Gene Signature Is Associated with Early Stage Rectal Cancer Recurrence

Matthew F Kalady, MD, FACS, FASCRS, Kathryn Dejulius, MS, James M Church, MB, CHB, FACS, FASCRS, Ian C Lavery, MD, FACS, FASCRS, Victor W Fazio, MD, FACS, FASCRS, Hemant Ishwaran, PhD

BACKGROUND: Despite expected excellent outcomes of surgical resection for early stage rectal cancers, 20% of stage I and II rectal cancers recur. Identifying biologic factors that predict the subset prone to recur could allow more directed therapy. This study identifies a tumor gene expression profile that accurately predicts disease recurrence.

STUDY DESIGN: Stage I/II rectal cancer patients treated by surgery alone at a single institution were included and classified as having recurrent or nonrecurrent cancer. Tumor mRNA was isolated from frozen tissue and evaluated for total genome gene expression by microarray analysis. Background-corrected and normalized microarray data were analyzed using BAMarray software. Selected genes were further analyzed using unsupervised clustering and nearest-centroid classification. A balanced K-fold scoring-pair algorithm using 1,000 independent replications was used for gene signature development.

RESULTS: Sixty-nine patients with disease-free survival and 31 patients with recurrent disease were included at a median follow-up of 105 months (interquartile range 114 months) and 32 months (interquartile range 25 months), respectively. Demographics and tumor characteristics between groups were similar. Fifty-two genes from 43,148 probes were differentially expressed, and a 36-gene signature was found to be statistically associated with recurrence using a scoring-pair algorithm. Accuracy to identify recurrence as measured by area under the receiver operating characteristic curve was 0.803.

CONCLUSIONS: Differential gene expression within rectal cancers is associated with recurrence of early stage disease. A 36-gene signature correlates with an increased risk of more or less aggressive tumor behavior. This information obtainable at biopsy may assist in determining treatment decisions.
(J Am Coll Surg 2010;211:187–195. © 2010 by the American College of Surgeons)

Despite clinical advances, rectal cancer remains a significant cause of cancer-related death.[1] Treatment strategies and clinical outcomes are determined by cancer stage as defined by local tumor penetration and spread to lymph nodes or distant organs. Although patients with early stage rectal cancers generally enjoy excellent outcomes with surgery as the sole treatment,[2] advanced tumors have a worse prognosis and are additionally treated with neoadjuvant or adjuvant chemotherapy and/or radiation.[3,4] In spite of established treatment protocols, a significant number of early stage rectal cancer patients still develop recurrent cancer and die from their disease.

Unfortunately, there is no accurate means to predict which patients with early stage disease will suffer recurrence, so there is no way of identifying which patients should be targeted for neoadjuvant treatment. An accurate prognostic model could identify which patients might benefit from neoadjuvant therapy while sparing risks for those who would not benefit. Although various molecular markers involved in colorectal cancer etiology have been identified,[5-12] the process of oncogenesis and cancer metastasis is likely a complex chain of events with multiple intertwined pathways, most of which remain unknown.

The search for new factors has been boosted by development of technologies capable of high-throughput analysis such as microarrays for gene expression.[13-15] Broad molecular and genetic analyses using these techniques have been performed and validated for various tumors including

Disclosure Information: Nothing to disclose.

This work was supported by an American Society of Colon and Rectal Surgeons Career Development Award (MFK).

Presented at the American Society of Colon and Rectal Surgeons, Hollywood, FL, May 2009.

Received January 6, 2010; Revised March 23, 2010; Accepted March 24, 2010.

From the Departments of Colorectal Surgery, Digestive Disease Institute (Kalady, Church, Lavery, Fazio), Cancer Biology (Kalady, DeJulius), and Quantitative Health Sciences (Ishwaran), Cleveland Clinic, Cleveland, OH. Correspondence address: Matthew F Kalady, 9500 Euclid Ave, A30, Cleveland Clinic, Cleveland, OH 44195. email: kaladym@ccf.org

ISSN 1072-7515/10/$36.00
doi:10.1016/j.jamcollsurg.2010.03.035

colorectal cancer.[16-19] This study uses a well-defined rectal cancer population along with microarray technology to develop a gene signature that accurately predicts recurrence or nonrecurrence of early stage rectal cancer.

METHODS

Patient selection and outcomes

The Cleveland Clinic Department of Colorectal Surgery has an IRB-approved database that collects clinical information and follow-up for colorectal cancer patients. This database was queried for patients with stage I or II rectal cancer who were treated by surgery alone. Any patients receiving pre- or postoperative chemotherapy and/or radiation were excluded to avoid the confounding influence on tumor composition and clinical outcomes. The study endpoint was disease recurrence. Only patients with recurrent disease or those without recurrence and at least 3-year follow-up were included. Time to recurrence or disease-free interval was defined as the time from the date of surgery to the date of confirmed tumor relapse for patients with recurrence, and from the date of surgery to the date of last follow-up for disease-free patients. Disease-free survival was defined as being alive without any evidence of recurrent disease as of the latest clinical follow-up. From these groups, patients with available fresh frozen tumor samples comprised the final study population. Charts were reviewed to validate the clinical endpoints of recurrent or nonrecurrent cancer from the database. Basic demographic, clinical, and tumor characteristics were analyzed.

Fresh frozen tissue samples

Tumor tissue was obtained according to Institutional Review Board-approved protocols using frozen tumor specimens from patients treated at the Cleveland Clinic. Tumor tissues were obtained through a dedicated tissue procurement team within the Department of Anatomic Pathology. A portion of the tumor was snap frozen and banked at −80°C. A gastrointestinal pathologist confirmed the histopathology diagnosis of each specimen independently. Specimens chosen for analysis contained at least 60% tumor cells.

RNA isolation from frozen tissue samples

RNA was extracted from fresh frozen tumor tissue. Frozen tissue blocks stored at -80°C were cut on a microtome into 5×10 μm-thick samples and resuspended in 100 μL of tissue lysis buffer, 16 μL 10% sodium dodecyl sulfate (SDS) and 80 μL Proteinase K. Samples were vortexed and incubated in a thermomixer set at 400 revolutions per minute for 3 hours at 55°C. Subsequent steps of sample processing were performed according to manufacturer's protocol. RNA samples were quantified by optical density 260/280 readings using a spectrophotometer and diluted to a final concentration of 50 ng/μL. To assure RNA quality, the mRNA of each specimen was run on a gel to assure lack of degradation before being hybridized for the microarray.

Total genome gene expression analysis

Isolated total genome RNA was tested for total genome expression using >46,000 transcript-specific sequences on the Sentrix Human-6 Expression BeadChip (Illumina). Briefly, 100 ng of total RNA was amplified by an in vitro transcription amplification kit (Ambion) and hybridized to the platform using commercially available kits (Illumina). Illumina BeadStation 500 software was used for imaging and normalization of data.

Statistical analysis

Quantitative variables are summarized by mean ± standard deviation or median with interquartile ranges. Categorical variables are summarized by frequency. Demographic and tumor differences between recurrent and nonrecurrent populations were assessed using chi-square or Fisher's exact test for categorical variables and Wilcoxon rank sum test for quantitative variables. Because recurrence occurred at various follow-up times and not all patients were observed with equal follow-up, factors associated with recurrence were best assessed using the log-rank estimates and Kaplan-Meier analyses for recurrence-free survival.

Microarray statistical analysis

Microarray data (43,148 probes per sample) were background corrected and median baseline normalized using the Beadarray R-software package for Bioconductor.[20] Normalized data were analyzed using Bayesian Analysis of Variance for Microarrays (BAM) methodology.[21,22] Computations were implemented using BAMarray 2.0 software[23] under the no-baseline option assuming unequal variances across cancer (phenotype) groups, with variance clustering[24] set to 2 clusters. To invoke the no-baseline option, normalized data were transformed by baseline centering.[25] For each sample (69 nonrecurrent and 31 recurrent tissues), expression values were subtracted from the corresponding probeset for all patients in the opposing phenotype class. So, each probeset for the 69 nonrecurrent samples had 31 baseline expression values, and each probeset for the 31 recurrent samples had 69 baseline expression values. This resulted in 4,278 observations ($69 \times 31 + 31 \times 69$) for each probeset, and a total of 184,587,144 ($4{,}278 \times 43{,}148$) data values.

Computations were implemented on an Altix 350 Silicon Graphics multiprocessor server. A total of 52 genes

from the 43,148 probes were found to be differentially expressed using the automatic thresholding rule used by BAMarray.[23] Subsequent analyses focused on these 52 genes. The first method used to further condense the gene signature included nearest shrunken centroid classification.[26] Using normalized expression data for the 52 genes, a nearest shrunken centroid classifier was derived. Computations were implemented using the pamr R-software package.[27] Misclassification error rate for the classifier was estimated using 5-fold cross-validation.

A second method to develop the gene signature using a scoring-pair algorithm was derived using control genes. Candidate control genes were defined by removing the 52 differentially expressing genes, as well as all genes with BAM test statistics exceeding a nominal cut-off value, from the 43,148 probes. All Illumina specific probesets that could not be annotated using data from National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) were also removed. This left a total of 2,325 candidate control probe sets.

Balanced K-fold validation (K = 6) was used in the gene signature development. The classifier was trained using 4 folds of the data (training folds). To train the classifier, a control gene from the candidate pool was found for each of the 52 differentially expressing genes. A control gene (Cg) for a gene (g) was defined as that gene with the maximum number of expression values lying between the mean expressions for the 2 phenotype groups for g. The classifier was defined by assigning the value +1 to a gene g if the mean expression for g for the recurrent tissues was larger than Cg, otherwise it was defined to be −1. The overall score for the classifier was the sum total over the 52 genes. The trained classifier was tuned on the fifth-fold (tuning-fold) of the data. This was done by varying the number of genes in the signature. The classifier with highest area under the receiver operator characteristic (ROC) curve was chosen. The accuracy of the tuned classifier was estimated by calculating its area under the ROC curve using the sixth-fold (test-fold) of the data. This process was repeated 1,000 times independently. The final classifier was defined by using only genes that appeared more than 70% of the time with the same +1 or −1 score. This resulted in a gene signature comprising 36 genes and 36 matched control genes ("scoring-pair" signature). Accuracy of this classifier was estimated by the area under the ROC curve using the 1,000 test-fold datasets.

RESULTS

Patient and tumor characteristics

Fresh frozen tumors from 100 patients were available to build the predictive model: 69 rectal cancers from patients with nonrecurrent disease and 31 rectal cancers from patients who subsequently developed recurrence. All cancers were from pathologic early stage, node-negative, rectal cancer patients who were treated by surgical resection alone with curative intent. Eighty patients underwent low anterior resection, 18 underwent abdominoperineal resection, and 2 underwent total proctocolectomy. Patients undergoing local excision were not included in this study. No patients received preoperative chemoradiation or adjuvant treatment before recurrence. Mean follow-ups for nonrecurrent and recurrent patients were 120 and 66.9 months, respectively. The median follow-up for nonrecurrent patients was 104.6 months, with 25th, 50th, and 75th percentiles of 58.3, 104.6, 172.7, respectively (interquartile range 114.4 months). The mean time to recurrence was 37.1 months. There were 24 patients with distant recurrence, 6 patients with local recurrence, and 1 patient with both distant and local recurrence.

Demographics and tumor characteristics are shown in Table 1. Patients with nonrecurrent cancer had higher mean and median numbers of lymph nodes evaluated than those with recurrent disease: 23 versus 16, and 20 versus

Table 1. Patient Demographics and Tumor Characteristics

Variable	Nonrecurrent	Recurrent	p Value
n	69	31	
Mean age (y), SD	64.5 ± 11	65.6 ± 9	0.79
Gender, male/female	49/20	19/12	0.34
Time to recurrence (mo), mean ±SD	NA	37 ± 26	NA
Median follow-up, mo (IQR)	105 (114)	32 (25)	<0.001
Distance from anal verge (cm), mean ±SD	9.3 ± 3.6	8.7 ± 4.0	0.47
Tumor size (cm), mean ±SD	4.3 ± 1.2	4.7 ± 1.8	0.27
Lymph nodes examined (n), mean ± SD	23 ± 23	16 ± 11	0.04
Cancer stage, n (%)			0.03
I	43 (62)	12 (39)	
II	26 (38)	19 (61)	
T stage, n (%)			0.04
T1	8 (12)	0 (0)	
T2	35 (51)	13 (42)	
T3	26 (38)	17 (55)	
T4	0 (0)	1 (3)	
Differentiation, n (%)			0.34
Well	9 (13)	1 (3)	
Moderately	52 (75)	27 (87)	
Poorly	8 (12)	3 (10)	

IQR, interquartile range.

12, respectively ($p = 0.04$, Wilcoxon rank sum test). There were 2 significant outliers in terms of number of lymph nodes evaluated. One patient with nonrecurrent disease had 180 lymph nodes evaluated. One patient with recurrent disease had no lymph nodes evaluated. This case was re-reviewed by Pathology at the time of resection and still no lymph nodes were found. If these 2 outliers are removed from the data purely to evaluate the number of nodes examined, the difference is no longer significant ($p = 0.10$). Regardless, evaluation of at least 12 lymph nodes has been shown to be accurate for staging rectal cancer[28] and both groups met this requirement. In addition, a log-rank estimate was performed to evaluate the influence of lymph node harvest on disease-free survival in our study population using 12 as the cut-off for number of nodes evaluated. There was no significant difference in recurrence-free survival ($p = 0.23$).

Not unexpectedly, there was a higher percentage of stage II rectal cancers among the group of patients who developed recurrence. However, this did not statistically influence recurrence-free survival in this study population based on the log-rank test ($p = 0.53$).

Gene expression

A total of 52 genes from the 43,148 probes were found to be differentially expressed using the automatic thresholding rule used by BAMarray.[23] Unsupervised hierarchical clustering of BAM test statistics for the 52 differentially expressing genes identified 2 distinct populations corresponding to patients with nonrecurrent or recurrent rectal cancer (Fig. 1). BAM test statistics (1 statistic for each gene and each sample) measured distance for a patient's gene expression to the gene expression for all other patients from the alternate phenotype. The large clustering patches of green and red in Figure 1 showed consensus among the 52 genes in delineating cancer outcome status.

Nearest shrunken centroid gene signature

In an attempt to further condense the 52 differentially expressed genes, nearest shrunken centroid classification was used.[26] Error rates using 5-fold validation for nearest centroid classification were flat as a function of threshold-shrinkage value (Supplemental Fig. 1, online only). This demonstrated that the 52 nearest centroid gene signature could not be improved by removing genes. Centroids for recurrent and nonrecurrent data were relatively large for all genes (Supplemental Fig. 2, online only). Five-fold validation error rate for the classifier was 29% (Supplemental Fig. 1, online only). Error rates were significantly lower for nonrecurrent data. Predicted class probabilities also showed that the nearest centroid classifier was better over nonrecurrent data (Supplemental Fig. 3, online only).

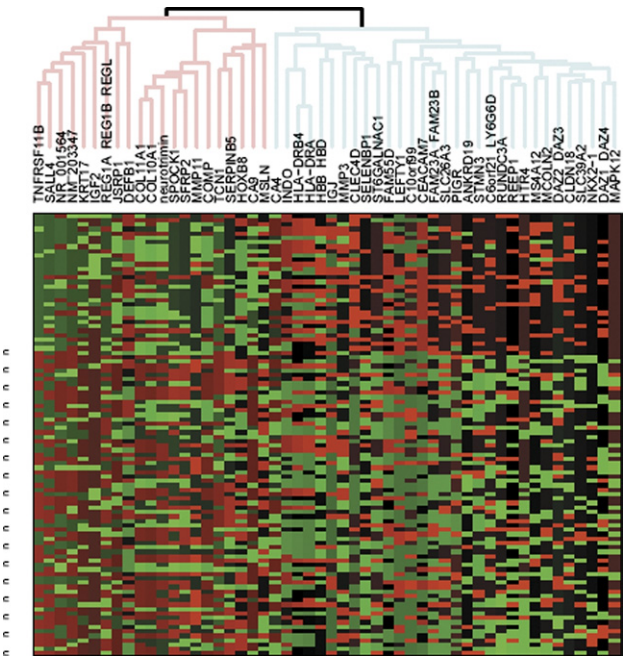

Figure 1. Unsupervised hierarchical clustering of genes and samples using 52 genes found differentially expressing from a Bayesian Analysis of Variance for Microarrays (BAM) analysis. Heatmap values are BAM test statistics measuring distance of a given patient to all patients in the opposing class. Rows correspond to patients, columns to genes. Nonrecurrent patients are indicated by an "n" to the left of the rows (some values obscured due to resolution). Large clustering patches of green and red delineate 2 distinct populations corresponding to patients with nonrecurrent or recurrent rectal cancer.

Scoring-pair gene signature

Using an alternative method, a scoring-pair algorithm, a 36-gene signature was derived. Hierarchical unsupervised clustering of the scoring-pair 36-gene signature identified 2 subsets of genes that delineated between recurrent and nonrecurrent cancer groups (Fig. 2). A complete gene list is shown in Table 2. Twenty-two genes were underexpressed and 14 were overexpressed in recurrent rectal cancers in relation to nonrecurrent rectal cancers. The direction of expression is shown in Table 2.

Overall accuracy of the 36 scoring-pair signature, as measured by area under the ROC curve, was 0.803 (Fig. 3). This was significantly better than the 52 nearest centroid gene signature. Distribution of the scores for the 36-gene signature over the 1,000 testfold datasets showed good separation between recurrent and nonrecurrent data (Fig. 4).

DISCUSSION

This study introduces a novel means to identify early stage rectal cancer patients who are at risk for recurrence by using

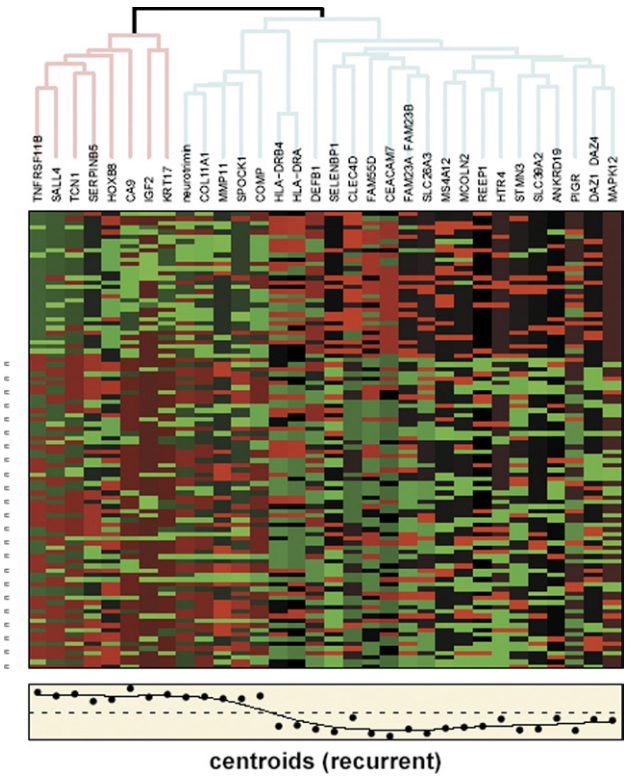

Figure 2. Hierarchical clustering of the 36-gene scoring pair signature. The pink and blue dendritograms represent clustering of recurrent and nonrecurrent genes, respectively. Rows correspond to patients and columns correspond to genes. Nonrecurrent patients are denoted by an "n" next to the row. Clustering was performed on genes (columns) but not samples (rows). Again, distinct patterns between cancer outcomes are seen. The bottom panel displays centroid values for recurrent patients from the nearest shrunken centroid classifier. Note that only 32 of the 36 genes are shown here. Four genes with Illumina-specific probesets were discontinued on the National Center for Biotechnology Information bank and these were not included.

a predictive gene signature that can be assessed by tumor tissue analysis at the time of diagnosis or resection. The model was built on a well-characterized patient population treated by total mesorectal excision and used robust statistical methods that yield an overall accuracy of 80%.

Despite best practices, the staging system and associated treatment protocols for rectal cancer still remain flawed. Although early stage cancers are treated by surgery alone, distal disease recurrence still occurs in approximately 20% of cases. Identifying this subset of patients could allow for an opportunity to intervene with neoadjuvant or adjuvant therapy. For example, patients with stage I rectal cancer with a high probability of recurrence based on the gene expression profile could theoretically be offered neoadjuvant or adjuvant chemotherapy. The signature may also assist in a more directed therapy for stage II tumors. Current standards dictate neoadjuvant chemoradiation for stage II or III adenocarcinoma in the middle or lower third of the rectum.[29] Patients with clinical or pathologic lymph node involvement usually also receive subsequent further adjuvant chemotherapy. However, the added benefit for adjuvant therapy for all stage II patients receiving neoadjuvant chemoradiation is controversial. Patients whose tumors fit the gene signature for recurrence could potentially benefit from adjuvant treatments. Conversely, patients with stage II rectal cancer whose gene profile is consistent with nonrecurrent disease could possibly forgo neoadjuvant chemoradiation because total mesorectal surgery alone provides long-term cure in the majority of cases.[2,30,31] So, directed selection of patients to receive neoadjuvant or adjuvant therapy could minimize unnecessary treatment.

One unique aspect of this signature is the paired-scoring concept. Unsupervised approaches (eg, hierarchical clustering) tend to find statistical separation in outcomes unrelated to biology; semisupervised approaches (eg, nearest-centroid classification) find valid biologic class separation but tend to be accurate over only select phenotype groups. Unlike these approaches, we sought to use a strongly supervised approach that separated both outcome groups equally well, both statistically and biologically. To do so, we used a scoring-pair approach that allowed each differentially expressed gene to be scored individually as -1, $+1$, or zero, relative to a control gene. A non-zero value was assigned only when one could find a control gene whose expression value consistently lay between expression values for the 2 outcome groups, ensuring that differential expression was not only statistically significant but also biologically consistent. Each of these values for the 36 genes was summed to derive a score for that particular patient. So, any one patient can have a score from -36 to $+36$ that is associated with a particular risk. The histogram of these scores is shown in Figure 4. Although there is overlap in the middle range of these scores, the model is particularly useful for patients scoring at the extremes of the scale, where there is a nearly 100% chance of either recurrence or nonrecurrence of disease.

This work is the first to report a gene signature to predict recurrence of rectal cancer treated by surgery alone. The Memorial Sloan Kettering group studied predictors of recurrence for early stage rectal adenocarcinoma by evaluating tissue microarray expression of several known molecular markers. However, the given set of known markers did not accurately correlate with recurrence,[32] underscoring the need for better ways of predicting outcomes. Other groups have used tumor microarray platforms to predict response to preoperative radiation in Japanese rectal cancer patients[33] and response to neoadjuvant chemoradiation in

Table 2. Genes Included in the 36-Gene Predictive Signature

Gen Bank ID	Gene name	Expression in recurrent	Description/annotation
NM_000111	SLC26A3	Decreased	Chloride anion exchanger (Protein DRA)
NM_006890	CEACAM7	Decreased	Carcinoembryonic antigen-related cell adhesion molecule 7 precursor (Carcinoembryonic antigen CGM2)
NM_001216	CA9	Increased	Carbonic anhydrase 9 precursor (EC 4.2.1.1) (Carbonic anhydrase IX) (Carbonate dehydratase IX) (CA-IX) (CAIX) (Membrane antigen MN) (P54/58N) (Renal cell carcinoma-associated antigen G250) (RCC- associated antigen G250) (pMW1).
NM_002644	PIGR	Decreased	Polymeric-immunoglobulin receptor precursor (Poly-Ig receptor) (PIGR) (Hepatocellular carcinoma-associated protein TB6)
NM_017678	FAM55D	Decreased	Protein FAM55D precursor
NM_000612	IGF2	Increased	Insulin-like growth factor II precursor (IGF-II) (Somatomedin A)
NM_001062	TCN1	Increased	Transcobalamin-1 precursor (Transcobalamin I)
XM_940969	LOC651751	Decreased	
NM_000422	KRT17	Increased	Keratin
NM_000095	COMP	Increased	Cartilage oligomeric matrix protein precursor (COMP)
NM_003944	SELENBP1	Decreased	Selenium-binding protein 1 (56 kDa selenium-binding protein) (SP56)
NM_015894	STMN3	Decreased	Stathmin-3 (SCG10-like protein)
XM_939003	LOC649923	Decreased	
NM_005218	DEFB1	Decreased	Beta-defensin 1 precursor (BD-1) (Defensin)
NM_020420	DAZ1 DAZ4	Decreased	Deleted in azoospermia protein 4
NM_021983	HLA-DRB4	Decreased	
NM_002546	TNFRSF11B	Increased	Tumor necrosis factor receptor superfamily member 11B precursor (Osteoprotegerin) (Osteoclastogenesis inhibitory factor).
NM_014579	SLC39A2	Decreased	Zinc transporter ZIP2 (Eti-1) (6A1) (hZIP2) (Solute carrier family 39 member 2)
NM_080629	COL11A1	Increased	Collagen alpha-1(XI) chain precursor
NM_019111	HLA-DRA	Decreased	major histocompatibility complex
NM_024016	HOXB8	Increased	Homeobox protein Hox-B8 (Hox-2D) (Hox-2.4)
NM_001013629	FAM23	Decreased	Protein FAM23A, Protein FAM23B
NM_005940	MMP11	Increased	Stromelysin-3 precursor (EC 3.4.24.-) (ST3) (SL-3) (Matrix metalloproteinase-11) (MMP-11)
NM_002969	MAPK12	Decreased	Mitogen-activated protein kinase 12 (EC 2.7.11.24) (Extracellular signal-regulated kinase 6) (ERK-6) (ERK5) (Stress-activated protein kinase 3) (Mitogen-activated protein kinase p38 gamma) (MAP kinase p38 gamma)
NM_020436	SALL4	Increased	Sal-like protein 4 (Zinc finger protein SALL4)
NM_153259	MCOLN2	Decreased	Mucolipin-2
NM_017716	MS4A12	Decreased	Membrane-spanning 4-domains subfamily A member 12
NM_004598	SPOCK1	Increased	Testican-1 precursor (Protein SPOCK)
XM_941444	LOC652113	Decreased	
NM_022912	REEP1	Decreased	Receptor expression-enhancing protein 1
NM_016522	HNT	Increased	Neurotrimin precursor (hNT)
XM_945536	LOC652470	Increased	
NM_002639	SERPINB5	Increased	Serpin B5 precursor (Maspin) (Protease inhibitor 5)
NM_080387	CLEC4D	Decreased	C-type lectin domain family 4 member D (C-type lectin superfamily member 8) (C-type lectin-like receptor 6) (CLEC-6)
NM_000870	HTR4	Decreased	5-hydroxytryptamine 4 receptor (5-HT-4) (Serotonin receptor 4) (5-HT4)
NM_001010925	ANKRD19	Decreased	

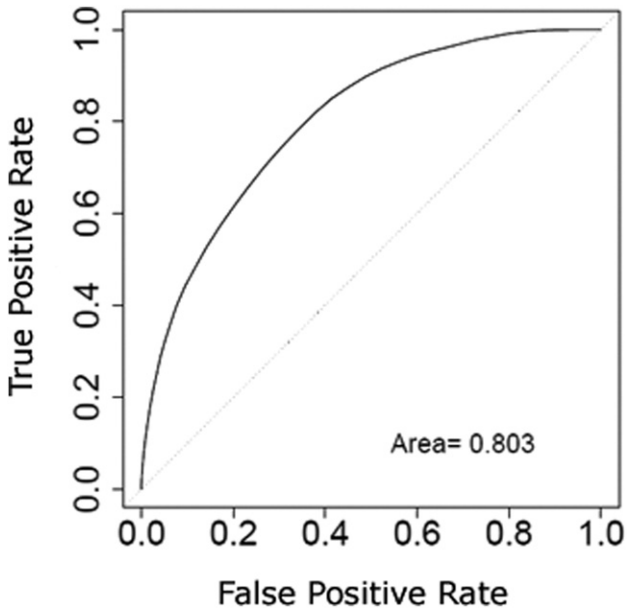

Figure 3. Receiver operator characteristic (ROC) curve for tuned scoring-pair classifier. ROC curve was calculated using test-fold data from K-fold validation. Data are based on 1,000 independent replications.

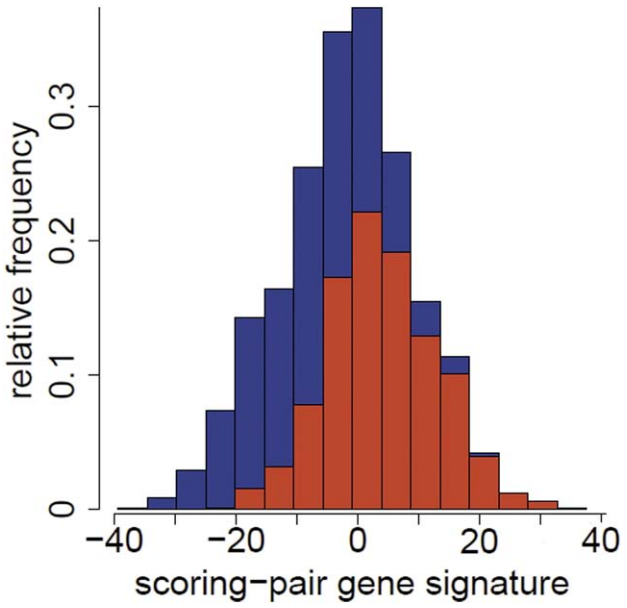

Figure 4. Histogram for individual patient score values for the 36-gene signature. Each of the 36 genes in the signature was compared with a control gene and given a score of −1, 0, or +1 depending on decreased, similar, or increased expression, respectively. The 36 values for each patient were summed, allowing a possible range of scores from -36 to +36. Red bars represent patients who developed recurrent disease and blue bars represent patients without recurrent disease. Data are based on test-fold data. Scores on the extremes of this histogram are highly predictive for either recurrence or nonrecurrence.

the German Rectal Cancer Trial.[34] Gene profiles in these studies are influenced by treatment interventions and cannot be extrapolated for patients treated by surgery alone. Gene signatures have been developed for predicting recurrence of colon cancers.[35,36] One group reported on 70 genes associated with recurrence of stage II and III colon cancers.[36] None of the genes overlapped with those reported in this study, but similar genes within a common family such as insulin-like growth factor and tumor necrosis factor were found to be increased in recurrent patients in both Barrier's work and this study. A multi-institutional group reported a 7-gene signature to predict stage II colon cancer recurrence.[35] Similarly, there was no overlap with the current study gene signature. These findings are not surprising due to genetic and molecular differences between colon and rectal cancers,[37] heterogeneity of patients used between the studies, and differences in methodology in developing the signatures.

Microarray technology is increasingly being used to identify and define genes associated with subclasses of disease. However, there are challenges to building an accurate gene signature model. Hurdles that impede the success of a clinically applicable signature include the lack of uniformity of patients studied, difficulty with quality procurement specimens, and lack of good clinical follow-up.[38] These potential pitfalls have been addressed in the design of this study. We included a stringently defined patient population of pathologically defined stage I and II rectal cancer patients who were treated by formal resection (no local excision). Patients receiving neoadjuvant therapy were excluded to avoid the influence of medical treatment on gene expression or recurrence. All operations were done with curative intent and performed by colorectal surgeons at a tertiary care center using techniques of vessel high ligation and sharp total mesorectal excision. So we have tried to we create a model for which tumor biology trumps any surgical factors on cancer recurrence. Although this study included 6 patients whose recurrence was local, adequate lymph node harvest, intact mesorectum on the surgical specimen, and clear margins minimize the impact of technique over biology on tumor recurrence. Each tumor sample was freshly frozen from the operating room and procured for tissue banking. Detailed clinical databases allowed for close clinical follow-up and the median follow-up for nonrecurrent patients was 8.7 years.

One may argue that excluding patients who received neoadjuvant or adjuvant therapy to build the model creates selection bias because theoretically, stage II patients who are considered higher risk based on clinical evaluation might receive additional therapy, so the signature would not be broadly applicable. However, reasons for not giving

neoadjuvant treatment to this group included the time period when the patient was treated, surgeon prevailing attitudes regarding the benefit of additional therapy, and patient comorbidities. Regardless of this, the signature is still associated with a group of patients who develop recurrence, and even more so in patients who would not be expected to have recurrence. A prospective signature validation study will eliminate any such bias.

In addition to identifying a gene signature, this work has identified individual genes that may be important to understanding the biologic process of cancer. Not all genes have a known biologic process nor are they linked to cancer. However, multiple genes in the signature are involved with cell adhesion and signaling, cellular proliferation, angiogenesis, and apoptosis, among others. A detailed discussion about each of the individual genes and their potential significance is beyond the scope of this article, but several of these genes are worth mentioning briefly.

Downregulation of CEA cellular adhesion molecule-7 (CEACAM-7), a regulator of normal cellular differentiation, has been demonstrated in aberrant crypt foci and adenomas.[39,40] Large decreases, as seen in this signature, could be associated with more aggressive disease. Loss of selenium-binding protein 1 (SELENBP1), which is decreased in our recurrent patients, is associated with a worse overall prognosis for stage II and III colorectal cancer patients.[41] Expression of collagen matrix protein COL11A1 is seen in adenomas and sporadic colon cancers, but not normal colonic epithelium.[42,43] This gene is overly expressed in recurrent patients compared with nonrecurrent patients in this study. Matrix metalloproteinase 11 (MMP11), a protein instrumental in degradation of extracellular matrix and whose increased expression portends metastatic disease,[44] is elevated in our signature for recurrent disease patients, as might be expected. Each of these genes is under further laboratory investigation.

The authors acknowledge that this is an early model that needs validation before serious discussion about it being used as a clinical tool. Expression of each gene identified as significant in the microarray will be validated separately using quantitative real-time polymerase chain reaction techniques, which are relatively inexpensive and readily available. Genes that pass this test can then be used to evaluate prognosis in an independent set of rectal cancers. Data gained from that validation could provide information to be used in consultation with patients in discussing clinical algorithms. Ideally, successful validation could provide the basis for a potential prospective randomized clinical trial regarding the use adjuvant therapy to prevent disease recurrence.

Appendix

Supplementary data

Supplementary data associated with this article can be found ,in the online version, at doi:10.1016/j.jamcollsurg.2010.03.035.

Author Contributions

Study conception and design: Kalady, Church, Lavery, Fazio, Ishwaran

Acquisition of data: DeJulius, Kalady, Ishwaran

Analysis and interpretation of data: Kalady, DeJulius, Ishwaran

Drafting of manuscript: Kalady, Ishwaran

Critical revision: Kalady, DeJulius, Church, Lavery, Fazio, Ishwaran

Acknowledgment: The authors recognize the following for their important contributions to this study: Pieter Faber, PhD, at the Cleveland Clinic Genomics Core Facility for his work in preparing the microarray assays; Dr Marek Skacel from Dahl-Chase Pathology Associates in Bangor, ME, for his confirmation of histologic diagnosis on rectal cancer specimens; Elena Manilich and Jeff Hammel at the Cleveland Clinic for their assistance with the clinical database.

REFERENCES

1. Parkin DM, Bray F, Ferlay J, et al. Global cancer statistics, 2002. CA: Cancer J Clin 2005;55:74–108.
2. Lavery IC, Lopez-Kostner F, Fazio VW, et al. Chances of cure are not compromised with sphincter-saving procedures for cancer of the lower third of the rectum. Surgery 1997;122:779–784; discussion 784–785.
3. Steele G Jr. Adjuvant therapy for patients with colorectal cancer. World J Surg 1995;19:241–245.
4. Holen KD, Saltz LB. New therapies, new directions: advances in the systemic treatment of metastatic colorectal cancer. Lancet Oncol 2001;2:290–297.
5. Robbins DH, Itzkowitz SH. The molecular and genetic basis of colon cancer. Medical Clin North Am 2002;86:1467–1495.
6. Haydon AM, Jass JR. Emerging pathways in colorectal-cancer development. Lancet Oncol 2002;3:83–88.
7. Gill S, Lindor NM, Burgart LJ, et al. Isolated loss of PMS2 expression in colorectal cancers: frequency, patient age, and familial aggregation. Clin Cancer Res 2005;11:6466–6471.
8. Hardingham JE, Butler WJ, Roder D, et al. Somatic mutations, acetylator status, and prognosis in colorectal cancer. Gut 1998; 42:669–672.
9. Ferraz JM, Zinzindohoue F, Lecomte T, et al. Impact of GSTT1, GSTM1, GSTP1 and NAT2 genotypes on KRAS2 and TP53 gene mutations in colorectal cancer. Int J Cancer 2004;110: 183–187.
10. Russo A, Bazan V, Iacopetta B, et al. The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type

of mutation, and adjuvant treatment. J Clin Oncol 2005;23:7518–7528.
11. Deng G, Bell I, Crawley S, et al. BRAF mutation is frequently present in sporadic colorectal cancer with methylated hMLH1, but not in hereditary nonpolyposis colorectal cancer. Clin Cancer Res 2004;10:191–195.
12. Lievre A, Bachet JB, Le Corre D, et al. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res 2006;66:3992–3995.
13. Thomas DC, Haile RW, Duggan D. Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 2005;77:337–345.
14. Jones C, Simpson P, Mackay A, et al. Expression profiling using cDNA microarrays. Methods Mol Med 2006;120:403–414.
15. Duggan DJ, Bittner M, Chen Y, et al. Expression profiling using cDNA microarrays. Nat Genet 1999;21(1 Suppl):10–14.
16. Potti A, Mukherjee S, Petersen R, et al. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. N Engl J Med 2006;355:570–580.
17. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 2002;346:1937–1947.
18. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–536.
19. Wang Y, Jatkoe T, Zhang Y, et al. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J Clin Oncol 2004;22:1564–1571.
20. Dunning M, Smith M, Camilier I, et al. Beadarray: Quality control and low-level analysis of BeadArrays. R package version 1.4.0. 2007.
21. Ishwaran H, Rao JS. Detecting differentially expressed genes in microarrays using Bayesian model selection. J Amer Stat Assoc 2003;98:438–455.
22. Ishwaran H, Rao J. Spike and slab gene selection for multigroup microarray data. J Amer Stat Assoc 2005;100:764–780.
23. Ishwaran H, Rao JS, Kogalur UB. BAMarraytrade mark: Java software for Bayesian analysis of variance for microarray data. BMC Bioinformatics 2006;7:59.
24. Papana A, Ishwaran H, Papana A, et al. CART variance stabilization and regularization for high-throughput genomic data. Bioinformatics 2006;22:2254–2261.
25. Ishwaran H, Rao JS. Clustering gene expression profile data by selective shrinkage. Stat Prob Letters 2008;78:1490–1497.
26. Tibshirani R, Hastie T, Narasimhan B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002;99:6567–6572.
27. Hastie T, Tibshirani R, Narasimhan N, et al. pamr: Pam: prediction analysis for microarrays. R package version 1.31. 2008.
28. Murphy J, Pocard M, Jass JR, et al. Number and size of lymph nodes recovered from Dukes B rectal cancers: correlation with prognosis and histologic antitumor immune response. Dis Colon Rectum 2007;50:1526–1534.
29. Sauer R, Becker H, Hohenberger W, et al. Preoperative versus postoperative chemoradiotherapy for rectal cancer. N Engl J Med 2004;351:1731–1740.
30. Improved survival with preoperative radiotherapy in resectable rectal cancer. Swedish Rectal Cancer Trial. N Engl J Med 1997; 336:980–987.
31. Folkesson J, Birgisson H, Pahlman L, et al. Swedish Rectal Cancer Trial: long lasting benefits from radiotherapy on survival and local recurrence rate. J Clin Oncol 2005;23:5644–5650.
32. Hoos A, Nissan A, Stojadinovic A, et al. Tissue microarray molecular profiling of early, node-negative adenocarcinoma of the rectum: a comprehensive analysis. Clin Cancer Res 2002;8: 3841–3849.
33. Kawakami K, Ruszkiewicz A, Bennett G, et al. DNA hypermethylation in the normal colonic mucosa of patients with colorectal cancer. Br J Cancer 2006;94:593–598.
34. Ghadimi BM, Grade M, Difilippantonio MJ, et al. Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy. J Clin Oncol 2005;23:1826–1838.
35. Jiang Y, Casey G, Lavery IC, et al. Development of a clinically feasible molecular assay to predict recurrence of stage II colon cancer. J Mol Diagn 2008;10:346–354.
36. Barrier A, Lemoine A, Boelle PY, et al. Colon cancer prognosis prediction by gene expression profiling. Oncogene 2005;24: 6155–6164.
37. Kalady MF, Sanchez JA, Manilich E, et al. Divergent oncogenic changes influence survival differences between colon and rectal adenocarcinomas. Dis Colon Rectum 2009;52:1039–1045.
38. Quackenbush J. Microarray analysis and tumor classification. N Engl J Med 2006;354:2463–2472.
39. Thompson J, Zimmermann W, Nollau P, et al. CGM2, a member of the carcinoembryonic antigen gene family is downregulated in colorectal carcinomas. J Biol Chem 1994;269: 32924–32931.
40. Scholzel S, Zimmermann W, Schwarzkopf G, et al. Carcinoembryonic antigen family members CEACAM6 and CEACAM7 are differentially expressed in normal tissues and oppositely deregulated in hyperplastic colorectal polyps and early adenomas. Am J Pathol 2000;156:595–605.
41. Kim H, Kang HJ, You KT, et al. Suppression of human selenium-binding protein 1 is a late event in colorectal carcinogenesis and is associated with poor survival. Proteomics 2006;6: 3466–3476.
42. Fischer H, Stenling R, Rubio C, et al. Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. Carcinogenesis 2001;22:875–878.
43. Fischer H, Salahshor S, Stenling R, et al. COL11A1 in FAP polyps and in sporadic colorectal tumors. BMC Cancer 2001; 1:17.
44. Asano T, Tada M, Cheng S, et al. Prognostic values of matrix metalloproteinase family expression in human colorectal carcinoma. J Surg Res 2008;146:32–42.