Supporting Information

Garman *et al.* 10.1073/pnas.0806674105

SI Methods

Development of Metagene Predictor. From publicly available gene expression data collections, all early stage colon cancer patients (stages I and II) with known survival outcomes were identified, which constituted the initial training data set ($n = 52$) for the development of a genomic predictor of disease recurrence. Two data sets were used for validation: an independent data set of 55 patients with stage I/II disease (1) and a plasmode data set ($n = 73$, GSE10402) representing consecutive patients with early stage colon cancer treated at the University Medical Center Göttingen, Germany, which was used to independently validate the 50-gene predictor in a blinded manner.

Before statistical modeling, data sets with appropriate clinical data were chosen for the training set. These include GSE5206 ($n = 100$) and GSE2138 ($n = 20$). Only early stage (stages I and II) patient samples were identified and isolated from these two data sets. These subsets of patients from both data sets were later merged using the DWD method. The entire merged data set was rearranged based on patient recurrence score, and appropriate class labels were assigned a zero for all of the patients with no recurrence ($n = 45$) and a one for all of the patients with recurrence ($n = 7$) to create an initial training set that represents two distinct biological states. This training set was further filtered based on P values obtained from multiple t tests, and 91 specific genes with significant survival effects ($P < 0.001$) were selected for Bayesian binary regression analyses. Bayesian fitting of binary probit regression models to the training data then permits an assessment of the relevance of the metagene signatures in within-sample classification, and estimation and uncertainty assessments for the binary regression weights mapping metagenes to probabilities. To guard against over-fitting given the disproportionate number of variables to samples, leave-one-out cross validation analysis was performed to test the stability and predictive capability of the model. Finally, a metagene predictor consisting of 50 genes was developed using the aforementioned methodologies. To understand the full meaning of the biology captured in the 50 genes, within the context of the entire biological system, GATHER (http://gather.genome.duke.edu/) was used. GATHER is a tool that integrates various forms of available metadata to elucidate the biological context within molecular signatures produced from high-throughput data. GATHER also has the capacity to discover novel functions of gene groups by integrating annotations from evolutionary homologs and other genes related through protein interactions or literature networks. GATHER further annotates the characteristics of the genes with respect to data sets from multiple systems, helping synthesize evidence to develop or reinforce hypotheses. Finally, the accuracy with which GATHER can infer novel functions of signatures is interpreted through a Bayesian statistical model.

Validating the 50-Gene Classifier. First, an optimal threshold recurrence score value of 0.76 was chosen based on a receiver operated characteristic (ROC) analysis, and was used as the predefined 'cut-point', to dichotomize samples into low risk (Recurrence Score <0.76) and high risk (Recurrence Score ≥ 0.76).

Given a training set of expression vectors (of values across metagenes) representing two biological states (in this case, patient samples with recurrence and with no recurrence), a binary probit regression model is estimated using Bayesian methods. Before applying these methods, the initial training set and the validation sets [E-MEXP-1224 (ref. 1) and GSE10402 (ref. 2)] were normalized using the DWD method. Standard Kaplan-Meier mortality curves and their statistical significance were generated from the predictive probability values of patients using GraphPad software. For the Kaplan–Meier survival analyses, the survival curves were compared using the log-rank test. This test generates a two-tailed P value testing the null hypothesis that the survival curves are identical in the overall populations. Therefore, the null hypothesis is that the populations have no differences in survival.

Univariate and Multivariate Analysis. In an effort to fully understand the prognostic significance of the 50-gene predictor for colon cancer recurrence, univariate and multivariate analyses were performed using Cox proportional hazard models. As seen in Table 3, only factors that were significant in a univariate analyses were used in the multivariate models. Analysis included continuous covariates for age, and dichotomous covariates for gender, stage of disease, and a prediction of recurrence (based on the 50-gene predictor). No adjustment for multiple testing was necessary. Hazard ratios and 95% confidence intervals are reported. P values are based on likelihood ratio tests, and analyses are performed using the statistical package R (3).

Cross-Platform Comparison. An in-house program that has been previously validated (4), *Chip Comparer,* was used to map probe sets (http://tenero.duhs.duke.edu/genearray/perl/chip/chipcomparer.pl) across various generations and platforms of Affymetrix GeneChip (http://www.affymetrix.com) and spotted arrays. Also, where needed, to reduce the likelihood of batch effects, a normalizing algorithm, ComBat (http://statistics.byu.edu/johnson/ComBat/) was applied (5). Because several different microarray platforms were used in the data sets (HG-U133A, HG-U133 two plus, and HG-U95Av2), the probe sets should be matched to the identical genes. Each probe set ID in given Affymetrix gene chips were first mapped to the corresponding LocusID by parsing local copies of LocusLink and UniGene databases to identify any inherent relationship between the GenBank accession number associated with each probe set sequence and its corresponding LocusID. This was followed by matching probe sets from different gene chips that share the same LocusID.

ComBat Method. When combining data sets from different platforms and different experiments, non-biological experimental variation or "batch effects" are most commonly faced by researchers. It is inappropriate to combine data sets without adjusting for batch effects. To reduce the systematic differences from different data sets and integrate gene expression from all data sets, ComBat method (http://statistics.byu.edu/johnson/ComBat/) was applied. ComBat method applies either parametric or non-parametric empirical Bayes framework for adjusting data for batch effects that is robust to outliers in a given data set. The location (mean) and scale (variance) model parameters are specifically estimated by pooling information across genes in each batch to shrink the batch effect parameter estimated toward the overall mean of the batch effect estimates. This method was applied to data sets consisting of normal colon samples and tumor samples separately.

Colon Cancer Cell Lines. Fourteen colon cancer cell lines (COLO-320 HSR, DLD-1, HCT115, HCT116, HT29, LS174T, LS180,

RKO, SW48, SW403, SW1116, SW1417, SW1463, and WiDr) that were commercially available were grown as recommended by the supplier (American Type Culture Collection). Culture media RPMI 1640 was used for COLO-320 HSR, HCT 15 and DLD-1; Leibovitz-15 (L-15) was used for SW1417, SW48, SW1116, SW403 and SW1463; Modified Essential Eagle Medium was used for RKO, LS174T, LS180, and WiDr; McCoy 5A was used for HT-29 and HCT-1116. All tissue culture media and were obtained from Sigma–Aldrich and was supplemented with 10% Fetal Bovine Serum (FBS). For the drug-sensitivity assays, celecoxib was obtained from LKT Laboratories Inc., the LY294002 was obtained from Cayman Chemical, and the retinol, and sulindac were obtained from Sigma–Aldrich. 5-FU and oxaliplatin were obtained from the Duke University pharmacy.

We hypothesized that cell lines with a high Recurrence Score would be more sensitive to treatment with targeted therapy. We also predicted that treatment would reverse the high risk phenotype in gene expression analysis. To classify cell lines, we measured genome-wide expression in the 14 colon cancer cell lines using the Affymetrix U133A Plus 2.0 GeneChip. Total RNA was extracted from the cells with RNeasy kits (Qiagen). The RNA quality was assessed with the use of a bioanalyzer (Agilent 2100 model). Hybridization targets were prepared from the total RNA according to standard Affymetrix protocols. 'Recurrence Scores' were generated for the cell lines and the predefined (from the training set) threshold value was then used to dichotomize the cell lines into low and high risk phenotypes. *In vitro* cell proliferation assays were used to demonstrate the mean percent sensitivity when the highest concentration of drug (celecoxib, retinol, LY294002, sulindac, 5-FU, and oxaliplatin) was used in each cell line as the basis for comparisons of sensitivity, between high and low recurrence risk groups, for each of the drugs tested. Finally, in the cell lines with high Recurrence Scores, 8 h after treatment with targeted drugs, gene expression profiles were re-assessed to determine if the high-risk phenotype had been reversed.

***In Vitro* Drug Sensitivity Assays.** The cell proliferation assays for the 14 colon cancer cell lines profiled by gene array analyses included growth inhibition measurements using standard colorimetric assays. Cells were plated in 96-well assay plates at a density of 5,000 cells per well. After incubating for 24 h at 37°C, drugs were added to each well at specific concentrations. Cells were grown in the presence of drugs for an additional 96 h. Celecoxib was used at concentrations of 0.1, 5, 25, 50, and 100 (μmol/l); LY294002, a PI3Kinase inhibitor was used at concentrations of 0.1, 1, 10, 20 and 200 (μmol/l); retinol was used at concentrations of 0, 0.1, 1, 5, 10, and 50 (μmol/l); sulindac was used at concentrations of 0, 0.1, 1, 10, 100, and 1,000 (μmol/l); 5-FU and oxaliplatin were used at concentrations of 0, 0.1, 1, 5, 10, and 20 (μmol/l). All concentrations were micromolar. Sensitivity to celecoxib, LY294002, retinol, sulindac, 5-FU, and oxaliplatin was determined by quantifying the percent reduction in growth at 96 h using the standard MTT Cell Proliferation Kit from Roche Applied Science. A Perkin–Elmer Victor 3 Multilabel Plate Reader was used to determine UV absorbance. All experiments were repeated at least three times.

Cell and RNA Preparation. Total RNA was extracted using the Qiashredder and Qiagen Rneasy Mini kits. Quality of the RNA was checked by an Agilent 2100 Bioanalyzer. The targets for Affymetrix DNA microarray analysis were prepared according to the manufacturer's instructions. Biotin-labeled cRNA, produced by *in vitro* transcription, was fragmented and hybridized to the Affymetrix HG-U133A Plus 2.0 GeneChip arrays at 45°C for 16 h and then washed and stained using the GeneChip Fluidics. The arrays were scanned by a GeneArray Scanner and patterns of hybridization were detected as light emitted from the fluorescent reporter groups incorporated into the target and hybridized to oligonucleotide probes. All analyses were performed in a MIAME (minimal information about a microarray experiment)-compliant fashion, as defined in the guidelines established by Microarray and Gene Expression Data (MGED) Society.

RT-PCR Analysis. The top 10 differentially expressed genes from among the 50-gene model were chosen for further validation using real-time PCR. Briefly, the methods involved Taqman (Applied Biosystems) custom arrays, a 384-well micro fluidic card that enables 384 real-time PCR reactions to be performed simultaneously without the use of liquid-handling robots or multichannel pipettes. The array is designed for a two-step RT-PCR. In the reverse transcription (RT) stage, total RNA extracted from each of the seven high risk and seven low risk cell lines are reverse transcribed into cDNA using random primers from the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). One thousand ng of total RNA were transcribed in a 25 μl reaction. After cDNA synthesis, 25 μl of RNase/DNase-free water was added. For each cell line, a total of four replicate samples were generated. For PCR, Taqman (Applied Biosystems) gene expression assays (including 18 s used as the manufacturing control) were preloaded into each of the wells of the array. Sample-specific PCR mix was generated by adding 50 μl of the Taqman Universal Master Mix (Applied Biosystems) to the 50 μl of cDNA plus water reactions. One hundred μl were pipetted into each port of the Taqman array and run on the 7900HT Fast Real-Time PCR System with the Low Density Array Block (Applied Biosystems). After PCR, gene targets were analyzed by assessing Ct values after normalization to GAPDH to compare quantitative expression values between the low risk and high risk cell lines.

1. Lin YH, *et al.* (2007) Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer. *Clin Cancer Res* 13:498–507.
2. Reid T, GSE10402.
3. Ihaka R, Gentleman RE (1996) A language for data analysis and graphics. *J Comp Graph Stats* 5:299–314.
4. Pittman J, Huang E, Nevins J, Wang Q, West M (2004) Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes. *Biostatistics* 5:587–601.
5. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8:118–127.

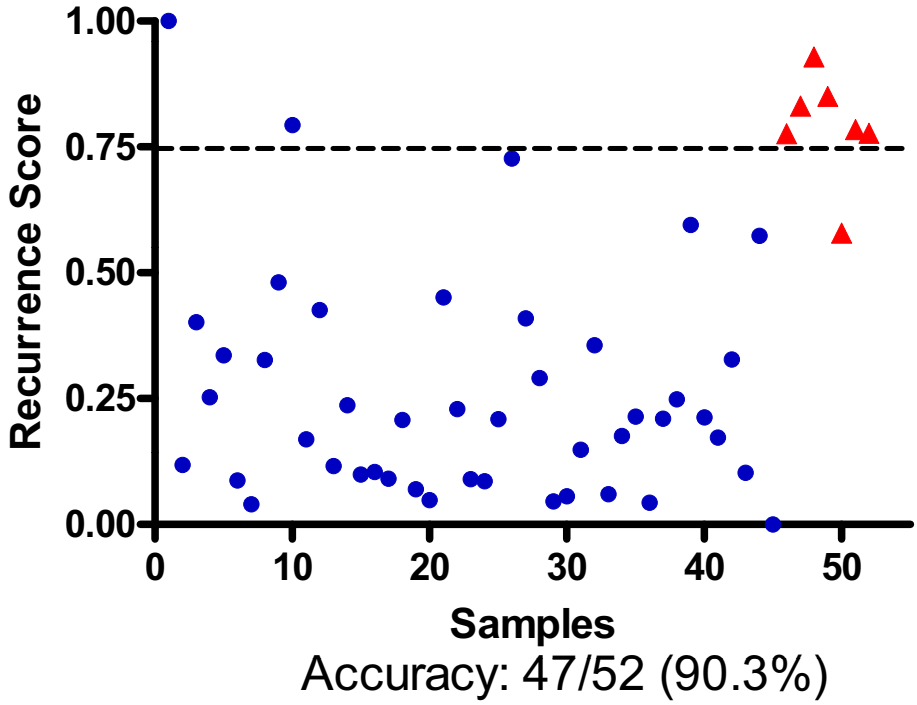


Fig. S1. Accuracy of the 50-Gene Model. The 50-Gene Model successfully predicted recurrence with an accuracy of 90.3%. In this leave-one-out validation, of the cases with actual recurrence (in red), all but one had a high recurrence score. Similarly, all but two of the patients who were disease free (in blue) were also predicted accurately. The cut-point for this analysis was based on the ROC analysis.

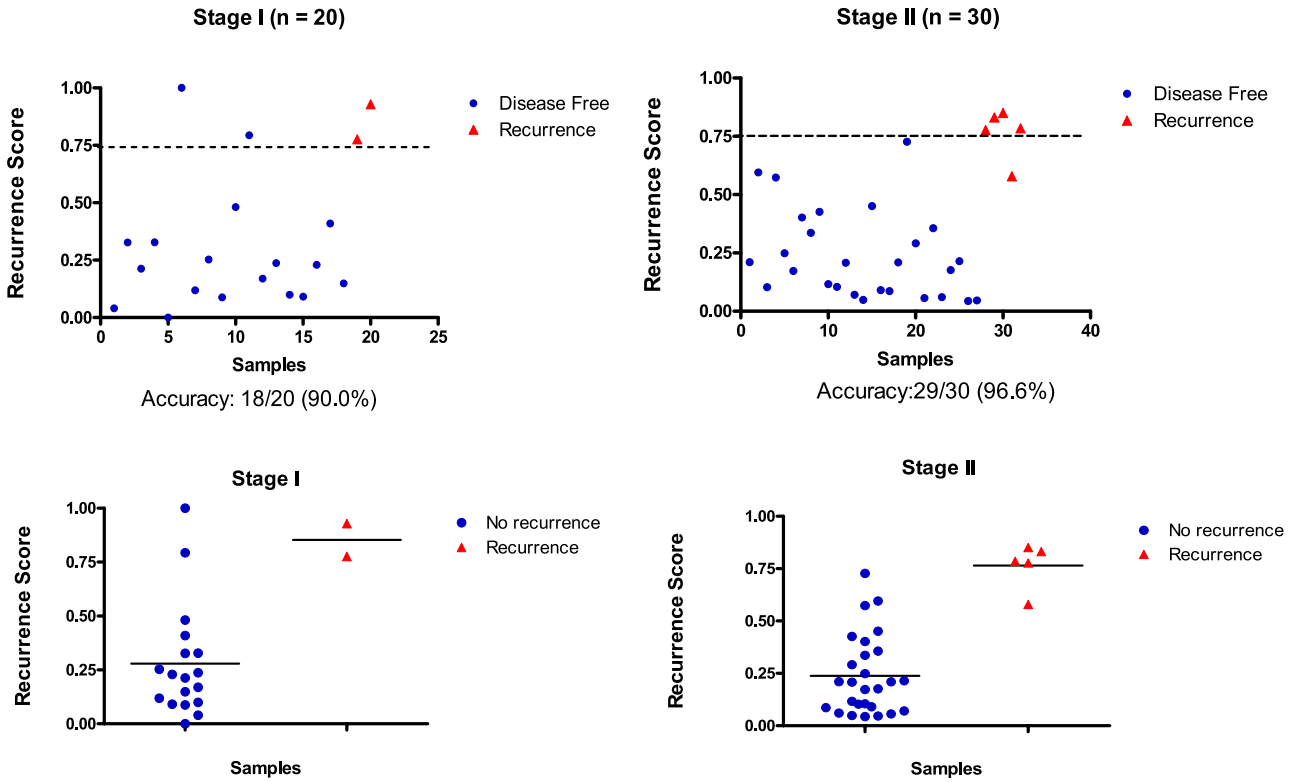
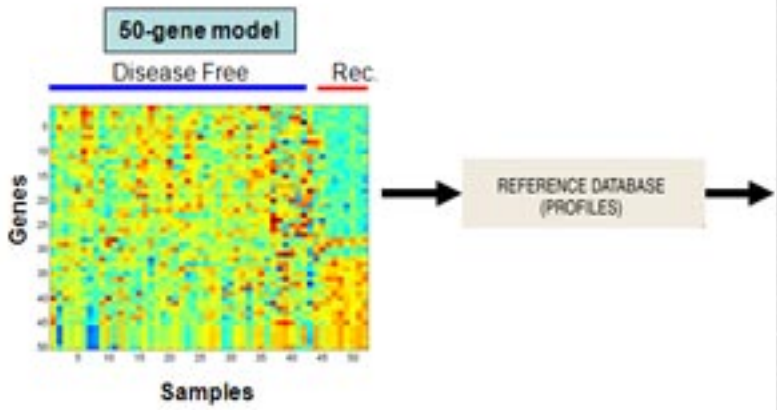


Fig. S2. The 50-Gene Model of recurrence by stage of disease. The 50-Gene predictor successfully predicted recurrence for both stage I and stage II colon cancer. When separated by stage, accuracy for stage I was 90.0% and for stage II was 96.6%.



Drug	Replicates	Concentration
raloxifene	38	100 nM
arachadonic acid	59	10 ÂµM
SC-58125	75	10 ÂµM
celecoxib	39	10 ÂµM
HNMPA-(AM)3	87	5 ÂµM
rofecoxib	18	10 ÂµM
SC-58215	73	10 ÂµM
estradiol	38	10 nM
imatinib	36	10 ÂµM
oligomycin	59	1 ÂµM
tamoxifen	38	1 ÂµM
genistein	111	10 ÂµM
tamoxifen	39	1 ÂµM
fulvestrant	23	10 nM
haloperidol	68	10 ÂµM
monensin	514	11 ÂµM
5253409	504	17 ÂµM
sulindac	23	50 ÂµM
5149715	505	10 ÂµM
LY-294002	53	10 nM
ikarugamycin	504	2 ÂµM
tretinoin	504	1 ÂµM

Fig. S3. The Connectivity Map computationally matches drugs likely to be effective based upon a core gene expression profile. Represented here is the list of drugs identified in such an analysis, based upon the 50-Gene Model.

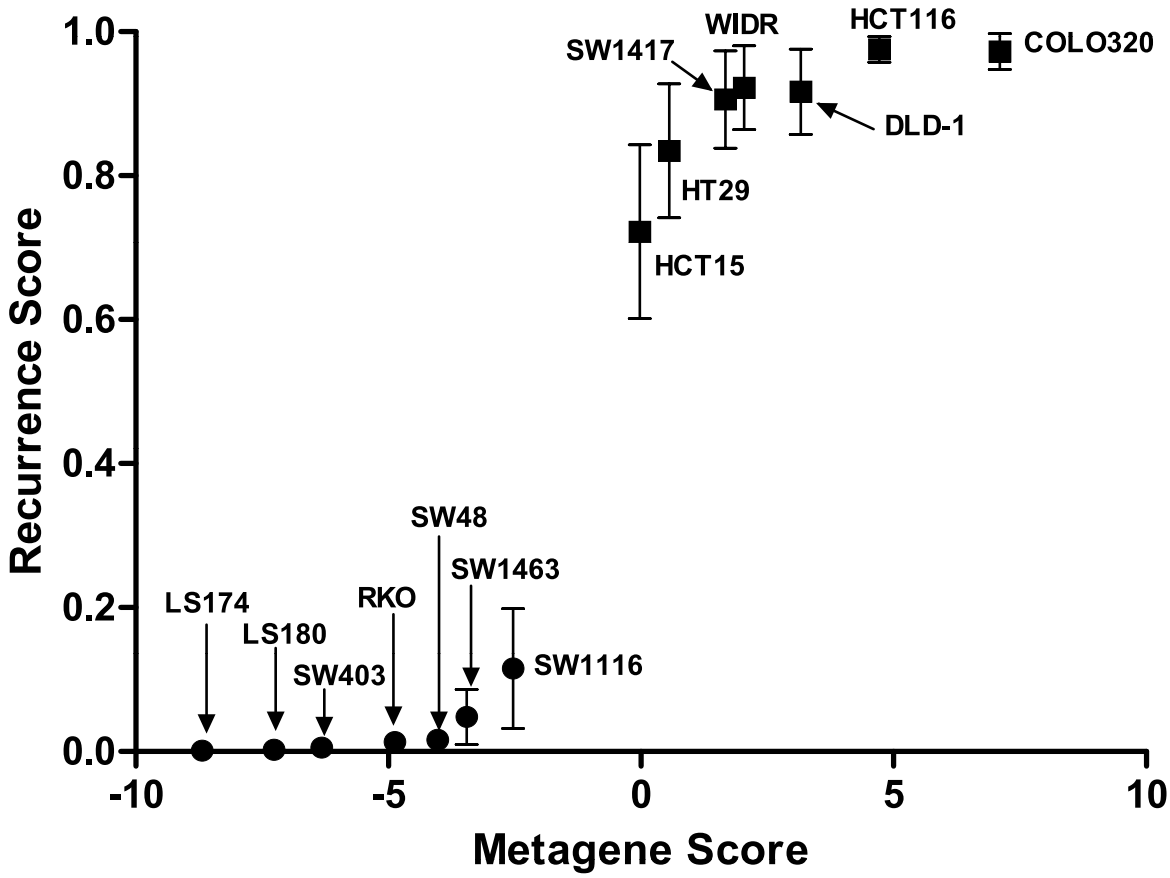


Fig. S4. Application of 50-Gene Model to 14 colon cancer cell lines. Gene expression data from the 14 cancer cell lines listed was used and Recurrence Scores estimated for each cell, using the 50-Gene Model. The cell lines with high Recurrence Scores are clustered on the upper right, and those with low Recurrence Scores are on the lower left.

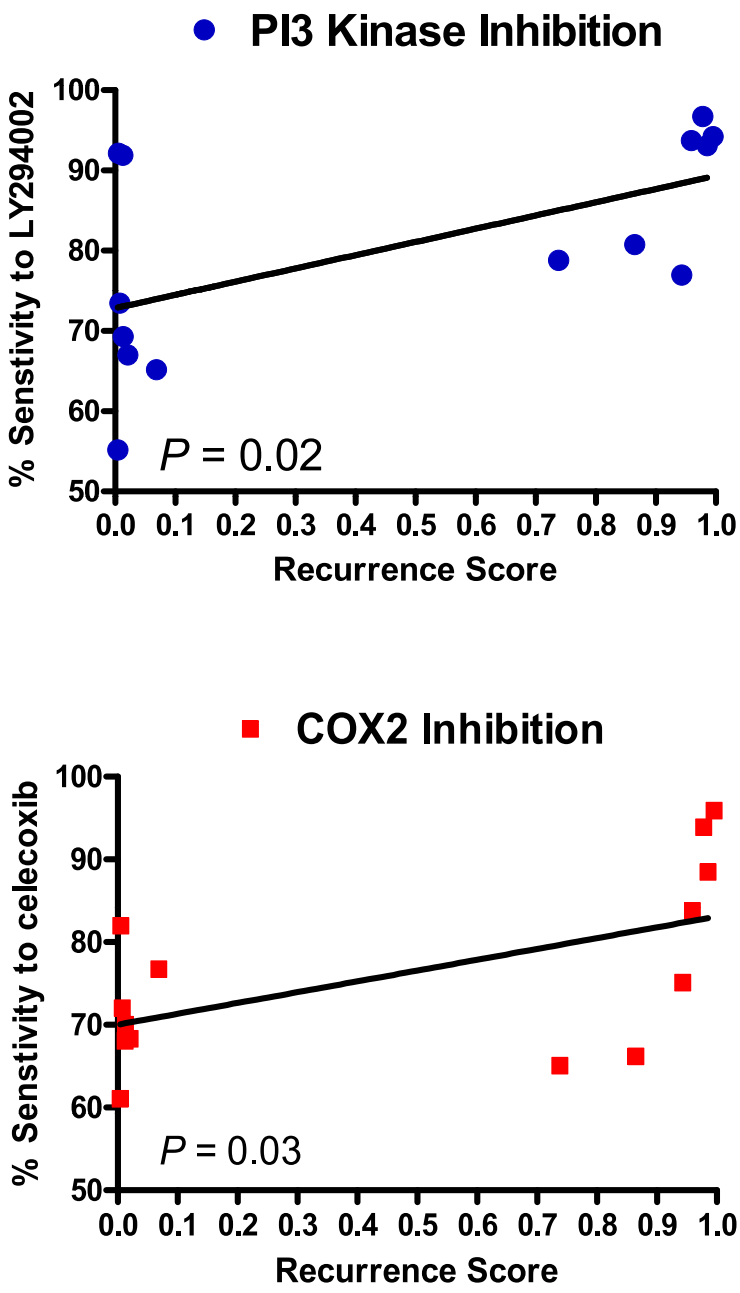


Fig. S5. Linear regression analyses of the probability of recurrence phenotype/recurrence scores in colon cancer cell lines and *in vitro* sensitivity to individual therapeutic agents reveal a significant correlation for COX2 (celecoxib) ($P = 0.03$) and PI3kinase inhibition (LY294002) ($P = 0.02$), suggesting that specific COX2 and PI3Kinase inhibitors could be valuable as initial agents in therapeutic intervention studies, involving early stage colon cancer patients at high risk for disease recurrence.

Table S1. The demographic and clinical characteristics of the patients included in the development and validation of the 50-gene predictor of disease recurrence

	Training set (n = 52)	Independent validation (n = 55)	Blinded validation (n = 73)
Age, years			
Median	68	64.5	69
Range	40–87	Unknown	28–92
Mean ± SD	66.5 ± 11.5	63.3 ± 11.5	67.0 ± 13.0
Sex			
Male	28 (54%)	33 (60%)	54 (74%)
Female	24 (46%)	22 (40%)	19 (26%)
Histologic type			
Adenocarcinoma	52	55	73
TNM Stage			
0/1	20 (38%)	11 (20%)	0
2	32 (62%)	44 (80%)	33 (45%)
3	0	0	40 (55%)
Recurrence	7 (13%)	26 (47.2%)	10 (14%)

Table S2. The genes that constitute the predictor of colon cancer recurrence

Probe ID	Gene symbol	Gene title
1598_g_at	GAS6	growth arrest-specific 6
200917_s_at	SRPR	signal recognition particle receptor ('docking protein')
201835_s_at	PRKAB1	protein kinase, AMP-activated, beta 1 non-catalytic subunit
202052_s_at	RAI14	retinoic acid induced 14
202619_s_at	PLOD2	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2
202620_s_at	PLOD2	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2
202723_s_at	FOXO1	forkhead box O1
202848_s_at	GRK6	G protein-coupled receptor kinase 6
203508_at	TNFRSF1B	tumor necrosis factor receptor superfamily, member 1B
204430_s_at	SLC2A5	solute carrier family 2 (facilitated glucose/fructose transporter), member 5
204629_at	PARVB	parvin, beta
205868_s_at	PTPN11	protein tyrosine phosphatase, non-receptor type 11 (Noonan syndrome 1)
207495_at	RAB28	RAB28, member RAS oncogene family
208424_s_at	CIAPIN1	cytokine induced apoptosis inhibitor 1
210405_x_at	TNFRSF10B	tumor necrosis factor receptor superfamily, member 10b
210495_x_at	FN1	fibronectin 1
210809_s_at	POSTN	periostin, osteoblast specific factor
211251_x_at	NFYC	nuclear transcription factor Y, gamma
211628_x_at	FTHP1	ferritin, heavy polypeptide pseudogene 1
211719_x_at	FN1	fibronectin 1
212113_at	LOC552889	hypothetical LOC552889
212437_at	CENPB	centromere protein B, 80kDa
212464_s_at	FN1	fibronectin 1
212656_at	TSFM	Ts translation elongation factor, mitochondrial
212775_at	OBSL1	obscurin-like 1
213256_at	MARCH3	membrane-associated ring finger (C3HC4) 3
213446_s_at	IQGAP1	IQ motif containing GTPase activating protein 1
213799_s_at	PTPRA	protein tyrosine phosphatase, receptor type, A
214048_at	MBD4	methyl-CpG binding domain protein 4
214279_s_at	NDRG2	NDRG family member 2
214827_at	PARD6B	par-6 partitioning defective 6 homolog beta (C. elegans)
215127_s_at	RBMS1	RNA binding motif, single stranded interacting protein 1
215210_s_at	DLST /// DLSTP	dihydrolipoamide S-succinyltransferase (E2 component of 2-oxo-glutarate complex)
216205_s_at	MFN2	mitofusin 2
216442_x_at	FN1	fibronectin 1
217625_x_at	—	Homo sapiens, clone IMAGE:3851018, mRNA
217818_s_at	ARPC4	actin related protein 2/3 complex, subunit 4, 20kDa
218382_s_at	U2AF2	U2 small nuclear RNA auxiliary factor 2
218433_at	PANK3	pantothenate kinase 3
219408_at	PRMT7	protein arginine methyltransferase 7
219736_at	TRIM36	tripartite motif-containing 36
219878_s_at	KLF13	Kruppel-like factor 13
221463_at	CCL24	chemokine (C-C motif) ligand 24
221518_s_at	USP47	ubiquitin specific peptidase 47
221899_at	PFAAP5	Phosphonoformate immuno-associated protein 5
221926_s_at	IL17RC	interleukin 17 receptor C
222020_s_at	HNT	neurotrimin
38340_at	HIP1R /// LOC728014	huntingtin interacting protein 1 related /// similar to huntingtin interacting protein 1 related
39835_at	SBF1	SET binding factor 1
47571_at	ZNF236	zinc finger protein 236

There are two probes (202619_s_at and 202620_s_at) that represent splice variants for PLOD2.