

LETTER TO EDITOR

Identification of a novel 15-gene expression signature predicting overall survival of human colorectal cancer

Dear Editor,

In this study, we developed a robust and clinically applicable 15-gene prognostic signature for colorectal cancer (CRC) patients using our multi-step bioinformatics analysis strategy. So far, no multigene expression signature is available for CRC in China where the tumor incidence is rapidly increasing.

Globally, this disease is the fourth most common of all cancer types and third leading cause of cancer mortality, with 1.8 million incident cases in 2017.¹ Currently, the 5-year relative survival is 65% for CRC patients, but much lower (12%) for stage IV. With the application of next-generation sequencing and microarray technologies, several genomic biomarkers have emerged to stratify CRC patients and predict clinical outcome. Of these, the RT-PCR-based 7-gene *Oncotype DX* colon cancer panel²⁻⁴ has been utilized as a clinical tool for the prediction of recurrence risk for stage II and III CRCs. However, there is still an urgent need to develop better and more robust gene signatures in this area. In this study, we developed a clinically applicable and robust prognostic signature for CRC patients using our multistep bioinformatics analysis strategy.⁵⁻⁸

For our gene expression signature development, we first identified 738 genes (971 gene probe IDs) that were consistently deregulated in CRC versus normal colon tissue in six transcriptome datasets (Figure 1A and B, Figure S1A, Table S1). Kaplan-Meier survival analysis together with Cox regression was used to evaluate whether their expression levels were associated with overall survival (OS) in GSE17536 dataset. Out of 738 genes, expression levels of 78 genes were associated with OS ($P < .05$) (Figure 1C, Table S2). These genes were significantly enriched for GO biological processes related to wound healing, collagen fibril organization, endothelial cell proliferation, and apoptosis (Figure S1C; $P < .05$).

We then employed a cross-validation method to resample the 373 patients in TCGA-COAD (where TCGA is the

cancer genome atlas) into 100 randomly selected training (248 cases) and test sets (125 cases) (Figure 2A). We utilized training sets to define a signature for prognosis, and to build a scoring system and prediction model, whereas test sets were used for validation. Multivariate Cox regression was carried out on all 100 training sets to identify which of the 78 genes were significantly associated with OS. Genes were ranked based on the frequency of selection in the Cox models. We next employed concordance statistics for Cox modeling to further refine the gene set with respect to their goodness-of-fit in survival models. Specifically, step-wised inclusion of candidate genes into the Cox regression model based on their rank order indicated a saturated concordance statistic using the top 15 genes (Figure 2B). A 15-gene prognostic score for a CRC patient was defined as the linear combination of logarithmically transformed gene expression levels weighted by average Cox regression co-efficient obtained from 100 training sets (Table S3). Based on the average cut point score from the training sets, patients were divided into three prognostic groups. The OS rates of good, intermediate, and poor groups for all samples in TCGA-COAD were significantly different based on Kaplan-Meier analysis and log-rank test ($P = .0001$; Figure 2D). In each test set, we compared the hazard ratios (HRs) of the intermediate versus good and poor versus good outcomes (Figure 2C). The distribution of the three survival outcome groups varied significantly across stage ($P = .0061$; Figure S2A), with increased numbers of patients in the “poor” group with increasing stage. Nevertheless, our signature significantly separated the good from the poor outcome groups for stage II, III, and IV CRC patients ($P < .05$; Figure S2B-E). Kaplan-Meier survival analysis in the TCGA-COAD dataset stratified by subtype annotations (chromosomal instability [CIN], genome stable [GS], microsatellite instability [MSI], and hypermutated-single nucleotide variant [HM-SNV])⁹ did not show an enhancement in a particular subtype (Figure S3). Taken together, these

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Clinical and Translational Medicine* published by John Wiley & Sons Australia, Ltd on behalf of Shanghai Institute of Clinical Bioinformatics

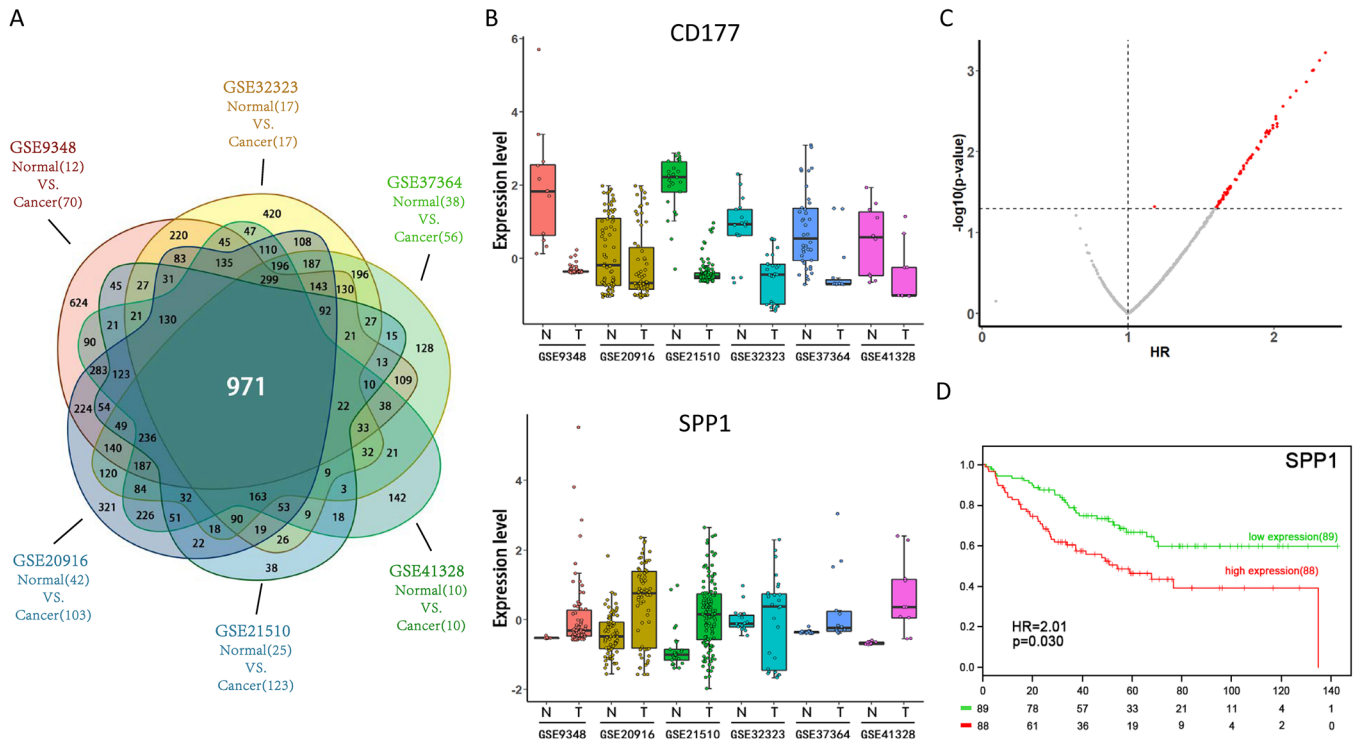


FIGURE 1 Identification of genes consistently deregulated in human colorectal cancers (CRCs) and significantly associated with overall survival (OS) of CRC patients. (A) Venn diagram of genes significantly and consistently deregulated (more than fivefold change and adjusted P -value $< .05$) in CRC compared to normal colon tissue gene expression across six publicly available gene transcript datasets. (B) Two representative genes significantly and consistently up- or downregulated in CRC (T) compared to normal colon (N) tissue gene expression across all six datasets. (C) Volcano plot of the hazard ratio (HR) and P -value of the association of 738 genes with OS. The 78 genes that were significantly associated with the OS of CRC ($HR, P < .05$) are highlighted in red. (D) Kaplan-Meier survival analysis for the representative gene SPP1 from the 78 genes significantly associated with OS in CRC patients. The CRC patient cohort was divided into two groups based on the median value. The log-rank P -value and HR of the curve comparison between the high expression (red) and low expression (green) groups is shown

findings demonstrate that our 15-gene signature has clear discriminative capability to stratify CRC patients based on good versus poor prognosis.

To independently validate this signature, we performed Cox regression analysis in GSE28722 and GSE39582 to calculate the Cox regression co-efficient for each of 15 genes, using the same method described above. As shown in Figure 2E and F, high prognostic score patients had a significantly shortened OS compared to low score patients (GSE28722: $P = .0033$; GSE39582: $P = .00058$). Moreover, the signature and the prognostic prediction model were tested in a cohort of 203 patients with stage I or II CRC from Nanjing Drum Tower Hospital (Figure S4). Gene expression in formalin-fixed paraffin-embedded (FFPE) specimens was measured using an mRNA hybridization-based assay.¹⁰ OS analysis demonstrated significantly different OS rates ($P < .0001$ by log-rank test) among the three prognostic score groups in the cohort (Figure 2G), with distribution of 32.0%, 35.0%, and 33.0% in good, intermediate, or poor prognostic score groups, respectively. These results strongly support the prognostic capability of the

15-gene signature and score system in an independent Chinese patient cohort with early-stage CRC.

To examine whether the prognostic effects of our signature is independent of clinicopathological factors potentially associated with patient outcomes, a multivariate Cox regression analysis was carried out on all available parameters and our signature in both TCGA-COAD (Table S5) and our own hospital cohort (Table 1 and Table S6). These data support that the prognostic effectiveness of the 15-gene signature was independent of clinical parameters, including molecular subtypes ($P < .05$).

Finally, with Cox regression analysis, the prognostic power of our signature was compared with the 7-gene panel in the *Oncotype DX* Colon Test using two datasets. In GSE17536, the median HR of our signature for poor versus good outcomes was 2.32-fold higher compared to the 7-gene signature, and in GSE28722, this fold difference was 1.58 (Figure S5) indicating that our 15-gene score outperformed the 7-gene signature in predicting CRC patient OS.

In conclusion, we identified a novel 15-gene prognostic signature and developed a score system that robustly and

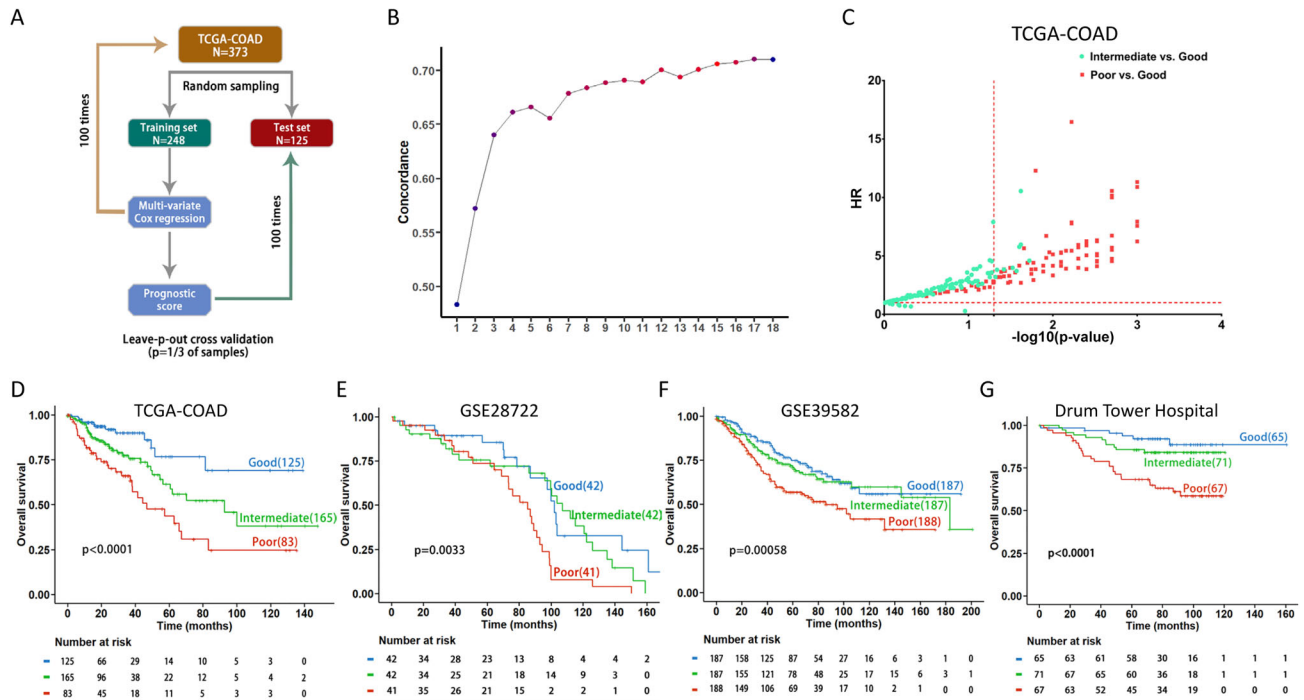


FIGURE 2 Development and validation of a multigene prognostic signature and scoring system for CRC. (A) Workflow used to generate a prognostic score system for CRC using a resampling/cross-validation approach with Cox regression analysis. (B) Concordance statistic: Step-wise inclusion of candidate genes into the Cox regression model. (C) HR analysis in the TCGA-COAD dataset; (D) Kaplan-Meier survival analysis for the TCGA-COAD test cohort using the 15-gene expression-based prognostic scores. The log-rank P -value of the curve comparison between the groups is shown. (E and F) Independent validation of the 15-gene signature in two CRC datasets GSE28722 and GSE39582, respectively. The log-rank P -value of the curve comparison between the groups is shown. (G) Independent validation of the 15-gene prognostic score system using 203 CRC patient samples from the Nanjing Drum Tower Hospital. The patients were divided into three groups based on the cutoff of prognostic score calculated from the 135 patient training cohort. The log-rank P -value of the curve comparison between the groups is shown

TABLE 1 Multivariate Cox regression analysis on Nanjing Drum Tower Hospital cohort

Variables	HR	95% CI		P-value
		Lower limit	Upper limit	
Baseline information				
T grade				.006
T3 vs T2	3.121	0.421	23.167	.266
T4 vs T2	15.619	1.696	143.832	.015
WHO classification				.077
Well differentiation vs moderate differentiation	1.019	0.355	2.921	.973
Poor differentiation vs moderate differentiation	0.381	0.042	3.426	.389
Mucinous adenocarcinoma vs moderate differentiation	2.529	0.771	8.303	.126
Group				.002
Intermediate vs good	1.802	0.627	5.178	.274
Poor vs good	4.206	1.693	10.446	.002

reliably predicts patient OS. The signature was validated in two independent public datasets and in our hospital cohort in Nanjing, China. The signature is independent of clinical factors as well as molecular classifiers.

ACKNOWLEDGMENT

We thank the pathologists at Nanjing Drum Tower Hospital for providing tissue samples and initial quality control. We also thank Ms Shuang Wu and Mr Jinlong Cui from the

Berkeley-Nanjing Research Center for technical support in the mRNA hybridization assay.

CONFLICT OF INTEREST

All authors declare that there is no conflict of interest.

ETHICS STATEMENT

This study was approved by the Ethics Committee of the Nanjing Drum Tower Hospital (document no: 2020-040-01), and written informed general consent was obtained from each patient.

FUNDING INFORMATION

National Natural Science Foundation of China; Grant Number: 81802388 (to Pin Wang); Natural Science Foundation from the Department of Science & Technology of Jiangsu Province; Grant Number: BK20180120 (to Pin Wang).

AUTHOR CONTRIBUTIONS

Study concept and design: Pin Wang, Xiaoping Zou, Jian-Hua Mao, Antoine M. Snijders, and Bo Hang; *data acquisition:* Chengfei Jiang, Yue Zhao, Pin Wang, and Binbin Yuan; *statistical data analysis:* Jian-Hua Mao, Pin Wang, and Chengfei Jiang; *manuscript drafting:* Pin Wang, Chengfei Jiang, Bo Hang, and Jian-Hua Mao; *funding and study supervision:* Pin Wang and Xiaoping Zou; *manuscript editing:* all authors.

DATA AVAILABILITY STATEMENT

Requests for the datasets utilized for the current study will be reviewed and considered by the corresponding authors.

Chengfei Jiang^{1,#}


Yue Zhao^{2,#}

Binbin Yuan¹


Hang Chang³

Bo Hang³

Antoine M. Snijders³

Jian-Hua Mao³ 

Xiaoping Zou¹

Pin Wang¹ 

¹ Department of Gastroenterology, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, China

² Department of Gynecology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China

³ Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California

Correspondence

Pin Wang and Xiaoping Zou, Department of Gastroenterology, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing 210008, China.

Email: pinwang729@126.com (P.W.); 13770771661@163.com (X.Z.)

#Both the authors contributed equally to this work.

ORCID

Jian-Hua Mao  <https://orcid.org/0000-0001-9320-6021>

Pin Wang  <https://orcid.org/0000-0001-8939-5413>

REFERENCES

1. GBD 2017 Colorectal Cancer Collaborators. The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol Hepatol.* 2019;12:913-933.
2. Clark-Langone KM, Sangli C, Krishnakumar J, Watson D. Translating tumor biology into personalized treatment planning: analytical performance characteristics of the Oncotype DX Colon Cancer Assay. *BMC Cancer.* 2010;10:691.
3. Srivastava G, Renfro LA, Behrens RJ, et al. Prospective multicenter study of the impact of oncotype DX colon cancer assay results on treatment recommendations in stage II colon cancer patients. *Oncologist.* 2014;19:492-497.
4. Brenner B, Geva R, Rothney M, et al. Impact of the 12-gene colon cancer assay on clinical decision making for adjuvant therapy in stage II colon cancer patients. *Value Health.* 2016;19:82-87.
5. Wang P, Wang Y, Hang B, et al. A novel gene expression-based prognostic scoring system to predict survival in gastric cancer. *Oncotarget.* 2016;7:55343-55351.
6. Chen E-G, Wang P, Lou H, et al. A robust gene expression-based prognostic risk score predicts overall survival of lung adenocarcinoma patients. *Oncotarget.* 2018;9:6862-6871.
7. Zhao Y, Yang S-M, Jin Y-L, et al. A robust gene expression prognostic signature for overall survival in high-grade serous ovarian cancer. *J Oncol.* 2019;2019:3614207.
8. Zhu L, Wang H, Jiang C, et al. Clinically applicable 53-gene prognostic assay predicts chemotherapy benefit in gastric cancer: a multicenter study. *EBioMedicine.* 2020;61:103023.
9. Liu Y, Sethi NS, Hinoue T, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell.* 2018;33:721-735.e8.
10. Canales RD, Luo Y, Willey JC, et al. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol.* 2006;24:1115-1122.

SUPPORTING INFORMATION

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