A Monotonic and Prognostic Genomic Signature from Fibroblasts for Colorectal Cancer Initiation, Progression, and Metastasis

Mireia Berdiel-Acer1,6, Daniel Cuadras4, Natalia Guillén Díaz-Maroto1, Xavier Sanjuan5, Teresa Serrano5, Antoni Berenguer2, Victor Moreno2, Samuel Gonçalves-Ribeiro1, Ramon Salazar3, Alberto Villanueva1, and David G. Mollevi1

Abstract
The differential gene expression patterns between normal colonic fibroblasts (NCF), carcinoma-associated fibroblasts from primary tumors (CAF-PT), and CAFs from hepatic metastasis (CAF-LM) are hypothesized to be useful for predicting relapse in primary tumors. A transcriptomic profile of NCF (n = 9), CAF-PT (n = 14), and CAF-LM (n = 11) was derived. Prediction Analysis of Microarrays (PAM) was used to obtain molecular details for each fibroblast class, and differentially expressed transcripts were used to classify patients according to recurrence status. A number of transcripts (n = 277) were common to all three types of fibroblasts and whose expression level was sequentially deregulated according to the transition: NCF → CAF-PT → CAF-LM. Importantly, the gene signature was able to accurately classify patients with primary tumors according to their prognosis. This capacity was exploited to obtain a refined 19-gene classifier that predicted recurrence with high accuracy in two independent datasets of patients with colorectal cancer and correlates with fibroblast migratory potential. The prognostic power of this genomic signature is strong evidence of the link between the tumor-stroma microenvironment and cancer progression. Furthermore, the 19-gene classifier was able to identify low-risk patients very accurately, which is of particular importance for stage II patients, who would benefit from the omission of chemotherapy, especially T4N0 patients, who are clinically classified as being at high risk.


Introduction
The huge outcome heterogeneity among patients with cancer, which hinders their clinical management, cannot be explained solely by the genetic background of the cancer cells. Some of this variation is due to intratumoral heterogeneity, whereby tumors contain multiple cell populations, each with distinct characteristics of karyotype, receptors, antigenicity, etc. The relative abundance of each cell type varies considerably between tumors and even between parts of the same tumor (1). However, the predominant cell type in desmoplastic tumors is the carcinoma-associated fibroblast (CAF; ref. 2).

To date, several genetic classifiers of colorectal cancer have been identified (3–6) but none of these is of relevance in relation to the stroma. In this article, we provide information about the importance of the transcriptomic status of fibroblasts to colorectal cancer prognosis.

Fibroblasts are ubiquitous mesenchymal cells with different vital functions during development, wound healing, and disease. Depending on their anatomic location, fibroblasts should be considered as different cell types as they have particular gene expression programs and different features according to their tissue specialization (7). However, a stereotyped gene expression program in response to serum was identified in fibroblasts from 10 different anatomic sites, revealing that in response to a wound many of the normal differentiated particularities of the fibroblasts in the wounded tissue are temporarily set aside in favor of a common response (8). In addition, this common serum response gene expression program was identified in many human tumors and was a powerful predictor of metastasis in different carcinomas. The transcriptomic program of serum response genes was used as a prototype of the fibroblasts’ activation...
pattern, and the correlation of this expression in fibroblasts with a centroid of the differential expression in tumor samples displayed a powerful prognostic information, generating a continuous score that can be scaled (9).

It is already known that as tumor evolves, malignant cells educate surrounding stroma to create an adequate microenvironment for their convenience (10). This education can be understood through the degree of activation of fibroblasts as a surrogate marker. In other words, as the fibroblasts interact with other cells in the tumor, fibroblasts progressively evolve, and this evolution could provide new biomarkers and valuable information for patient’s outcome.

Thus, according to the concept that there is a common serum response program in different fibroblasts, despite being considered different cells, we hypothesize that since the earliest start of tumorigenic process, when an oncogenic stress transforms an epithelial cell until the cell metastasizes, there is a common and monotonic response of the surrounding fibroblasts that can be measured and associated to prognosis. We call this response as the fibroblasts progression genes (FPG). We used this common response signature (277 genes) to develop a 25-gene classifier that characterized each fibroblast type with respect to location and that was of prognostic value for classifying patients with primary tumors. We also defined and validated a 19-gene classifier, being the best from the 277 genes predicting recurrence in patients with colorectal cancer. This simple signature score has remarkable predictive capabilities that confirm the biologic relevance of CAFs and their gene programs to cancer progression.

Materials and Methods

Fibroblast isolation

We obtained fibroblast cultures from fresh surgical specimens resected from patients with primary colorectal carcinoma: normal colonic fibroblasts (NCF, 9) from the normal colonic mucosa (NCM) at least 5 to 10 cm from the surgical margin, carcinoma-associated fibroblasts from the primary tumor (CAF-PT, 14), and carcinoma-associated fibroblasts (CAF-LM, 11) from fresh surgical specimens of liver metastases. All samples were collected under the supervision of the Ethics Committee of the Hospital Universitari de Bellvitge. These 34 samples comprise the MB dataset. Clinical baseline data of patients whose fibroblasts were isolated are in Supplementary Table S1. To establish primary cultures, we used a large panel of markers (vinculin, α-smooth muscle actin, vimentin, E-cadherin, N-cadherin, and VE-cadherin) to exclude epithelial and endothelial cell contamination (Supplementary Fig. S1).

Microarray procedures and signature definition

RNA from fibroblast cultures in passage #3 was used to hybridize with the Affymetrix GeneChip Human Gene 1.0 ST Array. The microarray data were read, normalized, and transformed to numerical expression values using the justRMA function in the simpleaffy package (11). Differentially expressed genes were identified using the multi-class Significance Analysis of Microarrays (SAM) technique, available in the same package (12).

We selected those probes that simultaneously displayed a robust multi-array average (RMA) normalized expression value of > ±4 in the three fibroblast populations (MB dataset). This yielded 10,779 probes common to NCF, CAF-PT, and CAF-LM. This rule was intended to exclude genes that had not normalized expression in at least one category, as well as background noise (Supplementary Fig. S2). The signature was defined as follows: fibroblast progression genes (FPG; NCF > CAF-PT > CAF-LM and NCF < CAF-PT < CAF-LM) contained genes with a log n-fold change (logFC) NCF versus CAF-PT of > 1.5 and CAF-PT versus CAF-LM of > 1.5 (179 genes), and CAF-PT versus NCF of > 1.5 and CAF-LM versus CAF-PT of > 1.5 (98 genes); q < 0.05. Therefore, the FPG signature comprises genes whose expression level increased or decreased during the NCM–primary tumor–liver metastasis sequence.

Definition of FPG class prediction

The Prediction Analysis of Microarrays (PAM; ref. 13) nearest shrunken centroid classifier was applied to the MB dataset to refine the 277 differentially expressed genes (FPGs) to derive a classifier that better characterized each class (normal colonic fibroblasts, NCFs; CAFs from the primary tumor, CAF-PTs; and CAFs from the liver metastasis, CAF-LMs). PAM estimates the predicted error rate based on misclassification error, which is calculated by averaging the errors from each of the 10-fold cross-validations (CV).

The classifier was then validated in external datasets. GSE31279 (14) comprised 24 samples obtained by laser capture microscopy (LCM); 10 normal colonic stroma, 10 carcinoma-associated stroma, and 4 metastasis-stroma; GSE22598 (15) comprised 17 whole-tissue samples of NCM and 17 paired colorectal carcinomas; the LP dataset comprised 19 colorectal carcinomas and 21 paired liver metastases from our own institution. Public datasets from GEO are detailed in Supplementary Table S2.

Finally, we computed a signature score for the genes of the classifier to assess its power in predicting patient survival in two independent datasets, GSE14333 (ref. 6; stage I, 31, stage II, and 71, stage III patients) and GSE33133 (16). The signature score was obtained by summarizing centered gene expression means and computing a z-score for each patient. The expression score for each patient was classified as high or
low according to the cutoff provided by Euclidian distance in the ROC curve analysis.

**Recurrence-prediction classifier development**

We developed a recurrence signature from the 277 genes comprising the FPG signature. PAM analysis identified relevant genes for the classifier from GSE14333 (RMA-normalized) defining the recurrent and nonrecurrent classes. Internal validation was done by the leave-one-out CV method. Validation was carried out with GSE17538 (ref; 5; RMA-normalized).

We also computed the z-score (signature score normalized by SD) and classified patients as high or low expression according to the cutoff obtained by Euclidian distance in the ROC curve analysis for recurrence prediction (cutoff, 0.187). Patients with a score >0.187 were classified as high risk and score <0.187 as low-risk. The same cutoff value (0.187) was applied to the validation datasets. Results were validated in the independent datasets GSE17538 (144 patients; 33, stage I; 50, stage II; and 61, stage III) and GSE33113 (ref; 16; 90 stage II patients; RMA-normalized) taking disease-free survival (DFS) as a primary endpoint. GSE14333 and GSE17538 cohorts partially overlapped. Cases duplicated in GSE14333 and GSE17538 were excluded from the validation.

We chose these datasets for training and validation because data came from whole-tumor samples without any enrichment in tumor cells, ensuring a minimum stromal percentage between 30% and 50%.

**Wound-healing assay**

Fibroblasts were grown on P60 dishes until confluence. We made a small scratch with a yellow pipette tip. After several PBS washes to remove floating cells, we began the assay by adding fresh medium. We obtained images of a predefined area of the scratch under a Leica microscope every 2 hours. The experiment was done twice, with three replicates on each occasion.

**Ontology analysis**

The Database for Annotation, Visualization and Integrating Discovery (DAVID) v6.7 was used to determine likely biologic processes in which differentially expressed genes (FPGs) were involved. Gene ontology (GO) attributes of biologic processes were listed, ordered by their FDR-adjusted q value of <0.05.

**Determination of the 19-gene signature cell-type specificity**

We used data from GSE39396 (17) to check the stromal specificity of the genes in our CAF signature. GSE39396 consisted of four cell type populations isolated by FACS from six different colorectal tumors: EPCAM⁺ epithelial cells, CD45⁻ inflammatory cells, CD31⁻ endothelial cells, and FAPt⁺ CAFs.

**Internal microarray validation**

We internally validated the genes comprising the 25-gene signature by means of RT-PCR, confirming the results by Western blotting. The antibodies used for this purpose were rabbit anti-human SLC7A2 (HPA009169; Sigma-Aldrich), rabbit anti-human TGFB2 (V; sc-90; Santa Cruz Biotechnology, Inc.), and rabbit anti-human ARHGDB (Ab15198; Abcam).

**Statistical analysis**

Kaplan–Meier and Cox regression analysis estimates for DFS and disease-specific survival (DSS) were calculated and compared by the likelihood ratio tests. Recurrence and death from disease were considered as events.

Positive and negative posterior probabilities were chosen to measure the power of the classifier when considering clinical decision making.

**Results**

**Differential gene expression of fibroblasts along cancer progression**

According to the overall design (Supplementary Fig. S2), we isolated fibroblasts (n = 34) from the three microenvironments in which colorectal epithelial cells reside (NCM, primary tumor, and liver metastasis). Because fibroblasts from different locations seem to have particular transcriptomic programs but share a stereotyped gene expression pattern in response to serum (8) and this response can be scaled to predict prognosis (9), we aimed to establish whether there is a set of fibroblastic genes whose expression arises gradually according to their anatomic demarcation involved in the steps of colorectal cancer progression (NCF, CAF-PT, and CAF-LM). After filtering the data, 10,779 probes met the characteristics imposed (cartoon, Supplementary Fig. S3A), most of which corresponded to development-related ontologies (Table 1). Interestingly, overexpressed genes in the sequence NCF-to-CAF-LM (low in NCF and high in CAF-LM) were also associated with wound-healing response, whereas in the opposite direction the biologic processes seem to be enriched in inflammatory response. From these probes, we selected those that fulfilled the criterion of a >1.5-fold change (normalized log2 value) in expression at each step at the same time (as illustrated in the cartoon, Supplementary Fig. S3B). We refer to this as monotonic gene expression during cancer progression, meaning that gene expression follows a decreasing or increasing monotonic function. By this means, we identified 98 and 179 genes that were upregulated and downregulated, respectively, from NCF to CAF-LM (Supplementary Table S3).

Together, these 277 genes were designated as the FPG signature (fibroblastic genes relevant to cancer progression; Supplementary Fig. S3C).

**Refining the FPG signature: developing a classifier**

We aimed to refine the FPG signature to determine the minimum number of genes whose expression levels characterize each class of fibroblast involved in colorectal cancer progression. Our 34 samples of fibroblasts (MB Dataset) were subjected to PAM using the nearest shrunk centroid method to define the classifier. Ten-fold cross-validation was...
used, randomly dividing the samples into 10 approximately equally sized parts. All parts were balanced to ensure that the three classes (NCF, CAF-PT, and CAF-LM) were distributed proportionally among each of the 10 parts. A minimum of 25 estimated genes predicted the lowest misclassification error (Supplementary Fig. S3D and S3E). Genes are listed in Table 2 and Supplementary Fig. S4A. Nine of these genes have been validated by means of RT-PCR (Supplementary Fig. S4B). Eleven of the 25 genes were downregulated, most of them [EVA4 (18), FBLN1 (19), (20), FBN2 (21–23), ADH1B (24), CYP39A1 (25), CHRD1L1 (26), and TCF21 (27, 28)] by methylation in tumorigenic processes. The other 14, most of which are already known to be involved in cancer (TGFBR2, NTF3 (29), UACA (30), SFRP4 (31), and ULBP2 (32), were overexpressed. Another gene like ARHGDIB has been associated with prognosis (33, 34) and cisplatin resistance (35). Selected genes were validated at the protein level (Supplementary Fig. S3F), and were found to have a greater level of expression in fibroblasts as the cancer progressed.

The overall error rate after 10-fold cross-validation was 0.174 in the training samples (Table 2). The CAF-PT class was the biggest contributor to the misclassification, some samples being mistaken for CAF-LM and for NCF (Fig. 1A). We validated the 25-gene signature in three independent datasets, two whole-tumor samples (Dataset LP and GSE22598) and one microdissected sample (GSE31279). The probability of correctly classifying a sample was very high (Fig. 1B, Table 2), enabling us to conclude that these 25 genes were those whose expression levels best characterize each fibroblast class. Therefore, depending on the average gene expression value we could discriminate whether a given sample was an NCM, a primary tumor, or a liver metastasis.

### Gene expression associated with risk of progression

The CAF-PT class exhibited the greatest variation of those considered. We attempted to check that misclassified samples could be associated with prognosis, on the assumption that CAF-PTs with a higher signature score (reaching a value typical of CAF-LM) were at greater risk (i.e., had a poorer

<p>| Table 1. Gene ontology Overexpressed along the sequence NCF, CAF-PT, CAF-LM |
|--------------------------|------------------|----------|----------|----------|</p>
<table>
<thead>
<tr>
<th>Category Term</th>
<th>Count</th>
<th>P</th>
<th>Fold enrichment</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009611~response to wounding</td>
<td>31</td>
<td>4.86E–07</td>
<td>2.841</td>
<td>8.42E–04</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0042060~wound healing</td>
<td>17</td>
<td>2.04E–06</td>
<td>4.319</td>
<td>3.54E–03</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009605~response to external stimulus</td>
<td>41</td>
<td>4.62E–06</td>
<td>2.173</td>
<td>8.00E–03</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0008544~epidermis development</td>
<td>16</td>
<td>5.95E–06</td>
<td>4.222</td>
<td>1.03E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0007398~ectoderm development</td>
<td>16</td>
<td>1.36E–05</td>
<td>3.94</td>
<td>2.36E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0048856~anatomical structure development</td>
<td>82</td>
<td>1.59E–05</td>
<td>1.568</td>
<td>2.75E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0048513~organ development</td>
<td>62</td>
<td>1.80E–05</td>
<td>1.723</td>
<td>3.11E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0048731~system development</td>
<td>77</td>
<td>1.81E–05</td>
<td>1.595</td>
<td>3.13E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006928~cell motion</td>
<td>26</td>
<td>1.81E–05</td>
<td>2.64</td>
<td>3.14E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0050879~regulation of body fluid levels</td>
<td>13</td>
<td>3.24E–05</td>
<td>4.491</td>
<td>5.61E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0001568~blood vessel development</td>
<td>17</td>
<td>5.09E–05</td>
<td>3.346</td>
<td>8.81E–02</td>
</tr>
</tbody>
</table>

<p>| Infraexpressed along the sequence NCF, CAF-PT, CAF-LM |
|--------------------------|------------------|----------|----------|----------|</p>
<table>
<thead>
<tr>
<th>Category Term</th>
<th>Count</th>
<th>P</th>
<th>Fold enrichment</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0050727~regulation of inflammatory response</td>
<td>12</td>
<td>9.41E–07</td>
<td>7.021</td>
<td>1.64E–03</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0051270~regulation of cell motion</td>
<td>18</td>
<td>1.62E–06</td>
<td>4.147</td>
<td>2.83E–03</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0042127~regulation of cell proliferation</td>
<td>40</td>
<td>2.51E–06</td>
<td>2.263</td>
<td>4.38E–03</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009611~response to wounding</td>
<td>30</td>
<td>7.01E–06</td>
<td>2.546</td>
<td>1.22E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0002526~acute inflammatory response</td>
<td>12</td>
<td>9.74E–06</td>
<td>5.559</td>
<td>1.70E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0045595~regulation of cell differentiation</td>
<td>28</td>
<td>1.60E–05</td>
<td>2.541</td>
<td>2.80E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0045597~positive regulation of cell differentiation</td>
<td>18</td>
<td>1.62E–05</td>
<td>3.495</td>
<td>2.82E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0030334~regulation of cell migration</td>
<td>15</td>
<td>2.72E–05</td>
<td>3.947</td>
<td>4.75E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0040012~regulation of locomotion</td>
<td>16</td>
<td>2.82E–05</td>
<td>3.706</td>
<td>4.92E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009605~response to external stimulus</td>
<td>41</td>
<td>2.93E–05</td>
<td>2.012</td>
<td>5.10E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006950~response to stress</td>
<td>63</td>
<td>3.49E–06</td>
<td>1.682</td>
<td>6.10E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006954~inflammatory response</td>
<td>21</td>
<td>3.98E–05</td>
<td>2.900</td>
<td>6.94E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0050793~regulation of developmental process</td>
<td>33</td>
<td>4.75E–05</td>
<td>2.184</td>
<td>8.29E–02</td>
</tr>
<tr>
<td>GOTERM_BP_ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0032101~regulation of response to external stimulus</td>
<td>14</td>
<td>5.65E–05</td>
<td>3.940</td>
<td>9.85E–02</td>
</tr>
</tbody>
</table>
### Table 2. Twenty-five–gene signature and prediction in training dataset MB and validation datasets LP, GSE22598, and GSE31279

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Group</th>
<th>NCF</th>
<th>CAF-PT</th>
<th>CAF-LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARHGDIB</td>
<td>UP</td>
<td>0.0179</td>
<td>0</td>
<td>0.9268</td>
</tr>
<tr>
<td>C5orf46</td>
<td>UP</td>
<td>0</td>
<td>0</td>
<td>0.8287</td>
</tr>
<tr>
<td>TMEM176B</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.5635</td>
</tr>
<tr>
<td>TSPAN2</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.5229</td>
</tr>
<tr>
<td>EYA4</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.4271</td>
</tr>
<tr>
<td>CYP39A1</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.327</td>
</tr>
<tr>
<td>ADH1B</td>
<td>DW</td>
<td>0.3203</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SRGN</td>
<td>UP</td>
<td>0</td>
<td>0</td>
<td>0.3089</td>
</tr>
<tr>
<td>CHRD1</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.304</td>
</tr>
<tr>
<td>PTX3</td>
<td>UP</td>
<td>0</td>
<td>0</td>
<td>0.1956</td>
</tr>
<tr>
<td>SLC7A2</td>
<td>UP</td>
<td>0</td>
<td>0</td>
<td>0.1849</td>
</tr>
<tr>
<td>TGFβ2</td>
<td>UP</td>
<td>-0.172</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SFRP4</td>
<td>UP</td>
<td>0</td>
<td>0</td>
<td>0.1665</td>
</tr>
<tr>
<td>FBXN1</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.1621</td>
</tr>
<tr>
<td>ULBP2</td>
<td>UP</td>
<td>-0.1598</td>
<td>0</td>
<td>0.0049</td>
</tr>
<tr>
<td>FBN2</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.1571</td>
</tr>
<tr>
<td>CCDC81</td>
<td>UP</td>
<td>0</td>
<td>0</td>
<td>0.1485</td>
</tr>
<tr>
<td>GPR1</td>
<td>UP</td>
<td>0</td>
<td>0</td>
<td>0.1256</td>
</tr>
<tr>
<td>TNFSF4</td>
<td>UP</td>
<td>-0.0732</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHL1</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.0609</td>
</tr>
<tr>
<td>UACA</td>
<td>UP</td>
<td>0</td>
<td>0</td>
<td>0.056</td>
</tr>
<tr>
<td>TCF21</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.0406</td>
</tr>
<tr>
<td>NTF3</td>
<td>UP</td>
<td>-0.0191</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HECW2</td>
<td>UP</td>
<td>0</td>
<td>0</td>
<td>0.0072</td>
</tr>
<tr>
<td>SLC38A4</td>
<td>DW</td>
<td>0</td>
<td>0</td>
<td>-0.0004</td>
</tr>
</tbody>
</table>

#### Dataset MB

<table>
<thead>
<tr>
<th>Real status</th>
<th>NCF</th>
<th>CAF-PT</th>
<th>CAF-LM</th>
<th>Class error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCF</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CAF-PT</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>0.428</td>
</tr>
<tr>
<td>CAF-LM</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
</tr>
</tbody>
</table>

Overall error rate = 0.174 (cross-validation)

#### Validation

<table>
<thead>
<tr>
<th>Real status</th>
<th>Primary tumor</th>
<th>Liver metastasis</th>
<th>Class error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary tumor</td>
<td>16</td>
<td>3</td>
<td>0.158</td>
</tr>
<tr>
<td>Liver metastasis</td>
<td>0</td>
<td>21</td>
<td>0</td>
</tr>
</tbody>
</table>

Overall error rate = 0.075 (cross-validation)

#### GSE22598

<table>
<thead>
<tr>
<th>Real status</th>
<th>Normal mucosa</th>
<th>Primary tumor</th>
<th>Class error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal mucosa</td>
<td>15</td>
<td>2</td>
<td>0.118</td>
</tr>
<tr>
<td>Primary tumor</td>
<td>0</td>
<td>17</td>
<td>0</td>
</tr>
</tbody>
</table>

Overall error rate = 0.059 (cross-validation)

(Continued on the following page)
Carcinoma-Associated Fibroblasts and Prognosis

Table 2. Twenty-five–gene signature and prediction in training dataset MB and validation datasets LP, GSE22598, and GSE31279 (Cont’d)

<table>
<thead>
<tr>
<th>GSE31279</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real status</td>
<td>Normal stroma</td>
</tr>
<tr>
<td>Normal stroma</td>
<td>9</td>
</tr>
<tr>
<td>Primary tumor stroma</td>
<td>3</td>
</tr>
<tr>
<td>Liver metastasis</td>
<td>1</td>
</tr>
</tbody>
</table>

Overall error rate = 0.203 (cross-validation)

NOTE: Shrunken centroids for each class of fibroblast (n = 34). The column “group” describes the direction of the expression in the sequence NCF-to-CAF-LM.

The method computes a standardized centroid for each class. Centroids are the average gene expression for each gene in each class divided by the within-class SD for that gene. Nearest centroid classification takes the gene expression profile of a new sample, and compares it with each of these class centroids. The class whose centroid that is closest to, in squared distance, is the predicted class for that new sample. Nearest shrunken centroid classification shrinks each of the class centroids toward the overall centroid for all classes by a threshold. This shrinkage consists of moving the centroid toward zero by threshold, setting it equal to zero if it hits zero. Methylation status of these 25 genes has been evaluated in 258 samples of primary tumors from the TCGA (The Cancer Genome Atlas). Results are depicted in Supplementary Table S5.

Abbreviations: UP, CAF-LM > CAF-PT > NCF; DW, NCF > CAF-PT > CAF-LM.

prognosis) than CAF-PTs with a lower signature score (Fig. 1C). We evaluated the prognostic power of the 25-gene classifier in GSE14333, which consisted of 166 colorectal primary tumors (stages I to III) and GSE33113 (90 patients, stage II). Assigning patients to high or low expression score (according to the 50th percentile of the 25-gene scores obtained from isolated fibroblasts; cutoff, −0.43), ROC curve (Fig. 1D), and DFS Kaplan–Meier analysis (Fig. 1E) indicated that these genes are associated with prognosis (HR, 3.5; P = 0.00029). An additional validation in GSE33113 (90 patients stage II) confirmed the tendency associated to prognosis (Fig. 1F and G). Therefore, this 25-gene score not only classified fibroblast sample correctly but also classified patients with primary tumors according to their prognosis. The FPG and the 25-gene signature were obtained regardless of the gene expression data of the tumor specimens and patients’ clinical data, so the prognostic power of this signature is strong evidence of the link between the tumor stroma and cancer progression.

Best genes associated with recurrence in patients with colorectal cancer

The 25-gene classifier was defined as the best genes whose monotonic expression characterized each fibroblast type. However, it was not specifically designed for predicting outcome. Nevertheless, given the prognostic power of the 25-gene classifier obtained from the FPG stereotyped gene expression pattern of fibroblasts, we aimed to define the best signature (from the 277 deregulated genes) for classifying patients with colorectal cancer, using recurrence as a primary endpoint. A PAM-optimized process yielded a 19-gene signature from GSE14333 (Supplementary Table S4). The score for each gene derived by PAM was associated with the status of the patient (recurrent or nonrecurrent) and the level of expression of the gene during cancer progression (UP or DW in Supplementary Table S4). Validation of PAM in the independent dataset GSE17538 showed a specificity of 79.7% and sensitivity of 54.9% after cross-validation. Of these 19 genes, 15 were overexpressed in the sequence NCF<CAF-PT<CAF-LM, and were thereby overexpressed during cancer progression; four genes were underexpressed.

To analyze the cell type–specific expression of the 19-gene signature further, we used data from GSE39396 (17), consisting of FACS-purified cell types from 6 patients with colorectal cancer. Relative levels of cell type–specific marker genes confirmed that most of the 19 genes are representative of the microenvironment compartment, especially from FAPα+ CAFs, with the exception of CFTR, which was expressed by EPCAM+ cells (Fig. 2A). In addition, because recruitment of fibroblasts is one of the characteristics of desmoplastic tumors, we performed a wound-healing assay to compare the migration of fibroblasts with high and low 19-gene signature scores. Fibroblasts with high signature scores had better migratory abilities, as illustrated in Fig. 2B.

To rank patients by their level of expression of the 19 genes, and assuming equal biologic relevance for each of the 19 genes of the signature, we derived a score by calculating the average of the z-scores, applying a weight of +1 to overexpressed genes and of −1 to underexpressed genes. For patient’s categorization, we used the Euclidian distance between sensitivity and specificity (ROC curve analysis; AUC, 0.79; cutoff, 0.187; Fig. 3A and B). The recurrence classifier identified 97 low-expression (58.4%) and 69 high-expression patients (41.6%; OR, 3.47; Fig. 3C). Low-risk and high-risk patients had significantly different survival outcomes, measured as DFS [HR, 5.49; 95% confidence intervals (CI), 2.86–10.53; P < 0.0001; Fig. 3D]. An
Figure 1. A, estimated probabilities for the fibroblast training set. Samples are partitioned by the true (top) and predicted (bottom) classes. NCF and CAF-LM are always well classified and some samples of CAF-PT have higher probabilities of being misclassified. B, validation of PAM analysis in the GSE22598 dataset consisting of 17 pairs of homogenized samples of whole-tumor or whole normal mucosa. (Continued on the following page.)
independent cohort of 141 patients (GSE17538) was used to evaluate the performance of the signature score of the 19-gene classifier. In the validation set, 55 patients were defined as high risk (39%) and 86 as low risk (61%; Fig. 3E; AUC, 0.81). The prevalence of this validation dataset was 24.1% and the positive and negative posterior probabilities were 50% (95% CI, 36%–64%) and 8% (95% CI, 3.5%–16.4%), respectively. Stratifying patients by stage, the prevalence for stage II patients was 18%, with positive and negative posterior probabilities of 37% (95% CI, 17%–61%) and 6% (95% CI, 1%–22%), respectively. Interestingly, the clinical decision margin was wider for stage III, with a prevalence of 39% and positive and negative posterior probabilities of 61% (95% CI, 42%–77%) and 16% (95% CI, 6%–34%), respectively, thus with a difference of 45 points. Univariate Cox regression analysis revealed that high-risk patients had an HR of 7.43 ($P < 0.0001$) for DFS (Fig. 3G; statistically significant also stratifying by stage).

Carcinoma-Associated Fibroblasts and Prognosis

Figure 2. A, unsupervised hierarchical clustering of the 19-gene signature with data from GSE39396, consisting of four cell type populations isolated by FACS from six different colorectal tumors: EPCAM$^+$ epithelial cells (green), CD45$^+$ inflammatory cells (gray), CD31$^+$ endothelial cells (yellow), and FAP$^+$ CAFs (orange). The heatmap shows that CAFs (orange) cluster in a separate group with respect to the expression values of the 19-gene signature. With the exception of CFTR, which is mainly expressed by epithelial cells, the other 18 genes are basically of stromal origin. B, representative wound-healing assay images of replicates of 19-gene signature low-score fibroblasts and 19-gene signature high-score fibroblasts. The micrographs show that high signature score fibroblasts (right) have a higher migratory capacity than those with a low signature score (left).
Figure 3. A, ROC curve for the 19-gene signature performance in the GSE14333 dataset (AUC, 0.79). The AUC corresponds to the probability of the 19-gene signature ranking a randomly selected positive example higher than a randomly selected negative example, where 1 is the perfect test value and the dotted diagonal line shows the performance of a random predictor. B, histogram for 19-gene classifier score values for the patients of the training set. The red line corresponds to the cutoff (0.187) obtained from the Euclidian distance between sensitivity and specificity for recurrence prediction. The same cutoff value (0.187) is applied to the validation datasets. C, waterfall plot for recurrence risk of 166 patients from GSE14333 according to the 19-gene signature score (OR, 3.47). D, DFS Kaplan–Meier curves for the performance of the 19-gene signature in the GSE14333 dataset for signatures classified as high (red) or low (black) expression. HR, 5.49; likelihood ratio test, \( P < 0.0001 \). Five-year DFS rates were 36% (high-expression patients) and 83% (low-expression patients). E, ROC curve for the 19-gene signature performance in the GSE17538 validation dataset (AUC, 0.81). F, waterfall plot for recurrence risk of 141 patients from validation dataset GSE17538 according to the 19-gene signature score (OR, 3.29; sensitivity, 78.8%; and specificity, 73.6%). G, DFS Kaplan–Meier curves for the performance of the 19-gene signature in the GSE17538 dataset for scores classified as high (red) or low (black) expression. HR, 7.43; likelihood ratio test, \( P < 0.0001 \). Five-year DFS rates were 32% (high-expression patients) and 83% (low-expression patients). H and I, stratifying for stages, patients with high expression scores have a risk of shorter DFS 5.71-times higher than low-expression patients for stage II (\( P = 0.015 \)) and 4.71 times for stage III. J, DSS Kaplan–Meier curves for the performance of the 19-gene signature in the GSE17538 validation dataset for scores classified as high (red) or low (black) expression. HR, 7.44; \( P < 0.0001 \). Five-year DSS rates were 41% (high-expression patients) and 87% (low-expression patients).
stages; Fig. 3H and I) and of 7.44 (P < 0.0001) for DSS (Fig. 3J). In multivariate analysis, after adjusting for significant variables in univariate analysis (stage and grade) the classifier score remained statistically significant for DFS (HR, 5.52; 95% CI, 2.23–13.65; P < 0.0001; Table 3). The 19-gene signature was also independently associated with DSS (P = 0.004; Table 3). In addition, the signature score as a continuous variable was also significant, increasing the risk of relapse (HR) 1.93 times per unit increase of the score (P < 0.0001; 95% CI, 1.47–2.53).

To confirm the prognostic value in stage II patients, we used a third independent cohort of 90 patients (GSE33113). Using the cutoff obtained from the training dataset (Fig. 4A), the classifier also performed well in this third validation cohort (AUC, 0.71; Fig. 4B and C). A high level of expression of the 19-gene signature was clearly associated with poor prognosis in this cohort (HR, 3.84; P = 0.0034; Fig. 4D). Interestingly, the prevalence in this cohort was 21% and the positive and negative posterior probabilities were 37% (95% CI, 22%–55%) and 10% (95% CI, 4%–23%). Moreover, the signature score as a continuous variable was also significant, increasing the risk of relapse (HR) 2.06 times per unit increase of the score (P = 0.003; 95% CI, 1.29–3.29).

Thus, for both validation datasets, the classifier clearly identified low-risk patients, particularly in the case of stage II patients, who would benefit from the omission of chemotherapy, especially those who are T4N0 and thus clinically classified as being at higher risk.

Unfortunately, we only know the age, gender, recurrence status, grade (only for GSE17538), stage, and survival time of the patients in these cohorts, as the authors who uploaded the data to the GEO did not provide information about other characteristics.

Discussion

Molecular biomarkers for predicting relapse of colorectal cancer are needed to improve the selection of patients who would benefit from an adjuvant treatment. Recently, the great importance of the stroma to the prognosis of various types of cancer, especially those with a high degree of desmoplasia, has become recognized. Our hypothesis is that differential gene expression between NCF, CAF-PT, and CAF-LM would be useful for predicting relapse in primary tumors. We identified a short list of genes of stromal fibroblasts from anatomic sites involved in colorectal carcinogenesis whose transcriptional level was correlated with cancer progression. These genes display a gradual increase/decrease in expression during cancer progression, from NCM to liver metastasis. The scalability observed for this response may link the regenerative capacities of different organs after an injury (i.e., liver) and the greater ability of CAF-LM to induce epithelial–mesenchymal transition (EMT) phenotypes in epithelial cells more efficiently than in other types of myofibroblasts (36). Many of these genes are associated with wound healing and correlate with the activation status of fibroblasts. As an example, SERPINE1, POSTN, HBEGF, PTX3 or TGFβ2, IL1β, and TGFβ1 responsive genes, increase from NCF to CAF-LM, with IL1β and TGFβ1 being the main inducers of fibroblast activation. This transcriptomic program for wound response seems to be more effective in the more aggressive liver and primary tumors, and could explain the greater degree of desmoplasia that is associated with poor outcome (37). About downregulated genes, there is an enrichment of genes involved in inflammatory processes. This fact might explain the better outcome of patients with an intense inflammation-related component infiltration. Probably the lack of such CAF-derived molecules in the more aggressive tumors would avoid the recruitment of inflammatory cells to the tumor. In summary, this result strongly suggests that changes in colorectal tumor stroma have a crucial role in disease progression and outcome. Obviously, other fibroblast-specific genes could have prognostic value, although their level of expression would not be associated with the transition from normal fibroblasts to highly activated fibroblasts, as those genes would be differentially expressed in a particular microenvironment (e.g., they could be expressed in hepatic metastases but be absent from primary tumors).

Although fibroblasts from different demarcations display particular transcriptomic programs (38), even considering each fibroblast type as a different cell, it is interesting to observe that the transcriptome is dynamic and plastic when

| Table 3. Multivariate Cox proportional hazard analysis for DFS and DSS in the validation set GSE17538 |
|------------------------------------------|-------------------------------|-------------------------------|-------------------------------|
| Multivariate analysis 19-gene signature GSE17538 |
| DFS | DSS |
| HR (95% CI) | P | HR (95% CI) | P |
| Grade | 1.59 (0.66–3.79) | 0.3 | 1.63 (0.64–4.12) | 0.3 |
| Stage | 2.06 (1.1–3.84) | 0.023 | 1.455 (1.07–5.51) | 0.033 |
| 19-gene classifier (H vs. L) | 5.52 (2.23–13.65) | <0.0001 | 6.029 (1.76–20.65) | 0.004 |

NOTE: We included in the analysis only variables that were statistically significant in univariate analysis, then gender and age >55 years were excluded (n = 20 clinical variables were available for the dataset at GEO).
confronted with a malignant cell, probably as a result of the cross-talk between the two compartments. This is particularly relevant if we consider that CAFs from the primary tumor are derived from local NCF when oncogenically stimulated (same anatomic demarcation), but CAF-LM are cells with a different origin.

Previous studies have not focused on the role of the three microenvironments in which colorectal cancer cells interact during tumor progression. In addition, different stromal classifiers exist for breast (39), prostate (40), lung (41), and hepatocellular carcinoma (42), but this issue has not so far been addressed with respect to colorectal cancer; instead, the focus has been on CAFs, the principal component of the stroma. We also identified a 19-gene signature that predicts disease outcome with remarkable accuracy in whole-tumor samples.

The scalability of the stereotyped gene expression pattern was obtained from isolated fibroblasts of colorectal tumor-associated demarcations, so the prognostic power of the expression pattern is strong evidence of the importance of the stroma in the oncogenesis of colorectal tumors. This was observed implicitly in a signature previously obtained from whole-tumor samples, in which many of the 48 most relevant genes were of stromal origin (4). Because the method has been performed and validated in whole-tumor samples, in which the stroma/epithelium ratio was not known (37), the expression signal detected emphasizes the role of stroma. Most of the 19 genes identified in the present study are strongly expressed by CAFs (with the exception of ULBP2 and CFTR) and to a lesser extent by other cells of the tumor stroma. In addition, many of these genes have been associated previously with tumorigenic processes. Stromal SERPINE1 (coding for PAI-1) has been involved in cell migration and angiogenesis promotion in different desmoplastic tumors (43, 44). INHBA (a TGFB family member) has been recently observed to be expressed in the sequence from NCM–colonic adenoma–colorectal adenocarcinoma (45) and has been associated with prognosis (46). Elevated expression of POSTN is reported to be associated with the invasion and anchorage-independent growth of the metastatic process of head and neck squamous cell carcinoma (47) and the invasive neoplasm of the pancreas (48). Furthermore, secreted POSTN was reported to trigger EMT through coactivation of EGFR and alpha 5 integrin (49). Other genes such as NTM (neurotrimin) and DENND2A have not been yet involved in cancer processes neither in the stroma nor epithelium.

![Graph showing the 19-gene classifier score values for the patients of the validation dataset GSE33113 (90 patients stage II). The red line corresponds to the cutoff (0.187) obtained from the Euclidian distance between sensitivity and specificity for recurrence prediction in the training dataset. B, ROC curve for 19-gene signature performance in the GSE33113 (AUC, 0.71) validation dataset. C, waterfall plot for recurrence risk of 90 patients from validation dataset GSE33113 according to the 19-gene signature score (OR = 2.21; Sensitivity 86.4% and specificity 69%). D, DFS Kaplan–Meier curves for the performance of the 19-gene signature in the GSE33113 validation dataset (stage II) for signatures classified as high (red) or low (black) expression. HR, 3.84; likelihood ratio test, \( P = 0.0034 \). Five-year DFS rates were 60% and 87% for high- and low-expression patients, respectively.](image-url)
According to previously reported gene signatures in colorectal cancer, we found overlap only between two genes (PDILIM3 and ULBP2) of our classifier and a recent molecular classification (50), contributing also these genes to characterize the group of worst outcome.

We present different, complementary ways of interpreting the results that could be of great relevance, depending on the clinical scenario. First, PAM analysis identifies the genes of relevance in terms of binary events (recurrence/nonrecurrence) that are highly specific after cross-validation. Second, the signature score (equal weight for all genes) reveals the biologic relevance of the genes selected by PAM, clearly demonstrating that this signature is an independent predictor of outcome, even for stage III patients. This step is crucial before trying to develop a model that is more mathematically accurate albeit without a biologic rationale. This reinforces the role of the stroma, and particularly fibroblasts, as important components of cancer progression, rather than as mere companions of malignant cells.

On the basis of our results, we provide accurate information for clinical decision making, as our classifier performs very well in identifying low-risk patients who would benefit from an omission of adjuvant chemotherapy if validated in independent clinical samples. This could be very significant for those stage II T4N0, perforated, obstructed, poorly differentiated, or microsatellite stable patients, who would otherwise be offered toxic adjuvant chemotherapy. In stage III patients, our signature also differentiates low-risk patients, but these results need to be corroborated in a larger series and in trials to convince oncologists to withhold chemotherapy completely from these patients, and even those of the T2N1 subgroup, whose members have a better clinical prognosis.

In conclusion, using the transcriptomic signature of fibroblasts that we defined and specifically the 19-gene classifier derived from it in a clinical setting, we can provide accurate information about the risk of recurrence, and may facilitate the selection of patients at risk of recurrence, especially for high-risk stage II patients who would benefit from adjuvant therapy.

The findings from the stratification of patients based exclusively on the CAF transcriptomic program indicate the need to develop therapeutic strategies focused on these cells, as they are the main components of desmoplasic tumors. Prospective studies are needed to determine whether treatment decisions based on our stromal 19-gene classifier could benefit patients.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Conception and design: D.G. Molleví
Development of methodology: N.G. Díaz-Murtra, D.G. Molleví
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M. Berdiel-Acer, X. Sanjuán, T. Serrano, V. Moreno, S. Gonçalves-Ribeiro, A. Villanueva, D.G. Molleví
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M. Berdiel-Acer, D. Cuadras, A. Berenguer, V. Moreno, A. Villanueva, D.G. Molleví
Writing, review, and/or revision of the manuscript: M. Berdiel-Acer, X. Sanjuán, A. Berenguer, V. Moreno, S. Gonçalves-Ribeiro, D.G. Molleví
Study supervision: R. Salazar, D.G. Molleví

Grant Support
This work was supported by projects of the Instituto de Salud Carlos III (grants P07/0657 and P11/1604) and the Agència de Gestió d’Ajuts Universitaris i de Recerca (grant number AGAUR SGR-2009-681). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 5, 2014; revised April 25, 2014; accepted April 25, 2014; published OnlineFirst May 14, 2014.

References


A Monotonic and Prognostic Genomic Signature from Fibroblasts for Colorectal Cancer Initiation, Progression, and Metastasis
