Oncology Research, Vol. 16, pp. 535–548
Printed in the USA.

0965-0407/07 $90.00 + .00
E-ISSN 1555-3906
www.cognizantcommunication.com

Analysis of Gene Expression Profiles Reveals Novel Correlations With the Clinical Course of Colorectal Cancer

Duccio Cavalieri,* Piero Dolara,* Enrico Mini,* Cristina Luceri,* Cinzia Castagnini,* Simona Toti,*† Karolina Maciag,* Carlotta De Filippo,* Stefania Nobili,* Maria Morganti,* Cristina Napoli,* Giulia Tonini,† Michela Baccini,† Annibale Biggeri,† Francesco Tonelli,‡ Rosa Valanzano,‡ Claudio Orlando,‡ Stefania Gelmini,‡ Fabio Cianchi,§ Luca Messerini,¶ and Lucio Luzzatto#

*Department of Pharmacology, University of Florence, Florence, Italy
†Department of Statistics, University of Florence, Florence, Italy
‡Department of Physiopathology, University of Florence, Florence, Italy
§Dipartimento di Area Critica Medico Chirurgica, University of Florence, Florence, Italy
¶Department of Human Pathology and Oncology, University of Florence, Florence, Italy
#Istituto Toscano Tumori, Florence, Italy

(Submitted August 30, 2007; accepted September 12, 2007)

In order to discover potential markers of prognosis in colorectal cancer (CRC) we have determined gene expression profiles, using cDNA microarrays in CRC samples obtained from 19 patients in Dukes stages C and D, with favorable clinical course (Dukes C patients, survival >5 years after surgery, group A, $n = 7$) or unfavorable clinical course (Dukes stage C and D patients, survival <5 years after surgery, group B, $n = 12$). Gene expression was measured in RNA from each tumor, using a pool of equal amounts of RNA from all tumors as a reference. To identify and rank differentially expressed genes we used three different analytical methods: (i) Significance Analysis of Microarrays (SAM), (ii) Cox's Proportional Hazard Model, and (iii) Trend Filter (a mathematical method for the assessment of numerical trends). The level of expression of a gene in an individual tumor was regarded as of interest when that gene was identified as differentially expressed by at least two of these three methods. By these stringent criteria we identified eight genes (*ITGB2, MRPS11, NPR1, TXNL2, PHF10, PRSS8, KCNK3, JAK3*) that were correlated with prolonged survival after surgery. Pathway analysis showed that patients with favorable prognosis had several activated metabolic pathways (carbon metabolism, transcription, amino acid and nitrogen metabolism, signaling and fibroblast growth factor receptor pathways). To further validate individual gene expression findings, the RNA level of each gene identified as a marker with microarrays was measured by real-time RT-PCR in CRC samples from an independent group of 55 patients. In this set of patients the Cox Proportional Hazard Model analysis demonstrated a significant association between increased patient survival and low expression of *ITGB2* ($p = 0.011$) and *NPR1* ($p = 0.023$) genes.

Key words: Colorectal cancer prognosis; Gene expression; Microarrays; Real-time RT-PCR

INTRODUCTION

Colorectal cancer (CRC) is heterogeneous at a molecular level (1,2); its clinical course is also quite variable and it correlates strongly with Dukes stage. However, even patients with the same Dukes stage and histological grading may have widely different clinical outcomes. Therefore, reliable prognostic markers would be of great value and would serve to guide treatment options.

In recent years microarray studies have been used for precisely this purpose. Profiles of gene expression have been compared with samples of tumors at different stages with the aim of identifying correlations with prognosis (3–24). In some studies the gene expression pattern of tumors was compared with that of normal reference tissues or cell lines. In another study the gene expression pattern of the apparently normal colon mucosa was tested as a predictor of clinical outcome (22). Although significantly different patterns of gene expression have emerged from these studies, no single marker or signature of gene expression is yet accepted as a prognostic parameter in CRC patients.

Address correspondence to Prof. Enrico Mini, Department of Pharmacology, Viale Pieraccini 6, 50139, Firenze, Italy. Fax: +39 055 4271280; E-mail: enrico.mini@unifi.it

We thought that the chance of identifying gene expression changes as markers of prognosis would increase if we were able to focus on differences between more invasive and less invasive tumors, rather than on changes from normal to tumor tissue. Therefore, using microarrays we analyzed tumors from individual patients with different clinical outcome against a reference pool consisting of RNA extracted from the entire set of tumors. In addition, the microarray gene expression data were processed using three independent statistical methodologies. We also used a novel statistical tool for studying the differential variation of genes involved in metabolic pathways between tumors with different prognosis. Finally, the genes emerging as potential markers in microarray data analysis were subjected to a validation procedure, consisting of real-time reverse transcriptase PCR assays (real-time RT-PCR) on a separate group of unselected patients with known clinical outcome.

MATERIALS AND METHODS

Patients and Samples

For the analysis of gene expression with microarrays we selected samples from 19 patients with Dukes stage C and D among the frozen colorectal cancer samples available in the tumor bank of Chemotherapy Unit, Department of Pharmacology, University of Florence, Italy, considering favorable (group A) or unfavorable (group B) survival (>5 or <5 years, respectively) (Table 1). G2 grade and adenocarcinoma histotype were other selection criteria. All patients had surgery and postoperative chemotherapy. Patients who died of causes unrelated to colorectal cancer were excluded except for one patient who died after 10 years of renal adenocarcinoma and who was included in the favorable prognosis group. These samples were collected before combination chemotherapy became standard practice both in the adjuvant setting and in advanced disease. Patients had been administered 5-fluorouracil-based chemotherapy. The mean follow-up duration was 4.8 years.

An additional 55 CRC patients were recruited after surgery with no selection criteria (Tables 2 and 5) and the variation in expression of a few marker genes, selected by microarray analysis on the first set of analysis, was analyzed by real-time RT-PCR in resected tumor samples. Primary tumors in both groups of patients were obtained at surgery and immediately after resection. The tumor samples were divided into two equal portions after washing and removal of necrotic tissues. One portion was fresh frozen in liquid nitrogen until RNA extraction and the other portion was embedded in paraffin to confirm that it did not contain significant contamination by normal tissues, necrotic tissues, and lymphocytes, according to standard histological practices.

Follow-up information was available from the Division of Clinical Chemotherapy Patient Database (Careggi University Hospital, Florence, Italy) and from the surgical units cooperating with this project. After surgery, patient follow-up was carried out according to standard practices for colon cancer patients. Patients were evaluated at 3-month intervals for the first postoperative year and at 6-month intervals thereafter.

Time to recurrence, or disease-free interval, was defined as the time elapsing from the date of surgery to confirmed cancer relapse date for relapsed patients and from the date of surgery to the date of last follow-up for disease-free patients. Overall survival was measured as difference between the date of the last check-up and date of surgery or between date of death and date of surgery.

Informed consent was obtained from patients regarding use of their specimens and clinical/pathological data for research purposes and all procedures followed the guidelines established by the local ethical committee.

RNA Analyses

The 19 samples of the first group of CRC were used for cDNA microarray analyses. Total RNA was extracted using the RNeasy Midi kit (Qiagen, Milan, Italy).

We constructed a reference RNA pool by mixing equal amounts of RNA extracted from each CRC tumor specimens; the RNA from each tumor was hybridized against the RNA reference pool.

Gene Expression Measurements With Microarrays

We used the human 1A Oligo Microarray Kit (V2) (Agilent Technologies, Palo Alto, CA, USA), containing 22,575 elements covering 18,000 genes for gene expression measurements; the indirect labeling method described by De Risi (http://derisilab.ucsf.edu/) was used. The incubation was performed at 63°C for 14–16 h in a humid chamber, using the Agilent 2X hybridization buffer (Agilent Technologies, Palo Alto, CA, USA). Fluorescent DNA bound to the microarray was detected with a GenePix 4000 microarray scanner (Axon Instruments, Foster City, CA, USA), using the GenePix 6.0 software to locate spots in the microarray. Each comparison was performed in duplicate. Laser scanner-acquired images were quantified by analyzing each spot. Data were analyzed according to stringent quality control procedures and consistency with a validated flow chart analysis.

Quality Control

The following features were required to control microarray spot quality: 60 pixels minimum spot diameter; 50% as the minimum percentage of pixels for which the

foreground intensity was greater than the background intensity + 2 SD; 80 pixels as the minimum number of pixels; 20% as the maximum percentage of saturated pixels.

We divided the data of each microarray into two clusters around "medoids" using the "CLARA" function (25) from the library "cluster" of R package (www.r-project.org). We eliminated the spots agglomerated around the lower medoid from the analysis, which presented a strong linear and artificial relationship between the average intensity and the differential intensity of the foreground medians of red (Cy5) and green channel (Cy3). After background adjustment for each gene we calculated the difference between the median foreground intensity and background intensity for each channel.

Normalization

We evaluated the differential expression of genes in CRC samples compared to the reference pooled tumor RNAs. To do so we used the log-ratio of the background-adjusted intensity on the red versus green channel. The genes that corresponded to a zero value of log-ratio were considered not differentially expressed. However, due to the technical steps of microarray production, such as the labeling and scanning, the log-ratio values were not centered around zero. We corrected for this anomaly by normalizing (i.e., recentering) the medians of log-ratios around zero for each microarray dataset. After normalization the log-ratios from different microarrays were comparable and we could express the information for each patient by averaging the log-ratio intensity of the two replicate microarrays. The number of genes analyzed in the 19 subjects ranged from a minimum of 4727 to a maximum of 8839.

Differential Analysis by SAM

We analyzed the normalized log-ratios by comparing the gene expression values. The groups analyzed were 7 subjects living after 5 years of follow-up (favorable group) and 12 subjects who died within the same period (unfavorable group). We determined which genes were differentially expressed for the two groups using the

Table 1. Summary of the Clinical Characteristics of Patients Analyzed for Gene Expression in Microarray Experiments

Patient ID	Prognosis Group	Sex	Age	Dukes Stage	TNM	Grading	Histotype	Location	Survival Time (Months)	Censoring Status
C01	A	M	63	C1	T3N1M0	G2	ADK	transverse colon	102.97	censored
C03	A	M	68	C1	T3N1M0	G2	ADK	sigmoid colon	109.13	censored
C04	A	M	46	C2	T3N2M0	G2	ADK*	sigmoid colon	77.47	censored
C05	A	F	58	C2	T3N2M0	G2	ADK	rectum	62.93	censored
C07	A	F	52	C1	T3N1M0	G2	ADK	sigmoid colon	135.73	censored
C08	A	F	52	C1	T3N1M0	G2	ADK	rectum	127.17	not censored
C09	A	M	71	C1	T3N1M0	G2	ADK	rectum	125.00	censored
C11	B	F	61	C1	T3N1M0	G2	ADK	sigmoid colon	28.20	not censored
C12	B	F	70	C2	T3N2M0	G2	ADK	sigmoid colon	27.47	not censored
C13	B	M	66	C2	T3N2M0	G2	ADK	left colon	16.40	not censored
C14	B	M	57	D	T3N1M1	G2	ADK	rectum	39.33	not censored
C15	B	F	59	C2	T3N2M0	G2	ADK	right colon	34.20	not censored
C16	B	M	47	C1	T3N1M0	G2	ADK	transverse colon	48.77	not censored
C17	B	F	71	D	T3N1M1	G2	ADK	rectum	38.63	not censored
C21	B	M	57	D	T3N0M1	G2	ADK	rectum–sigmoid colon junction	33.20	not censored
C22	B	F	73	D	T3N2M1	G2	ADK	rectum	31.30	not censored
C26	B	F	62	D	T3N0M1	G2	ADK	sigmoid colon	29.90	not censored
C27	B	M	63	D	T3N2M1	G2	ADK	right colon	12.93	not censored
C28	B	M	70	D	T3N2M1	G2	ADK	rectum–sigmoid colon junction	12.53	not censored

Favorable prognosis group (A): patients alive at 5 years after surgery; median age: 58 (range 46–71); median disease-free survival: 102 months (range 72–134); median overall survival: 108 months (range 72–134+). Patient C08 died for a non-colon-related pathological process. Unfavorable prognosis group (B): patients deceased within 5 years from surgery; median age: 63 (range 47–73); median disease-free survival: 16 months (range 11–27); median overall survival: 30 months (range 12–48). Censoring status indicates whether a patient was alive (censored) when the study was terminated or not alive (not censored).
*Colloid.

Table 2. Characteristics of the 55 Patients and Tumors Studied With RT-PCR

Patient ID	Sex	Age	Dukes Stage	Location	TNM	Mucinous	Grading	Survival Time (Months)	Censoring Status
80	M	80	A	sigmoid colon	T1N0M0	yes	G1	63.65	censored
82	F	56	D	rectum	T4N1M1	no	G2	97.77	not censored
84	M	54	A	rectum & sigmoid colon	T1N0M0	yes	G1	96.99	censored
85	M	68	C	right colon	T3N1M0	no	G2	96.46	censored
86	M	59	A	right colon	T1N0M0	yes	G1	100.60	censored
87	M	71	A	rectum & sigmoid colon	T1N0M0	yes	G1	95.44	censored
88	M	68	C	right colon	T2N1M0	no	NA	95.34	censored
91	M	69	A	cecum	T1N0M0	yes	G1	94.26	censored
92	F	70	B	cecum	T3N0M0	yes	G3	94.09	censored
93	M	75	B	rectum	T3N0M0	no	G2	93.96	censored
94	M	74	B	rectum	T1N0M0	no	G2	88.67	not censored
95	M	84	C	sigmoid colon	T3N1M0	yes	G3	40.54	not censored
96	F	60	C	rectum	T2N1M0	no	G2	92.71	censored
98	M	73	A	right colon	T1N0M0	yes	G1	3.35	not censored
100	M	76	B	right colon	T3N0M0	yes	G3	51.88	censored
101	F	57	D	rectum	T4N1M1	no	G2	17.96	not censored
102	M	80	D	cecum	T3N1M1	no	G2	5.82	not censored
104	F	70	C	cecum	T3N1M0	yes	G3	90.18	censored
105	M	54	A	rectum	T2N0M0	yes	G2	88.60	censored
106	M	63	B	rectum	T3N0M0	yes	G3	89.95	censored
107	F	75	B	right colon	T3N0M0	no	G2	18.87	not censored
108	M	73	B	sigmoid colon	T3N0M0	yes	G2	36.39	censored
112	M	59	C	left colon	T3N0M0	no	G2	84.66	censored
113	F	74	A	rectum	T2N0M0	yes	G2	59.66	censored
114	M	75	D	left colon	T3N1M1	yes	G2	25.61	censored
116	M	72	A	sigmoid colon	T1N0M0	no	G2	85.64	censored
117	F	70	A	rectum	T2N0M0	yes	G3	16.61	censored
118	M	63	C	right & transverse colon	T3N1M0	no	G2	49.25	censored
120	F	64	B	sigmoid colon	T3N0M0	yes	G2	15.00	censored
121	F	59	C	right colon	T3N1M0	no	G2	29.19	censored
122	M	71	B	right colon	T3N0M0	no	G2	48.20	censored
123	M	75	C	rectum & sigmoid colon	T3N1M0	no	G2	71.51	censored
126	M	75	B	left colon	T3N0M0	no	G2	71.28	censored
128	M	66	C	rectum	T3N1M0	no	G2	12.99	not censored
130	M	70	C	rectum	T3N1M0	no	G2	80.55	censored
131	M	65	B	rectum	T3N0M0	no	G2	32.68	not censored
132	M	67	C	rectum	T3N1M0	no	G2	79.04	censored
136	M	71	C	right colon	T3N1M0	no	G2	61.78	censored
137	M	66	B	right colon	T3N1M0	yes	G1	84.07	censored
138	M	68	B	sigmoid colon	T3N0M0	no	G2	54.30	censored
141	M	63	C	right colon	T2N1M0	yes	G1	83.08	censored
142	F	73	C	sigmoid colon	T3N1M0	yes	G3	76.83	censored
143	F	60	B	sigmoid colon	T3N0M0	no	G2	40.96	censored
144	F	66	D	cecum	T3N0M1	no	G2	7.74	censored
145	M	72	B	rectum	T3N0M0	no	G2	42.81	censored
146	F	56	B	sigmoid colon	T3N0M0	no	G2	77.29	censored
151	M	68	C	left colon	T3N2M0	yes	G3	16.64	censored
156	M	79	D	rectum	T3N2M1	no	G2	16.54	censored
159	M	63	D	right colon	T3N2M1	no	G2	15.89	not censored
160	F	62	B	rectum	T3N0M0	no	G2	75.13	censored
161	M	74	B	right colon	T3N0M0	no	G2	15.20	censored
162	M	68	A	rectum	T2N0M0	no	G2	47.93	censored
167	M	62	C	left colon	T3N1M0	no	G2	13.59	censored
169	M	76	B	right colon	T3N0M0	no	G2	47.28	censored
177	F	61	B	right colon	T3N0M0	yes	G3	51.90	censored

Table 3. Genes Differentially Expressed in Association With the Prognosis With Different Statistical Approaches, According to Microarray Analysis of Tumor Samples From Patients of Table 1 With Favorable or Unfavorable Prognosis

Gene Name	Gene Description	Cox's Proportional Hazard Analysis	SAM	Trend Filter	Up-/Downregulation
ITGB2	*Homo sapiens* integrin, beta 2 (antigen CD18 (p95), lymphocyte function-associated antigen 1; macrophage antigen 1 (mac-1) beta subunit) (ITGB2), mRNA	yes	yes	no	Upregulated in patients with unfavorable prognosis
NPR1	*Homo sapiens* natriuretic peptide receptor A/ guanylate cyclase A (atrionatriuretic peptide receptor A) (NPR1), mRNA	yes	yes	no	Upregulated in patients with unfavorable prognosis
TXNL2	*Homo sapiens* thioredoxin-like 2 (TXNL2), mRNA	yes	yes	no	Upregulated in patients with unfavorable prognosis
MRPS11	*Homo sapiens* mitochondrial ribosomal protein S11 (MRPS11), nuclear gene encoding mitochondrial protein, transcript variant1, mRNA	yes	no	yes	Upregulated in patients with favorable prognosis
JAK3	*Homo sapiens* Janus kinase 3 (a protein tyrosine kinase, leukocyte) (JAK3), mRNA	no	yes	yes	Downregulated in patients with favorable prognosis
KCNK3	*Homo sapiens* potassium channel, subfamily K, member 3 (KCNK3), mRNA	yes	yes	no	Downregulated in patients with unfavorable prognosis
PRSS8	*Homo sapiens* protease, serine, 8 (prostasin) (PRSS8), mRNA	yes	yes	no	Upregulated in patients with unfavorable prognosis
PHF10	*Homo sapiens* PHD finger protein 10 (PHF10), transcript variant 1, mRNA	yes	yes	no	Upregulated in patients with unfavorable prognosis

Significant Analysis of Microarrays (SAM), as specified in the "one-class method" (26), which tests the hypothesis of over- or underexpression for every gene relative to the reference gene, controlling for the False Discovery Rate (FDR) (27).

Trend Filter

We used a mathematical method called Trend Filter, derived from the Trend Capture feature of the commercial package "Rosetta Resolver" developed by Rosetta Biosoftware Inc to analyze microarray data (http://www.rosettabio.com/products/resolver/default.htm). In the routine trend filter a gene was said to exhibit a "trend" in an experimental group (e.g., "favorable prognosis patients"), if it was over- or underexpressed in a set proportion of the patients in the group. Conversely, a gene did not exhibit a "trend" if its over- or underexpression was limited to a small proportion of the patients in the group.

We checked for genes that exhibited a trend of over- or underexpression in the group of patients with favorable prognosis and did not exhibit the same trend in the group with unfavorable prognosis, and reverse. When comparing the favorable prognosis versus unfavorable prognosis groups, we called a gene over- or underexpressed when the minimum ratio for the favorable prognosis patients was at least 1.2. The minimum number of patients exhibiting a trend in the favorable prognosis patients ranged between 4 and 6; the maximum number of patients exhibiting a trend in the unfavorable prognosis patients ranged between 0 and 4. When comparing unfavorable versus favorable prognosis groups, we called a gene over- or underexpressed when the minimum ratio value for the unfavorable prognosis patients was at least 1.2. The minimum number of patients exhibiting a trend in the unfavorable prognosis patients ranged between 6 and 8; the maximum number of patients exhibiting a trend in the favorable prognosis patients ranged between 1 and 2.

Cox's Survival Analysis Model

Survival analysis was determined using the Cox model (28), where the hazard function is defined as proportional to different covariates. The model assumes that

Table 4. Gene Set Enrichment Analysis (GSEA) Results, Obtained With the Eu.Gene Program, Using Values of Microarray Gene Expression in the Favorable Versus Unfavorable Patient Group

Cover (%)	e-Scores	*p*-Value	Pathway
Carbon metabolism			
23	0.91173	0	Oxidative decarboxylation of pyruvate to acetyl CoA by PDH
20	0.83994	0	Hs_Krebs-TCA_cycle
21	0.66524	0.03	Hs_glycolysis_and_gluconeogenesis
23	0.71363	0.02	Glycolysis/gluconeogenesis
34	0.79091	0.04	Ascorbate and aldarate metabolism
26	0.82439	0.03	Methane metabolism
25	0.75378	0.01	Pyruvate metabolism
27	0.74551	0.01	Propanoate metabolism
24	0.84012	0.01	Citrate cycle (TCA cycle)
22	0.72421	0.03	Butanoate metabolism
35	0.86629	0.02	Cholesterol biosynthesis
35	0.86629	0.02	Steroid metabolism
20	0.70509	0.03	Hs_Fatty_acid_beta_oxidation_1_BiGCaT
24	0.67662	0.04	Hs_adipogenesis
50	0.83753	0.04	ChREBP activates metabolic gene expression
Aminoacid and nitrogen			
26	0.83321	0	beta-Alanine metabolism
23	0.85149	0	Thiamine metabolism
37	0.75676	0.01	Hs_Tryptophane_metabolism_KEGG
22	0.81284	0.01	Limonene and pinene degradation
24	0.79237	0.01	Histidine metabolism
21	0.7649	0.04	Valine, leucine, and isoleucine degradation
26	0.72852	0.04	Lysine degradation
Transcription			
34	0.81019	0.01	RNA polymerase III chain elongation
34	0.81019	0.01	RNA polymerase III transcription termination
31	0.75265	0.01	Hs_RNA_transcription_reactome
30	0.72917	0.02	Transcription of the HIV genome
23	0.72449	0.03	RNA polymerase II transcription
23	0.72209	0.04	Transcription
Signaling			
28	0.76452	0.01	Hs_Hedgehog_Netpath_10
24	0.75676	0.02	Hs_Delta-notch_NetPath_3
17	0.79047	0.05	notch Signaling pathway
22	0.7685	0.05	Fibroblast growth factor receptor (FGFR) signaling
43	0.82723	0.02	FGFR3b ligand binding and activation
34	0.80403	0.04	FGFR1c ligand binding and activation
34	0.80403	0.04	FGFR2c ligand binding and activation
40	0.79135	0.02	FGFR3 ligand binding and activation
40	0.79135	0.02	FGFR3c ligand binding and activation
31	0.78715	0.05	FGFR1 ligand binding and activation
24	0.78715	0.05	FGFR2 ligand binding and activation
22	0.7685	0.05	FGFR ligand binding and activation
Infiltration			
24	0.67617	0.02	Tight junction
29	0.76262	0	Hematopoietic cell lineare
16	**–0.8347**	**0.02**	Dissolution of fibrin clot
23	0.79047	0.05	A third proteolytic cleavage releases NICD
Others			
25	0.76427	0.03	Amyotrophic lateral sclerosis (ALS)
24	0.64059	0.04	Neurodegenerative disorders
38	0.68547	0.03	Taste transduction
23	0.82173	0.02	Hs_Id_NetPath_5

The first column (% cover) indicates the percentage of genes in the pathway which are present on the microarray used for the evaluation of gene expression. The second column (E-scores) is an enrichment score. The third column shows the empirical p-value. The last column indicates the pathway names according to human databases (KEGG, Reactome, GenMAPP). The positive signs indicate an expression level of a pathway higher in the favourable outcome group. The negative value (bold character) indicates that the expression level was lower in the favourable prognosis group.

Table 5. Clinical Characteristics of Patients and Tumors Analyzed by Real-Time RT-PCR ($N = 55$)

Characteristic	*N*
Age (years) (mean 68 ± 7.1)	
≤60	10
>60	45
Sex	
Male	39
Female	16
Dukes stage	
A	11
B	20
C	17
D	7
Tumor grading	
G1	8
G2	36
G3	10
NA	1
Pathology	
Colloid adenocarcinoma	22
Adenocarcinoma	33
Tumor location	
Right colon	15
Left colon	5
Cecum	5
Sigma	10
Rectum	16
rectum–sigma junction	3
Right and transverse colon junction	1

the underlying hazard rate is a function of several independent variables. The model may be written as:

$$h\{(t), (z_1,z_2, \ldots,z_m)\} = h_0(t)*\exp(b_1*z_1 + \ldots + b_m*z_m \quad (1)$$

where $h(t, \ldots)$ denotes the resultant hazard, given the values of the m covariates for the respective case ($z_1, \ldots, z_m$) and the respective survival time (t). The term $h_0(t)$ is the baseline hazard (i.e., the hazard for each individual when all independent variable values are equal to zero). In this analysis we did not consider the genes with missing values in any of the arrays. The dataset analyzed with this approach was comprised of 2587 genes overall.

We considered the following covariates: age at surgery (reference class <65), sex (reference class: male), Dukes stage (reference class: C1), tumor localization (reference class: transverse-sigma). Finally, we inserted each value of gene expression as calculated above. We then obtained one regression coefficient for each gene. We performed a probabilistic clustering, using the EM algorithm (29,30). We then considered the genes grouped in the more extreme clusters. With this approach we identified two groups of genes, in the highest and in the lowest clusters.

To analyze the effect of gene expression on survival as measured by real-time RT-PCR, we used the Cox regression model (as defined by the function above), considering age, sex, and Dukes stage as covariates.

Real-Time PCR

Gene expression was measured by quantitative real-time PCR (TaqMan™). The amount of target, normalized to an endogenous reference (18S) and relative to a calibrator (Quantitative PCR human reference total RNA; Stratagene, La Jolla, CA, USA) was expressed as $2^{-\Delta\Delta Ct}$. For each sample, 12.5 ng of cDNA was added to 10 μl of PCR mix containing each primers/probe mix ("Assay-on-demand," Applied Biosystems, Foster City, CA, USA) and 1× Universal Master Mix (Applied Biosystem, USA). The samples were then subjected to 40 cycles of amplification at 95°C for 15 s and 60°C for 60 s in the ABI Prism 7700 Sequence Detector (Applied Biosystems, USA).

Pathway Analysis

Pathway analysis is an attempt of identifying variations of activity of metabolic pathways through the study of the expression of genes attributed to these pathways in public databases (KEGG, Reactome, GenMAPP). Pathways analysis was carried out on the data of microarray gene expression using an original program (31) of our group freely available on the net (Eu Gene, http://www.ducciocavalieri.org/bio.htm) and based on a model described by Subramanian et al. (32). This model, known as Gene Set Enrichment Analysis (GSEA), orders microarray gene expression values in a vector and evaluates the statistical probability of a set of gene expression values in a pathway to be differentially expressed (either up- or downregulated) between experimental groups (in this case in the favorable or unfavorable CRC prognosis group).

RESULTS

The characteristics of the first set of patients are summarized in Table 1. This case series consisted of 19 selected CRC patients: (A) 7 patients with a favorable course and (B) 12 patients with an unfavorable course (Table 1). The histology and grading were the same in the two groups (Table 1). All patients received 5-fluorouracil-based chemotherapy after surgery (as adjuvant therapy in 12 Dukes C cases and as palliative therapy in 7 Dukes D cases). All patients in the favorable course group were disease free 5 years after surgery.

We searched for genes that showed significantly different expression in subjects with a favorable course

Table 6. Values of RT-PCR Measurements Expressed as $2^{-\Delta\Delta Ct}$ for Each Gene Studied

ID	*ITGB2*	*JAK3*	*KCNK3*	*NPR1*	*PHF10*	*PRSS8*	*MRPS11*	*TXNL2*
80	1.07	0.31	0.19	0.08	0.29	6.5	0.81	0.57
82	0.62	0.62	0.18	0.11	0.23	7.46	0.47	0.44
84	0.27	0.08	0.09	0.01	0.09	1.23	0.14	0.13
85	0.13	0.05	0.05	0.01	0.16	1	0.14	0.07
86	0.62	0.31	0.16	0.03	0.14	1.62	0.22	0.13
87	0.22	0.07	0.07	0.01	0.15	1.52	0.38	0.27
88	0.62	0.57	0.16	0.04	0.16	2.3	0.2	0.33
91	0.38	0.13	0.05	0.05	0.2	4	0.19	0.31
92	1	0.29	0.07	0.03	0.1	2	0.2	0.2
93	1	0.66	3.73	0.14	0.38	6.96	0.5	0.35
94	0.62	0.35	0.16	0.1	0.07	0.81	0.22	0.2
95	0.2	0.05	1.23	0.04	0.44	4.29	0.27	0.31
96	2.83	0.93	2.46	0.13	0.29	4.92	0.25	0.27
98	1.52	0.87	1.52	0.11	0.87	12.13	1.23	0.93
100	1.41	0.62	0.93	0.29	1.15	16	1.62	1.15
101	0.71	0.47	8	0.44	0.15	2.64	0.15	0.22
102	0.81	0.71	8	0.09	0.14	3.25	0.38	0.57
104	2.83	1.07	0.5	0.09	0.29	1.62	0.33	0.47
105	0.5	0.57	1.15	0.02	0.25	4.59	0.15	0.23
106	1.23	1	1	0.05	0.62	8	0.29	0.41
107	0.81	0.5	11.31	0.07	0.08	1.07	0.29	0.33
108	0.57	1	0	0.03	0.35	5.28	0.23	0.35
112	1.74	2	1.74	0.11	0.38	9.85	0.29	0.47
113	2.3	0.93	1.87	0.06	0.35	7.46	0.31	0.62
114	0.19	0.22	0.01	0.01	0.23	3.25	0.23	0.33
116	0.13	0.41	0.22	0.01	0.06	1.07	0.04	0.08
117	0.47	0.5	0.76	0.04	0.33	7.46	0.31	0.25
118	0.13	0.13	0.04	0.02	0.06	2.64	0.16	0.2
120	0	3.25	0.22	0.04	0.54	6.96	0.47	0.62
121	2.14	1.41	4.59	0.04	0.33	13.93	0.66	0.66
122	0.47	0.47	0.1	0.01	0.19	4.92	0.29	0.41
123	0.06	0.12	27.86	0.13	9.19	36.76	0.33	0.41
126	1.15	1.74	1.07	0.1	0.09	22.63	0.44	1.07
128	0.71	0	13.93	0.18	1.74	0.71	0.57	1.74
130	0.03	22.63	4.29	0.06	0.15	3.03	0.2	0.35
131	0.12	0.13	0.5	0.11	0.04	2.64	0.07	0.08
132	0.19	0.23	0.13	0.04	0.01	1.41	0.05	0.25
136	0.35	0.93	0.33	0.04	0.13	4	0.11	0.22
137	0.57	0.93	0.14	0.03	2.14	6.5	0.38	0.38
138	1	1.07	0.81	0.09	1.62	8.57	0.47	0.44
141	0.41	0.71	0.01	0.02	0.44	4	0.66	0.62
142	0.29	0.5	1.07	0.08	0.08	6.96	0.76	0.5
143	1.07	2	0.38	0.35	1.32	16	1.87	1.32
144	7.46	8	0.87	0.09	0.71	2	0.62	0.41
145	1.23	2.83	1	0.19	0.29	11.31	0.71	0.87
146	0.19	0.18	2.46	0.06	0.02	0.01	0.12	0.07
151	1.23	0.76	0.23	0.04	0.31	6.06	0.62	0.81
156	0.31	0.29	0.13	0.01	0.01	2.83	0.23	0.29
159	0.5	1	0.14	0.22	0.1	1.74	0.31	0.44
160	1.62	3.25	2.46	0.35	0.57	5.66	0.54	0.38
161	1.15	3.25	0.11	0.05	0.76	0.19	0.47	0.81
162	1.62	8	59.71	0.35	0.66	6.96	0.66	0.66
167	0.71	2.83	0.1	0.1	0.22	6.06	0.31	0.62
169	0.93	2.3	1.62	0.22	0.08	4.29	0.22	1
177	3.03	1.74	0.02	0.09	0.25	21.11	2.46	1.62

Table 7. Association of RT-PCR Gene Expression Data With Survival Analyzed With Cox's Proportional Hazard Model in the 55 CRC Cases of Table 5

Gene Name	Gene Description	Low or High Risk	*p*-Value
ITGB2	*Homo sapiens* integrin, beta 2 (antigen CD18 (p95), lymphocyte function-associated antigen 1; macrophage antigen 1 (mac-1) beta subunit)	+	**0.011**
NPR1	*Homo sapiens* natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)	+	**0.023**
TXNL2	*Homo sapiens* thioredoxin-like 2	+	0.085
MRPS11	*Homo sapiens* mitochondrial ribosomal protein S11, nuclear gene encoding mitochondrial protein, transcript variant 1	+	0.237
JAK3	*Homo sapiens* Janus kinase 3 (a protein tyrosine kinase, leukocyte)	–	0.61
KCNK3	*Homo sapiens* potassium channel, subfamily K, member 3	+	0.35
PRSS8	*Homo sapiens* protease, serine, 8 (prostasin)	–	0.76
PHF10	*Homo sapiens* PHD finger protein 10, transcript variant 1	–	0.85

Low gene expression associated with low risk (+) or high risk (–).
Values of $p < 0.05$ shown in bold.

versus patients with a unfavorable course. As specified in the Materials Methods section, three independent statistical methods were used: SAM, Cox's Proportional Hazard Model, and Trend Filter. For each of the two groups and for each gene we analyzed whether the expression of that gene was significantly higher or significantly lower relative to the reference pool using SAM. In the favorable prognosis group we found one overexpressed gene and 29 underexpressed genes; in the unfavorable prognosis group we found 114 over expressed genes and 423 underexpressed genes (i.e., by SAM analysis a total of 567 genes appeared to be associated with prognosis).

By using Cox's Proportional Hazard Model we quantitatively tested for each gene measurable with microarrays to what extent the level of expression was correlated with survival in the entire set of 19 patients. We controlled for the effect of age, sex, Dukes stage, and CRC localization; of these covariates, only Dukes stage was correlated with survival ($p < 0.05$). The 2587 scored genes were arranged by probabilistic clustering into eight classes, based on the coefficient of association between the level of expression and survival. In the two classes with the highest association values, 35 genes significantly associated with survival.

By our third method of analysis, the Trend Filter, we tested the number of patients showing a differential expression of each gene and belonging to one of the two groups (favorable and unfavorable course); we found 96 genes associated with a favorable prognosis and 40 genes associated with an unfavorable prognosis, for a total of 136.

In order to improve our chances to find good predictors of clinical course, we reasoned that the best candidates would be those genes that were picked up by at least two of the three methods of analysis used (Table 3). Seven genes were identified by both SAM and Cox analyzes and one gene by the Cox and Trend Filter analysis, for a total of eight genes.

Using pathway analysis on the microarray gene expression data we found that 48 pathways (Table 4) were differently regulated in the favorable or unfavorable prognosis ($p \leq 0.05$). Among the pathways significantly varied, most were overexpressed in patients with good prognosis and only one pathway ("Dissolution of fibrin clot") showed a lower expression level in the good prognosis group.

To validate the microarrays' analysis obtained on this relatively small number of samples, we performed real-time RT-PCR on each of these eight genes on samples from a separate, larger group of 55 patients (Tables 2, 5, and 6). These samples had been collected from consecutive surgical procedures without any selection.

We analyzed these data with Cox's proportional hazard analysis using the continuous values obtained by RT-PCR; moreover, we took into account in the survival model of covariates, such as Dukes stage and age.

In our compilation of the results (Table 7) the plus (+) sign of column 3 indicates that low gene expression of a specific gene was associated with lower risk; the

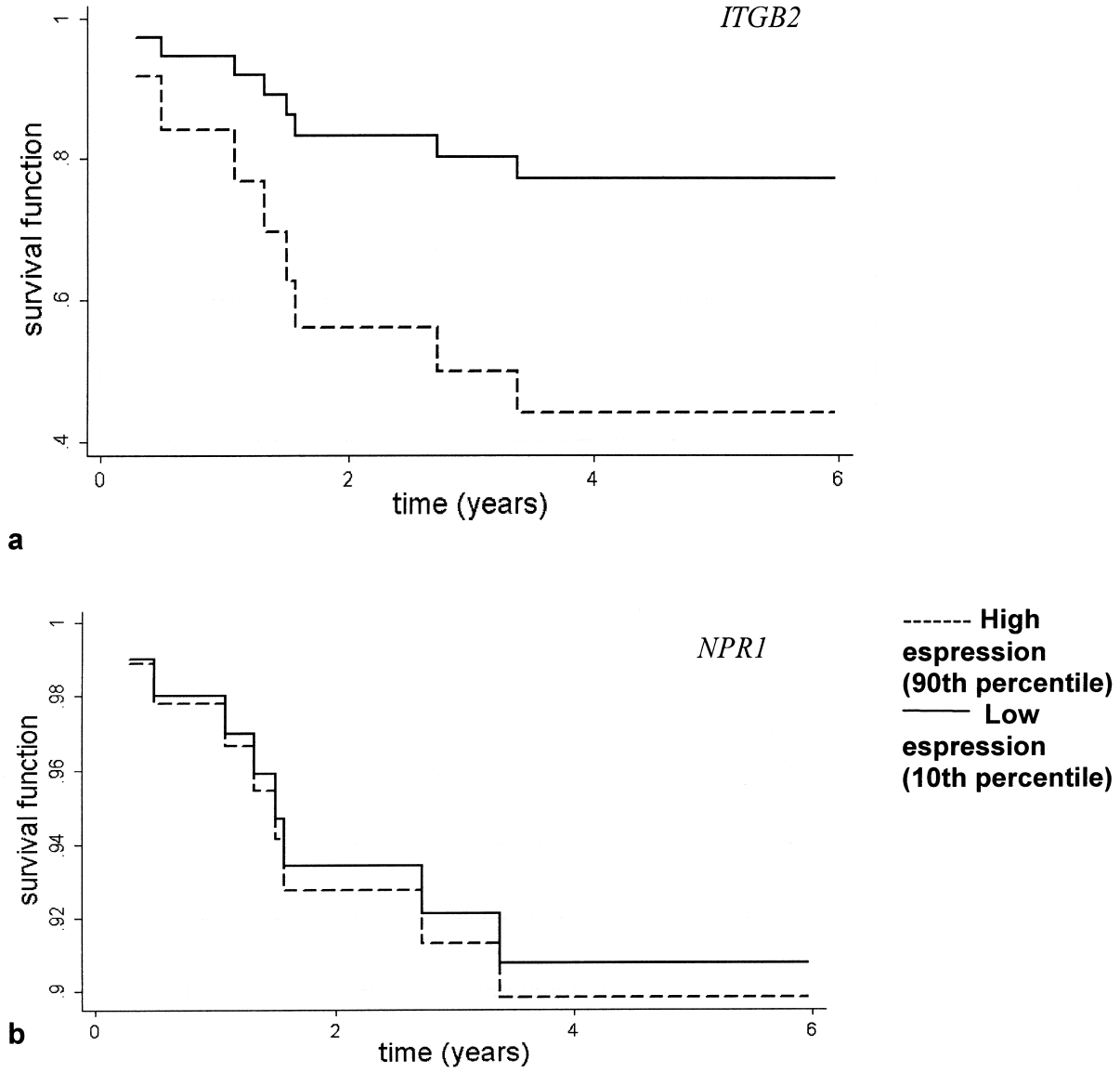

Figure 1. (a) Survival hazard function $h(t)$ (see the equation of the Cox model in Materials and Methods) for patients having a low expression (10th percentile of expression, upper continuous line) or a high expression (90th percentile of expression, lower dotted line) of gene *ITGB2*. (b) Survival hazard function $h(t)$ for patients having a low expression (10th percentile of expression, upper continuous line) or a high expression (90th percentile of expression, lower dotted line) of gene *NPR1*.

minus (–) sign indicates that low gene expression was associated with a higher risk. We found that low expression of *ITGB2* ($p = 0.011$) and, to a lesser extent, of *NPR1* ($p = 0.023$) were correlated significantly with prognosis ($p < 0.05$).

The survival hazard function $h(t)$ (see the equation of the Cox model above) for these last two genes are reported in Figure 1. The survival was significantly better for patients having a low expression (10th percentile of expression) versus a high expression (90th percentile of expression) of gene *ITGB2*. The effect was still significant, but much less striking, in the case of *NPR1*.

DISCUSSION

The genetic and pathological heterogeneity of colorectal cancer may be one of the reasons why it has been difficult to find a molecular marker that can serve as a reliable prognostic parameter for this type of tumor, despite the numerous efforts carried out through large-

Table 8. Gene Expression Values (as Ratio) of Genes Associated With Prognosis as Analyzed in Table 3

Patient ID	*ITGB2*	*JAK3*	*KCNK3*	*MRPS11*	*NPR1*	*PHF10*	*PRSS8*	*TXNL2*
C01	1.2177	NA	1.2714	1.306	1.1363	0.9613	1.2127	1.1925
C03	1.4641	0.5555	0.7317	1.2521	1.1243	1.4541	1.396	1.4302
C04	1.7751	0.63	0.7908	1.3301	1.2089	1.4781	1.3323	1.527
C05	1.3717	0.488	0.3395	1.2731	1.3552	1.1074	1.6318	1.7091
C07	1.2263	0.8247	1.0309	0.9626	1.2194	1.1611	1.3345	1.1259
C08	1.1573	0.508	0.4876	0.791	0.8725	0.9844	1.4498	1.2155
C09	1.264	0.9874	0.3338	1.3841	1.1278	1.414	2.6177	1.1414
C11	1.2124	0.8461	0.6436	0.9799	1.0418	1.1847	1.1944	1.248
C12	1.1707	NA	0.6727	0.7447	1.1068	1.268	1.4159	1.2401
C13	1.4049	NA	NA	1.2718	1.3311	1.3153	1.3934	1.1785
C14	1.7128	0.5476	0.9083	1.1147	1.3552	1.3869	1.8024	1.3731
C15	1.6583	1.4546	1.0687	1.1704	1.1	1.2229	1.5406	1.5095
C16	1.2715	NA	0.6548	1.1917	1.2632	1.2708	1.2736	1.3333
C17	1.2197	0.5505	0.8728	1.1854	1.1896	1.241	1.3455	1.4654
C21	1.4487	0.6961	0.5041	1.6015	1.3084	1.4589	1.5518	1.3097
C22	1.1244	NA	NA	1.1224	1.2936	1.2867	1.5372	1.1723
C26	0.9239	1.323	0.9425	0.9667	0.9573	0.784	0.868	1.0304
C27	1.1097	1.071	0.8557	1.306	1.2499	1.2458	1.2459	1.3131
C28	1.0967	0.7006	0.7524	1.1551	1.2091	1.5094	1.1981	1.274

scale functional genomics screenings. Studies of CRC using two-color microarrays or Affymetrix arrays (5–11) have shown that the number of genes with significant changes in expression relative to the normal colon epithelium range from a few hundreds to several thousands. Wang et al. (19), using Affymetrix U133a oligonucleotide microarrays, identified a set of 23 genes as a marker signature of 5-year survival for patients with Dukes B CRC. Eschrich et al. (18) found 43 genes differentially expressed in Dukes B and C, while Arango et al. (20) suggested 17 out of a total of 218 genes as prognostic markers genes out of a total of 218 with a significant difference in expression in Dukes C patients with good prognosis. Very recently Barrier et al. (22,23) analyzed patients with oligonucleotide microarrays and found a cluster of differentially expressed genes correlated with prognosis; Johnston et al. (24) suggested a prognostic signature of 48 genes. Overall, there is limited overlap among the genes identified in different studies. As a consequence, it has been suggested (33) that it may be necessary to test much larger numbers of patients.

We chose to bypass the problems associated with comparing normal tissue to tumor tissue by testing each

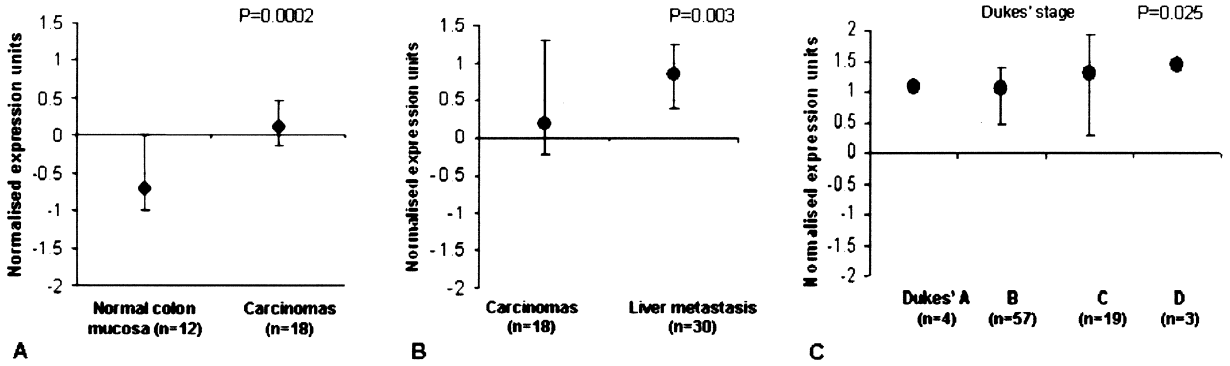

Figure 2. Plot of medians, 90th and 10th percentiles of normalized *ITGB2* expression units (log ratio values, transformed setting the array median as 0, the array standard deviation as 1). (A, B) Data from Graudens; (C) data from Bittner. Adapted from ONCOMINE.

tumor sample versus a pool of RNA from the same samples. Although we expected that the differences in gene expression would be smaller (Table 8), we also expected to increase our chances to identify genes whose expressions correlates with invasive or aggressive tumor characteristics.

First we focused our analysis on a small set of selected samples, homogeneous by stage and histology, but with known different outcomes. In this part of the study we used a panel of three numerical and statistical techniques, aiming to reduce the number of false-positive signals (a common problem in microarray studies). Thus, even if each method produces a long list of genes, partly due to considerable heterogeneity in the expression profiles of the various tumors, we would be able, at least in principle, to single out the genes of more general significance by combining the three methods of analysis.

This approach enabled us to carry out a validation study on just eight genes. We verified the association of the expression of these eight genes with prognosis, by using a Cox survival analysis model on real-time RT-PCR expression data of an independent set of 55 CRC cases. This analysis demonstrated a statistically significant association of survival with low expression of *ITGB2* ($p = 0.011$) and a less strong association with *NPR1* gene expression ($p = 0.023$). A high expression of these two genes was significantly associated with an unfavorable prognosis.

ITGB2 codes for a protein named integrin beta 2 [antigen CD18 or p95; lymphocyte function-associated antigen 1; macrophage antigen 1 (mac-1) beta subunit]. Integrin beta 2 is a cell surface receptor mediating adhesion to the extracellular matrix or to other cells, and connecting the cytoskeleton to various signaling molecules. This protein has been associated with leukocyte adhesion, inflammation, and necrosis, and it is highly expressed in cells belonging to the immune system. As shown in Figure 2, *ITGB2* was reported to be overexpressed in the colon mucosa compared to other normal tissues, in adenocarcinomas versus the normal mucosa, and in liver metastasis versus adenocarcinomas; moreover, *ITGB2* expression was reported to increase progressively from Dukes stage A to D.

NPR1 encodes a natriuretic peptide receptor: in the ONCOMINE database it was reported to be significantly downregulated in serrated versus conventional colon adenocarcinomas.

Pathway analysis has never been used for the study of gene expression variations in metabolic pathways possibly associated with tumor aggressivity. Patients with favorable prognosis appeared to have many activated pathways relative to patients with unfavorable prognosis [i.e., carbon metabolism, transcription, amino acid and nitrogen metabolism, signaling, and fibroblast growth factor receptor (FGFR) pathways]. The meaning of the activation of these pathways is at present difficult to understand; however, FGFR signaling has recently been associated with cancer progression and tumor cell motility (34). It is also interesting to note the only downregulated pathway in the favorable prognosis group was the "dissolution of fibrin cloth pathway," the effect of which on tumor aggressiveness was not previously documented.

In conclusion, we found that in patients with CRC overexpression of *ITGB2* and, to a lesser extent, of *NPR1* were associated with higher risk of succumbing to the disease. The data also show that numerous gene networks associated with specific metabolic pathways are differentially expressed in patients within different prognostic groups. If these data are confirmed, measuring the expression of a few individual genes or variations of gene networks in tumor tissues at the time of surgery may help to predict the prognosis in CRC patients. It will be important to investigate the mechanisms whereby *ITGB2* and *NPR1* and possibly other genes may modulate the invasive properties of CRC.

ACKNOWLEDGMENTS: This study was supported by Istituto Toscano Tumori (ITT) Florence, Italy; IARC Regional Grants, Milan, Italy; Network of Excellence NUGO (Food-CT-2004-506360 NUGO, NOE); grants from the University of Florence, Italy; Cassa di Risparmio di Firenze (2005, to E.M); and from the Ministero dell'Istruzione, dell'Università e della Ricerca, Rome, Italy (2005, to E.M.).

REFERENCES

1. Vogelstein, B.; Fearon, E. R.; Hamilton, S. R.; Kern, S. E.; Preisinger, A. C.; Leppert, M.; Nakamura, Y.; White, R.; Smits, A. M.; Bos, J. L. Genetic alterations during colorectal-tumor development. N. Engl. J. Med. 319:525–532; 1988.
2. Fearon, E. R.; Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell 61:759–767; 1990.
3. Alon, U.; Barkai, N.; Notterman, D. A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96:6745–6750; 1999.
4. Backert, S.; Gelos, M.; Kobalz, U.; Hanski, M. L.; Bohm, C.; Mann, B.; Lovin, N.; Gratchev, A.; Mansmann, U.; Moyer, M. P.; Riecken, E. O.; Hanski, C. Differential gene expression in colon carcinoma cells and tissues detected with a cDNA array. Int. J. Cancer 82:868–874; 1999.
5. Hegde, P.; Qi, R.; Gaspard, R.; Abernathy, K.; Dharap, S.; Earle-Hughes, J.; Gay, C.; Nwokekeh, N. U.; Chen, T.; Saeed, A. I.; Sharov, V.; Lee, N. H.; Timothy, J.; Yeatman, T. J.; Quackenbush, J. Identification of tumor markers in models of human colorectal cancer using a 19.200-element complementary DNA microarray. Cancer Res. 61:7792–7797; 2001.
6. Takemasa, I.; Higuchi, H.; Yamamoto, H.; Sekimoto, M.;

Tomita, N.; Nakamori, S.; Matoba, R.; Monden, M.; Matsubara, K. Construction of preferential cDNA microarray specialized for human colorectal carcinoma: Molecular sketch of colorectal cancer. Biochem. Biophys. Res. Commun. 285:1244–1249; 2001.
7. Williams, N. S.; Gaynor, R. B.; Scoggin, S.; Verma, U.; Gokaslan, T.; Simmang, C.; Fleming, J.; Tavana, D.; Frenkel, E.; Becerra, C. Identification and validation of genes involved in the pathogenesis of colorectal cancer using cDNA microarrays and RNA interference. Clin. Cancer Res. 9:931–946; 2003.
8. Notterman, D. A.; Alon, U.; Sierk, A. J.; Levine, A. J. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res. 61:3124–3130; 2001.
9. Birkenkamp-Demtroder, K.; Christensen, L. L.; Olesen, S. H.; Frederiksen, C. M.; Laiho, P.; Aaltonen, L. A.; Laurberg, S.; Sorensen, F. B.; Hagemann, R.; Orntoft, T. F. Gene expression in colorectal cancer. Cancer Res. 62:4352–4363; 2002.
10. Croner, R. S.; Peters, A.; Brueckl, W. A.; Matzel, K. E.; Klein-Hitpass, L.; Brabletz, T.; Papadopoulos, T.; Hohenberger, W.; Reingruber, B.; Lausen, B. Microarray versus conventional prediction of lymph node metastasis in colorectal carcinoma. Cancer 104:395–404; 2005.
11. Croner, R. S.; Foertsch, T.; Brueckl, W. M.; Guenther, K.; Siebenhaar, R.; Stremmel, C.; Matzel, K. E.; Papadopoulos, T.; Kirchner, T.; Behrens, J.; Klein-Hitpass, L.; Stuerzl, M.; Hohenberger, W.; Reingruber, B. Common denominator genes that distinguish colorectal carcinoma from normal mucosa. Int. J. Colorectal Dis. 20:353–362; 2005.
12. Kitahara, O.; Furukawa, Y.; Tanaka, T.; Kihara, C.; Ono, K.; Yanagawa, R.; Nita, M. E.; Takagi, T.; Nakamura, Y.; Tsunoda, T. Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumor tissues and normal epithelia. Cancer Res. 61:3544–3549; 2001.
13. Lin, Y. M.; Furukawa, Y.; Tsunoda, T.; Yue, C. T.; Yang, K. C.; Nakamura, Y. Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas. Oncogene 21:4120–4128; 2002.
14. Agrawal, D.; Chen, T.; Irby, R.; Quackenbush, J.; Chambers, A. F.; Szabo, M.; Cantor, A.; Coppola, D.; Yeatman, T. J. Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. J. Natl. Cancer Inst. 94:513–521; 2002.
15. Zou, T. T.; Selaru, F. M.; Xu, Y.; Shustova, V.; Yin, J.; Mori, Y.; Shibata, D.; Sato, F.; Wang, S.; Olaru, A.; Deacu, E.; Liu, T. C.; Abraham, J. M.; Meltzer, S. J. Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon. Oncogene 21:4855–4862; 2002.
16. Frederiksen, C. M.; Knudsen, S.; Laurberg, S.; Orntoft, T. F. Classification of Dukes' B and C colorectal cancers using expression arrays. J. Cancer Res. Clin. Oncol. 129: 263–271; 2003.
17. Tureci, O.; Ding, J.; Hilton, H.; Bian, H.; Ohkawa, H.; Braxenthaler, M.; Seitz, G.; Raddrizzani, L.; Friess, H.; Uchle, M.; Sahin, U.; Hammer, J. Computational dissection of tissue contamination for identification of colon cancer-specific expression profiles. FASEB J. 17:376–385; 2003.
18. Eschrich, S.; Yang, I.; Bloom, G.; Kwong, K. J.; Boulware, D.; Cantor, A.; Coppola, D.; Kruhøffer, M.; Aaltonen, L.; Orntoft, T. F.; Quackenbush, J.; Yeatman, T. J. Molecular staging for survival prediction of colorectal cancer patients. J. Clin. Oncol. 23:3526–3535; 2005.
19. Wang, Y.; Jatkoe, T.; Zhang, Y.; Mutch, M. G.; Talantov, D.; Jiang, J.; McLeod, H. L.; Atkins, D. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J. Clin. Oncol. 22:1564–1571; 2004.
20. Arango, D.; Laiho, P.; Kokko, A.; Alhopuro, P.; Sammalkorpi, H.; Salovaara, R.; Nicorici, D.; Hautniemi, S.; Alazzouzi, H.; Mecklin, J. P.; Jarvinen, H.; Hemminki, A.; Astola, J.; Schwarz, S.; Aaltonen, L. Gene-expression profiling predicts recurrence in Dukes' C colorectal cancer. Gastroenterology 129:874–884; 2005.
21. Bertucci, F.; Salas, S.; Eysteries, S.; Nasser, V.; Finetti, P.; Ginestier, C.; Charafe-Jauffret, E.; Loriod, B.; Bachelart, L.; Montfort, J.; Victorero, G.; Viret, F.; Ollendorff, V.; Fert, V.; Giovaninni, M.; Delpero, J-R.; Nguyen, C.; Viens, P.; Monges, G.; Birnbaum, D.; Houlgatte, R. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. Oncogene 23:1377–1391; 2004.
22. Barrier, A.; Boelle, P. Y.; Lemoine, A.; Tse, C.; Brault, D.; Chiappini, F.; Lacaine, F.; Houry, S.; Huguier, M.; Flahault, A.; Dudoit, S. Gene expression profiling of nonneoplastic mucosa may predict clinical outcome of colon cancer patients. Dis. Colon Rectum 48:2238–2248; 2005.
23. Barrier, A.; Boelle, P. Y.; Roser, F.; Gregg, J.; Tse, C.; Brault, D.; Chiappini, F.; Lacaine, F.; Houry, S.; Huguier, M.; Franc, B.; Flahault, A.; Lemoine, A.; Dudoit, S. Stage II colon cancer prognosis prediction by tumor gene expression profiling. J. Clin. Oncol. 24:4685–4691; 2006.
24. Johnston, P. G.; Mulligan, K.; Kai, E.; Black, J.; Moore, S.; Ma Dermott, U.; Wilson, R.; Harkin, D. A genetic signature of relapse in stage II colorectal cancer derived from formalin fixed embedded tissue (FFPA) using a unique disease specific colorectal array. J. Clin. Oncol. 24:3519; 2006.
25. Kaufman, L.; Rousseuw, P. J. Finding groups in data. New York: John Wiley & Sons; 1990.
26. Tusher, V. G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA 98:5116–5121; 2001.
27. Storey, J. D. A direct approach to false discovery rates. J. R. Stat. Soc. B 64:479–498; 2002.
28. Cox, D. R. Regression models and life tables. J. R. Stat. Soc. B 34:187–220; 1972.
29. Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. B 39:1–38; 1977.
30. McLachlan, G.; Peel, P. Finite mixture models. New York: John Wiley & Sons; 2000.
31. Cavalieri, D.; Castagnini, C.; Toti, S.; Maciag, K.; Kelder, T.; Gambineri, L.; Angioli, S.; Dolara, P. Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases. Bioinformatics 23:2631–2632; 2007.
32. Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102(43):15545–15550; 2005.

33. Ein-Dor, L.; Zuk, O.; Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc. Natl. Acad. Sci. USA 103(15): 5923–5928; 2006.
34. Bange, J.; Prechtl, D.; Cheburkin, Y.; Specht, K.; Harbeck, N.; Schmitt, M.; Knyazeva, T.; Müller, S.; Gärtner, S.; Sures, I.; Wang, H.; Imyanitov, E.; Häring, H. U.; Knayzev, P.; Iacobelli, S.; Höfler, H.; Ullrich, A. Cancer progression and tumor cell motility are associated with the FGFR4 Arg(388) allele. Cancer Res. 62(3):840–847; 2002.