

# A rank-based transcriptional signature for predicting relapse risk of stage II colorectal cancer identified with proper data sources

Wenyuan Zhao<sup>1</sup>, Beibei Chen<sup>1</sup>, Xin Guo<sup>1</sup>, Ruiping Wang<sup>1</sup>, Zhiqiang Chang<sup>1</sup>, Yu Dong<sup>1</sup>, Kai Song<sup>1</sup>, Wen Wang<sup>1</sup>, Lishuang Qi<sup>1</sup>, Yunyan Gu<sup>1</sup>, Chenguang Wang<sup>1</sup>, Da Yang<sup>3,4,5</sup> and Zheng Guo<sup>1,2</sup>

<sup>1</sup> College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

<sup>2</sup> Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Department of Bioinformatics, Fujian Medical University, Fuzhou, China

<sup>3</sup> Department of Pharmaceutical Sciences, University of Pittsburgh, Pittsburgh, PA, USA

<sup>4</sup> Women's Cancer Research Center, University of Pittsburgh Cancer Institute, Pittsburgh, PA, USA

<sup>5</sup> Department of Computational & Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA

**Correspondence to:** Zheng Guo, **email:** guoz@ems.hrbmu.edu.cn

Da Yang, **email:** dyang@pitt.edu

**Keywords:** gene expression profiles, prognostic signatures, gene pairs, experimental batch effect, relative expression

**Received:** December 11, 2015

**Accepted:** February 25, 2016

**Published:** March 07, 2016

## ABSTRACT

The irreproducibility problem seriously hinders the studies on transcriptional signatures for predicting relapse risk of early stage colorectal cancer (CRC) patients. Through reviewing recently published 34 literatures for the development of CRC prognostic signatures based on gene expression profiles, we revealed a surprising phenomenon that 33 of these studies analyzed CRC samples with and without adjuvant chemotherapy together in the training and/or validation datasets. This data misuse problem could be partially attributed to the unclear and incomplete data annotation in public data sources. Furthermore, all the signatures proposed by these studies were based on risk scores summarized from gene expression levels, which are sensitive to experimental batch effects and risk compositions of the samples analyzed together. To avoid the above-mentioned problems, we carefully selected three qualified large datasets to develop and validate a signature consisting of three pairs of genes. The within-sample relative expression orderings of these gene pairs could robustly predict relapse risk of stage II CRC samples assessed in different laboratories. The transcriptional and functional analyses provided clear evidence that the high risk patients predicted by the proposed signature represent patients with micro-metastases.

## INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer and the fourth leading cause of cancer death worldwide [1]. The main factor for therapeutic decisions and prognostic estimates is based on AJCC tumor stage. Patients with stage II disease used to be treated with surgery only. However, retrospective analyses on historical trials show that approximately 25-30% of stage II CRC patients undergoing curative surgery will experience relapse, and only these patients need adjuvant CTX to

reduce the relapse risk [2, 3]. Currently used clinical and pathologic parameters, such as intestinal perforation/obstruction, tumor size and tumor grade [4], molecular markers, such as mutations in KRAS and BRAF as well as chromosome and microsatellite instability (MSI) [5-10] cannot adequately assess relapse risk to guide the clinical adjuvant CTX after surgery [11, 12].

High-throughput gene expression profiling has emerged as a powerful tool to identify stage II CRC patients with potential relapse risk [13, 14]. However, previously reported prognostic transcriptional signatures

often fail to be validated in independent datasets [15-18]. Therefore, it is necessary to analyze the major factors, besides the commonly claimed problem of small sample sizes [18], that may lead to the irreproducibility of the reported prognostic signatures. A surprising phenomenon is that most of current studies for the development of CRC prognostic signatures based on gene expression profiles have the data misuse problem, which could be partially attributed to the unclear and incomplete data annotation in public data sources (see Results). Besides, most of current gene expression signatures, including those signatures based on functional categories [19-21], are based on risk scores calculated as some summaries of expression levels of signature genes, which are sensitive to experimental batch effects and could lead to irreproducibility of prognostic signatures [22]. For the applications of this type of prognostic signatures, the requirement of presetting risk score thresholds and data normalization would result in the risk classification of patients depends on the risk composition of the samples adopted for normalization together [22]. This could produce substantial uncertainty for patient risk classification especially when the sample sizes are insufficient to represent the disease populations [22]. It has been revealed that within-sample relative expression orderings (REOs) of genes are overwhelmingly stable in a particular type of normal human tissues, which could reflect the concerted correlations of gene expression in normal states, but widely disrupted in the corresponding cancer tissues [23]. This biological phenomenon provides a basis for analyses based on REOs of gene pairs to characterize cancer subtypes [24-26]. Because REOs of genes are insensitive to experimental batch effects of gene expression profiling and invariant to monotonic data normalization [27], it is worthwhile to apply the rank-based approach to find robust prognostic signatures for clinical application.

In this work (shown in the Figure 1), firstly, we showed that different CRC sample sets have heterogeneous risk compositions, which could produce substantial uncertainty for patient risk classification based on risk scores summarized from gene expression levels. Then, we extracted a rank-based prognostic signature for relapse risk of stage II CRC using three large datasets.

## RESULTS

### Problems of data usage and data annotation of public data sources

By searching the PubMed medical literature database, we collected a total of 34 studies published in English between January 1, 2010, and June 1, 2015 for developing prognostic signatures of CRC based on gene expression profiles (shown in Table 1). Surprisingly,

except for the study of ColoGuideEx [14], all the other 33 studies used a mixture of samples with and without CTX in the training and/or validation datasets, although it has been well recognized that pooling cancer samples with and without CTX for prognostic signatures discovery is unreasonable as adjuvant CTX could affect the relapse-free survival (RFS) of the patients [28, 29]. This surprising phenomenon of data misuse could be partially attributed to incomplete and inaccurate clinical data annotation of CRC samples collected in public databases. First, the clinical annotations of many datasets are incomplete. For example, the dataset GSE17536 in the GEO database [30] provides no information indicating whether the collected samples are for patients treated with CTX or not. Some researchers used this dataset for identifying the relapse risk signature, wrongly considering that the samples in this dataset after surgery did not undergo CTX without discriminating patients [31]. However, we found that 42 stage III CRC patients' samples documented in this dataset did accept CTX by tracing the original research papers [32, 33]. Second, some samples are repeatedly documented in different datasets. For example, 35 stage II CRC samples that were repeatedly measured are documented in two series of the GEO database, GSE14333 and GSE17538, but this information is not provided in the data description [30]. Consequently, these two datasets with technical replicates were used as both the training and validation sets in some studies such as the study for ColoGuidePro [13], resulting in non-independent verification of the signature. Due to these problems, we have to check the original papers to avoid inappropriate use of expression profiles documented in the public data sources.

### Heterogeneous risk compositions of independent datasets

CRC is clinically and pathologically highly heterogeneous with a large variation in 5-year survival rates in different countries and even different cities in the same country [1]. As shown in Figure 2, the stage II CRC patients without CTX after surgery from six datasets (Supplementary Table S1) had significantly different RFS ( $p = 0.0026$ , log-rank test). The heterogeneous risk compositions could be due to many factors such as the differences in diagnosis criteria, surgery quality and location of CRC. For signatures based on risk scores summarized from gene expression measurements of a set of signature genes, this problem would induce spurious risk classification and difficulty in clinical settings because the risk classification of a sample would change when different samples are adopted for analysis together [25].

For example, ColoGuideEx [14], a 13-gene prognostic classifier, assigned patients to a poor prognosis group when at least 5 genes in the 13-gene signature indicated poor prognosis. If the high or low expression of a gene included in the 13-gene signature was associated

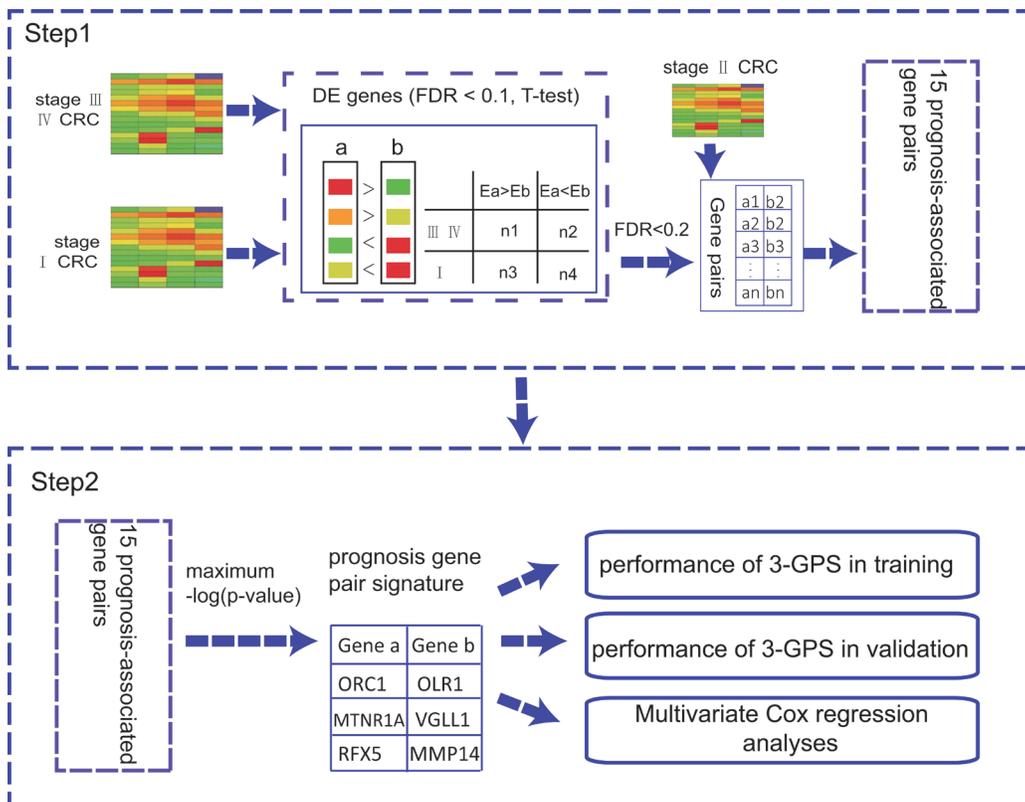


Figure 1: The flowchart for the development of the rank-based prognostic GPS.

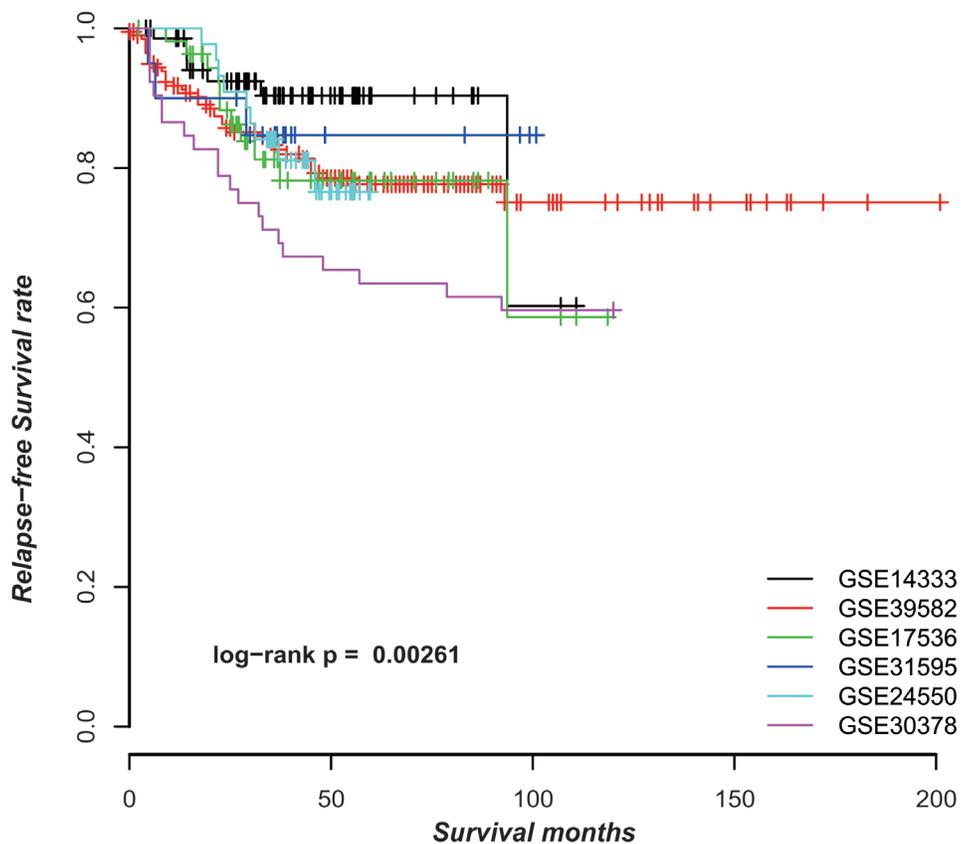


Figure 2: The Kaplan-Meier curves of RFS for samples in six datasets.

**Table 1: Proposed gene expression signatures for prognostic assessment of CRC**

Date	Datasets	Mixed	Tumor stage	Prognostic endpoint	References(PubMed index)
2015	GSE12945, GSE41258, GSE14333, GSE17538, GSE29623, GSE33113, GSE39582, GSE24549, GSE24550, GSE30378, GSE28722	yes	I-IV	OS	25853550
2015	GSE17536	yes	I-IV	DSS	25622900
2015	GSE24549, GSE24550, GSE39582	yes	I-IV	OS	25894381
2014	GSE13294, GSE5206, GSE17536, GSE17537	yes	II-III	DFS	24486594
2014	GSE14333, GSE17538, GSE33113, GSE31595, GSE14095, GSE26892	yes	II-III	Relapse	25115384
2014	GSE14333, GSE33113, GSE17538	yes	I-III	Relapse	24829396
2014	GSE17536, GSE17537, GSE38832	yes	II-III	OS, DSS and DFS	25320007
2014	GSE17536, GSE30378	yes	I-IV	DSS	25000257
2014	GSE17538, GSE14333	yes	II-III	RFS	24728738
2014	GSE17538, GSE14333	yes	II-III	RFS, OS	25504183
2014	GSE39582, GSE14333, GSE17536	yes	I-IV	DFS	24809982
2013	GSE14333, GSE17538	yes	I-IV	RFS	23372686
2013	GSE14333, GSE17538, GSE12032	yes	I-IV	DFS, DSS	23658834
2013	GSE17536	yes	I-IV	OS	24247253
2013	GSE17536	yes	I-IV	DSS	23799978
2013	GSE17536, GSE14333	yes	I-III	DFS	22859720
2013	GSE17536, GSE14333	yes	I-IV	OS	23807160
2013	GSE17536, GSE14333, GSE12945	yes	I-IV	OS	24140838
2013	GSE17536, GSE17537	yes	I-III	OS	23922772
2013	GSE17536, the training data was not provided	yes	II-III	OS	24170546
2013	GSE17537	yes	I-IV	RFS	24360964
2013	GSE17538	yes	I -IV	OS	24052018
2013	GSE17538, GSE14333, GSE37892	yes	I-IV	DFS	23626670
2012	GSE12032, GSE17538, GSE17181, GSE4526	yes	II-III	Relapse	22348113
2012	GSE14333, GSE17538, GSE30378, GSE24550	yes	II-III	RFS	22991413
2012	GSE17536, GSE14333	yes	I-IV	RFS	22710688
2012	GSE17536, GSE14333	yes	I-III	DFS	22859720
2012	GSE17536, GSE14333	yes	II-III	RFS	22844451
2012	GSE17537, GSE14333	yes	I-IV	RFS	23153532

2012	GSE29638, GSE24550, GSE30378	no	II	RFS	22213796
2011	GSE5206, GSE10402	yes	I-IV	RFS	21098318
2011	GSE5206, GSE17537	yes	I-IV	OS	22977525
2010	GSE17538	yes	II-III	OS, RFS	19914252
2010	GSE17538, GSE14333	yes	II-III	DFS, DSS	21119668

**Abbreviations:** RFS, relapse-free survival, also called DFS, disease-free survival. OS, overall survival. DSS, disease specific survival.

**Table 2: The CRC datasets used in this work generated on GPL570 platform**

Dataset	Stage I CRC#	Stage II CRC# without CTX	Stage II CRC# with CTX	Stage III CRC#	Stage IV CRC#
GSE39582	33	203	56	205	60
GSE14333	44	72	22	91	61
GSE17536	24	55	0	57	39

**Table 3: Multivariate Cox regression analyses of the 3-GPS**

Clinical Characteristic	HR	Cox p value	95% CI
GPS(High Risk vs Low Risk)	7.5479	$7.28 \times 10^{-6}$	[3.121, 18.257]
Age	2.7269	0.0423	[1.034, 7.182]
Sex (Male vs Femal)	0.9944	0.7477	[0.961, 1.029]
Localization (distal vs proximal)	1.9770	0.1794	[0.731, 5.348]
MSI	0.6845	0.6922	[0.105, 4.471]
Braf mut	3.8353	0.2086	[0.472, 31.178]
Kras mut	0.8843	0.8322	[0.284, 2.757]
Tp53 mut	1.2155	0.6301	[0.549, 2.690]

**Table 4: The consistency of the Risk-DE genes detected from three datasets**

Dataset1	Dataset2	DE genes 1	DE genes2	overlap	consistency
GSE39582	GSE14333	3599	2540	836	98.09%
GSE39582	GSE17536	3599	505	364	99.45%
GSE14333	GSE17536	2540	505	247	100%

with the high risk of relapse, and its expression level in a sample was above the 80th or below the 20th percentile of its expression levels among all the samples, then it was considered to indicate poor prognosis for this particular sample [14]. As the 80th and 20th percentile of a gene's expression values in a set of samples are dependent on the samples analyzed together, the risk classification of a sample by ColoGuideEx may change when it is analyzed together with different samples. We analyzed the 52 stage II CRC samples of the GSE30378 dataset to illustrate this problem. ColoGuideEx classified 45 of the 52 samples into the low-risk group. Applying ColoGuideEx to reanalyze these 45 low-risk samples, 8 samples were reclassified into the high risk group, indicating the uncertainty of this classifier for the risk classification of patients [22].

### The gene pair signature for the relapse risk of stage II CRC

We used the GSE39582 ( $n = 203$ ) dataset with the largest sample size to train a GPS of the relapse risk for stage II CRC and validated it in the GSE14333 and GSE17536 datasets (shown in the Table 2). Because GSE17536 ( $n = 55$ ) included 35 samples, which were technical replicates of the samples of GSE14333, we considered it as a validation dataset for technical reproducibility of the signature.

Based on the hypothesis that the stage II CRC at high risk of relapse could be attributed to micro-metastasis, we firstly extracted 174 and 278 Metastatic-DE genes (Student's *t*-test, FDR < 0.1) from the GSE39582 and GSE14333 datasets, respectively, and then found 149472 and 1154605 significantly reversed gene pairs (Fisher exact test, FDR < 0.2), each consisting at least one of the Metastatic-DE genes, between the metastatic samples (stage III and IV) and the non-metastatic samples (stage

**Table 5: The consistency between the Risk-DE genes and the Metastatic-DE genes**

Dataset1	Risk-DE genes#	Metastatic -DE genes#	Overlap	$p_1$	Consistency	$p_2$
GSE39582	3599	174	41	0.0307	100%	$4.55 \times 10^{-13}$
GSE14333	2540	278	118	$<2.2 \times 10^{-16}$	100%	$<2.2 \times 10^{-16}$
GSE17536	505	45	12	$6.79 \times 10^{-10}$	100%	$2.4 \times 10^{-4}$

Notes: #, the number of Risk-DE genes and Metastatic-DE genes;  $p_1$ , the  $p$  value of overlaps between Risk-DE genes and Metastatic-DE genes;  $p_2$ , the  $p$  value of the concordance score of the overlapped DE genes.

**Table 6: The KEGG function enrichment analysis results**

Pathway name	Adjusted $p$ -values	References(PubMed index)
ECM-receptor interaction	$2.22 \times 10^{-14}$	9854310
Focal adhesion	$7.99 \times 10^{-10}$	15246682
Protein digestion and absorption	$4.97 \times 10^{-5}$	21490305
PI3K-Akt signaling pathway	$2.98 \times 10^{-3}$	7558426
Glycosaminoglycan biosynthesis-chondroitin sulfate/dermatan sulfate	$5.29 \times 10^{-3}$	24035453
Regulation of actin cytoskeleton	$4.38 \times 10^{-2}$	11709869

I) in the two datasets, respectively. The two lists of gene pairs had 6386 overlaps and 99.86% of them had the same reversal patterns in the two datasets, which was unlikely to be observed by chance ( $p < 2.2 \times 10^{-16}$ , the binomial distribution model). Finally, from the 6377 metastasis-associated gene pairs consistently detected in the two datasets, we extracted 15 prognosis-associated gene pairs based on 203 stage II CRC samples from GSE39582 dataset, by univariate Cox proportional-hazards regression model with  $p < 0.01$ . The 15 prognosis-associated gene pairs are listed in Supplementary Table S2. Among these 15 prognosis-associated gene pairs, using the gene pair ORC1-OLR1 with the smallest log-rank  $p$ -value as a seed, we performed a forward selection procedure and obtained an optimal set consisting of three gene pairs reached the smallest  $p$ -value (C-index = 0.625, log-rank  $p = 8.09 \times 10^{-8}$ , HR = 5.209, shown in Figure 3A). We selected these three gene pairs as the final prognostic signature, referred to as 3-GPS (ORC1-OLR1, MTNR1A-VGLL1 and RFX5-MMP14 shown in Supplementary Table S3). For each of the three gene pairs, the  $E_a < E_b$  REO was associated with worse survival. Thus, a simple rule was used to classify patients: a sample was classified into the high-risk group only if at least two gene pairs suggested that this sample was at high risk. A multivariate Cox analysis showed that, after adjusting for age, gender, MSI and localization, the 3-GPS remained significantly associated with patient RFS (log-rank  $p = 7.28 \times 10^{-6}$ , HR = 7.5479, 95% CI, 3.121-18.257, shown in Table 3).

The 3-GPS was validated in the stage II CRC samples without CTX included in the GSE14333 ( $n = 72$ ) dataset. The RFS of the patients in the low-risk groups was significantly longer than that of the patients in the high-risk group (Figure 3B, C-index = 0.7424, log-rank  $p = 0.003$ , HR = 8.270). The 3-GPS was also validated in the stage II CRC samples without CTX included in the GSE17536 ( $n = 55$ ) dataset (Figure 3C, C-index = 0.618, log-rank  $p = 0.059$ , HR = 3.446).

### 3-GPS as a micro-metastatic signature

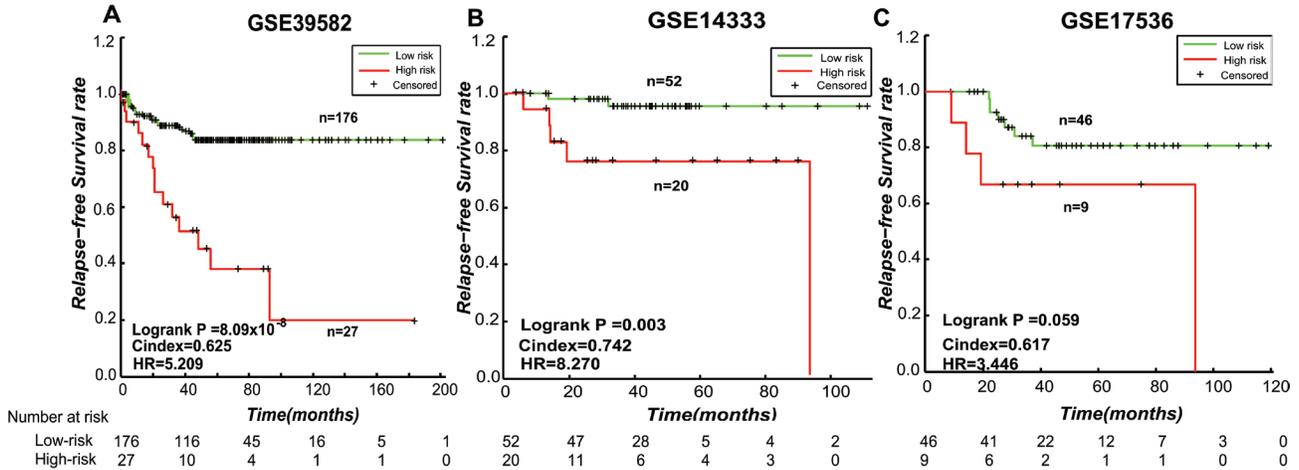
Using the Student's  $t$ -test with 10% FDR control, we extracted DE genes between the high- and low-risk groups identified from the training and validation datasets, denoted as Risk-DE genes, respectively. We found 98.09-100% of the Risk-DE genes commonly detected in any two of the three datasets were consistent in dysregulation directions in the high-risk group compared with the low-risk group, which was unlikely to happen by chance (binomial distribution test,  $p < 2.2 \times 10^{-16}$ , shown in Table 4). This result proved that the 3-GPS could robustly categorize stage II CRC patients into the high- and low-risk groups with distinct transcriptional characteristics. Finally, we obtained 1003 Risk-DE genes by integrating the Risk-DE genes from the three datasets according to the following criterion: DE genes selected in at least two of the three datasets were included in the list, after excluding those DE genes that had inconsistent dysregulation directions in any two datasets.

Moreover, for each of the three datasets, Risk-DE genes were significantly overlapped with the Metastatic-DE genes detected between the stage III-IV samples and stage I samples ( $p = 0.0307$  for GSE39582,  $p < 2.2 \times 10^{-16}$  for GSE14333 and  $p = 6.79 \times 10^{-10}$  for GSE17536, hypergeometric distribution model, shown in Table 5). For the overlapped Risk-DE genes and Metastatic-DE genes, all the dysregulation directions in the high-risk samples compared with the low-risk samples had concordant dysregulation directions in the metastatic samples compared with the non-metastatic samples ( $p = 4.55 \times 10^{-13}$  for GSE39582,  $p < 2.2 \times 10^{-16}$  for GSE14333 and  $p = 2.4 \times 10^{-4}$  for GSE17536, binomial distribution test). Functional enrichment analysis showed that the 1003 Risk-DE genes were significantly enriched in six KEGG pathways (FDR  $< 0.05$ , hypergeometric distribution, shown in Table 6). Five of these six pathways are well-

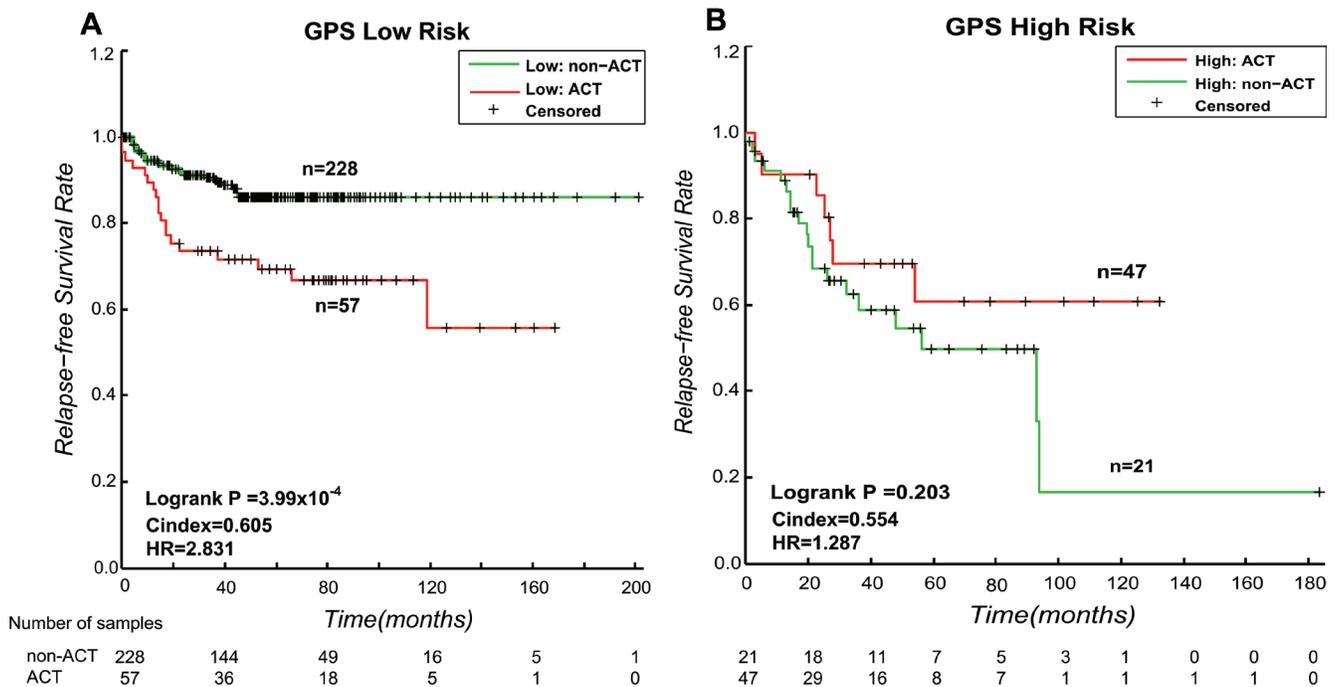
**Table 7: The distribution of stage II CRC predicted by the 3-GPS**

Datasets	L-risk CRC with CTX	L-risk CRC without CTX	H-risk CRC with CTX	H-risk CRC without CTX
GSE39582	46	176	10	27
GSE14333	11	52	11	20
All	57	228	21	47

Abbreviations: CTX, the patients with completely resected tumors who received adjuvant chemotherapy. L-risk or H-risk represents the patients with low or high relapse risk.



**Figure 3: The Kaplan-Meier curves of RFS for stage II CRC samples stratified by the 3-GPS in the training and validation datasets. A.** The training dataset GSE39582; **B.** The independent validation dataset GSE14333; **C.** The validation dataset GSE17536. A sample was classified into high-risk group (red line) only if at least two gene pairs in the 3-GPS voted for high-risk.



**Figure 4: Kaplan-Meier estimates of the RFS of GSE39582 and GSE14333 patients with CTX and non-CTX patients. A.** Kaplan-Meier curves for stage II CRC patients in the low relapse risk group. **B.** Kaplan-Meier curves for stage II CRC patients in the high relapse risk group.

known metastasis-associated pathways, including “ECM-receptor interaction” [35], “Focal adhesion” [36], “PI3K-Akt signaling pathway” [37-39], “Glycosaminoglycan biosynthesis-chondroitin sulfate/dermatan sulfate” [40] and “Regulation of actin cytoskeleton” [41]. The sixth pathway, “Protein digestion and absorption” [42], has also been reported to be associated with CRC development [42]. These results provided evidence that the high risk patients predicted from the stage II CRC patients by the 3-GPS represented patients with micro- metastases.

### 3-GPS risk stratification for potential benefit from CTX

The GSE39582 and GSE14333 datasets also included 56 and 22 samples of stage II CRC patients treated with CTX (shown in Table 7), besides the 203 and 72 samples of stage II CRC patients without CTX treatment, respectively. For each of the dataset, we used the 3-GPS signature to divide all stage II CRC samples with and without CTX into the high- and low-risk groups and compared the RFS between samples with and without CTX in the high- and low-risk groups, respectively.

Due to the small size of the stage II CRC patients with adjuvant CTX, we integrated the samples in these two datasets. The 3-GPS signature assigned 68 (21 samples with CTX and 47 samples without CTX) and 285 (57 samples with CTX and 228 samples without CTX) patients to the high- and low-risk groups, respectively. For the patients predicted to the low relapse risk group, the RFS of the patients with CTX was significantly shorter than that of the patients without CTX (Figure 4A, C-index = 0.6049, log-rank  $p = 3.99 \times 10^{-4}$ , HR = 2.831), indicating that the CTX was not beneficial to patients in the low risk group and even shortened their RFS. In contrast, for the patients predicted to the high relapse risk group, the patients with CTX tended to have significantly longer RFS than the patients without CTX (Figure 4B, C-index = 0.5536, log-rank  $p = 0.203$ , HR = 1.827), indicating that the CTX tended to increase the RFS of patients at high relapse risk.

## DISCUSSION

As shown in this study, most of current studies for the development of CRC prognostic signatures have the data misuse problem and the public data sources need improvement with clear and complete data annotation. Previously reported risk-score based transcriptional signatures tend produce spurious risk classification due to the requirement of data normalization to tackle experimental batch effects [22]. To address this problem, we developed a robust gene pair signature for predicting relapse risk of stage II CRC based on the within-sample REOs of gene pairs. The rank-based signature is robust

against batch effects and data normalizations of gene expression profiling experiments and can be easily applied to samples at the individual level [24, 27] and thus it merits further validation in clinical trial.

In this work, we proposed a hypothesis that the stage II CRC patients at high risk of relapse could be largely attributed to micro-metastases, given that most curative resection surgeries for stage II CRC patients are successful (i.e., no cancer cells are remained after surgeries) and that the random chance of a person to develop cancer ab initio is negligible. Our results supported this hypothesis, as evidenced by the observation the Risk-DE genes between the predicted high- and low-risk samples significantly overlapped with the Metastatic-DE genes detected between the metastatic samples (the stage III-IV samples) and non-metastatic samples (stage I samples) and the deregulation directions of the overlapped DE genes in the high-risk samples compared with the low-risk samples were significantly concordant with the deregulation directions in the metastatic samples compared with the non-metastatic samples. Additionally, the Risk-DE genes were significantly enriched in six well-known metastasis-associated pathways, including “ECM-receptor interaction” [35], “Focal adhesion” [36], “PI3K-Akt signaling pathway” [37-39], “Glycosaminoglycan biosynthesis-chondroitin sulfate/dermatan sulfate” [40], “Regulation of actin cytoskeleton” [41] and “Protein digestion and absorption” [42].

Currently, adjuvant chemotherapy (CTX) is the most common therapeutic regimen prescribed for patients with stage III CRC after surgical resection [43], while its routine use in patients with stage II CRC remains controversial [4, 44]. Our result showed that CTX may reduce RFS of those stage II CRC patients predicted into the low-risk group by 3-GPS. This indicated that the predicted low-risk patients would be indeed at the low risk and thus could not benefit but suffer injury from excessive CTX treatment.

## MATERIALS AND METHODS

### Selection of gene expression-based prognostic studies

We searched the PubMed medical literature database (<http://www.ncbi.nlm.nih.gov/pubmed/>) to identify articles on the analyses of gene expression data for developing prognostic signatures of CRC, published in English between January 1, 2010 and June 1, 2015.

## Microarray data and preprocessing

Three large datasets generated on the GPL570 platform were analyzed in this work (Table 2), which separately included at least 50 samples of stage II CRC patients without CTX treatment. The GSE39582 dataset included 56 samples of stage II CRC patients treated with fluorouracil-based CTX (only fluorouracil and folinic acid after surgery), besides the 203 samples of stage II CRC patients without CTX treatment. The GSE14333 dataset included 22 samples of stage II CRC patients treated with fluorouracil-based CTX (either single agent 5-fluorouracil/capecitabine or 5-fluorouracil and oxaliplatin after surgery) and 72 samples of stage II CRC patients without CTX treatment. All the data were downloaded from the Gene Expression Omnibus database (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) [30].

The raw data (.CEL files) for each dataset was processed using the RMA algorithm for background adjustment without quantile normalization [45]. Then, each probeset ID was mapped to Entrez gene ID with the custom CDF file. If multiple probesets were mapped to the same gene, the expression value for the gene was summarized as the arithmetic mean of the values of multiple probesets (on the log<sub>2</sub> scale).

## Survival analysis

Survival curves were estimated using the Kaplan-Meier method and were compared using the log-rank test [46]. The univariate Cox proportional-hazards regression model was used to evaluate the correlation of gene pairs with the RFS, and the multivariate Cox proportional-hazards regression model was used to evaluate the independent prognostic value of the signature after adjusting for clinical factors including age, gender, stage, MSI status and localization of the tumor (distal or proximal). We adopted the C-index proposed by Harrell et al. [47, 48] to evaluate the overall concordance between the predicted risk classification and the observed RFS.

## Development of the prognostic gene pair signature

Firstly, using Student's *t*-test, we selected differential expression genes between metastatic samples (stage III and IV CRC) and non-metastatic samples (stage I CRC), denoted as Metastatic-DE genes. From gene pairs including at least one Metastatic-DE genes, we selected gene pairs whose REOs were associated with metastasis. Let  $E_a$  and  $E_b$  represent the expression levels of gene *a* and gene *b*, respectively, we compared the frequency of samples with the REO pattern  $E_a > E_b$  between the metastatic CRC samples and the stage I CRC samples using Fisher exact test with 20% FDR control. The

*p*-values were adjusted using the Benjamini-Hochberg (BH) procedure [49]. The overlapped gene pairs between the two lists of significant gene pairs identified from GSE39582 and GSE14333 were defined as metastasis-associated gene pairs.

Then, for each of the metastasis-associated gene pairs, we classified stage II CRC samples without CTX into two groups according to the REO of this gene pair in each sample and compared RFS between the two groups using the univariate Cox proportional-hazards regression model. A gene pair was defined as a prognosis-associated gene pair if the two groups of samples had significantly different RFS. If the  $E_a > E_b$  REO was associated with worse outcome, then we considered that this REO and the reversal REO ( $E_a < E_b$ ) in a cancer sample votes for high and low risk, respectively. Gene pairs with *p* values less than 0.01 were considered as candidate prognosis-associated gene pairs. We chose the gene pair with the smallest *p*-value as a seed and added a prognosis-related gene pair at each iteration until the *p*-value did not decrease based on the classification rule as follows: a sample was classified into the high-risk group if the majority of the REOs of a set of gene pairs within this sample voted for high risk; otherwise, into the low risk group. The optimal set was defined as the gene pair signature (GPS).

The GPS was validated by two independent datasets GSE14333 and GSE17536. Figure 1 describes the flowchart for developing and validating the rank-based GPS for the risk of the relapse on stage II CRC.

## Concordance scores

If two lists of DE genes between the high- and low-risk groups detected separately from two datasets had *k* overlapped genes, among which *s* genes showed the same deregulation directions (up- or down-regulation) in the two DE gene lists, then the concordance score was calculated as *s/k*. The probability of observing a concordance score of *s/k* by chance was evaluated by the cumulative binomial distribution model as following [50]:

$$P = 1 - \sum_{i=0}^{s-1} \binom{k}{i} (P_e)^i (1 - P_e)^{k-i}$$

where  $P_e$  is the probability of one gene having the concordant relationship between the two lists of genes by chance (here,  $P_e = 0.5$ ).

The significance of a score indicated that DE genes extracted from independent datasets were significantly consistent.

## Functional enrichment analysis

The gene categories for functional enrichment analysis were downloaded from KEGG (<http://www.>

[genome.jp/kegg/](http://genome.jp/kegg/)) in July, 2014. The hypergeometric distribution model was used to test whether the number of DE genes annotated in a functional category was significantly more than what expected by random chance:

$$P_k = 1 - \sum_{i=0}^{m-1} \frac{\binom{n}{i} \binom{N-n}{M-i}}{\binom{N}{M}}$$

where  $N$  is the total number of the measured genes with functional annotation;  $n$  is the number of DE genes with functional annotation;  $M$  is the number of the measured genes in a functional category and  $m$  is the number of the DE genes in the functional category. The  $p$ -values were adjusted using the Benjamini-Hochberg (BH) procedure [49].

## ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China (grant numbers: 81372213, 81572935 and 81201822). This work was partially supported by a grant from the Elsa U. Pardee Foundation (Dr. Yang), the Career Development Award of RPCI-UPCI Ovarian Cancer SPORE (P50 CA159981 to Dr. Yang).

## CONFLICTS OF INTEREST

The authors declare that they have no conflict of interests.

## REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C and Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*. 2010; 127 :2893-2917.
2. Schmoll HJ, Van Cutsem E, Stein A, Valentini V, Glimelius B, Haustermans K, Nordlinger B, van de Velde CJ, Balmana J, Regula J, Nagtegaal ID, Beets-Tan RG, Arnold D, Ciardiello F, Hoff P, Kerr D, et al. ESMO Consensus Guidelines for management of patients with colon and rectal cancer. a personalized approach to clinical decision making. *Ann Oncol*. 2012; 23:2479-2516.
3. Gill S, Loprinzi CL, Sargent DJ, Thome SD, Alberts SR, Haller DG, Benedetti J, Francini G, Shepherd LE, Francois Seitz J, Labianca R, Chen W, Cha SS, Heldebrant MP and Goldberg RM. Pooled analysis of fluorouracil-based adjuvant therapy for stage II and III colon cancer: who benefits and by how much? *J Clin Oncol*. 2004; 22:1797-1806.
4. O'Connell MJ, Lavery I, Yothers G, Paik S, Clark-Langone KM, Lopatin M, Watson D, Baehner FL, Shak S, Baker J, Cowens JW and Wolmark N. Relationship between tumor gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin. *J Clin Oncol*. 2010; 28:3937-3944.
5. Sanchez JA, Krumroy L, Plummer S, Aung P, Merkulova A, Skacel M, DeJulius KL, Manilich E, Church JM, Casey G and Kalady MF. Genetic and epigenetic classifications define clinical phenotypes and determine patient outcomes in colorectal cancer. *Br J Surg*. 2009; 96:1196-1204.
6. Markowitz SD and Bertagnolli MM. Molecular origins of cancer: Molecular basis of colorectal cancer. *N Engl J Med*. 2009; 361:2449-2460.
7. Farina-Sarasqueta A, van Lijnschoten G, Moerland E, Creemers GJ, Lemmens VE, Rutten HJ and van den Brule AJ. The BRAF V600E mutation is an independent prognostic factor for survival in stage II and stage III colon cancer patients. *Ann Oncol*. 2010; 21:2396-2402.
8. Roth AD, Tejpar S, Delorenzi M, Yan P, Fiocca R, Klingbiel D, Dietrich D, Biesmans B, Bodoky G, Barone C, Aranda E, Nordlinger B, Cisar L, Labianca R, Cunningham D, Van Cutsem E, et al. Prognostic role of KRAS and BRAF in stage II and III resected colon cancer: results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00 trial. *J Clin Oncol*. 2010; 28:466-474.
9. Walther A, Houlston R and Tomlinson I. Association between chromosomal instability and prognosis in colorectal cancer: a meta-analysis. *Gut*. 2008; 57:941-950.
10. Popat S, Hubner R and Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol*. 2005; 23:609-618.
11. O'Connor ES, Greenblatt DY, LoConte NK, Gangnon RE, Liou JI, Heise CP and Smith MA. Adjuvant chemotherapy for stage II colon cancer with poor prognostic features. *J Clin Oncol*. 2011; 29:3381-3388.
12. Van Schaeuybroeck S, Allen WL, Turkington RC and Johnston PG. Implementing prognostic and predictive biomarkers in CRC clinical trials. *Nat Rev Clin Oncol*. 2011; 8:222-232.
13. Sveen A, Agesen TH, Nesbakken A, Meling GI, Rognum TO, Liestol K, Skotheim RI and Lothe RA. ColoGuidePro: a prognostic 7-gene expression signature for stage III colorectal cancer patients. *Clin Cancer Res*. 2012; 18:6001-6010.
14. Agesen TH, Sveen A, Merok MA, Lind GE, Nesbakken A, Skotheim RI and Lothe RA. ColoGuideEx: a robust gene classifier specific for stage II colorectal cancer prognosis. *Gut*. 2012; 61:1560-1567.
15. Sveen A, Nesbakken A, Agesen TH, Guren MG, Tveit KM, Skotheim RI and Lothe RA. Anticipating the clinical use of prognostic gene expression-based tests for colon cancer stage II and III: is Godot finally arriving? *Clin Cancer Res*. 2013; 19:6669-6677.
16. Kelley RK and Venook AP. Prognostic and predictive markers in stage II colon cancer: is there a role for gene expression profiling? *Clin Colorectal Cancer*. 2011; 10:73-80.
17. Nannini M, Pantaleo MA, Maleddu A, Astolfi A, Formica

- S and Biasco G. Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives. *Cancer Treat Rev.* 2009; 35:201-209.
18. Tabernero J and Baselga J. Multigene assays to improve assessment of recurrence risk and benefit from chemotherapy in early-stage colon cancer: has the time finally arrived, or are we still stage locked? *J Clin Oncol.* 2010; 28:3904-3907.
  19. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA, Jr., Marks JR, Dressman HK, West M and Nevins JR. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature.* 2006; 439:353-357.
  20. Dry JR, Pavey S, Pratilas CA, Harbron C, Runswick S, Hodgson D, Chresta C, McCormack R, Byrne N, Cockerill M, Graham A, Beran G, Cassidy A, Haggerty C, Brown H, Ellison G, et al. Transcriptional pathway signatures predict MEK addiction and response to selumetinib (AZD6244). *Cancer Res.* 2010; 70:2264-2273.
  21. Gatz ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, Datto MB, Kelley M, Mathey-Prevot B, Potti A and Nevins JR. A pathway-based classification of human breast cancer. *Proc Natl Acad Sci U S A.* 2010; 107:6994-6999.
  22. Qi L, Chen L, Li Y, Qin Y, Pan R, Zhao W, Gu Y, Wang H, Wang R, Chen X and Guo Z. 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.* 2015.
  23. Marusyk A, Almendro V and Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer.* 2012; 12:323-334.
  24. Wang H, Sun Q, Zhao W, Qi L, Gu Y, Li P, Zhang M, Li Y, Liu SL and Guo Z. Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics.* 2015; 31:62-68.
  25. Xu L, Tan AC, Winslow RL and Geman D. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics.* 2008; 9:125.
  26. Zhou X, Li B, Zhang Y, Gu Y, Chen B, Shi T, Ao L, Li P, Li S, Liu C and Guo Z. A relative ordering-based predictor for tamoxifen-treated estrogen receptor-positive breast cancer patients: multi-laboratory cohort validation. *Breast Cancer Res Treat.* 2013; 142:505-514.
  27. Geman D, d'Avignon C, Naiman DQ and Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol.* 2004; 3:Article19.
  28. Boutros PC, Pintilie M, John T, Starmans MH, Der SD, Shepherd FA, Tsao MS and Jurisica I. Re: Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst.* 2010; 102:1677-1678; author reply 1678-1679.
  29. Subramanian J and Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst.* 2010; 102:464-474.
  30. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 2013; 41:D991-995.
  31. Kuo TY, Hsi E, Yang IP, Tsai PC, Wang JY and Juo SH. Computational analysis of mRNA expression profiles identifies microRNA-29a/c as predictor of colorectal cancer early recurrence. *PLoS One.* 2012; 7:e31587.
  32. Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, Lu P, Johnson JC, Schmidt C, Bailey CE, Eschrich S, Kis C, Levy S, Washington MK, Heslin MJ, Coffey RJ, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology.* 2010; 138:958-968.
  33. Freeman TJ, Smith JJ, Chen X, Washington MK, Roland JT, Means AL, Eschrich SA, Yeatman TJ, Deane NG and Beauchamp RD. Smad4-mediated signaling inhibits intestinal neoplasia by inhibiting expression of beta-catenin. *Gastroenterology.* 2012; 142:562-571 e562.
  34. Zheng Y, Zhou J and Tong Y. Gene signatures of drug resistance predict patient survival in colorectal cancer. *Pharmacogenomics J.* 2015; 15:135-143.
  35. Lukashev ME and Werb Z. ECM signalling: orchestrating cell behaviour and misbehaviour. *Trends Cell Biol.* 1998; 8:437-441.
  36. Wozniak MA, Modzelewska K, Kwong L and Keely PJ. Focal adhesion regulation of cell behavior. *Biochim Biophys Acta.* 2004; 1692:103-119.
  37. Gonzalez-Angulo AM, Ferrer-Lozano J, Stemke-Hale K, Sahin A, Liu S, Barrera JA, Burgues O, Lluch AM, Chen H, Hortobagyi GN, Mills GB and Meric-Bernstam F. PI3K pathway mutations and PTEN levels in primary and metastatic breast cancer. *Mol Cancer Ther.* 2011; 10:1093-1101.
  38. Osaki M, Oshimura M and Ito H. PI3K-Akt pathway: its functions and alterations in human cancer. *Apoptosis.* 2004; 9:667-676.
  39. Nakanishi K, Sakamoto M, Yasuda J, Takamura M, Fujita N, Tsuruo T, Todo S and Hirohashi S. Critical involvement of the phosphatidylinositol 3-kinase/Akt pathway in anchorage-independent growth and hematogenous intrahepatic metastasis of liver cancer. *Cancer Res.* 2002; 62:2971-2975.
  40. Timar J, Diczhazi C, Bartha I, Pogany G, Paku S, Raso E, Tovari J, Ladanyi A, Lapis K, Kopper L and et al. Modulation of heparan-sulphate/chondroitin-sulphate ratio by glycosaminoglycan biosynthesis inhibitors affects liver metastatic potential of tumor cells. *Int J Cancer.* 1995; 62:755-761.
  41. Shibue T, Brooks MW and Weinberg RA. An integrin-

- linked machinery of cytoskeletal regulation that enables experimental tumor initiation and metastatic colonization. *Cancer Cell*. 2013; 24:481-498.
42. Hughes R, Magee EA and Bingham S. Protein degradation in the large intestine: relevance to colorectal cancer. *Curr Issues Intest Microbiol*. 2000; 1:51-58.
  43. Wong AC, Stock S, Schrag D, Kahn KL, Salz T, Charlton ME, Rogers SO, Jr., Goodman KA and Keating NL. Physicians' beliefs about the benefits and risks of adjuvant therapies for stage II and stage III colorectal cancer. *J Oncol Pract*. 2014; 10:e360-367.
  44. Meropol NJ. Ongoing challenge of stage II colon cancer. *J Clin Oncol*. 2011; 29:3346-3348.
  45. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4:249-264.
  46. FT. HD. A class of rank test procedures for censored survival data. *Biometrika*. 1982; 69:553-566.
  47. Harrell FE, Jr., Lee KL and Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996; 15:361-387.
  48. Pencina MJ, D'Agostino RB, Sr. and Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. *Stat Med*. 2012; 31:1543-1553.
  49. Benjamini A HY. Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*. 1995:289-300.
  50. Gong X, Wu R, Wang H, Guo X, Wang D, Gu Y, Zhang Y, Zhao W, Cheng L, Wang C and Guo Z. Evaluating the consistency of differential expression of microRNA detected in human cancers. *Mol Cancer Ther*. 2011:752-760.