A novel gene-pair signature for relapse-free survival prediction in colon cancer

Peng-fei Chen1–3, *
Fan Wang1,2,*
Zi-xiong Zhang4,*
Jia-yan Nie1,2
Lan Liu1,2
Jue-rong Feng1,2
Rui Zhou1,2
Hong-ling Wang1,2
Jing Liu1,2
Qiu Zhao1,2

1Department of Gastroenterology, Zhongnan Hospital of Wuhan University, Wuhan 430071, China; 2Hubei Clinical Center & Key Lab of Intestinal & Colorectal Diseases, Wuhan 430071, China; 3Department of Gastroenterology, The Central Hospital of Enshi Autonomous Prefecture, Enshi 445000, China; 4Department of Otolaryngology, The Central Hospital of Enshi Autonomous Prefecture, Enshi 445000, China

*These authors contributed equally to this work

Background: Colon cancer (CC) patients with early relapse usually have a poor prognosis. In this study, we aimed to identify a novel signature to improve the prediction of relapse-free survival (RFS) in CC.

Methods: Four microarray datasets were merged into a training set (n=1,045), and one RNA-sequencing dataset was used as a validation set (n=384). In the training set, microarray meta-analysis screened out 596 common RFS-related genes across datasets, which were used to construct 177,310 gene pairs. Then, the LASSO penalized generalized linear model identified 16 RFS-related gene pairs, and a risk score was calculated for each sample according to the model coefficients.

Results: The risk score demonstrated a good ability in predicting RFS (area under the curve [AUC] at 5 years: 0.724; concordance index [C-index]: 0.642, 95% CI: 0.615–0.669). High-risk patients showed a poorer prognosis than low-risk patients (HR: 3.519, 95% CI: 2.870–4.314). Subgroup analysis reached consistent results when considering multiple confounders. In the validation set, the risk score had a similar performance (AUC at 5 years: 0.697; C-index: 0.696, 95% CI: 0.627–0.766; HR: 2.926, 95% CI: 1.892–4.527). When compared with a 13-gene signature, a 15-gene signature, and TNM stage, the score showed a better performance (P<0.0001; R2=0.988, P<0.0001).

Conclusion: Our study identified a novel gene-pair signature for prediction of RFS in CC.

Keywords: colon cancer, relapse-free survival, gene pair, prognosis

Introduction

Colorectal cancer (CRC) is one of the most common cancers worldwide, with an estimated 1.4 million new cases and 693,900 deaths per year. Despite great improvements in curative resection and adjuvant chemotherapy, the relapse rate remains high. Approximately 40–50% of patients relapse within the first year after initial resection. The TNM staging system is widely used in relapse prediction. A higher TNM stage is usually associated with a higher relapse rate and shorter interval. However, this is controversial among a proportion of patients. Stage Ib patients tend to have a higher relapse rate than those with stage IIIa. Thus, in clinical practice, there is a critical need to improve the prediction of relapse in colon cancer (CC).

With the development of high-throughput gene detection technology, gene expression profiles became available to identify more reliable molecular subtypes,
and diagnostic and prognostic molecular signatures in multiple diseases. In 2013, Marisa et al classified CC into six molecular subtypes (C1–C6) by clustering gene expression data of 433 samples and found that the subtypes of C4 and C6 were associated with shorter relapse-free survival (RFS). In 2015, Guinney et al coalesced six CRC subtyping algorithms (including those from Marisa et al) into four consensus molecular subtypes (CMSs) with distinguishing features: CMS1 (microsatellite instability immune), CMS2 (canonical), CMS3 (metabolic), and CMS4 (mesenchymal). CMS4 tumors displayed worse overall survival and RFS. Alderdice et al compared the CRC transcriptional signatures of CMSs and colorectal cancer intrinsic subtypes (CRISs), and found that CRISs provided a better classification of molecular subtypes than CMSs. More reliable molecular subtyping system will be detected in large-scale prospective designed studies.

Several studies have also proposed gene-expression-based prognostic signatures in CRC. However, those signatures are limited in sample size and sufficient cross-validation. In this study, we adopted a method based on relative ranking of gene expression levels, which could integrate samples from different datasets without the need for eliminating batch effects. Thus, we reached a large sample size and developed and validated a novel gene-pair signature for RFS prediction in CC.

Materials and methods

Data collection

The Gene Expression Omnibus (GEO) and European Bioinformatics Institute (EMBL-EBI) databases were searched for related microarray datasets from inception to April 1, 2018. A dataset was included if it fulfilled the following criteria: 1) it focused on gene expression profiles of CC tissues; 2) it was based on the chip platform of Affymetrix Human Genome U133 Plus 2.0 Array (GPL570); 3) the follow-up duration and RFS status (RFS was defined as the time from surgery to the first confirmed relapse) were available; and 4) the sample size was >50. Normalized gene expression profiles from the GEO database were downloaded as the training set, while RNA-sequencing data from The Cancer Genome Atlas (TCGA) database (TCGA Colon Cancer [COAD]) was selected as the validation set.

Data quality control

Microarray quality was assessed by sample clustering according to the distance between different samples in Pearson’s correlation matrices. A height cut-off of 0.2 was chosen to identify potential microarray outliers, which were removed from the subsequent analysis.

RFS-related gene screening

To screen out the common RFS-related genes between datasets, microarray meta-analysis was conducted on the included datasets using R software and the MetaOmics package. In survival analysis, we took the log-rank test and permutation method (n=300) for an individual dataset, and then both Fisher and maxP methods were adopted to combine the P-values between datasets. A P-value <0.05 by both methods was considered statistically significant for the RFS-related probes, which were subsequently converted to gene symbols using the Affymetrix annotation files.

Development of gene-pair signature

The expression level of RFS-related genes in a specific sample underwent pairwise comparison to generate a score for each gene pair. If the first gene of one gene pair had a higher expression level than the second one, a gene pair score of 1 was assigned; otherwise, the gene pair score was 0. The profile of gene-pair scores was used to build the signature. To minimize the risk of over-fitting, we used a Cox proportional hazards regression model combined with the least absolute shrinkage and selection criteria operator (R glmnet package). The penalty parameter was estimated by 10-fold cross-validation at 1 standard error beyond the minimum partial likelihood deviance.

Signature evaluation and validation

The coefficients of significant gene pairs in the model were extracted to calculate a risk score for each sample. The concordance index (C-index) was used to assess the performance of the risk scores in RFS prediction. Then, the nearest neighbor estimation method was adopted to estimate the time-dependent receiver operating characteristics (ROC) curve of the risk scores (R survcomp package). The area under the ROC curve (AUC) was also used to assess the predictive ability of the signature risk score. The score corresponding to the shortest distance between the ROC curve and the point of 100% true positive and 0% false positive was selected as the cut-off value. The patients were divided into high- and low-risk groups according to the cut-off. Survival difference between the groups was assessed by the Kaplan–Meier estimate and compared using the log-rank test (R survival package). To validate the signature, we used the same methods in the RNA-sequencing dataset. Furthermore, univariate and multivariate Cox regression survival analyses were conducted on the risk score.
Functional enrichment analysis
To identify potential biological pathways related to the signature, gene set enrichment analysis (GSEA) was performed to detect whether a series of a priori defined biological pathways was enriched in the gene rank derived from differentially expressed genes between high- and low-risk groups. In addition, the gene-pair signature consisted of single genes, and gene ontology (GO) analysis was conducted on these genes to identify potential mechanism associated with the signature. A false discovery rate of <0.05 was chosen as the cut-off criterion.

Statistical analysis
All statistical analyses were performed using R 3.3.2 software. For use with GSEA software, the collection of annotated gene sets of c2.cp.kegg.v5.2.symbols.gmt in the Molecular Signatures Database was chosen as the reference gene sets. GO analysis was conducted with the online tool ToppGene (https://toppgene.cchmc.org/). A two-sided P-value <0.05 was considered statistically significant.

Results
Characteristics of included datasets
Three datasets (GSE17536, GSE17537, and GSE39582) were excluded owing to duplication, while one dataset (GSE33113) was excluded owing to incomplete expression data. Thirty cases of rectal cancer were excluded in GSE14333. Finally, four datasets (GSE14333 [n=196], GSE17538 [n=200], GSE38832 [n=92], and GSE40967 [n=557]) were included, with a total number of 1,045 patients. Those patients had a mean follow-up of 46.2 (0–201) months and a relapse rate of 37.7% (394/1,045). The validation set had a total of 384 patients, with a mean follow-up of 29.7 (0.2–150.1) months and a relapse rate of 20.1% (81/384). No samples were detected as outliers in the datasets (Figure 1).

Signature construction, evaluation, and validation
In microarray meta-analysis, 647 probes were significant (P<0.05) in both Fisher and maxP methods (Figure 2). These 647 probes could be mapped to 596 genes and constructed 177,310 gene pairs. The LASSO penalized generalized linear model identified 16 significant gene pairs (consisting of 24 genes) (Table 1). Then, a risk score was calculated for each sample according to the model coefficients as follows:

Risk score = Score$_{\text{CYP1B1/LIPE}}$ * 0.016 + Score$_{\text{CYP1B1/GOLGA2P10}}$ * 0.027 + Score$_{\text{SCG2/LYPD5}}$ * 0.234 + Score$_{\text{ACADSB/SDHAF1}}$ * 0.350 + Score$_{\text{ACADSB/L2C25A19}}$ * 0.042 + Score$_{\text{MATN3/RHDM1}}$ * (–0.209) + Score$_{\text{RGL2/L2C25A28}}$ * 0.223 + Score$_{\text{RGL2/TMEM30H}}$ * 0.409 + Score$_{\text{GRB10/GOLGA2P10}}$ * 0.280 + Score$_{\text{MAP1B/SLC25A22}}$ * 0.334 + Score$_{\text{ETS1/MEF2C1/HOOK3}}$ * (–0.418) + Score$_{\text{GOLGA2P10/ANGPT2}}$ * 0.088 + Score$_{\text{ICF21/ISM1}}$ * (–0.059)

The time-dependent ROC curve analysis indicated an optimal cut-off value of 1.168 for the classification of high-risk and low-risk groups. The high-risk group had an obviously higher incidence of relapse in the follow-up (Figure 3).
High-risk patients showed a poorer prognosis than low-risk patients (HR: 3.519, 95% CI: 2.870–4.314; \( P < 0.0001 \)) (Figure 4). In the validation dataset, high-risk patients also had a poorer prognosis (HR: 2.926, 95% CI: 1.892–4.527; \( P < 0.0001 \)).

In subgroup analysis, we evaluated the effects of age, gender, tumor location, TNM stage, genetic mutation (TP53, KRAS, and BRAF), molecular subtype (C1–C6), and adjuvant chemotherapy (Table 2). The results were consistent except for stage I and C1 subtype. We think that this may contribute to the excellent prognosis after curative surgery in these two subtypes.

As a continuous variable, the risk score showed a comparative ability in predicting RFS (AUC at 5 years: 0.724; C-index: 0.642, 95% CI: 0.615–0.669) (Figure 5). In the validation dataset, it had a similar performance (AUC at 5 years: 0.697; C-index: 0.696, 95% CI: 0.627–0.766). In multivariate Cox regression analysis, the risk score also showed a significant association with RFS (HR: 2.416, 95% CI: 1.824–3.198; \( P < 0.0001 \)) (Table 3).

### Comparison with other signatures

To further investigate the predictive power of the risk score, we compared it with a 13-gene signature, a 15-gene signature, and TNM stage. The risk score in the 13-gene signature and 15-gene signature was calculated according to the coefficients provided by the primary studies.\(^8,9\) In ROC curve analysis, our signature showed a better predictive ability (AUC at 5 years: 0.724) than the 13-gene signature, 15-signature, and TNM stage (AUC at 5 years: 0.575, 0.509 and 0.703, respectively) (Figure 5). Then, we calculated the C-indices according to minimum follow-up duration. We found that in the patients with a longer follow-up, the risk score showed a better performance in RFS prediction (\( R^2 = 0.988, P < 0.0001 \)) (Figure 6). In those with a follow-up of >5 years, the C-index was up to 0.869 (95% CI: 0.816–0.922), and high-risk patients had a significantly poorer prognosis (HR: 13.55, 95% CI: 7.409–24.78) (Figure 7). In comparison, our signature had a higher C-index than the 13-gene signature (\( P \) for paired \( t \)-test <0.0001), 15-gene signature (\( P = 0.0004 \)), and TNM stage (\( P \) for paired \( t \)-test =0.0125). Graphically, our signature also had better astuteness and stability.

### Table 1 Gene-pair signature information

<table>
<thead>
<tr>
<th>Gene-pair 1</th>
<th>Gene-pair 2</th>
<th>Probe 1</th>
<th>Probe 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP1B1</td>
<td>LIPE</td>
<td>202435_s_at</td>
<td>208186_s_at</td>
</tr>
<tr>
<td>CYP1B1</td>
<td>GOLGA2P10</td>
<td>202435_s_at</td>
<td>223327_x_at</td>
</tr>
<tr>
<td>SG2</td>
<td>FOXP3</td>
<td>204035_at</td>
<td>224211_at</td>
</tr>
<tr>
<td>SG2</td>
<td>MED26</td>
<td>204035_at</td>
<td>231724_at</td>
</tr>
<tr>
<td>SG2</td>
<td>LYPD3</td>
<td>204035_at</td>
<td>236039_at</td>
</tr>
<tr>
<td>ACADSB</td>
<td>SDHAF1</td>
<td>205355_at</td>
<td>221855_at</td>
</tr>
<tr>
<td>ACADSB</td>
<td>SLC25A19</td>
<td>205355_at</td>
<td>223222_at</td>
</tr>
<tr>
<td>MATN3</td>
<td>R3HDMI</td>
<td>206091_at</td>
<td>234774_at</td>
</tr>
<tr>
<td>RGL2</td>
<td>SLC25A28</td>
<td>209110_s_at</td>
<td>221432_s_at</td>
</tr>
<tr>
<td>RGL2</td>
<td>TMEM201</td>
<td>209110_s_at</td>
<td>227617_at</td>
</tr>
<tr>
<td>GRB10</td>
<td>GOLGA2P10</td>
<td>210999_s_at</td>
<td>223327_x_at</td>
</tr>
<tr>
<td>MAP1B</td>
<td>SLC25A22</td>
<td>212233_at</td>
<td>218725_at</td>
</tr>
<tr>
<td>ETS1</td>
<td>MED26</td>
<td>214447_at</td>
<td>231724_at</td>
</tr>
<tr>
<td>WDFC1</td>
<td>HOOK3</td>
<td>219478_at</td>
<td>226395_at</td>
</tr>
<tr>
<td>GOLGA2P10</td>
<td>ANGPT2</td>
<td>223327_x_at</td>
<td>236034_at</td>
</tr>
<tr>
<td>TCF21</td>
<td>ISMI</td>
<td>229529_at</td>
<td>235182_at</td>
</tr>
</tbody>
</table>

Figure 3 Distribution of relapse-free survival status and risk scores of the gene-pair signature.
Biological pathways associated with the gene-pair signature

In the validation set, 374 patients were divided into high- and low-risk groups according to the risk score. Three KEGG pathway-related gene sets were enriched in the low-risk group, namely “peroxisome”, “pyruvate metabolism”, and “citrate cycle (TCA cycle)” (where TCA is tricarboxylic acid) (Figure 8).

In GO analysis of the 24 genes comprising the signature, these genes were associated with multiple biological processes, such as angiogenesis, endothelial cell migration, and T-cell toleration.
Discussion

Tumor relapse within the first year after initial resection occurred in almost 50% patients, which was closely related to survival outcomes. Currently, the TNM staging system is widely used to predict RFS. However, CRC patients with the same stage and treatment varied considerably in the RFS prediction, which may be caused by the genetic differences between individuals.16 Thus, genomic biomarkers for relapse
A novel gene-pair signature for relapse-free survival prediction in CC
detection could make up for the deficiency of TNM classification, and thereby identify high-risk patients to improve personalized cancer care at an earlier stage.17,18

With the development of gene microarray and sequencing technologies, we are faced with a large amount of biological data. Genome and transcriptome information is available in more and more samples, prompting a search for new biomarkers for the discrimination of high-risk individuals in disease diagnosis or prognosis. A genomic signature is a biomarker in which the genomic data are combined in a defined manner to provide either a continuous score or a categorical classifier for clinical decision-making.

Several CRC prognostic signatures have been reported using the public microarray data from the GEO database. In

Figure 6 C-indices according to minimum follow-up duration.
Note: Dashed line: 95% CI.
Abbreviation: C-index, concordance index.

Figure 7 Kaplan–Meier curves of relapse-free survival in patients with a follow-up of >5 years.
Abbreviations: C-index, concordance index; RFS, relapse-free survival.
In this study, we constructed an RFS-related signature consisting of 16 gene pairs using a penalized Cox proportional hazards regression model. With specific coefficients, a risk score was calculated for each patient based on the gene expression profile, and this could be used without considering batch effects. Thus, we achieved a large sample size which also helped us to identify a robust signature, which showed a better performance than other single-gene signatures, especially among the patients with a longer follow-up. A study by Tauriello et al used a k-top scoring pair (k-TSP) approach to classify CRCs between CMS4 and CMS1–3. It also adopted the method of gene pairs and had a good performance.22 In our study, CMS information was unavailable, but C1–C6 subtyping information was available. High-risk individuals in our signature had a poor prognosis in C2–C6, indicating a signature stability of different molecular subtypes across patients.

To investigate the potential mechanism associated with the signature, GSEA and GO analyses were conducted. Three gene sets were significantly enriched in low-risk patients. The role of the peroxisome pathway in cancer is controversial. Its inactivation leads to metabolic stress, mTOR inhibition, and lethality in liver cancer cells.23 However, it protects lymphoma cells from histone deacetylase inhibitor-mediated apoptosis.24 Fransen et al present potential therapeutic opportunities for targeting the TCA cycle in cancer cells.25 Moreover, in GO analysis, 24 genes in the signature were associated with multiple biological processes, including angiogenesis, endothelial cell migration, and T-cell tolerance, which have been reported in tumor carcinogenesis, progression, and metastasis.26–28

Several limitations to our study should be acknowledged. First, this was a retrospective study, rather than a prospective cohort study. Second, in the training set, we considered the microarray data based only on GLP570 and ignored other unusual platforms, which may have led to selection bias. Third, we ignored batch effects caused by other confounders, such as genetic background and environmental factors, which may have affected the model coefficients and cut-off value.

**Conclusion**
Our study identified a novel gene-pair signature for RFS prediction in CC.

**Acknowledgment**
This study was supported by the Natural Science Foundation of Hubei Province (QZ, No. 2016CFA101).

**Disclosure**
The authors report no conflicts of interest in this work.

**References**


