Development and validation of an individualized gene expression-based signature to predict overall survival in metastatic colorectal cancer

Shu-Biao Ye¹,², Yi-Kan Cheng³, Jian-Cong Hu¹,², Feng Gao¹,², Ping Lan¹,²

¹Department of Colorectal Surgery, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, China; ²Guangdong Institute of Gastroenterology, Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Guangzhou 510655, China; ³Department of Radiation Oncology, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510655, China

Contributions: (I) Conception and design: P Lan, F Gao; (II) Administrative support: P Lan; (III) Provision of study materials or patients: JC Hu, P Lan; (IV) Collection and assembly of data: SB Ye, YK Cheng, JC Hu; (V) Data analysis and interpretation: SB Ye, F Gao; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

These authors contributed equally to this work.

Correspondence to: Ping Lan, MD; Feng Gao, PhD. Department of Colorectal Surgery, The Sixth Affiliated Hospital, Sun Yat-sen University; Guangdong Provincial Key Laboratory of Colorectal and Pelvic Floor Diseases, Guangzhou 510655, China. Email: lanping@mail.sysu.edu.cn; gaoF57@mail.sysu.edu.cn.

Background: Metastatic colorectal cancer (mCRC) is a heterogeneous disease. Predictive biomarkers are in great demand to optimize patient selection at high risk for death and to provide a novel insight into potential targeted therapy.

Methods: The present study retrospectively analyzed the gene expression profiles of tumor tissue samples from 4 public CRC cohorts, including 1 RNA-Seq data set from The Cancer Genome Atlas (TCGA) CRC cohort and 3 microarray data sets from GEO. Prognostic analysis was performed to test the predictive value of prognostic gene signature.

Results: Of 192 patients, 108 patients (56.3%) were men and median age was 65 years. A prognostic gene signature that consisted of 15 unique genes was generated in the discovery cohort. In the meta-validation cohorts, the signature significantly classified patients into high-risk and low-risk groups with regard to overall survival (OS) in mCRC patients with advanced stage disease and remained as an independent prognostic marker in multivariable analysis (1.57; 95% CI: 1.16–2.11; P=0.003) after adjusting for clinical parameters and molecular types. Gene Set Enrichment Analysis showed that several biological processes, including angiogenesis (P<0.001), epithelial mesenchymal transit (P<0.001) and inflammatory response (P=0.001), were enriched among this prognostic gene signature.

Conclusions: The proposed prognostic gene signature is a promising prognostic tool to estimate OS in mCRC. Prospective larger studies to examine the clinical utility of the biomarkers to guide individualized treatment of mCRC are warranted.

Keywords: Metastatic colorectal cancer (mCRC); TCGA; gene signature; prognostic

Submitted Sep 20, 2019. Accepted for publication Dec 03, 2019.
doi: 10.21037/atm.2019.12.112
View this article at: http://dx.doi.org/10.21037/atm.2019.12.112
Introduction

Metastatic colorectal cancer (mCRC) ranks the third most common cause of death from cancer worldwide (1), and the incidence of mCRC is growing, especially among younger patients (2). mCRC is a highly heterogeneous disease, which can show a wide range of clinical behavior, from curable oligometastatic disease to rapidly developing lethal disease (3).

Biomarkers which can reliably evaluate disease progression and survival outcome would have great value in guiding the treatment of mCRC. For instance, only a small proportion of mCRC patients are responsive to EGFR receptor (EGFR)-targeted or VEGF receptor (VEGFR)-targeted therapies (4-6), while no appropriate targeted therapies have been shown efficacy to the remaining patients. Thus, identification of new biomarkers to optimize patient selection at highest risk for death and to provide a novel insight into target therapy is warranted.

Current prognostic models use histoclinical parameters for prognostication of patients have limitation in capturing molecular heterogeneity of this disease. Previous studies have provided some prognostic mRNAs for mCRC patients (7-10). However, none has been applied into clinical use due to issues such as lack of consideration of other gene expressions and adequate validation. The availability of public, large-scale gene expression data profiling provides the opportunity to define reliable mCRC markers. Multiple gene expression data sets were combined to develop and validate an individualized gene-expressed signature for the survival of mCRC. The aim of this study is to identify the potential prognostic gene expression-based biomarkers of metastatic tumors.

Methods

Patients

Gene expression profiles of frozen colorectal cancer tumor tissue samples were from 4 public cohorts, including 73 stage IV CRC patients from TCGA CRC (11) as discovery cohort and 3 microarray data sets (GSE39582, GSE39084, and GSE17536) (12-14) obtained from Gene Expression Omnibus (GEO) database that were merged into a meta-validation cohort. We retrospectively analyzed these profiles. TCGA CRC cohort was downloaded from Broad GDAC Firehose (http://gdac.broadinstitute.org/) and transcripts per million (TPM) of level 3 RNA-Seq data in log2 scale were used. Other data sets were obtained directly in its processed format from GEO database through Bioconductor package ‘GEOquery’. Overall, 192 mCRC patients with valid survival information were included in this study. The batch effects were corrected using ‘combat’ algorithm implemented in R package ‘sva’ and z-scores for each gene were used for the following analyses. Data were collected from Dec 03 to Feb 04, 2018. We carefully reviewed both paper charts and electronic medical records when necessary. The present study obtained ethics approval from Sun Yat-sen University, Sixth Affiliated Hospital.

Construction and validation of mCRC prognostic gene signature

In order to construct a mCRC Prognostic Gene Signature (PGS), we first identified a list of candidate genes with relatively large variation [(median absolute deviation (MAD) >0.5]. Furthermore, to increase the robustness of the identification for the limited sample size, prognostic signature genes were further selected using the log-rank test with 100 randomizations (80% portion of samples each time) to assess the correlation between each candidate gene and patients’ overall survival (OS) in the discovery cohort. The genes showed significance repeatedly were selected as the candidates of the mCRC prognostic signature. For minimize over-fitting risk, we applied a Cox proportional hazards regression model on all advanced stage samples (stage III/IV) in combination of the least absolute shrinkage and selection operator (LASSO) (glmnet, version 2.0-16). The penalty parameter was calculated by 10-fold cross-validation in the training data set at the minimum partial likelihood deviance.

In order to stratify patients into low-risk or high-risk subgroups, the optimal cutoff value was determined by a time-dependent receiver operating characteristic (ROC) curve (survival ROC, version 1.0.3) at 3 years in the training dataset. The ROC curve was estimated by the Kaplan-Meier estimation method. We used the shortest distance between point representing the 100% true positive rate and 0% false-positive rate and the ROC curve as the cutoff value.

The prognostic value of the PGS was assessed in mCRC patients in the training and independent validation cohorts in univariable analyses respectively. Then we compared the risk scores derived from the gene signature with available clinicopathologic parameters in multivariable analyses.

Functional annotation and analysis

To investigate the biological characteristic of the gene
signature, we conducted enrichment analysis for interested biological pathways, Bioconductor package ‘HTSanalyzeR’ was used to perform by Gene Set Enrichment Analysis (GSEA) (15).

### Statistical Analysis

R software (version 3.5.1; http://www.Rproject.org) was used to conduct statistical analysis. Statistical description was analyzed for all variables. These included frequencies for categorical parameters, and means and standard deviations (SD) or medians and interquartile ranges (IQR) for continuous parameters. Continuous values were compared using Student-t tests between different groups. Univariable analysis of the correlation of PGS and other clinicopathologic features with OS was estimated using log-rank test. For features significantly correlated with OS in univariable analyses, the Cox proportional hazards regression model was used to perform multivariable analysis. P value less than 0.05 was defined as statistical significance in all tests.

### Results

#### Construction and definition of the gene expression signature

A total of 192 CRC patients were included in the analysis. The baseline clinical characteristics of training and validation data sets were shown in Table 1. From TCGA dataset, 18,113 genes were measured by all platforms and 4,172 genes were filtered by the conditions with MAD more than 0.5 and expression level more than median level. Using Cox regression to resample the discovery cohort, the association of 4,172 genes with OS was analyzed, resulting in 197 robust prognostic genes (>75 times showed significant during resampling). Then a prognostic gene signature (PGS) consisting of 15 genes was constructed with the use of LASSO Cox proportional hazards regression on the training cohort (Figure 1 and Table S1). The optimal cutoff from ROC curve analysis for the PGS to classify patients into the high or low risk group was 0.971 for 3-year OS (Figure 2).

#### Validation of the gene expression signature

Univariate analysis showed that the PGS classified patients into low-risk and high-risk groups in terms of OS in mCRC patients from TCGA dataset (Figure 3A) and meta-validation (Figure 3B). After adjusting for clinical features such as age, gender, tumor location and molecular types,
Figure 1 Identification and selection of prognostic genes by LASSO Cox proportional hazards regression. (A) LASSO coefficient profiles of the 15 robust prognostic genes; (B) clustering of the top 15 robust prognostic genes (rows) was identified by LASSO Cox proportional hazards regression in the training dataset from TCGA. The heatmap reflects relative mRNA expression levels.

PGS still resulted as an independent prognostic factor in multivariate analyses (11.54; 95% CI: 4.44–29.99; P<0.001). A higher GPI was associated with significantly poorer prognosis in the independent meta-validation cohorts (1.57; 95% CI: 1.16–2.11; P=0.003). Overall, the PGS may estimate OS independently of clinical parameters in mCRC (Table 2).

**Functional annotation of the prognostic genes**

We further investigate the potential functional mechanisms between the high risk and low risk divided by PGS. Gene Set Enrichment Analysis (GSEA) was performed between predicted high-risk vs. low-risk groups for cancer hallmark pathways, and identified several cancer-related biological processes gene sets including angiogenesis (P<0.001), epithelial mesenchymal transit (P<0.001), inflammatory response (P=0.001), TNF-α-NF-κB (P<0.001), IL6-JAK-
Table 2 Univariate and multivariable analysis of molecular, clinical and prognostic gene signature in training and validation cohorts

<table>
<thead>
<tr>
<th>Univariate</th>
<th>Training cohort</th>
<th>Meta-cohort microarray validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate analysis</td>
<td>Multivariable analysis</td>
</tr>
<tr>
<td></td>
<td>HR (95% CI)</td>
<td>P</td>
</tr>
<tr>
<td>PGS</td>
<td>13.17 (5.20–33.39)</td>
<td>0.001*</td>
</tr>
<tr>
<td>Sex</td>
<td>1.90 (0.60–6.00)</td>
<td>0.26</td>
</tr>
<tr>
<td>Age</td>
<td>1.03 (0.98–1.09)</td>
<td>0.18</td>
</tr>
<tr>
<td>Tumor location</td>
<td>0.25 (0.09–0.70)</td>
<td>0.0045</td>
</tr>
<tr>
<td>T stage</td>
<td>2.39 (0.85–6.73)</td>
<td>0.095</td>
</tr>
<tr>
<td>N stage</td>
<td>0.58 (0.13–2.62)</td>
<td>0.47</td>
</tr>
<tr>
<td>MMR status</td>
<td>1.13 (0.36–3.60)</td>
<td>0.83</td>
</tr>
<tr>
<td>KRAS mutation</td>
<td>0.46 (0.05–4.46)</td>
<td>0.49</td>
</tr>
</tbody>
</table>

*, P<0.001. PGS, prognostic gene signature; HR, hazard ratio; CI, confidence interval.

Figure 3: Univariate analyses of prognostic gene signature in terms of OS in mCRC patients from TCGA dataset (A) and meta-validation (B). OS, overall survival; mCRC, metastatic colorectal cancer.

**Discussion**

Patients with mCRC are at substantial risk for death. The treatment for mCRC is complicated and the outcome is not promising. Therefore, reliable prognostic markers are in great need to select patients at the highest risk for death and require more intensive treatment. Here a prognostic gene expression signature was developed for mCRC and was further validated in independent multiple datasets. This prognostic gene signature stratifies stage IV CRC patients into subsets with different survival outcomes.

Recently several independent studies have proposed prognostic subtypes based on distinct global gene expression profiles to improve the stratification and treatment of patients with mCRC.
CRC patients (16-19). However, none has investigated the role of gene expression in mCRC that harbor tumor heterogeneity. Thus, an individualized approach to stratify patients and guide treatment of mCRC is in great need. To provide a more accurate calculation of OS, we integrated clinicopathological characteristics and gene expression-based signatures from multiple public data sets and applied methods which are exclusively designed across different platforms with RNA-Seq or microarray technologies. Moreover, our prognostic gene signature was validated by multiple independent data sets, which may provide opportunity to translate into clinical routine practice.

Prognostic or predictive biomarkers which was associated with tumor microenvironment may hold great promise to identify new molecular targets and improve patient individualized management. Our analysis showed that gene expressions from angiogenesis signal pathway were associated with survival outcome of mCRC, which was consistent with the previous findings. The only antiangiogenic agent—bevacizumab, is approved by the US Food and Drug Administration for the first-line treatment of mCRC and shows improvement in response rate (RR) and progression-free survival (PFS) (20-22). Furthermore, some genes contained in signature played an important role in the pathways of epithelial mesenchymal transition (EMT), which is responsible for the development and aggressiveness of mCRC. Similarly, Calon et al. identified stromal gene expressions for EMT which defined the poor-prognosis subtype in CRC (23). Our proposed gene signatures also implied the crucial role of inflammatory...
response in mCRC. As many studies showed, an increased inflammatory microenvironment has been demonstrated to be an important element of neoplastic process and tumor progression (24-27). Our GSEA identified TNF-α-NF-κB and IL6-JAK-STAT3 pathways, which were well-known to play a crucial role in the progression and proliferation of mCRC in numerous studies (28-30). Above all, our proposed gene signatures included the molecules from various crucial biological processes.

Limitations of the present study include its retrospective nature, although we validated the signatures in independent data sets. Furthermore, gene expression-based signatures that are subject to the samples from primary tumor or metastatic disease may have inconsistent genetic heterogeneity. Although we investigated as many genes as possible, future studies are need to explore different biological processes that could provide a more comprehensive molecular landscape of mCRC.

Conclusions

In summary, the proposed gene expression-based signature is a promising prognostic tool to predict survival in mCRC. Further prospective studies are in need to validate its feasibility of analysis for assessing prognoses and to exam its clinical utility in personized treatment of mCRC.

Acknowledgments

Funding: This work was supported by National Natural Science Foundation of China (Grant No. 81703060, 81802441), Natural Science Foundation of Guangdong Province (Grant No. 2017A030310644), China Postdoctoral Science Foundation funded project (Grant No. 2018T110911, 2019B020229002), the National Key Research and Development Program of China (N0.2017YFC1308800), Science and Technology Planning Project of Guangdong Province (No. 20160916, 2014SC111, 201902020009), and National Key Clinical Discipline.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Data can be obtained from Gene Expression Omnibus (GEO) database and Broad GDAC Firehose (http://gdac.broadinstitute.org/), therefore the ethical review is exempted.

References

11. Comprehensive molecular characterization of human

<table>
<thead>
<tr>
<th>Gene</th>
<th>Name</th>
<th>Coefficient</th>
<th>Average P value</th>
<th>Frequency in resampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACOT11</td>
<td>Acyl-CoA thioesterase 11</td>
<td>-0.07332</td>
<td>0.004561</td>
<td>0.98</td>
</tr>
<tr>
<td>C12orf45</td>
<td>Chromosome 12 open reading frame 45</td>
<td>0.151998</td>
<td>0.005777</td>
<td>1</td>
</tr>
<tr>
<td>CHDH</td>
<td>Choline dehydrogenase</td>
<td>-0.17884</td>
<td>0.000235</td>
<td>1</td>
</tr>
<tr>
<td>COX17</td>
<td>Cytochrome c oxidase copper chaperone COX17</td>
<td>0.518366</td>
<td>0.016143</td>
<td>0.99</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>Catenin beta 1</td>
<td>-0.05411</td>
<td>0.009236</td>
<td>0.99</td>
</tr>
<tr>
<td>CYP2S1</td>
<td>Cytochrome P450 family 2 subfamily S member 1</td>
<td>-0.06485</td>
<td>0.003946</td>
<td>0.99</td>
</tr>
<tr>
<td>DDTL</td>
<td>D-dopachrome tautomerase like</td>
<td>-0.00135</td>
<td>0.016232</td>
<td>0.94</td>
</tr>
<tr>
<td>DUSP18</td>
<td>Dual specificity phosphatase 18</td>
<td>-0.15372</td>
<td>0.001322</td>
<td>1</td>
</tr>
<tr>
<td>FAM221A</td>
<td>Family with sequence similarity 221-member A</td>
<td>0.021363</td>
<td>0.040838</td>
<td>0.77</td>
</tr>
<tr>
<td>FGFR4</td>
<td>Fibroblast growth factor receptor 4</td>
<td>-0.09593</td>
<td>0.001142</td>
<td>1</td>
</tr>
<tr>
<td>KLC4</td>
<td>Kinesin light chain 4</td>
<td>-0.25503</td>
<td>0.007203</td>
<td>0.97</td>
</tr>
<tr>
<td>LARS2</td>
<td>Leucyl-tRNA synthetase 2 mitochondrial</td>
<td>-0.14444</td>
<td>0.027633</td>
<td>0.84</td>
</tr>
<tr>
<td>PFDN6</td>
<td>Prefoldin subunit 6</td>
<td>0.128019</td>
<td>0.015842</td>
<td>0.94</td>
</tr>
<tr>
<td>SLC27A3</td>
<td>Solute carrier family 27 member 3</td>
<td>-0.0091</td>
<td>0.032447</td>
<td>0.78</td>
</tr>
<tr>
<td>TNFRSF11A</td>
<td>TNF receptor superfamily member 11a</td>
<td>-0.11054</td>
<td>0.043234</td>
<td>0.76</td>
</tr>
</tbody>
</table>