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.