biomarker of CRC [34]. Besides, it is reported that overexpression of *AMFR* is significantly related to poor survival for CRC [35]. Additionally, the differential expression of these genes could cause reversal REOs of the selected gene pairs between MSI and MSS samples, and thus these gene pairs could have the ability to classify the MSI status.

In the process of screening MSI-related gene pairs of RCC, we extracted gene pairs by adjusting different FD thresholds. Then, it was found that when FD > 0.9, there were no gene pairs remaining. When FD > 0.7, there were 65 gene pairs were extracted. The classification performance reached the largest F-score (0.9630) of sensitivity (0.9649) and specificity (0.9610) in the training data according to the following decision rule: a sample is predicted as MSI when the REOs of at least 39 gene pairs vote for MSI; otherwise the MSS. Compared with the result of $FD > 0.8$, the performance of classification was slightly worse and it had more gene pairs. So we chose 0.8 as the threshold for FD. Similarly, based on the same considerations, we chose 0.9 as the threshold for FD in the process of screening MSI-related gene pairs of LCC.

The REO-based method was first proposed by Donald Geman et al. in 2004 [36]. The method has been proposed as a simple, accurate and easily interpretable decision rule for classification of gene expression profiles [37]. What's more, it is robust against the experimental batch effects and avoid the need of inter-sample data normalization and can be applied at individualized level [23, 24, 36]. So, there were many studies by others and us developing several prognostic and predictive biomarkers based on this method for different cancers [38–51]. It indicated that the clinical applicability of the signatures based on the robust qualitative REO information extracted from the quantitative measurements of gene expression, rather than the "exact" quantitative measurements themselves [52]. Given cost considerations and the often-limited quantity of tumor material available for testing in many cancer patients, NGS-based tumor profiling, which provides the basis for the concept of "a sequence for all" [53]. So, we have been focusing on developing qualitative transcriptional signatures to form the "a sequence for all" for CRC. All these signatures can be assessed in a single NGS assay, facilitating the optimum treatment of stage II-III CRC patients. In summary, we developed qualitative signatures for predicting MSI status of RCC and LCC, as a part of "a sequence for all" for CRC.

Conclusions

Currently, common methods for detecting MSI status of CRC such as PCR and IHC-based methods, exist high measurement variations between different laboratories, which have limited clinical utility. Herein, we developed robust qualitative transcriptional signatures to classify MSI status of RCC and LCC at the individualized level, as a part of "a sequence for all" for CRC. The simplicity and robustness of the signature would make it convenient in clinical settings and worthy to further validate in a prospective clinical trial.

Methods

Data sources and data preprocessing

The gene expression datasets used in this study were downloaded from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) database and The Cancer Genome Atlas data portal (TCGA, http://cancergenome.nih.gov/) (Table 4). As MSI-L patients are usually treated in a way similar to MSS patients in clinical practice, it is reasonable to group MSI-L with MSS [1]. In this study, we grouped MSI-L with MSS. The training data for extracting a REOs-based signature of RCC was GSE39582, including 57 MSI and 154 MSS of 211 RCCs. The GSE39582 dataset recording survival information of patients were used as the test for survival analyses.

The Robust Multi-array Average algorithm [54] was used for preprocessing the raw data measured by the Affymetrix platform. Using the corresponding platform files, probes were mapped to genes. For each sample, the expression measurements of several probes mapping to a gene were averaged to obtain a single measurement. Probes were discarded if they did not match any gene or matched multiple genes. The RNA-seq expression data were downloaded from the Broad Firehose webpage (http://gdac.broadinstitute.org/). For RNA-seq data derived from Illumina HiSeq 2000 RNA Sequencing Version 2, we directly downloaded the RSEM-normalized format and log2-transformed. We also downloaded the somatic mutation data of CRC from the Broad Firehose webpage.

Currently, most of the CRC data we collected were microarray datasets. With the development of the NGS technology, the RNA-seq data are increasing. In order to apply our signature to RNA-seq data, we used the expression data of overlapping Gene IDs of Affymetrix and Illumina platforms.

Signature development for predicting MSI status of RCC

We identified MSI-related gene pairs from the DE genes between MSI and MSS samples. For a gene pair, i and j, with expression values of E_i and E_j, whether the frequency of a specific REO pattern ($E_i > E_j$ or $E_i < E_j$) was significantly higher in the MSI samples than the frequency in the MSS samples was evaluated by Fisher's exact test [55]. The MSI-related gene pairs detected with $FDR < 0.01$. Then, the frequency difference (FD) was calculated for each MSI-related gene pair between the two groups (Formula 1). The larger FD of a REO pattern, the stronger classified ability of the REO pattern can classify the status of MSI and avoid the effect of degradation. It is more likely to be applied to multiple data sources produced by different laboratories. So, considering that some genes are influenced by RNA degradation, for a gene which appeared in multiple gene pairs, we kept only the gene pair with the largest FD value and discarded others.

$p_{ij}(c) = P(E_i > E_j \mid c)$, c = 1,2, the probabilities of observing $E_i > E_j$ in each class.

$FD_{ij} = p_{ij}(1) - p_{ij}(2)$, the FD value of gene pair (i, j) [Formula 1].

After that, the gene pairs with the FD value at least 0.8 were identified as the signature to predict MSI status of RCC. A sample was labeled as MSI if the REOs of at

Table 4 The datasets analyzed in this study from GEO and TCGA

	GSE39582 (*n* = 566)	GSE39084 (*n* = 70)	GSE18088 (*n* = 53)	GSE75317 (*n* = 59)	GSE13067 (*n* = 74)	GSE13294 (*n* = 155)	TCGA (*n* = 457)
Stage							
I	33	8	–	6	–	–	75
II	264	23	53	24	–	–	178
III	205	16	–	17	–	–	130
IV	60	22	–	12	–	–	64
Microsatellite status							
MSI	75	16	19	11	11	78	11
MSS	444	54	34	48	63	77	81
Location							
Right	224	31	28	26	–	–	261
Left	342	30	25	33	–	–	177
MSI_proportion							
Right	57:154[a] (27.0%)[b]	13:18 (41.9%)	14:14 (50.0%)	9:17 (34.6%)	–	–	10:43 (18.9%)
Left	18:290[c] (5.8%)[d]	3:27 (10.0%)	5:20 (20.0%)	2:31 (6.1%)	–	–	1:33 (2.9%)
MSI detection	PCR	PCR or IHC	PCR	PCR	PCR	PCR	PCR
Adjuvant chemotherapy							
Yes	233	–	–	–	–	–	–
No	316	–	53	–	–	–	–

Notes: The data from GEO were produced by the same gene expression profiling platform (GPL570, Affy-HG-U133_Plus_2). a represents the number of MSI and MSS of RCCs, respectively; b represents the proportion of MSI in RCCs; c represents the number of MSI and MSS of LCCs, respectively; d represents the proportion of MSI in LCCs

least k gene pairs in the signature are consistent with the specific patterns ($E_i > E_j$) of the training samples, and vice versa. For each k ranging from one to the number of gene pairs in the signature, we could compute a F-score. Finally, we selected the k which could reach the largest F-score.

$$F\text{-}score = 2 \times sensitivity \times specificity \div (sensitivity + specificity) \qquad (2)$$

Sample clustering

Student's t-test was performed to identify DE genes between the MSI and MSS patients confirmed with the original MSI status by the signature. Complete linkage hierarchical clustering was performed to stratify CRC samples into subgroups. The similarity between samples was evaluated by the Euclidean distance based on the expression measurements of DE genes.

Statistical analyses

For microarray data, we selected DE genes between two classes of samples using Student's t-test. The MSI rate in different groups was evaluated by Fisher's exact test. The Benjamini-Hochberg procedure for multiple testing was used to adjust p values in order to control the false discovery rate (FDR) [56]. Log-rank test was used to assess the difference between the Kaplan-Meier estimates of the RFS in the two different groups [57]. Hazard ratios (HRs) and 95% confidence intervals (CIs) were generated using the univariate Cox proportional hazards model. All statistical analyses were performed using the R 3.5.1 (http://www.r-project.org/).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6129-8.

Additional file 1: Table S1. The DE genes (FDR<0.01) between the 57 MSI RCC samples and the 154 MSS RCC samples. **Table S2.** The DE genes (FDR<0.01) between the 18 MSI LCC samples and the 290 MSS LCC samples.(XLS 692 kb)

Additional file 2: Table S1. The mutation status of the four MMR genes of the five signature-disconfirmed RCC in the TCGA dataset. **Table S2.** The performances of the 10-GPS in LCCs and CRCs of the independent datasets. **Table S3.** The Composition of 6-GPS. **Table S4.** The performances of the 6-GPS in LCCs of the independent datasets.(DOCX 18 kb)

Additional file 3. The Kaplan-Meier survival curve for the prediction of 10-GPS and original MSI status, respectively. (TIF 524 kb)

Abbreviations

5-Fu: 5-fluorouracil; ACT: adjutant chemotherapy; AUC: area under the curve; CIMP: CpG island methylator phenotype; CRC: colorectal cancer; DE: differentially expressed; dMMR: deficient mismatch repair; FD: frequency difference; FDR: false discovery rate; GPS: gene pairs; IHC: immunohistochemical; LCC: left-sided colon cancer; MSI: microsatellite instability; MSS: microsatellite stability; NGS: next-generation sequencing;

PCR: polymerase chain reaction; RCC: right-sided colon cancer; REOs: relative expression orderings; ROC: receiver operating characteristic

Acknowledgements
Not applicable.

Authors' contributions
WZ and ZG conceived the idea, YF conceived and designed the experiments, wrote the manuscript, LQ designed the experiments, WG and LJ analyzed the data, KS and TY performed the experiments, SZ and YG helped in writing the manuscript. All authors approved the final version.

Funding
This work was supported by the National Natural Science Foundation of China [grant numbers: 61601151, 61701143, 61673143, 81872396 and 81572935], the Natural Science Foundation of Heilongjiang Province [grant number: C2016037] and the Joint Scientific and Technology Innovation Fund of Fujian Province [grant number: 2016Y9044].

Availability of data and materials
All data analyzed in this study were downloaded from the public database: GEO and TCGA.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Systems Biology, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China. [2]Department of Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, School of Basic Medical Sciences, Fujian Medical University, Fuzhou 350122, China. [3]Key Laboratory of Medical Bioinformatics, Fujian Province, Fuzhou 350122, China.

Received: 17 June 2019 Accepted: 23 September 2019

References

1. Vanderwalde A, Spetzler D, Xiao N, Gatalica Z, Marshall J. Microsatellite instability status determined by next-generation sequencing and compared with PD-L1 and tumor mutational burden in 11,348 patients. Cancer medicine. 2018;7(3):746–56.
2. Gelsomino F, Barbolini M, Spallanzani A, Pugliese G, Cascinu S. The evolving role of microsatellite instability in colorectal cancer: a review. Cancer Treat Rev. 2016;51:19–26.
3. Cao B, Luo L, Feng L, Ma S, Chen T, Ren Y, Zha X, Cheng S, Zhang K, Chen C. A network-based predictive gene-expression signature for adjuvant chemotherapy benefit in stage II colorectal cancer. BMC Cancer. 2017;17(1):844.
4. Sargent DJ, Marsoni S, Monges G, Thibodeau SN, Labianca R, Hamilton SR, French AJ, Kabat B, Foster NR, Torri V, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. J Clin Oncol. 2010;28(20):3219–26.
5. Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, Hamilton SR, Laurent-Puig P, Gryfe R, Shepherd LE, et al. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. N Engl J Med. 2003;349(3):247–57.
6. Tian S, Roepman P, Popovici V, Michaut M, Majewski I, Salazar R, Santos C, Rosenberg R, Nitsche U, Mesker WE, et al. A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. J Pathol. 2012;228(4):586–95.
7. Hempelmann JA, Lockwood CM, Konnick EQ, Schweizer MT, Antonarakis ES, Lotan TL, Montgomery B, Nelson PS, Klemfuss N, Salipante SJ, et al. Microsatellite instability in prostate cancer by PCR or next-generation sequencing. J Immunother Cancer. 2018;6(1):29.
8. Boyle TA, Bridge JA, Sabatini LM, Nowak JA, Vasalos P, Jennings LJ, Halling KC, College of American Pathologists Molecular Oncology Comiittee. Summary of microsatellite instability test results from laboratories participating in proficiency surveys: proficiency survey results from 2005 to 2012. Arch Path Lab Med. 2014;138(3):363–70.
9. Waalkes A, Smith N, Penewit K, Hempelmann J, Konnick EQ, Hause RJ, Pritchard CC, Salipante SJ. Accurate Pan-Cancer molecular diagnosis of microsatellite instability by single-molecule molecular inversion probe capture and high-throughput sequencing. Clin Chem. 2018;64(6):950–8.
10. Chen G, Yang Z, Eshleman JR, Netto GJ, Lin MT. Molecular diagnostics for precision medicine in colorectal Cancer: current status and future perspective. Biomed Res Int. 2016;2016:9850690.
11. Trusky CL, Sepulveda AR, Hunt JL. Assessment of microsatellite instability in very small microdissected samples and in tumor samples that are contaminated with normal DNA. Diagn Mol Pathol. 2006;15(2):63–9.
12. Ryan E, Sheahan K, Creavin B, Mohan HM, Winter DC. The current value of determining the mismatch repair status of colorectal cancer: a rationale for routine testing. Crit Rev Oncol Hematol. 2017;116:38–57.
13. Zeinalian M, Hashemzadeh-Chaleshtori M, Salehi R, Emami MH. Clinical aspects of microsatellite instability testing in colorectal Cancer. Adv Biomed Res. 2018;7:28.
14. Pena-Diaz J, Rasmussen LJ. Approaches to diagnose DNA mismatch repair gene defects in cancer. DNA repair. 2016;38:147–54.
15. Shia J, Holck S, Depetris G, Greenson JK, Klimstra DS. Lynch syndrome-associated neoplasms: a discussion on histopathology and immunohistochemistry. Familial Cancer. 2013;12(2):241–60.
16. Kirkegaard T, Edwards J, Tovey S, McGlynn LM, Krishna SN, Mukherjee R, Tam L, Munro AF, Dunne B, Bartlett JM. Observer variation in immunohistochemical analysis of protein expression, time for a change? Histopathology. 2006;48(7):787–94.
17. Fitzgibbons PL, Murphy DA, Hammond ME, Allred DC, Valenstein PN. Recommendations for validating estrogen and progesterone receptor immunohistochemistry assays. Arch Pathol Lab Med. 2010;134(6):930–5.
18. Stadler ZK, Battaglin F, Middha S, Hechtman JF, Tran C, Cercek A, Yaeger R, Segal NH, Varghese AM, Reidy-Lagunes DL, et al. Reliable detection of mismatch repair deficiency in colorectal cancers using mutational load in next-generation sequencing panels. J Clin Oncol. 2016;34(18):2141–7.
19. Salem ME, Puccini A, Grothey A, Raghavan D, Goldberg RM, Xiu J, Korn WM, Weinberg BA, Hwang JJ, Shields AF, et al. Landscape of tumor mutation load, mismatch repair deficiency, and PD-L1 expression in a large patient cohort of gastrointestinal cancers. Mol Cancer Res. 2018;16(5):805–12.
20. Le Flahec G, Uguen M, Uguen A. Detection of mismatch repair deficiency in colorectal cancers: is it really time to eliminate immunohistochemistry? J Clin Oncol. 2017;35(3):376–7.
21. Uguen A, Gueguen P, Legoupil D, Bouvier S, Costa S, Duigou S, Lemasson G, Lede F, Sassolas B, Talagas M, et al. Dual NRASQ61R and BRAFV600E mutation-specific immunohistochemistry completes molecular screening in melanoma samples in a routine practice. Hum Pathol. 2015;46(11):1582–91.
22. Guan Q, Yan H, Chen Y, Zheng B, Cai H, He J, Song K, Guo Y, Ao L, Liu H, et al. Quantitative or qualitative transcriptional diagnostic signatures? A case study for colorectal cancer. BMC Genomics. 2018;19(1):99.
23. 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.
24. Ao L, Zhang Z, Guan Q, Guo Y, Guo Y, Zhang J, Lv X, Huang H, Zhang H, Wang X, et al. A qualitative signature for early diagnosis of hepatocellular carcinoma based on relative expression orderings. Liver Int. 2018;38(10):1812–9.
25. 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.
26. 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.
27. 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. 2017;8(4):6652–62.
28. Tejpar S, Stintzing S, Ciardiello F, Tabernero J, Van Cutsem E, Beier F, Esser R, Lenz HJ, Heinemann V. Prognostic and predictive relevance of primary tumor

location in patients with RAS wild-type metastatic colorectal Cancer: retrospective analyses of the CRYSTAL and FIRE-3 trials. JAMA oncology. 2016.

29. Loupakis F, Yang D, Yau L, Feng S, Cremolini C, Zhang W, Maus MK, Antoniotti C, Langer C, Scherer SJ *et al*: Primary tumor location as a prognostic factor in metastatic colorectal cancer. J Natl Cancer Inst 2015, 107(3).
30. Salem ME, Weinberg BA, Xiu J, El-Deiry WS, Hwang JJ, Gatalica Z, Philip PA, Shields AF, Lenz HJ, Marshall JL. Comparative molecular analyses of left-sided colon, right-sided colon, and rectal cancers. Oncotarget. 2017; 8(49):86356–68.
31. Cha Y, Kim KJ, Han SW, Rhee YY, Bae JM, Wen X, Cho NY, Lee DW, Lee KH, Kim TY, et al. Adverse prognostic impact of the CpG island methylator phenotype in metastatic colorectal cancer. Br J Cancer. 2016;115(2):164–71.
32. Mensenkamp AR, Vogelaar IP, van Zelst-Stams WA, Goossens M, Ouchene H, Hendriks-Cornelissen SJ, Kwint MP, Hoogerbrugge N, Nagtegaal ID, Ligtenberg MJ. Somatic mutations in MLH1 and MSH2 are a frequent cause of mismatch-repair deficiency in lynch syndrome-like tumors. Gastroenterology. 2014;146(3):643–6 e648.
33. Choi MR, An CH, Yoo NJ, Lee SH. Frameshift mutations of CAB39L, an activator of LKB1 tumor suppressor, in gastric and colorectal cancers. Pathol Oncol Res. 2016;22(1):225–6.
34. Ding W, Hu W, Yang H, Ying T, Tian Y. Prognostic correlation between MTA2 expression level and colorectal cancer. Int J Clin Exp Pathol. 2015;8(6):7173–80.
35. Chiu CG, St-Pierre P, Nabi IR, Wiseman SM. Autocrine motility factor receptor: a clinical review. Expert Rev Anticancer Ther. 2008;8(2):207–17.
36. Geman D, d'Avignon C, Naiman DQ, Winslow RL: Classifying gene expression profiles from pairwise mRNA comparisons. *Statistical applications in genetics and molecular biology* 2004, 3:Article19.
37. Kagaris D, Khamesipour A, Yiannoutsos CT. AUCTSP: an improved biomarker gene pair class predictor. BMC bioinformatics. 2018;19(1):244.
38. 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.
39. Wu J, Zhao Y, Zhang J, Wu Q, Wang W. Development and validation of an immune-related gene pairs signature in colorectal cancer. Oncoimmunology. 2019;8(7):1596715.
40. Shu P, Wu J, Tong Y, Xu C, Zhang X. Gene pair based prognostic signature for colorectal colon cancer. Medicine. 2018;97(42):e12788.
41. Yang Y, Zhang T, Xiao R, Hao X, Zhang H, Qu H, Xie B, Wang T, Fang X. Platform-independent approach for cancer detection from gene expression profiles of peripheral blood cells. Brief Bioinform. 2019.
42. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. J Natl Cancer Inst. 2015;107(1):357.
43. Li B, Cui Y, Diehn M, Li R. Development and validation of an individualized immune prognostic signature in early-stage nonsquamous non-small cell lung Cancer. JAMA oncology. 2017;3(11):1529–37.
44. 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.
45. Song K, Zhao W, Wang W, Zhang N, Wang K, Chang Z. Individualized predictive signatures for 5-fluorouracil-based chemotherapy in right- and left-sided colon cancer. Cancer Sci. 2018;109(6):1939–48.
46. Song K, Guo Y, Wang X, Cai H, Zheng W, Li N, Song X, Ao L, Guo Z, Zhao W. Transcriptional signatures for coupled predictions of stage II and III colorectal cancer metastasis and fluorouracil-based adjuvant chemotherapy benefit. FASEB J. 2019;33(1):151–62.
47. 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.
48. 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.
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. Consortium SM-I. 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. Cheng ML, Berger MF, Hyman DM, Solit DB. Clinical tumour sequencing for precision oncology: time for a universal strategy. Nat Rev Cancer. 2018;18(9):527–8.
54. 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.
55. Crans GG, Shuster JJ. How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial. Stat Med. 2008; 27(18):3598–611.
56. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med. 1990;9(7):811–8.
57. Bland JM, Altman DG. The logrank test. Bmj. 2004;328(7447):1073.

Publisher's Note