Noninvasive Detection of Candidate Molecular Biomarkers in Subjects with a History of Insulin Resistance and Colorectal Adenomas

Chen Zhao,[1] Ivan Ivanov,[2,3] Edward R. Dougherty,[1] Terryl J. Hartman,[6] Elaine Lanza,[7] Gerd Bobe,[7] Nancy H. Colburn,[7] Joanne R. Lupton,[3,4,5] Laurie A. Davidson[3,4,5] and Robert S. Chapkin[3,4,5]

Abstract We have developed novel molecular methods using a stool sample, which contains intact sloughed colon cells, to quantify colonic gene expression profiles. In this study, our goal was to identify diagnostic gene sets (combinations) for the noninvasive classification of different phenotypes. For this purpose, the effects of a legume-enriched, low glycemic index, high fermentable fiber diet was evaluated in subjects with four possible combinations of risk factors, including insulin resistance and a history of adenomatous polyps. In a randomized crossover design controlled feeding study, each participant (a total of 23; 5–12 per group) consumed the experimental diet (1.5 cups of cooked dry beans) and a control diet (isocaloric average American diet) for 4 weeks with a 3-week washout period between diets. Using prior biological knowledge, the complexity of feature selection was reduced to perform an exhaustive search on all allowable feature (gene) sets of size 3, and among these, 27 had (unbiased) error estimates of 0.15 or less. Linear discriminant analysis was successfully used to identify the best single genes and two- to three-gene combinations for distinguishing subjects with insulin resistance, a history of polyps, or exposure to a chemoprotective legume-rich diet. These results support our premise that gene products (RNA) isolated from stool have diagnostic value in terms of assessing colon cancer risk.

Introduction

Colon cancer is one of the leading causes of cancer-related deaths in the United States. Early detection is one of the proven strategies resulting in a higher cure rate (1). Unfortunately, the currently adopted screening procedures for early detection are often invasive (e.g., colonoscopy), and discomfort associated with such procedures generally leads to resistance toward the screening process. Thus, adoption of noninvasive methods designed to reduce anxiety over colorectal cancer screening and improve overall acceptance of the screening process would be highly desirable.

We recently showed that a high level of dry bean intake reduced tumor formation in carcinogen-injected mice (2, 3) and decreased the risk of advanced colorectal adenoma recurrence among participants in the Polyp Prevention Trial (4). Based on these studies, we hypothesized that a legume-enriched diet may reduce the rate of absorption of carbohydrates, lowering the postprandial glycemic index and insulinemic response, leading to a suppression in the level of inflammatory mediators and markers of insulin resistance (IR; ref. 5). In addition, the high level of fermentable fibers in beans would enhance the production of butyrate, an anti-inflammatory, antineoplastic short-chain fatty acid (6, 7). Although further studies are warranted to characterize the molecular features of chemoprotective diets, rigorous analysis of the effects of diet on transcriptome profiling has been limited thus far, largely due to difficulties in obtaining appropriate samples. Therefore, the development of noninvasive molecular methods using stool for the purpose of quantifying colonic gene expression profiles would be highly desirable.

Approximately one sixth to one third of normal adult colonic epithelial cells are shed daily (8). Exploiting this fact, we have developed novel noninvasive methods using feces, containing exfoliated colonocytes, to quantify colonic mRNAs (9–11). Although RNA is generally less suitable than DNA because it is readily degraded, we and others have shown that intact fecal eukaryotic mRNA can be isolated because of the presence of viable exfoliated colonocytes in the fecal stream (9, 11–14).

Using exfoliated colonocytes, we have previously reported the discriminative mRNA expression signatures between inflammatory bowel disease versus normal and between adenoma versus normal (11). These data suggest that mRNA

Authors' Affiliations: [1]Department of Electrical Engineering, [2]Department of Veterinary Pharmacology and Physiology, [3]Center for Environmental and Rural Health, [4]Intercollegiate Faculty of Nutrition, and [5]Program in Integrative Nutrition and Complex Diseases, Texas A&M University, College Station, Texas; [6]Department of Nutritional Sciences, Pennsylvania State University, University Park, Pennsylvania; and [7]National Cancer Institute, Bethesda, Maryland

Received 12/10/08; revised 2/27/09; accepted 3/23/09; published OnlineFirst 5/26/09.

Grant support: National Cancer Institute and NIH grants CA59034, CA129444, DK071707, and P30ES09106.

Note: Supplementary data for this article are available at Cancer Prevention Research Online (http://cancerprevres.aacrjournals.org/).

Requests for reprints: Robert S. Chapkin, Texas A&M University, 218 Kleberg Center, TAMU 2253, College Station, TX 77843-2253. Phone: 979-845-0419; Fax: 979-862-2378. E-mail: r-chapkin@tamu.edu

doi:10.1158/1940-6207.CAPR-08-0233

isolated from exfoliated human colonocytes can be used to detect early stages of colon cancer and possibly chronic inflammation. However, the microarray gene expression profile–based classification of colonic diseases for diagnostic purposes has yet to be solved. Therefore, in this study, we further determined the feasibility of the noninvasive mRNA procedure in patients at high risk for colorectal adenoma recurrence. Specifically, the effect of a legume-enriched, low glycemic index, high fermentable fiber diet on subjects exhibiting a combination of risk factors including IR and history of adenomatous polyps was evaluated. To our knowledge, this is the first controlled feeding study to examine the effects of legumes or a low glycemic index diet on changes in intestinal gene expression profiles using exfoliated colonocytes. Our goal was to develop diagnostic gene sets (combinations) for the objective classification of different phenotypes. Applying this approach to a test set of 23 subjects, we have identified the best single genes and two- to three-gene combinations for distinguishing polyps, IR, and exposure to a legume diet. We also report that using combinations of genes, the classification error rate can be significantly lowered. Two- and three-gene combinations thus provide robust classifiers with potential to noninvasively identify discriminative signatures for differential diagnostic purposes.

Materials and Methods

Experimental design

After obtaining informed consent of the subjects, a controlled feeding study was conducted, comparing the effects of a legume-enriched, low glycemic index diet to the average American diet (control) in four different groups of male participants: (*a*) previous history of adenomas and IR; (*b*) previous history of adenomas with no IR; (*c*) IR with no history of adenomas; and (*d*) no-IR and no history of adenomas. Subjects were enrolled into a two-period crossover study in which all four groups were randomly allocated to each of two diets: (*i*) a control diet); (*ii*) a high-legume, low glycemic index diet. The subjects (a total of 23; 5-12 per group) consumed the experimental diets for 4 wk with a 3-wk washout period before crossing over to the other diet. The overall study design is shown in Fig. 1. Baseline samples were collected before commencing each diet period, and additional samples collected at the end of each diet period. All procedures used in the study were reviewed and approved by the human subjects' committees at the Pennsylvania State University (PSU), Texas A&M University, and the NIH. Study procedures are briefly summarized below.

Subject recruitment

Subjects were recruited with the assistance of gastroenterologists performing colonoscopies at the Mount Nittany Medical Center in State College, Pennsylvania. Nursing staff reviewed all colonoscopy records for eligible participants. Eligible patients were mailed a letter signed by their endoscopist inviting them to participate in the study and asking them to return a postcard to indicate that they would like to be contacted or to call the PSU study coordinator (toll-free), if they would like to learn more about the study. A preliminary telephone eligibility screening was completed by the coordinator, and subjects eligible according to the phone interview were invited to the PSU General Clinical Center Research Center (GCRC). After receiving informed consent, the participant's height, weight, and blood pressure were checked by study staff or the nurses at the clinic, and a fasting blood sample was taken to determine overall health (including fasting insulin and glucose to determine insulin sensitivity and cholesterol levels and lab tests for heart and liver function). A physician at the GCRC reviewed the results to determine eligibility for participation. All eligible consented participants were asked to return to the GCRC to assess their resting metabolic rate. Demographic, health, and lifestyle questionnaires were completed, and the participants were given instructions for completing a 4-day food record for the purpose of estimating pre-study baseline dietary intake.

Inclusion and exclusion criteria

Eligible participants for the study were males between 35 and 75 y of age, with a body mass index of 25.0 to 34.9 kg/m^2, and having undergone a screening colonoscopy within the past 2 y. Only male subjects were recruited because in our previous study, males had a greater response to diet (e.g., beans) and exhibited a higher risk for polyp recurrence (4). Participants could not be diagnosed with a serious medical condition such as cancer, heart disease, kidney disease, diabetes, or other serious medical condition including a history of colorectal cancer, bowel resection, polyposis syndrome, or inflammatory bowel disease. Subjects were not permitted to take any medication that would alter inflammation markers, insulin, glucose, or blood lipids. The +Polyp group had polyps removed within the past 2 y. No subjects reported the development of colon polyps during the study.

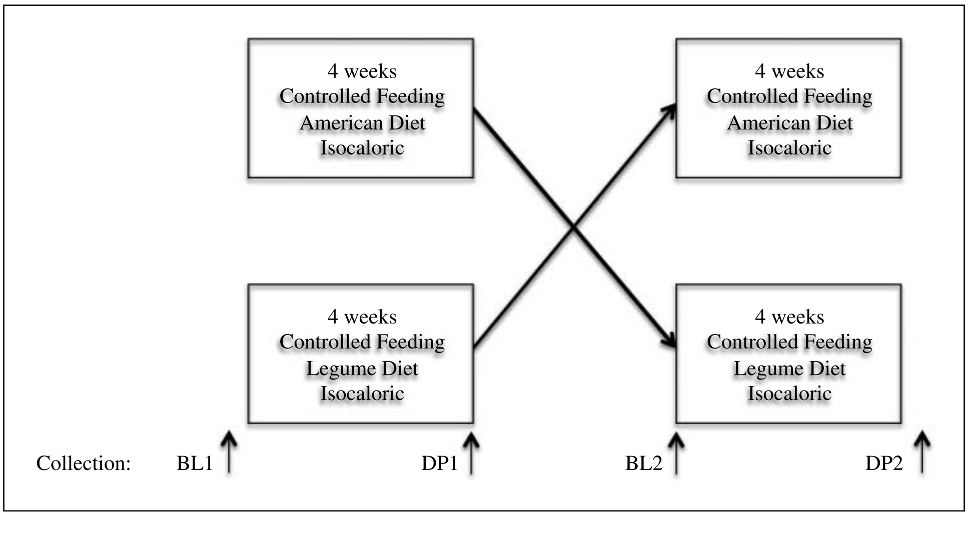


Figure 1. Overall study design. BL, baseline measurement; DP, diet period.

Dietary intervention

Subjects consumed one meal per day, breakfast or dinner, on site during the weekdays and consumed a packed lunch, snack, and other meal at a time and place of convenience. Weekend meals were prepared and packed for carryout. No foods other than those provided by the study kitchen were permitted. At each visit, subjects were weighed and asked to return any uneaten foods. Thus, compliance was monitored on a daily basis by assessment of body weight, direct observation of the consumption of one "in-house" meal per day, and by daily review of uneaten foods. Subjects were also asked to record any food items not provided by the study. Alcohol consumption was limited to no more than two drinks per week during the controlled feeding period. In addition to the monitoring of the dietary records, subjects were queried daily about alcohol consumption to ensure compliance. No subjects reported the consumption of non-study foods or excessive alcohol during the week of fecal and blood collections.

A 7-day menu cycle was developed with a standard set of legumes of the *Phaseolus vulgaris* species, such as navy beans, pinto beans, and kidney beans, to limit nutrient and phytochemical differences in the 7-day diet cycle. The diet contained ~250 g of legumes per day (1.5 cups). This level added ~20 g of total dietary fiber and 8 g of soluble fiber per day. The diet was modified to provide other high glycemic index foods in the control diet so that the glycemic index of the control diet was ~70, compared with a glycemic index of 30 in the legume diet. Each daily menu was designed to maintain a constant level of fat (32-33 energy %), whereas the high-legume low glycemic index diet had a total dietary fiber intake of ~40 g/d, compared with 20 g/d for the high glycemic index diet. The protein level of both diets was ~18 energy %. To maintain the same level of red meat and fish (foods that have been associated with colon cancer) in both diets, the protein in legumes was substituted for protein from poultry. All nutrients were provided in amounts to meet the recommended dietary allowances for men of the same age groups. A food composite for each of the 6 days was freeze-dried and analyzed for macronutrient and fiber levels. Individual food items were purchased at the same time from the same supplier to ensure uniformity of the diet.

Table 1. Classification of (+IR, +Polyps) subjects versus (−IR, −Polyps) subjects at BL1

Gene names	$\varepsilon_{\text{bolstered}}$	$\Delta(\varepsilon_{\text{bolstered}})$
IGF1R	0.1094	
CDK4	0.1200	
BECN1	0.1223	
NOS3	0.1436	
ALOX12B	0.1477	
NOS3, WNT1	0.1277	0.2656
HOXA3, UCP2	0.1415	0.3467
IGF1R, WNT1	0.1484	0.2449
ID2, IGF1R	0.1486	0.3139
HOXA3, YWHAZ	0.1503	0.3379
HOXA3, IGF1R	0.1513	0.3369
BECN1, HOXA3, MAPK11	0.0891	0.3991
BECN1, HOXA3, IGF1R	0.0907	0.3975
HOXA3, MAPK11, YWHAZ	0.0935	0.3947
HOXA3, HOXC6, MAPK11	0.0941	0.3941
HOXA3, MAPK11, NOS3	0.0987	0.3895
HOXA3, UCP2, YWHAZ	0.1001	0.3881
HOXA3, IGF1R, YWHAZ	0.1006	0.3876
BECN1, DAPK1, IGF1R	0.1012	0.3768
HOXA3, HOXC6, TJP1	0.1023	0.3859
HOXA3, HOXC6, IGF1R	0.1079	0.3803

NOTE: Single-gene, pair-wise, and triplet-wise LDA classifiers are shown. $\varepsilon_{\text{bolstered}}$ denotes the bolstered resubstitution error for the respective classifier; $\Delta(\varepsilon_{\text{bolstered}})$ denotes the largest decrease in error for the feature set relative to all of its subsets.

mRNA expression microarray analysis

The overall structure of the microarray data set is shown in Supplementary Table S1. Stool samples were collected by the subject into a sterile cup, sealed, and placed at 4°C storage for up to 12 h. Samples were then coded by the Research Assistant, homogenized in a guanidinium-based solution, and stored at −80°C until polyA RNA was isolated. From each subject, poly A^+ RNA was isolated from feces as we have previously described (11). Due to the high level of bacterial RNA in fecal samples, poly A^+ RNA was isolated to obtain a highly enriched mammalian poly A^+ RNA population. We have previously shown that with the isolation of poly A^+ RNA, contamination with bacterial RNA is undetectable (9). In addition, an Agilent 2100 Bioanalyzer was used to assess integrity of fecal poly A^+ RNA. Samples were processed in strict accordance to the CodeLink Gene Expression Assay manual (Applied Microarray) and analyzed using the Human Whole Genome Expression Bioarray as we have previously described (15). Each array contained the entire human genome derived from publicly available, well-annotated mRNA sequences.

Arrays were inspected for spot morphology. Marginal spots were flagged as background contaminated (C), irregularly shaped (I), or saturated (S) in the output of the scanning software. Spots that passed the quality control standards were categorized as good (G). In addition, spots marked with (L) indicated the reading was "near background." The low (L) measurements reflect either true low gene expression levels or may have been caused by degradation of the mRNA resulting in a low signal. Typically, samples collected from colonic mucosa (15) exhibited a relatively low proportion (5-8%) of L spots. In contrast, the proportion of L spots obtained from fecal samples was significantly higher (65-83%).

Microarray data normalization

For the purpose of interarray normalization, a set of housekeeping genes was used. These were determined in the following manner.

Housekeeping gene preparation. Common good probes (2,584) across all 86 microarrays were identified. A good probe is defined as having, at most, two low measures across all 86 microarrays. Using a list of 575 housekeeping genes (16), 18 genes were identified from the 2,584 probes found in the previous step. Subsequently, the raw intensity of each of the 18 housekeeping genes was quantified, and those with missing values were excluded. As a result, there were a total of 18 housekeeping genes used for normalization. Refer to Supplementary Methods and Supplementary Fig. S1 for details.

Additive normalization procedure. Arrays were grouped across time and the average values of 18 housekeeping genes were calculated (Supplementary Fig. S1). Median values of the averages were also calculated for the first 67 arrays. Subsequently, a robust piecewise linear regression was done and the corresponding regression value for each array was calculated. Following this step, the difference between the median and regression values for each array was determined, and the raw expression values of the genes on each array were shifted by the corresponding discrepancies.

Development of an algorithm for identifying feature (gene) sets

Details related to the development of an algorithm for identifying feature (gene) sets are described in Supplementary Methods. Because our main goal was to determine if mRNA data from exfoliated colonocytes have the potential to classify different colon cancer risk factors, we compared the obtained array data sets (termed A) with a set of 529 putative human colonic markers (termed B; refer to Supplementary Table S2). Using such prior biological knowledge, we investigated the set of genes common to the microarrays and putative colonic markers ($A_j^k \cap B$). The number of common genes for various values of analysis parameters is given in Supplementary Table S3. Based on these results, we used a conservative approach that provided us with a subset of putative colonic biomarkers that have strong signal ($k = 2$), compared with the CodeLink weaker default condition ($k = 1.5$), and no more than one low signal spot ($j = 1$) in the entire data set. It is possible, therefore, to group microarray data into various combinations of two different classes. This is due to the experimental design that lists risk factors: (+IR) and (−IR); four time points: baseline 1 (BL1), diet period 1 (DP1), baseline 2 (BL2), diet period 2 (DP2); and two diets: high legume/low glycemic index and control. These different groupings produced their respective sets of genes, which could be larger or smaller depending on the microarrays that were included in the corresponding groups or classes (Supplementary Table S4).

For the purpose of identifying feature sets, we designed classifiers that categorize samples based on the expression values of the genes from the intersection of the array gene set and the colon biomarker list ($A_1^2 \cap B$). An important consideration is that the number of genes in the feature sets should be sufficiently small. Hence, we constructed the classifiers for feature sets of sizes 1, 2, and 3. Generally, there are two reasons why it is desirable to design classifiers involving small numbers of genes: (*a*) the limited number of samples often available in clinical studies makes classifier design and error estimation problematic for large feature sets (17); and (*b*) small gene sets facilitate design of practical immunohistochemical diagnostic panels. For similar reasons, simple classifiers are preferable for small samples; indeed, for small samples, if good classification is possible, then a simple classifier such as linear discriminant analysis (LDA) using a small number of genes will typically outperform a complex classifier (18).

Given a set of features on which to base a classifier, one has to address not only the classifier design from sample data but also the estimation of its error. When the number of potential feature sets is large, the key issue is whether a particular feature set provides good classification. Therefore, a concern is the precision with which the error of the designed classifier estimates the error of the optimal classifier. When data are limited, an error estimator may have a large variance and therefore may often be low even if it is approximately unbiased. This can produce many feature sets and classifiers with low error estimates. The algorithm used in this study mitigated this problem by applying the bolstered error estimation (19). It has advantages with respect to commonly used error estimators such as resubstitution, cross-validation, and bootstrap methods for error estimation in terms of speed and accuracy (bias and variance). Basically, this approach "bolsters" the original empirical distribution of the available data by means of suitable bolstering kernels placed at each data point location. The error can be computed analytically in some cases, such as in the case of LDA. The relatively small size of the gene set ($A_1^2 \cap B$) allows for the comparison of the errors of all of the possible feature sets of sizes 1, 2, and 3, thereby avoiding feature selection, which can be highly unreliable in small sample settings (20). The result of the overall approach is a list of "best" feature sets from among all possible feature sets. Hence, the best feature set is the one possessing minimum classification error. Because we only have data and not the underlying feature-label distributions, the errors have been estimated from the data. This approach takes into account that, in small-sample settings, we do not have much confidence in any single feature set and that it is much more likely that, if there is an adequate sized collection of good-performing feature sets, then there are likely to be some that perform well on the overall population (21).

Results and Discussion

Classification analysis

Stool-based molecular diagnostic tests are emerging as important new approaches that have the potential of providing cost-effective, sensitive early detection of colorectal neoplasia. Details of many of the currently used and novel approaches have been recently reviewed (22). Because a single genetic product is unlikely to have sufficient detection sensitivity and specificity to be used as a "stand-alone" diagnostic test, a fecal-based DNA detection system that exploits the concept of chromosomal instability with mutations progressively accumulating in the adenomatous polyposis coli, p53 tumor suppressor genes, and the K-*ras* oncogene has been recently developed (23, 24). Publications in small trials (16-65 subjects) reported test sensitivity ranging from 62% to 91% for adenocarcinoma detection and 27% to 82% for adenoma detection, with specificity ranging from 93% to 98%. Validation of these preliminary data in a large (4,404 evaluated subjects) prospective colorectal cancer screening trial resulted in a sensitivity of 52% (95% confidence interval, 35-68%) for detection of adenocarcinoma and 15% (95% confidence interval, 12-19%) for detection of adenomas ≥1 cm, with double the sensitivity

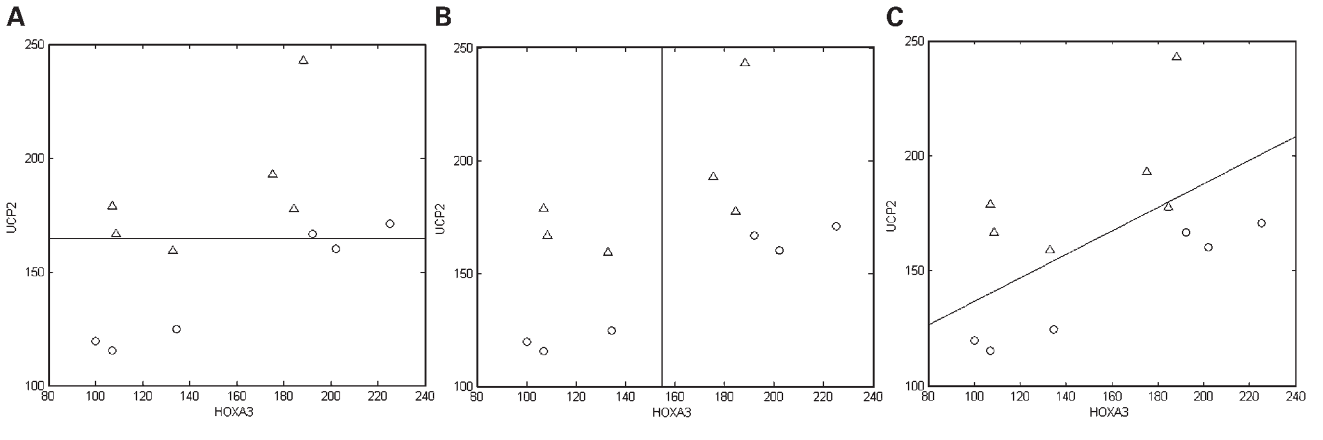


Figure 2. The concept of intrinsically multivariate predictive (*IMP*) genes is shown where expression profiles of a group of genes predict the phenotype. Results represent a linear classification of (+IR, +Polyps) subjects (○) versus (−IR, −Polyps) subjects (△) at BL1. UCP2 and HOXA3 were used as individual one-feature sets (*A* and *B*) as compared with both genes together as a two-feature set (*C*). The bolstered error is 0.2784, 0.4882, and 0.1415 for *A*, *B*, and *C*, respectively.

when the adenoma had dysplasia. Specificity for the fecal DNA test was 94% (23). Very recently, stool DNA test 2 and a novel digital melt curve assay, which targets more broadly informative markers, detected significantly more screen-relevant neoplasms compared with occult blood testing (25, 26). From these data, it is logical to assume that fecal DNA tests could serve as an intermediate, noninvasive screening tool for colorectal adenocarcinoma.

A major disadvantage of DNA-based methods is that it is inherently limited to a small number of hybridizing oligonucleotides, which reduces the likelihood that a neoplasia-associated mutation will be found in the large number and heterogeneity of mutational events occurring in human neoplasia. In addition, a fecal DNA testing panel using nucleotide probes will not detect important epigenetic events associated with human carcinogenesis. For example, epigenetic modifications of DNA (i.e., aberrant promoter hypermethylation) of multiple tumor suppressor genes lead to loss of expression (27). DNA-based methods do not detect these important molecular events. This severely limits the utility of current DNA-based assays. Recently, several attempts have been made to use DNA from stool to detect aberrant CpG island methylation (28, 29). Thus, it is possible that methylated genes may be effective early detection markers for colon adenomas, and offer another mechanistic approach that may increase performance characteristics of stool markers based on mutation detection alone.

To enhance current colon cancer molecular detection assays, our laboratory was first to develop a novel noninvasive molecular method using feces containing intact viable exfoliated colonocytes to quantify colonic mRNAs and determine gene expression profiles (9). Because "global" changes in patterns of gene expression occur throughout the colon well before macroscopic tumors are apparent (30, 31), these data suggest that "diagnostic" gene expression profiles are associated with a large number of shed cells, and hence, recovered cell number should not be a limiting factor (13).

In this feasibility study, our goal was to identify mRNA expression patterns that may establish the basis of a new noninvasive molecular diagnostic method. For this purpose, we applied an algorithm to 12 different pairs of classes arising from the experimental design as described in Fig. 1 and Supplementary Table S1. The number of genes/features for each linear classifier was limited to 3, which allowed for an exhaustive search. The use of small (three-gene) classifiers is not new in the classification of cancer. It goes back some number of years (21, 32). As an initial step within the context of classification, we identified the best single genes (single-gene classifiers) to distinguish phenotype. To illustrate how this approach compares to the traditional statistical analysis, we considered the classes (+IR, + Polyps) versus (−IR, −Polyps) at BL1. The top 10 feature sets of size 1 were compared with the differentially expressed genes in the colonic biomarker set ($A_1^2 \cap B$), where t tests were done using normalized and log-transformed gene intensity values. The comparison revealed that 7 of the 10 top one-feature sets (genes) identified by the linear (LDA) classifier also had P values <0.05. This is not surprising because individual differentially expressed genes have been traditionally used to discriminate between phenotypes (33). Interestingly, the results show that there are several cases where single genes can provide good (in terms of the error estimate) classification (Table 1). However, when comparing these results to the two-feature classification for the same two classes, a phenomenon was observed that has been recently documented in the context of gene network modeling (34). Specifically, the expression profiles of a group of genes predicted the target (either a gene or a phenotype) with greater accuracy relative to any proper subset of these genes. For example, single-gene classifiers (one-feature) based on either the Homeobox protein-A3 (*HOXA3*) or uncoupling protein-2 (*UCP2*) performed very poorly when discriminating between (+IR, + Polyps) and (−IR, −Polyps) at BL1 (Table 1; Fig. 2A and B). Interestingly, HOXA3 was close to the worst predictor of all of the available 97 genes (ranked 94). In comparison, when combined as a two-feature set, *UCP2* and *HOXA3* provided one of the best two-feature classifiers (one misclassified data point only) among all of the 4,656 possible two-gene sets (Table 1; Fig. 2C). These data clearly illustrate why complex phenotypes can be explained better by multivariate feature sets.

Table 2. Classification of (−IR, −Polyps) subjects on control diet versus (−IR, −Polyps) subjects on the legume diet

Gene names	$\varepsilon_{\text{bolstered}}$	$\Delta(\varepsilon_{\text{bolstered}})$
TGFB3	0.2350	
FOXP4	0.2586	
TP53	0.2970	
BAD	0.3009	
FOXO1A	0.3033	
DAPK1, HOXA3	0.1829	0.3760
BAD, LYZL6	0.2275	0.2321
IGF1R, LEF1	0.2315	0.2488
DAPK1, FOXM1	0.2371	0.2336
IGF2, TGFB3	0.2455	0.2814
LEF1, TGFB3	0.2459	0.2344
DAPK1, TP53	0.2642	0.2426
APC, CDC42	0.2650	0.2564
DAPK1, HOXA3, TGFB3	0.1675	0.3914
DAPK1, LEF1, TGFB3	0.1799	0.3004
DAPK1, HOXA3, LEF1	0.1854	0.3735
DAPK1, HOXA3, SELP	0.1887	0.3702
CAMK2A, DAPK1, HOXA3	0.1922	0.3667
DAPK1, HOXA3, SPARC	0.1944	0.3645
DAPK1, HOXA3, PRKACG	0.1969	0.3620
DAPK1, HOXA3, SFRP5	0.1982	0.3607
BAD, FOXE3, PTK2	0.2003	0.3018
CA5B, DAPK1, HOXA3	0.2028	0.3561
CD44, DAPK1, HOXA3	0.2052	0.3537
BAD, FOXP4, GSS	0.2056	0.3112
BAD, FOXE3, PTK2B	0.2072	0.3187
APC2, DAPK1, HOXA3	0.2117	0.3472

NOTE: Single-gene, pair-wise, and triplet-wise LDA classifiers are shown. Refer to Table 1 for legend details.

To identify sets of genes that perform in a multivariate manner to provide strong classification, we specifically looked for pairs of genes that performed better than either of the genes individually, and triplets of genes that performed well and substantially better than the best-performing pair among the three, and so on. To estimate the improvements of the classification performance, we introduced two quantities for each feature set: $\varepsilon_{\mathrm{bolstered}}$ and $\Delta(\varepsilon_{\mathrm{bolstered}})$. $\varepsilon_{\mathrm{bolstered}}$ denotes the bolstered resubstitution error for the LDA classifier for the respective feature set, and $\Delta(\varepsilon_{\mathrm{bolstered}})$ denotes the largest decrease in error for the full feature set relative to all of its subsets. The feature sets were initially ranked based on the value of $\varepsilon_{\mathrm{bolstered}}$, and subsequently ranked again based on the improvement $\Delta(\varepsilon_{\mathrm{bolstered}})$. For multiple-gene classifiers, we focused on feature sets with high rank in both lists. Along these lines, we designed two-feature classifiers for the classification of (+IR, +Polyps) versus (−IR, −Polyps) data at baseline BL1; (−IR, −Polyps, control diet) versus (−IR, −Polyps, legume diet) data at the end of the two diet periods DP1 and DP2; (+IR, + Polyps) versus (−IR, −Polyps) at baselines BL1 and BL2; (+Polyps) versus (−Polyps) at baselines BL1 and BL2; and (+IR) versus (−IR) at all of the time points. Tables 1 and 2 describe the best (according to this ranking procedure) feature sets identified for the first two of these classification categories, and Fig. 3A and B shows representative multivariate classifiers.

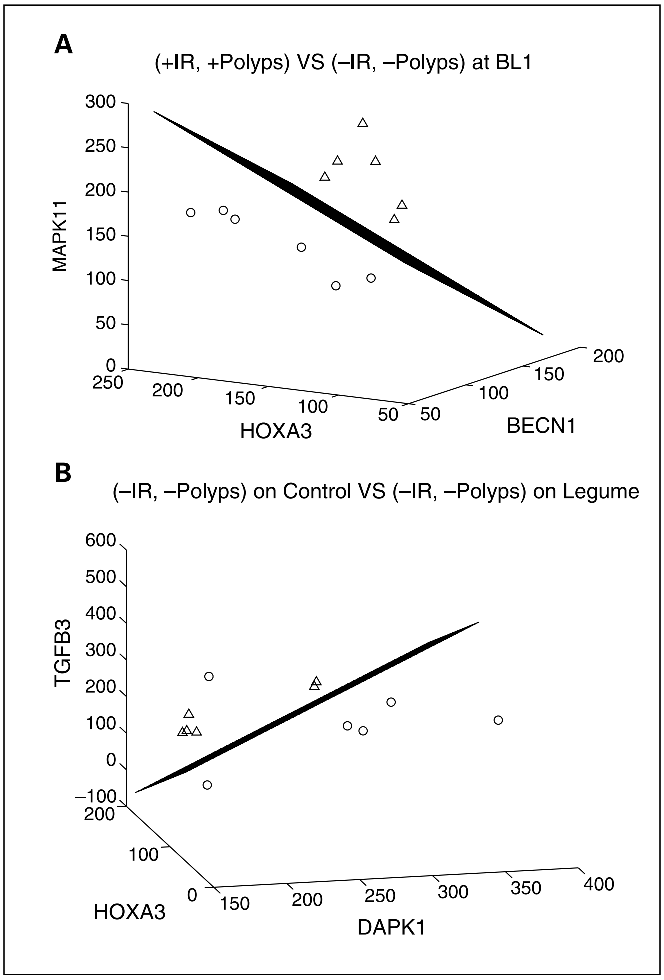


Figure 3. Effective classification of clinical phenotype or diet. *A*, linear (LDA) classification of (+IR, +Polyps) subjects (○) versus (−IR, −Polyps) subjects (△) at BL1; *B*, linear (LDA) classification of (−IR, −Polyps) subjects on the control diet (○) versus (−IR, −Polyps) subjects on the legume diet (△) using the crossover design and combining the microarrays from samples collected at the end of the two diet periods DP1 and DP2.

The results in Fig. 4 show that the two factors, IR and history of adenomas, should be considered in tandem when determining the risk for the patient. For example, combining baseline samples (BL1 and BL2) increased the classification error, indicating complications related to the crossover design (Fig. 4A). Similarly, the three-feature set LDA classifiers performed poorly when the classification was considered separately with respect to either one of the two experimental factors (IR) or (Polyps; Fig. 4B and C). The advantage of reporting the results in this way is that multivariate discriminatory power is revealed. This is clearly shown in Table 1 with regard to *HOXA3*. The gene did not appear on the single-gene list, indicating that the error of the respective classifier exceeded 0.3 ($\varepsilon_{\mathrm{bolstered}}$ = 0.4882). However, it appeared with *UCP2*, 14-3-3ζ (*YWHAZ*), insulin growth factor receptor-I (*IGF1R*), beclin-1 (*BECN1*), and mitogen-activated protein kinase-11 (*MAPK11*) genes in the two-gene and three-gene lists, which improved classification error. Interestingly, members of the homeoprotein family of transcription factors (*HOXA3* and *HOXC6*) are developmental regulators of gastrointestinal growth, patterning, and differentiation (35). It is also noteworthy that *YWHAZ* and *IGF1R* are capable of regulating apoptosis and cell adhesion (36, 37); *UCP2* promotes chemoresistance in cancer cells and mitochondrial Ca^{2+} sequestration (38, 39); *BECN1* stimulates autophagy and inhibits tumor cell growth (40); and *MAPK11* (*p38β*) mediates response to inflammatory cytokines and cellular stress (41). For comparative purposes, fold changes in select genes are presented in Supplementary Table S5.

Legumes and pulses are a rich source of fermentable dietary fibers, which are precursors to luminal butyrate (4). Butyrate has well-known anti-inflammatory and antineoplastic actions (6, 7). In addition, pulses have a low glycemic index (5). Some studies suggest that diets high in fiber and with a lower glycemic index may reduce risk of colorectal cancer and decrease inflammatory markers (4, 42, 43). Therefore, it was important to note that the approach applied in this study can be used to identify genes that are modulated by the consumption of a legume-rich diet (Table 2). Our data show that although transforming growth factor β (TGFβ), which plays a permissive role in cancer progression and wound repair (44, 45), is by itself a reasonable discriminator, when it is combined with HOXA3 and death-associated protein kinase (DAPK1), the error is significantly improved. These observations are worth noting in view of the fact that DAPK1 is an extremely pleiotropic molecule capable of influencing the propensity of cells to undergo autophagy (46). Moreover, it has been recently shown that dietary fiber (butyrate) can enhance TGFβ/Smad3-tumor suppressor signaling in the colon (47, 48). Considering that dietary legumes promote short-chain fatty acid production in the colonic lumen, it is probable that butyrate may have altered TGFβ expression. Clearly, additional studies are needed to elucidate the effect of legume consumption on TGFβ-dependent signaling.

The objective of this proof-of-principle study was to develop diagnostic gene sets for the noninvasive identification of different phenotypes. As opposed to using expression levels of either significantly increased or decreased genes, we applied novel mRNA-based noninvasive methods to identify

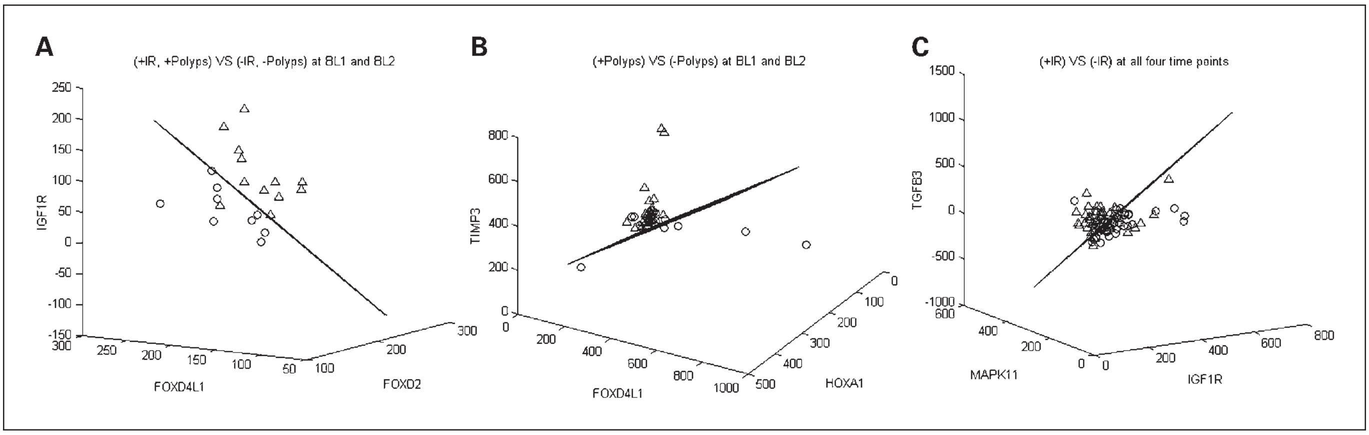


Figure 4. Potential design problems and importance of the experimental design factors IR and history of adenomas. *A,* increased error in the LDA classification of (+IR, +Polyps) subjects (○) versus (–IR, –Polyps) subjects (△) when both baselines BL1 and BL2 were included. *B,* (+Polyps) subjects (○) versus (–Polyps) subjects (△) at baselines BL1 and BL2. *C,* (+IR) subjects (○) versus (–IR) subjects (△) at all time points.

the best single genes and two- to three-gene combinations for distinguishing polyps, IR, and exposure to a chemoprotective legume-enriched diet. Similar to previous studies (20, 21, 32, 49), we report that by using combinations of genes, the classification error rate can be significantly lowered. Two- and three-gene combinations thus provide robust classifiers with potential to noninvasively identify discriminative molecular signatures for differential diagnostic purposes. These findings provide insight into a new paradigm and support the development of noninvasive methods using exfoliated colonocytes to quantify colonic mRNAs. This strategy can be a complementary, and likely useful, approach to enhance current efforts to define colon cancer risk. In addition, because of a lack of genomic precision in defining clinically relevant phenotypes, two- and three-gene combinations may have application in personalized genomic medicine (e.g., the stratification of patients according to response to risk of recurrence in trials of adjuvant treatment of the disease). Further studies are needed to validate the prognostic power and reliability of this molecular diagnostic approach.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Jennifer Goldsby (Genomics Core Facility, Texas A&M) for technical assistance.

References

1. Rutter MD, Saunders BP, Wilkinson KH, et al. Thirty-year analysis of a colonoscopics surveillance program for neoplasia in ulcerative colitis. Gastroenterology 2006;13:1030–8.
2. Bobe G, Barrett KG, Mentor-Marcel RA, Saffiotti U. Dietary cooked navy beans and their fractions attenuate colon carcinogenesis in azoxymethane-induced ob/ob mice. Nutr Cancer 2008;60:373–81.
3. Mentor-Marcel RA, Bobe G, Barrett KG, et al. Inflammation-associated serum and colon markers as indicators of dietary attenuation of colon carcinogenesis in *ob/ob* mice. Cancer Prev Res 2009;2:60–9.
4. Lanza E, Hartman TJ, Alberts PS, et al. High dry bean intake and reduced risk of advanced colorectal adenoma recurrence among participants in the polyp prevention trial. J Nutr 2006;136:1896–903.
5. Jenkins DJ, Wolever T, Jenkins AL. Starchy foods and glycemic index. Diabetes Care 1988;11:149–59.
6. Rose DJ, DeMeo MT, Keshavarzian A, Hamaker BR. Influence of dietary fiber on inflammatory bowel disease and colon cancer: importance of fermentation pattern. Nutr Rev 2007;65:51–62.
7. Bordonaro M, Lazarova DL, Sartorelli AC. Butyrate and Wnt signaling. Cell Cycle 2008;7:1178–83.
8. Potten CS, Schofield R, Lajtha LG. A comparison of cell replacement in bone marrow, testis and three regions of epithelium. Biochim Biophys Acta 1979;560:281–99.
9. Davidson LA, Jiang YH, Lupton JR, Chapkin RS. Non-invasive detection of putative biomarkers for colon cancer using fecal mRNA. Cancer Epidemiol Biomarkers Prev 1995;4:643–7.
10. Davidson LA, Aymond CM, Jiang YH, Turner ND, Lupton JR, Chapkin RS. Non-invasive detection of fecal protein kinase C βII and ζ messenger RNA: putative biomarkers for colon cancer. Carcinogenesis 1998;19:253–7.
11. Davidson LA, Lupton JR, Miskovsky E, Fields AP, Chapkin RS. Quantification of human intestinal gene expression profiles using exfoliated colonocytes: a pilot study. Biomarkers 2003;8:51–61.
12. Albaugh GP, Iyengar V, Lohani A, Malayeri M, Bala S, Nair P. Isolation of exfoliated colonic epithelial cells, a novel, non-invasive approach to the study of cellular markers. Int J Cancer 1992;52:347–50.
13. Santiago ML, Bibollet-Roche F, Bailes E, et al. Amplification of a complete simian immunodeficiency virus genome from fecal RNA of a wild chimpanzee. J Virol 2003;77:2233–42.
14. Kanaoka S, Yoshida KI, Miura N, Sugimura H, Kajimura M. Potential usefulness of detecting cyclooxygenase 2 messenger RNA in feces for colorectal screening. Gastroenterology 2004;127:422–7.
15. Davidson LA, Nguyen DV, Hokanson RM, et al. Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. Cancer Res 2004;64:6797–804.
16. Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends Genet 2003;19:362–5.
17. Dougherty ER. Small sample issues for microarray-based classification. Comp Funct Genomics 2001;2:28–34.
18. Attoor SN, Dougherty ER. Classifier performance as a function of distributional complexity. Pattern Recognit 2004;37:1629–40.
19. Braga-Neto UM, Dougherty ER. Bolstered error estimation. Pattern Recognit 2004;37:1267–81.
20. Sima C, Dougherty ER. What should be expected from feature selection insmall-sample settings. Bioinformatics 2006;22:2430–6.
21. Kim S, Dougherty ER, Shmulevich I, et al. Identification of combination gene sets for Glioma classification. Mol Cancer Ther 2002;1:1229–36.
22. Davies RJ, Miller R, Coleman N. Colorectal cancer screening: prospects for molecular stool analysis. Nat Rev Cancer 2005;5:199–209.
23. Imperiale TF, Ransohoff DF, Itzkowitz SH, Turnbull BA, Ross ME. Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average-risk population. N Engl J Med 2004;351:2704–14.
24. Itzkowitz SH, Jandorf L, Brand R, et al. Improved fecal DNA test for colorectal cancer screening. Clin Gastroenterol Hepatol 2007;5:111–7.
25. Ahlquist DA, Sargent DJ, Loprinzi CL, et al. Stool DNA and occult blood testing for screen detection of colorectal neoplasia. Ann Intern Med 2008;149:441–50.

26. Zou H, Taylor WR, Harrington JJ, et al. High detection rates of colorectal neoplasia by stool DNA testing with a novel digital melt curve assay. Gastroenterology 2008;136:459–70.

27. Esteller M. Epigenetics and cancer. N Engl J Med 2008;358:1148–59.

28. Chen WD, Han ZJ, Skoletsky J, et al. Detection of fecal DNA of colon cancer-specific methylation of the nonexpressed vimentin gene. J Natl Cancer Inst 2005;97:1124–32.

29. Zou H, Harrington J, Rego RL, Ahlquist DA. A novel method to capture methylated human DNA from stool: implications for colorectal cancer screening. Clin Chem 2007;53:1646–51.

30. Jiang YH, Lupton JR, Chapkin RS. Dietary fish oil blocks carcinogen-induced down-regulation of colonic protein kinase C isozymes. Carcinogenesis 1997;18:351–7.

31. Ahlquist DA, Skoletsky JE, Boynton KA, et al. Colorectal cancer screening by detection of altered human DNA in stool: feasibility of a multitarget assay panel. Gastroenterology 2000;119:1219–27.

32. Kobayashi T, Yamaguchi M, Kim S, et al. Gene expression profiling identifies strong feature genes that classify *de novo* CD5$^+$ and CD5$^-$ diffuse large B-cell lymphoma and Mantle cell lymphoma. Cancer Res 2003;63:60–6.

33. Potti A, Dressman HK, Bild A, et al. Genomic signatures to guide the use of chemotherapeutics. Nat Med 2006;11:1294–300.

34. Martins D, Braga-Neto U, Hashimoto R, Bittner ML, Dougherty ER. Intrinsically multivariate predictive genes, IEEE. J Select Topics Signal Processing 2008;2:424–39.

35. Fujiki K, Duerr E, Kikuchi H, et al. Hoxc6 is overexpressed in gastrointestinal carcinoids and interacts with JunD to regulate tumor growth. Gastroenterology 2008;135:907–16.

36. Sekharam M, Zhao H, Sun M, et al. Insulin-like growth factor 1 receptor enhances invasion and induces resistance to apoptosis of colon cancer cells through the Akt/Bcl-x$_L$ pathway. Cancer Res 2003; 63:7708–16.

37. Niemantsverdriet M, Wagner K, Visser M, Backendorf C. Cellular functions of 14-3-3ζ in apoptosis and cell adhesion emphasize its oncogenic character. Oncogene 2008;27:1315–9.

38. Trenker M, Malli R, Fertschai I, Levak-Frank S, Graier WF. Uncoupling proteins 2 and 3 are fundamental for mitochondrial Ca^{2+} uniport. Nat Cell Biol 2007;9:445–52.

39. Derdak Z, Mark NM, Beldi G, Robson SC, Wands JR, Baffy G. The mitochondrial uncoupling protein-2 promotes chemoresistance in cancer cells. Cancer Res 2008;68:2813–19.

40. Pattingre S, Espert L, Biard-Piechaczyk M, Codogno P. Regulation of macroautophagy by mTOR and Beclin 1 complexes. Biochimie 2008; 90:313–23.

41. Beardmore VA, Hinton HJ, Eftychi C, et al. Generation and characterization of p38β (MAPK11) gene-targeted mice. Mol Cell Biol 2005;25: 10454–64.

42. Lin J, Zhang SM, Cook NR, et al. Dietary intakes of fruit, vegetables, and fiber, and risk of colorectal cancer in a prospective cohort of women (United States). Cancer Causes Control 2005; 16:225–33.

43. Michels KB, Giovannucci E, Chan AT, Singhania R, Fuchs CS, Willett WC. Fruit and vegetable consumption and colorectal adenomas in the Nurses' Health Study. Cancer Res 2006;66:3942–53.

44. McKaig BC, Makh SS, Hawkey CJ, Podolsky DK, Mahida YR. Normal human colonic subepithelial myofibroblasts enhance epithelial migration (restitution) via TGF-β3. Am J Physiol 1999;276: G1087–93.

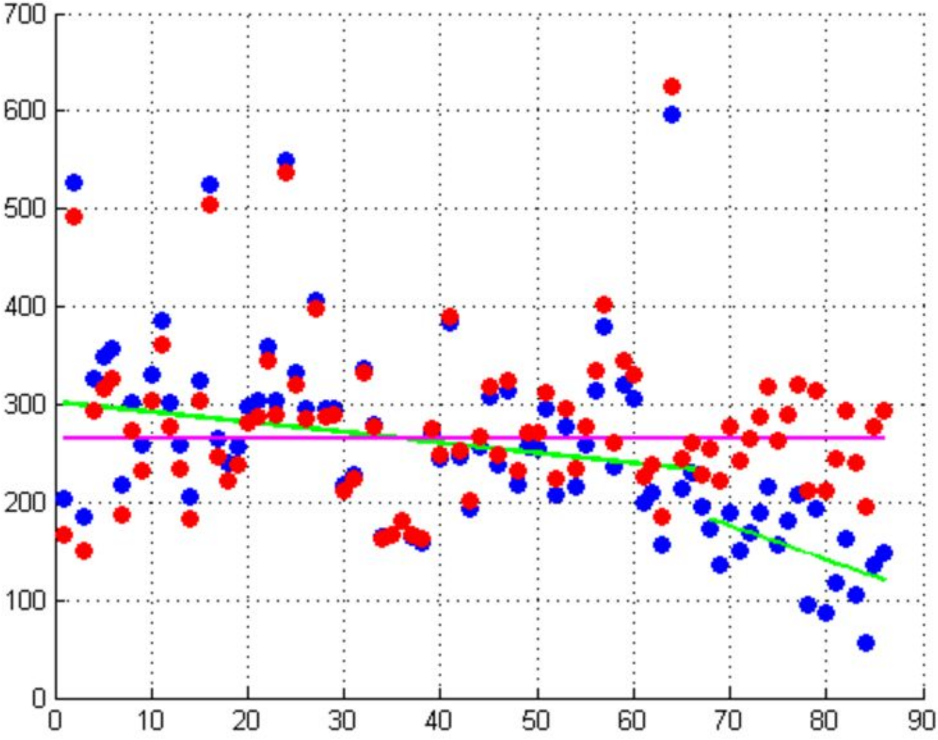
45. Bellone G, Carbone A, Tibaudi D, et al. Differential expression of transforming growth factors-β1, -β2 and -β3 in human colon carcinoma. Eur J Cancer 2001;37:224–33.

46. Maiuri MC, Tasdemir E, Criollo A, et al. Control of autophagy by oncogenes and tumor suppressor genes. Cell Death Differ 2009;16:87–93.

47. Nguyen KA, Cao Y, Chen JR, Townsend CM, Ko TC. Dietary fiber enhances a tumor suppressor signaling pathway in the gut. Ann Surg 2006; 243:619–27.

48. Daniel C, Schroder O, Zahn N, Gaschott T, Steinhilber D, Stein JM. The TGFβ/Smad 3-signaling pathway is involved in butyrate-mediated vitamin D receptor (VDR)-expression. J Cell Biochem 2007;102:1420–31.

49. Morikawa J, Li H, Kim S, et al. Identification of signature genes by microarray for acute myeloid leukemia without maturation (FAB-M1) and AML with t(15;17)(q22;q12)(*PML/RARα*). Int J Oncol 2003;23:617–25.


700
600
500
400
300
200
100
0
0
10
20
30
40
50
60
70
80
90

Supplemental Table 1. *Overall structure of the microarray data set.*

x = array was processed; m=missing sample

L=Legume diet;
C=American control diet

Subject ID	Study Group	BL 1	End DP 1	BL 2	End DP 2	Subject ID	DP1	DP2
LEG 01	3	x	x	x	x	LEG 01	L	C
LEG 02	2	x	x	x	x	LEG 02	C	L
LEG 03	1	x	x	x	x	LEG 03	C	L
LEG 04	3	x	x	x	x	LEG 04	L	C
LEG 05	2	x	x	x	x	LEG 05	L	C
LEG 06	4	x	x	x	x	LEG 06	C	L
LEG 08	3	x	x	x	x	LEG 08	C	L
LEG 09	2	x	x	x	x	LEG 09	L	C
LEG 10	2	x	x	m	x	LEG 10	C	L
LEG 11	4	x	x	x	x	LEG 11	L	C
LEG 13	1	x	x	m	x	LEG 13	C	L
LEG 14	3	x	x	x	x	LEG 14	C	L
LEG 18	3	x	x	x	x	LEG 18	L	C
LEG 19	4	x	x	x	x	LEG 19	C	L
LEG 24	4	x	x	x	x	LEG 24	L	C
LEG 26	4	x	x	x	x	LEG 26	C	L
LEG 27	4	x	x	x	x	LEG 27	L	C
LEG 33	2	x	x	m	x	LEG 33	C	L
LEG 44	1	x	x	m	x	LEG 44	L	C
LEG 47	1	x	x	x	x	LEG 47	C	L
LEG 49	1	x	x	m	x	LEG 49	L	C
LEG 54	1	x	x	x	m	LEG 54	L	C
LEG 65	3	x	x	x	x	LEG 65	C	L

Study Group : 1 = + insulin resistance/ + polyps;   2 = - insulin resistance/ + polyps

3 = + insulin resistance/ - polyps;   4 = - insulin resistance/ - polyps

*Refer to Figure 1 for details

Supplemental Table 2. Final classifier gene list. Refer to attached 529 genes - XLS file.

Supplemental Table 3. $A_j^k \cap B$ represents the number of genes that are common between the set B of established colonic biomarkers and the spots A_j^k on the microarray set that passed quality threshold set by the parameters k and j. The value k=1.5 is the default value for the CodeLink image processing software, and j represents the number of accepted low (L) spots for a gene across all of the microarrays in the experiment.

$A_j^k \cap B$	*k = 1.5*	*k = 2*	*k = 2.5*	*k = 3*
j = 0	50	36	23	10
j = 1	65	54	35	18
j = 2	84	61	46	29
j = 3	94	70	51	37

Supplemental Table 4. Classification groups, sample size and number of common genes in each data set. BL1, baseline 1; BL2, baseline 2; +IR and –IR indicate presence or absence of insulin resistance, respectively. +Polyps and –polyps indicate the presence or absence of polyps, respectively.

Classification Groups	*Sample Size*	*Common Genes in* $A_1^2 \cap B$
(+IR, +Polyps) VS (-IR, -Polyps) at BL1	12	97
(+IR, +Polyps) on Control VS (+IR, +Polyps) on Legume	11	103
(-IR, -Polyps) on Control VS (-IR, -Polyps) on Legume	12	145
(+IR, +Polyps) on Control VS (-IR, -Polyps) on Control	11	121
(+IR, +Polyps) on Legume VS (-IR, -Polyps) on Legume	12	114
(+IR, +Polyps) VS (-IR, -Polyps) at BL1 & BL2	21	92
(+Polyps) VS (-Polyps) at BL1	23	64
(+IR) VS (-IR) at BL1	23	64
(+Polyps) VS (-Polyps) at BL1 & BL2	41	59
(+Polyps) on Control VS (+Polyps) on Legume	21	87
(+IR) on Control VS (+IR) on Legume	23	74
(+IR) VS (–IR) at all time points	86	54

Supplemental Table 5. Relative exfoliated cell gene expression levels in (+IR, +Polyps) vs (-IR, -Polyps) subjects at baseline 1 (BL1). Fold change represents the relative expression level in (+IR, +Polyps) subjects divided by (–IR, -Polyps) subjects for individual genes described in Table 1. p-values were computed using t-tests applied to the normalized data.

Gene name	*p-value*	*Fold change*
ALOX12B	0.1841	0.6486
BECN1	0.0580	0.5140
CDK4	0.0370	0.5787
DAPK1	0.0639	1.1258
HOXA3	0.0202	1.0712
HOXC6	0.0134	0.4352
ID2	0.0626	0.9413
IGF1R	0.0040	0.4537
MAPK11	0.6291	0.7521
NOS3	0.0285	0.4451
TJP1	0.0168	0.6092
UCP2	0.6330	0.7669
WNT1	0.7147	0.8290
YWHAZ	0.0298	0.4901

Supplemental Methods:

Data Normalization. Two normalization issues were addressed. First, there was a large number of low-quality spots and second, while the microarray intensities showed no aberrant trend up to a certain point in time (relative to when microarray was performed), after a certain point there was a somewhat linear decline in intensity. Data points (blue dots) in **Supplemental Figure 1** show the average values of the 18 housekeeping genes across microarrays, ordered from earliest to latest with respect to the time of processing.

Development of an Algorithm for Identifying Feature (Gene) Sets. We first examined how the parameters used by the CodeLink scanning software affected the number of G spots on the arrays. Specifically, genes denoted by A_j^k, i.e., the set of genes x_i that have at most j raw mean spot intensity values less than $\mu_{i,l} + k\sigma_{i,l}$, where $\mu_{i,l}$ is the value of local background median for the spot representing the gene x_i on the l-th array, and $\sigma_{i,l}$ is the corresponding standard deviation for that background signal, were identified. For example, $A_0^{1.5}$ is the set of (G) spots that are common for all of the arrays in the data set (by default $k = 1.5$ in the CodeLink software). Spots that were flagged as (C) were not considered when the sets A_j^k were formed. Notice that $A_j^k \subseteq A_r^s$ if $s \le k$ and $j \le r$ (s and r are defined similarly to k and j). In particular, $A_j^k \subseteq A_j^s$, $s \le k$ indicates a lower number of common good spots if one requires stronger signal, compared to the background. Also, $A_j^k \subseteq A_r^k$, $j \le r$ demonstrates that the number of common genes increases if one allows more (L) spots per gene.