Am J Physiol Gastrointest Liver Physiol 298: G582–G589, 2010.
First published March 4, 2010; doi:10.1152/ajpgi.00004.2010.

TRANSLATIONAL PHYSIOLOGY

Noninvasive stool-based detection of infant gastrointestinal development using gene expression profiles from exfoliated epithelial cells

Robert S. Chapkin,[1,2] Chen Zhao,[3] Ivan Ivanov,[2,4] Laurie A. Davidson,[1,2] Jennifer S. Goldsby,[1,2] Joanne R. Lupton,[1,2] Rose Ann Mathai,[5] Marcia H. Monaco,[5] Deshanie Rai,[6] W. Michael Russell,[6] Sharon M. Donovan,[5] and Edward R. Dougherty[2,3]

[1]*Program in Integrative Nutrition and Complex Diseases,* [2]*Center for Environmental and Rural Health, and Departments of* [3]*Electrical Engineering and* [4]*Veterinary Physiology and Pharmacology, Texas A & M University, College Station, Texas;* [5]*Division of Nutritional Sciences, University of Illinois, Urbana-Champaign, Illinois; and* [6]*Mead Johnson Nutrition, Evansville, Indiana*

Submitted 6 January 2010; accepted in final form 25 February 2010

Chapkin RS, Zhao C, Ivanov I, Davidson LA, Goldsby JS, Lupton JR, Mathai RA, Monaco MH, Rai D, Russell WM, Donovan SM, Dougherty ER. Noninvasive stool-based detection of infant gastrointestinal development using gene expression profiles from exfoliated epithelial cells. *Am J Physiol Gastrointest Liver Physiol* 298: G582–G589, 2010. First published March 4, 2010; doi:10.1152/ajpgi.00004.2010.—We have developed a novel molecular methodology that utilizes stool samples containing intact sloughed epithelial cells to quantify intestinal gene expression profiles in the developing human neonate. Since nutrition exerts a major role in regulating neonatal intestinal development and function, our goal was to identify gene sets (combinations) that are differentially regulated in response to infant feeding. For this purpose, fecal mRNA was isolated from exclusively breast-fed ($n = 12$) and formula-fed ($n = 10$) infants at 3 mo of age. Linear discriminant analysis was successfully used to identify the single genes and the two- to three-gene combinations that best distinguish the feeding groups. In addition, putative "master" regulatory genes were identified using coefficient of determination analysis. These results support our premise that mRNA isolated from stool has value in terms of characterizing the epigenetic mechanisms underlying the developmentally regulated transcriptional activation/repression of genes known to modulate gastrointestinal function. As larger data sets become available, this methodology can be extended to validation and, ultimately, identification of the main nutritional components that modulate intestinal maturation and function.

systems biology; breast feeding; biomarkers; microarray

TRANSLATIONAL HIGHLIGHTS *Given the need to better understand gastrointestinal health and development, the use of noninvasive tests is expected to become increasingly relevant tools in tailoring diet to meet the nutritional needs of the growing infant. Chapkin and colleagues developed a novel molecular methodology that utilizes stool samples containing intact sloughed epithelial cells, to noninvasively quantify intestinal gene expression profiles in the developing human neonate. Linear discriminant analysis was successfully used to identify the best single genes and two- to three-gene combinations for distinguishing the feeding groups.*

Address for reprint requests and other correspondence: R. S. Chapkin, Texas A & M Univ., 218 Kleberg Center, 2253 TAMU, College Station, TX 77843-2253 (e-mail: r-chapkin@tamu.edu).

Gastrointestinal (GI) maturation involves a continuous cascade of growth, differentiation, and renewal of epithelial cells. In humans, the intestinal tract is functionally immature and immunologically naïve at birth (18). Evidence from animal models (48) and human infants (7, 12) demonstrates that the intestinal tract of the newborn undergoes marked structural and functional adaptation in response to feeding. The trophic response to human milk exceeds that to formula, suggesting that bioactive components in human milk are important in this response (19, 20). Nutrition during early postnatal development can modulate intestinal development in at least three ways: *1*) directly through stimulation of transcription and/or translation by milk-derived nutrients (5), *2*) indirectly through modulation of the intestinal microbiota (9) or the mucosal immune system (29), and/or *3*) through induction of maturation of gene-specific methylation, thereby permanently altering the expression of developmental genes in the intestine (46, 47). Therefore, it is essential to determine the effects of postnatal nutrition on intestinal global gene expression profiles in the infant. However, because of the ethical constraints associated with obtaining tissue biopsies from healthy infants, no investigators have comprehensively profiled intestinal gene expression during early postnatal development.

Recent interest has focused on exfoliated cells as a tool to investigate the impact of therapeutic and nutritional regimens on the maturation of GI functions (17, 30). Approximately one-sixth to one-third of normal adult colonic epithelial cells are shed daily (37). This number corresponds to the daily exfoliation of $\sim 10^{10}$ cells. Because the number of intact cells that can be isolated from fecal material is low (1, 14), we have developed a patented noninvasive mRNA-based method as a highly sensitive technique for detecting molecular markers of intestinal development and function. This methodology has the advantage of using exfoliated cell mRNA directly isolated from feces, which contain sloughed small intestinal and colon cells, and, therefore, does not result in any discomfort to the subject. The method is capable of isolating and quantifying specific mRNAs under various intestinal conditions, including mode of feeding, and has been tested in a rat experimental colon carcinogenesis model (8, 13) and in adult humans (15, 51). We propose that the ability to use exfoliated cell mRNA, instead of biopsy or autopsy material, would be highly advan-

tageous to document the impact of nutrition on the continuum of intestinal development and maturation in the healthy newborn. Therefore, in this study, we noninvasively examined intestinal transcriptional profiles in infants exclusively fed formula or breast milk *1*) to validate that this technique can be applied to infant fecal samples and *2*) to identify gene sets (combinations) in response to mode of feeding, i.e., human milk or infant formula. We demonstrate for the first time that two- and three-gene combinations provide classifiers with potential to noninvasively identify discriminative signatures for the development of molecular markers to explore nutritional effects on maturation of intestinal function. In addition, we utilized complementary systems biology approaches, e.g., multivariate measurement of gene expression relationships, to identify "master" genes, which regulate the transcriptional states of other "slave" genes in the infant intestine.

METHODS

Subject Recruitment and Inclusion and Exclusion Criteria

Mothers of infants were recruited into the study between the third trimester of pregnancy and 1 mo postpartum. Healthy, full-term infants, who were exclusively breast-fed (BF) or fed commercially available infant formula (FF; Enfamil LIPIL, Mead Johnson Nutrition, Evansville, IN) and medically certified as healthy, i.e., asymptomatic and with no clinical indication of disease, were eligible for enrollment into the study. Infants who were diagnosed with intolerance to cow's milk or who were receiving both breast milk and formula, soy or casein hydrolysate formula, juice, or solid foods were excluded from the study. Enrolled infants who became clinically ill (fever, contagious diseases, or active diarrhea) or had received antibiotic treatment within 2 wk of the sample collection were excluded from the study. Mothers of the FF infants were provided with Enfamil LIPIL formula for the duration of the study, and mothers of the BF infants received a $25.00 gift card for each stool sample collection. Fecal samples were coded with a unique numerical identification to maintain confidentiality of the participants. The experimental protocol was approved by the University of Illinois and Texas A & M Institutional Review Boards, and informed consent was obtained from parents prior to participation in the study.

Sample Collection

Stool specimens were collected by the parent from exclusively BF and FF infants at 3 mo postpartum. Instructions on sample collection were provided to the parents. A sterile spoon was used to place ~10 g of freshly voided fecal material into a sterile 50-ml conical tube containing 20 ml of guanidinium denaturation solution (Ambion, Austin, TX). Samples were mixed by hand to create a homogenous sample, which was immediately frozen at −20°C until it was transported on ice to the laboratory by the research staff. Samples were held at −80°C until they were shipped on dry ice to Texas A & M University for analysis.

Breast Milk and Formula Intake

For a full 24-h period preceding the stool sample collection, parents weighed the infant before and after each feeding on an electronic scale (Medela, McHenry, IL) without a change of clothing or diaper. The change in body weight was determined as an estimate of breast milk or formula intake. All data are presented as means ± SD. Maternal age and infant birth weight and length were compared using a Student's *t*-test (SAS version 6.0, SAS, Cary, NC). Differences in breast milk and formula intake, Z scores, and body weight at 1, 2, and 3 mo of age were determined using a repeated-measures ANOVA (SAS). Statistical significance was set at $P \leq 0.05$.

mRNA Expression Microarray Analysis

From each subject, polyA$^+$ RNA was isolated from feces as previously described (15). Because of the high level of bacterial RNA in fecal samples, polyA$^+$ RNA was isolated to obtain a highly enriched mammalian polyA$^+$ RNA population (14). In addition, an Agilent 2100 Bioanalyzer was used to assess integrity of fecal polyA$^+$ RNA, and quantification was performed by spectrophotometer (NanoDrop, Wilmington, DE). Samples were processed in strict accordance to the CodeLink Gene Expression Assay manual (Applied Microarray, Tempe, AZ) and analyzed using the Human Whole Genome Expression Bioarray, as we previously described (16, 51). 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, irregularly shaped, or saturated in the output of the scanning software. Spots that passed the quality-control standards were categorized as good (G). In addition, a reading of L indicated "near background." The low-L measurements reflect 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 (16) exhibit a relatively low proportion (30–45%) of L spots. In comparison, we previously reported that the proportion of L spots obtained from adult fecal samples is significantly higher (65–83%) (51). In the present study, the proportion of L spots was 45–77%; therefore, we performed statistical and classification analyses using only the common G spots (4,250) for all 22 samples.

Microarray Data Normalization

For the purpose of interarray normalization, a set of housekeeping genes was used. These were determined as follows.

Housekeeping gene preparation. Common G probes (4,250) across all 22 microarrays were identified. Using a list of 575 housekeeping genes (24), we identified 33 housekeeping genes from the 4,250 common G probes found in the previous step (see supplemental methods, supplemental Fig. 1, and supplemental Table 1 in the online version of this article).

Additive normalization procedure. Arrays were grouped across the type of feeding, and the average values of the 33 housekeeping genes were calculated (see supplemental Fig. 1). Median values of the averages were also calculated. Subsequently, a robust piecewise linear regression was performed, and the corresponding regression value for each array was calculated. Then the difference between the median and regression values for each array was determined, and the raw expression values of the common 4,250 genes on each array were shifted by the corresponding discrepancies.

Identifying Multivariate Discriminators (Feature Gene Sets) for Diet Classification

We used a previously described algorithm for feature set identification (51; also see supplemental methods). Estimation of the classification error is of critical importance when the number of potential feature sets is large. When sample size is 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. We mitigate this problem by applying bolstered error estimation (3). This procedure places a kernel (density) at each data point and computes the error by integrating the kernels over their misclassification regions, rather than simply by counting incorrectly classified points, as is done in resubstitution error estimation, thereby giving more weight to points near the classification boundary (see supplemental material for details on bolstering). Bolstered error estimation performs especially well compared with other error estimation methods in ranking feature sets, which was important in this analysis (41). The bolstered error estimated can be computed analytically in some cases, such as linear discriminant

analysis (LDA; see supplemental material). Using prior knowledge consisting of a set of 529 biomarkers (genes) known to be involved in intestinal biology (51), we identified 146 potential features (genes) that were biomarkers and exhibited a G-flagged expression value for all the arrays. The relatively small size of this gene set allows for the comparison of the errors of all the possible single-, two-, and three-gene feature sets, thereby avoiding feature selection, which can be highly unreliable in small sample settings (42). The result of the overall approach is a list of "best" feature sets among all possible feature sets, i.e., those possessing minimum classification error.

Identifying Potential Master and Slave Genes

Mathematical modeling of "master-slave" relationships in a gene regulatory network was originally described by Dougherty et al. (22). While examining a portion of a gene regulatory network, one can view genes at various positions in that cascade as regulators/masters or regulated/slaves. This is a relative characterization: in certain situations, a gene might act as a master; in other situations, it might act as a slave. The intuitive notion of regulatory cascade can be formalized using the coefficient of determination (CoD) (22, 31). The CoD measures the relative increase in prediction accuracy by using the optimal predictor for the target based on the predictors compared with the best estimate of the target in the absence of predictors, i.e., the increase in prediction accuracy vs. only knowledge of the target distribution. The CoD has previously been used to measure multivariate nonlinear gene interaction (22, 31) and is defined as follows: $(\varepsilon_0 - \varepsilon.)/\varepsilon_0$, where $\varepsilon.$ and ε_0 are the errors of the optimal predictor and the predictor for the target using only statistics relating to the target itself, respectively. The key concept in defining master genes is the expectation that this type of regulator, when expressed, will be responsible for activating one or more cascades of subsequent changes in the expression of other genes. Thus one can achieve a more accurate prediction of the state of a master gene in a specific microarray sample if inferences are based on the observation of many of its slaves, rather than only on the behavior of a single slave. This modeling approach postulates probabilistic relations between masters and slaves and derives the CoD for the master relative to the slaves from the postulated relations (21). The validation of the model was performed using expression profiles from a series of 40 cell lines from a study of the response of benchmark cancer cell lines, the National Cancer Institute 60 Anti-Cancer Drug Screen cell line (27). The results suggest that the shapes of the histograms of the CoD values, when all the possible triples of genes are used as predictors of a given gene, can identify both master and slave genes in a particular cell context. In general, CoD histograms for master genes are skewed toward larger (closer to 1) CoD values, while the CoD histograms for the slaves tend to exhibit the opposite behavior (skewed toward 0) (21, 32).

RESULTS

Subject Demographics, Birth Weight, and Length

A total of 12 mothers of BF infants and 10 mothers of FF infants were recruited for the study. There were no differences in mean age or parity between mothers of BF and FF infants (Table 1). The sex distribution of the infants was similar in both groups, and most infants enrolled in the study were Caucasian. Birth length and weight were similar in both groups (Table 1).

Breast Milk and Formula Intake

Breast milk and formula intake ($ml \cdot kg^{-1} \cdot day^{-1}$) were unaffected by sex but differed over time ($P = 0.001$) and between diets ($P = 0.038$). Intake was affected by an interaction of feeding mode and time ($P = 0.004$). BF infant intake decreased between 1 and 2 mo of age and then remained stable (Table 1). FF infant intake was similar at 1, 2, and 3 mo of age. Intake differed between BF and FF infants at 2 mo of age, with FF infants having significantly higher intake per kilogram of body weight than BF infants ($P = 0.007$).

Table 1. *Demographics and growth of study subjects*

	Breast-Fed	Formula-Fed
Sample size	12	10
Maternal age, yr	28.0 ± 5.4	30.5 ± 5.0
Parity	2.0 ± 0.0	2.1 ± 0.6
Infant sex	8 male, 4 female	7 male, 3 female
Infant ethnicity	10 Caucasian, 1 Hispanic, 1 African-American	9 Caucasian, 1 African-American
Length at birth, cm	52.7 ± 3.6	51.0 ± 3.1
Body weight, kg		
Birth	3.8 ± 0.6*	3.5 ± 0.6*
1 mo	4.9 ± 0.7†	4.6 ± 0.7†
2 mo	5.9 ± 0.7‡	5.7 ± 0.8‡
3 mo	6.5 ± 0.7§	6.4 ± 0.9§
Breast milk or formula intake, $ml \cdot kg^{-1} \cdot day^{-1}$		
1 mo	168.8 ± 36.4*	175.4 ± 30.0*
2 mo	128.0 ± 20.6*	162.9 ± 43.2*†
3 mo	123.0 ± 35.7†	141.4 ± 14.2†

Values are means ± SD. Symbols (*,†‡§) indicate significant difference at $P \leq 0.001$.

Postnatal Growth

Body weights (kg) of BF and FF infants at birth and 1, 2, and 3 mo of age are shown in Table 1. Body weight was significantly different across all time points, but no differences were detected between males and females or by feeding method. To compare the growth of infants in the study with reference standards, weight-for-age Z scores for individual infants were calculated in reference to the 2000 Centers for Disease Control growth standards for male and female infants (www.cdc.gov/GrowthCharts/). Z scores describe how far a child's weight is from the average weight of a child of the same age in the reference data. It is expressed in multiples of the standard deviation, with 0 reflecting the 50th percentile. Weight-for-age of all infants was within +2 to −1 Z scores, supporting normal growth and good nutritional status of the study population.

Classification of Diet

Although our primary goal was to validate this methodology for use in infants, we also wanted to identify if mode of feeding differentially influenced mRNA expression patterns. We applied a linear discriminator algorithm initially described by Zhao et al. (51). The number of genes (features) for each linear classifier was limited to 3, which allowed for an exhaustive search and, thus, avoided errors associated with small sample setting feature extraction. We identified the 10 best single-, two-, and three-gene classifiers to distinguish between the two groups of infants at 3 mo of age (Table 2). The results show several cases where single genes can provide good (in terms of the error estimate) classification (Table 2; see supplemental Table 2). However, when considering these features as part of two- or three-gene classifiers, we observed a significant decrease in the classification error. A similar phenomenon was recently documented in the context of gene network modeling (35). Specifically, the target (a gene or a phenotype) was

Table 2. *Classification of breast-fed vs. formula-fed infants*

Gene			$\varepsilon_{bolstered}$	$\Delta\varepsilon_{bolstered}$
EPAS1			0.1214	
NR5A2			0.1356	
NR3C1			0.1364	
PCDH7			0.1367	
ITGB2			0.1374	
FGF5			0.1385	
TJP1			0.1400	
MYB			0.1456	
EPIM			0.1478	
BAD			0.1496	
EPAS1	UCP2		0.0869	0.0345
CTDSPL	NR3C1		0.0975	0.0381
NR3C1	TNFRSF10B		0.0987	0.0377
FOXP4	NR3C1		0.1026	0.0341
CDK4	EPAS1		0.1039	0.0335
EPAS1	SYP		0.1045	0.0340
NR3C1	SLC26A2		0.1057	0.0343
GPR41	TJP1		0.1060	0.0396
FOXP1	NR3C1		0.1064	0.0414
HSPA1A	NR3C1		0.1064	0.0432
EPAS1	FOXE3	SYP	0.0778	0.0091
CTDSPL	FOXE3	NR3C1	0.0785	0.0190
EPAS1	TLR5	UCP2	0.0785	0.0202
EPAS1	REG4	UCP2	0.0802	0.0224
EPAS1	LIFR	UCP2	0.0808	0.0231
EPAS1	NODAL	UCP2	0.0821	0.0224
EPAS1	HIF3A	UCP2	0.0823	0.0234
EPAS1	HOXD10	UCP2	0.0823	0.0237
EPAS1	KIT	UCP2	0.0824	0.0240
ALOX5	EPAS1	UCP2	0.0827	0.0237

Top 10 one-, two-, and three-gene linear discriminant analysis classifiers are shown. $\varepsilon_{bolstered}$, bolstered resubstitution error for the respective classifier (classifiers are ranked according to that error measurement); $\Delta\varepsilon_{bolstered}$, decrease in error for each feature set relative to its highest ranked subset of features. (See supplemental Tables 2–4 for a complete listing of classifiers based on 1-, 2-, and 3-gene sets.)

predicted with greater accuracy by the expression profiles of a group of genes than by any proper subset of these genes (Table 2; see supplemental Tables 2–4). For example, the best single-gene classifier (1-feature) based on endothelial PAS domain-containing protein (EPAS1; ranked 1 in the list of single-gene classifiers) had an estimated classification error of 0.1214. Interestingly, conbination of this gene with a poorly performing single-gene classifier, uncoupling protein 2 (UCP2; ranked 33 in the list of single-gene classifiers), resulted in a two-gene classifier with a much lower estimated error of 0.0869 (Table 2). Moreover, UCP2 appeared frequently in the 10 best three-gene classifiers. These data clearly illustrate why complex phenotypes can be explained better by multivariate feature sets. 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: $\varepsilon_{bolstered}$ and $\Delta\varepsilon_{bolstered}$. $\varepsilon_{bolstered}$ denotes the bolstered resubstitution error for the LDA classifier for the respective feature set of size n, and $\Delta\varepsilon_{bolstered}$ denotes the decrease in error with respect to the highest ranked of its subsets of features (in the list of features of size $n - 1$, $n = 2,3$). The feature sets were ranked on the basis of the value of $\varepsilon_{bolstered}$. Figures 1 and 2

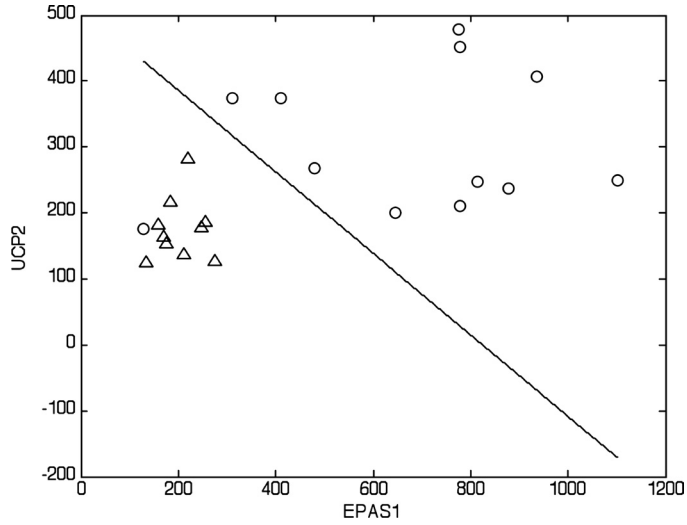


Fig. 1. Linear discriminant analysis (LDA) phenotype classification using the genes endothelial PAS domain-containing protein 1 (hypoxia-inducible factor 2α, EPAS1) and uncoupling protein 2 (UCP2) provides the best-performing 2-gene feature set. Classification is between breast-fed (○) and formula-fed (△) infants (see Table 2 for additional details). Axes represent normalized intensity values of the indicated genes. Note clear separation between the groups, except for 1 outlier.

show LDA classifiers for the best two- and three-gene feature sets listed in Table 2. The segregation of the two groups of infants can be readily seen in Figs. 1 and 2.

To evaluate how this classification approach relates to differential expression, we compared the 10 best one-feature sets and the genes showing differential expression between the two groups at $P < 0.05$ (Table 3), where *t*-tests are performed using normalized and logarithmically transformed gene intensity values. The comparison revealed that 9 of the 10 best one-feature (gene) sets identified by the linear (LDA) classifier also have $P < 0.05$. This is not surprising, because individual differentially expressed genes have been traditionally used to discriminate between phenotypes (38). We emphasize that these P values are only for comparison purposes; therefore, the cutoff of 0.05 has no bearing on classification in this study.

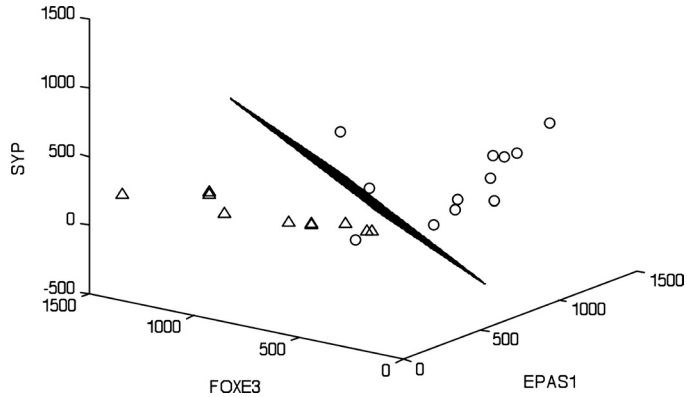


Fig. 2. LDA phenotype classification using the genes EPAS1, forkhead box protein E3 (FOXE3), and synaptophysin (SYP) provides the best-performing 3-gene feature set. The 3-dimensional LDA hyperplane discriminates between breast-fed (○) and formula-fed (△) infants (see Table 2 for additional details). Axes represent normalized intensity values of the corresponding genes. Note clear separation between the groups, except for 1 outlier (the same infant outlier in Fig. 1).

Table 3. *Relative exfoliated cell gene expression levels in breast-fed vs. formula-fed infants following a 3-mo feeding period*

Gene Name	*P* Value	Fold Change	NCBI Accession No.
EGFR	7.94E-06	0.3339	CA313202.1
DAPK1	9.99E-06	0.7359	AF074988.1
FOXE3	1.75E-05	0.3287	NM_012186.1
EPAS1	5.36E-05	3.3001	AA838165.1
SPARC	1.08E-04	0.8228	NM_003118.2
REG4	3.81E-04	2.4223	NM_032044.2
NR3C1	4.60E-04	5.5126	X03225.1
WNT7B	6.12E-04	0.5819	NM_058238.1
NEUROD1	6.15E-04	0.8177	NM_002500.1
HIF3A	6.34E-04	0.7088	CF593418.1
NR5A2	6.65E-04	2.8371	U80251.1
LYZL6	7.38E-04	0.6997	NM_020426.1
PCDH7	1.06E-03	3.9210	AA252063.1
GATA6	1.51E-03	2.4767	BF002339.1
BAD	2.80E-03	4.0367	AI360092.1
MAP17	3.16E-03	2.1919	NM_005764.3
CAMK2D	3.25E-03	0.8352	AA701319.1
TDGF1	3.62E-03	0.8000	NM_003212.1
ITGB2	3.77E-03	2.5813	BX113382.1
APC	3.82E-03	0.7725	T83934.1
UCP3	4.42E-03	0.8683	NM_022803.1
FOXM1	4.97E-03	2.0341	AI801441.1
ICAM1	5.23E-03	2.6462	NM_000201.1
PLCB2	5.38E-03	2.7635	NM_004573.1
WNT8B	5.52E-03	0.7948	NM_003393.2
PLCD3	6.02E-03	2.1256	NM_133373.2
PPAP2B	6.66E-03	0.9621	AI242906.1
FOXP1	6.97E-03	2.7249	BE676551.1
TJP1	8.60E-03	1.0345	BG546714.1
DDX11	8.79E-03	0.9640	NM_030655.2
GNL1	1.04E-02	0.9432	NM_005275.2
SYP	1.04E-02	1.9512	NM_003179.2
MYB	1.11E-02	2.7804	NM_005375.2
TLR5	1.18E-02	0.7598	NM_003268.3
TJP1	1.26E-02	2.1625	AL109707.1
HOXC5	1.29E-02	0.9778	NM_018953.2
ALOX5	1.33E-02	1.0729	AL110200.1
BAX	1.41E-02	2.0993	NM_004324.3
FOXJ3	1.41E-02	0.9458	AW086233.1
REG4	1.49E-02	0.6170	AW389191.1
MAML1	2.11E-02	1.0164	BE674445.1
HOXA1	2.12E-02	1.7642	NM_005522.3
TGFB2	2.21E-02	2.5229	BI495496.1
HSPA1A	2.31E-02	6.4201	NM_005345.4
PDE4D	2.45E-02	3.7358	U50157.1
HOXC6	2.63E-02	1.0272	NM_014620.2
FGF5	2.79E-02	1.9989	AA461028.1
LIFR	2.79E-02	0.9364	NM_002310.2
CD9	2.93E-02	2.4451	AK025016.1
NOTCH3	3.03E-02	0.9989	NM_000435.1
SELP	3.35E-02	1.0456	NM_003005.2
GSTM4	3.71E-02	3.0954	NM_147149.1
TP53	3.98E-02	0.9947	AA928725.1
NOS3	4.27E-02	0.7267	NM_173681.2
FOXP1	4.43E-02	1.8863	BX538242.1
IL22	4.44E-02	1.0646	NM_020525.4
TCF7L2	4.79E-02	2.6630	CK905987.1
NODAL	4.96E-02	1.1373	BF967917.1
PTGER4	4.96E-02	0.7172	NM_000958.2
PLD2	5.01E-02	1.0786	NM_002663.2
TFCP2L1	9.54E-02	0.9853	NM_014553.1
IGF1R	2.04E-01	1.7147	AF020763.1

Fold change represents relative expression level in breast-fed divided by formula-fed infants for individual genes. Data are ranked by *P* value, computed using *t*-tests applied to normalized data.

Identification of Master Regulatory Genes

Table 2 lists the 10 best single-gene classifiers among the 146 genes considered. We subsequently posed the following question: Do these genes provide good discrimination between the classes because they regulate large numbers of other genes? Although this question cannot be answered directly using the CoD method, a master gene that controls many slaves will possess a CoD histogram, the mass of which is concentrated to the right (toward 1). Thus we examined whether the best discriminators have master-like CoD histograms (see supplemental methods). After binarization, genes exhibiting less than 4 changes in their binary profiles were excluded from further CoD analyses, leaving only 55 of the original 146 biomarker genes. Of the 10 best single-gene classifiers in Table 2, all but tight junction protein 1 (TJP1) are among the 55 genes. TJP1 is missing because its binary profile exhibited less than 4 changes over the 22 infant samples. Rather than performing the CoD analysis with this small set, which includes only genes selected on the basis of putative biological relevance, we randomly selected 100 of the remaining 764 genes to provide a total of 155 genes to construct CoD histograms. To compute the CoD distributions, each gene was treated as the target in turn, and triples of the remaining 154 genes were queried to predict the target. This approach was used to examine a total of $C(154,3) = 596{,}904$ different combinations of triple predictors for each target gene and its CoD distribution histograms.

The CoD histograms for the single-classifier genes in Table 2, excluding TJP1, are shown in Fig. 3. CoD-histogram means >0.7 are considered strong, and those >0.8 are considered extremely strong. CoD means are >0.8 for six and very close to 0.7 for two of the nine genes from Table 2 (see supplemental Fig. 2 for CoD histograms of the 9 genes with the lowest CoD means).

DISCUSSION

The interaction between the major components of the intestinal ecosystem, including nutrients, microflora, epithelium, and aspects of nonspecific and specific immunity, has important implications for GI development and function, as well as overall health (6). Nutrition, in particular, is believed to play a major role in the modulation of the evolving infant gut. Recent evidence suggests that host factors in amniotic fluid and breast milk contribute to gut maturation (19, 44). The regulatory role of nutrition is particularly crucial during the early postnatal period, impacting GI barrier function, gut motility, mucosal immunity, digestive and absorptive capacity, and microbial colonization (10, 44). With respect to underlying mechanisms, data indicate that epigenetic regulation plays an important role in intestinal development and pathology (18, 46, 47). Unfortunately, because of the lack of tissue biopsies, no investigators have performed a global transcriptional analysis of the developing human intestine in the early postnatal period. Therefore, the overall objectives of this proof-of-principle study were *1*) to validate the use of methodology to extract mRNA from infant fecal samples and *2*) to noninvasively fingerprint and compare intestinal transcriptional profiles in exclusively BF or FF infants. As part of this effort, gene sets or combinations were identified in response to human milk or infant formula feeding. As opposed to using expression levels of significantly up- or downregulated genes, we applied novel mRNA-based noninvasive methods to

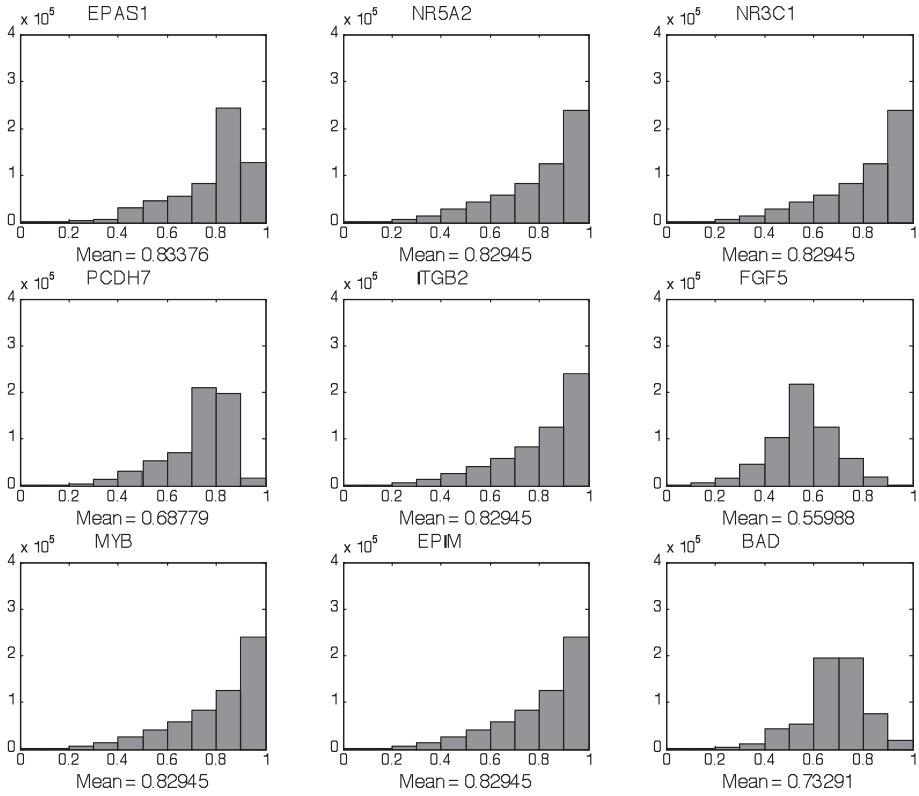


Fig. 3. Association between expression patterns of genes (triple predictors) was determined using coefficient of determination (CoD). CoD measures the degree to which the transcriptional levels of an observed gene set can be used to improve the prediction of the transcriptional state of a target gene relative to the best-possible prediction in the absence of observations. Examples for the strong single-gene classifiers in Table 2, excluding tight junction protein 1 (TJP1), are shown. NR5A2, nuclear receptor subfamily 5, group A, member 2; NR3C1, nuclear receptor subfamily 3, group C, member 1; PCDH7, protocadherin 7; ITGB2, integrin-β_2; EPIM, epimorphin; BAD, Bcl2 antagonist of cell death.

identify the best single-gene and two- to three-gene combinations for distinguishing treatment effects. Similar to previous studies, we demonstrate that three-gene combinations provide classifiers with potential to noninvasively identify discriminative signatures for diagnostic purposes (32, 36). Collectively, these data indicate for the first time that genes that are modulated during neonatal gut epithelial differentiation are differentially represented in BF and FF infants.

It is likely that some of the variability among the BF infants results from potential differences in milk composition, namely, oligosaccharides and, potentially, lipids, as well as differences in the intestinal microbiota. Human milk contains >200 different oligosaccharides, which vary by maternal genetics (25). Human milk oligosaccharides have direct interactions with intestinal epithelial cells (33), as well as the microbiota (45), and, thus, may directly or indirectly affect host gene expression. Lipids are the most variable components of human milk and are directly impacted by maternal diet (4). This may be relevant, because several long-chain polyunsaturated fatty acids can directly regulate transcriptional activation of gene expression via activation of nuclear transcription factors (43). Lastly, the microbiota of BF and FF infants are reported to differ (2), and it is known that the microbiota and soluble metabolites of the microbiota can directly influence host gene expression (49).

From a biological context, the best-performing gene, on the basis of its ranking of single-, two-, and three-gene classifiers, was EPAS1. The EPAS1 protein is a basic helix-loop-helix PAS transcription factor involved in cellular response to hypoxia, whereby its activation is stimulated by hypoxia, with subsequent induction of vascular development (26, 40). Interestingly, EPAS1 was significantly (3-fold) upregulated in BF infants. This is noteworthy, because hypoxia-triggered angiogenesis may affect predisposition to proinflammatory states, including necrotizing enterocolitis (50). Although the role of EPAS1 in the evolving human neonatal intestine is not well defined, CoD analyses indicated its role as a master regulator of signaling cascades. This novel method finds associations between the expression patterns of individual genes by determining whether knowledge of the transcriptional levels of a small gene set can be used to predict the associated transcriptional state of another gene (21, 31). The power of this approach lies in its ability to find unexpected links between processes not previously known to be coordinated. Interestingly, the majority of master genes identified in this study are transcription factors associated with angiogenesis and wound repair, e.g., EPAS1, nuclear receptor subfamily 5, group A, member 2 (NR5A2), MYB, and nuclear receptor subfamily 3, group C, member 1 NR3C1 (11, 23, 26, 28, 34, 39). Other genes have also been implicated in GI morphogenesis: integrin-β_2 (ITGB2), Bcl2 antagonist of cell death (BAD), protocadherin 7 (PCDH7), and fibroblast growth factor 5 (FGF5). It is noteworthy that several genes exhibited the same binary profile; i.e., their CoD distributions were identical (Fig. 3). This represents the interplay of three factors: *1*) the sample size is small; *2*) binarization has removed small differences, which indeed is a salient purpose of binarization; and *3*) we conjecture that these genes are reacting concordantly in response to cellular conditions, which is supported by their strong CoD values and discriminatory power. We also must keep in mind that identical profiles might indicate that the genes are part of a tightly controlled regulated pathway in which one of them is actually the master over the pathway and that this master, or the pathway as a whole, controls a large family of genes. The

validity and reliability of these molecular master-slave gene relationships may be better defined at the protein expression level and, therefore, warrant further studies to confirm physiological relevance.

Although the precise origin of exfoliated cells is not known, results from this study indicate that a number of genes associated with discrete epithelial cell types are detectable. Examples include absorptive enterocytes (lactase, sucrase-isomaltase), goblet cell (mucin-2), enteroendocrine cells (chromogranin A, intestinal mucin 3a), and paneth cell (lysozyme; data not shown). Thus it is likely that transcriptome signatures of the small and large intestine can be monitored over time. Clearly, additional studies are needed to elucidate the contribution of small and large intestine exfoliated cells to the pool of mRNA used in our screening test. We previously tested the feasibility of the noninvasive mRNA procedure in *1*) a rat experimental colon carcinogenesis model (13) and *2*) human subjects at high risk for colonic adenoma recurrence (15, 51). In a significant advancement, we reported that two- and three-gene combinations provide classifiers with potential to noninvasively identify discriminative molecular signatures. These findings support our hypothesis that it is possible to noninvasively detect molecular markers (combination gene sets) in exfoliated cells, which may provide an alternative approach to evaluating intestinal development and function in infants.

In conclusion, we demonstrate that fecal samples containing exfoliated cells hold potential for evaluating intestinal exposure to food and possibly other nondietary components in infants. Response to type of infant feeding based on one-, two-, and three-gene feature sets was accomplished using mRNA directly isolated from infant feces. In addition, on the basis of these preliminary data, the concordance between classification and CoD analyses emphasizes the potential usefulness of exfoliated cells in the identification of master controlling genes, many of which may be associated with mucosal morphogenesis. Given the need to better understand the role of nutrition in promoting GI health and development, noninvasive tests are expected to become increasingly relevant tools in tailoring diet to meet the nutritional needs of the growing infant. By continuing to build on this data set, this methodology may be used as a molecular tool to noninvasively evaluate the performance of infants, at the intestinal level, in response to different dietary regimens.

GRANTS

This work was supported by Mead Johnson and by National Institutes of Health Grants CA-59034, CA-129444, DK-71707, and P30 ES-09106.

DISCLOSURES

No conflicts of interest are declared by the authors.

REFERENCES

1. **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* 52: 347–350,1992.
2. **Adlerberth I.** Factors influencing the establishment of the intestinal microbiota in infancy. *Nestle Nutr Workshop Ser Pediatr Program* 62: 13–29, 2008.
3. **Braga-Neto UM, Dougherty ER.** Bolstered error estimation. *Pattern Recognition* 37: 1267–1281, 2004.
4. **Brenna JT, Salem N Jr, Sinclair AJ, Cunnane SC.** α-Linolenic acid supplementation and conversion to n-3 long-chain polyunsaturated fatty acids in humans. *Prostaglandins Leukot Essent Fatty Acids* 80: 85–91, 2009.
5. **Burrin DB, Shulman RG, Reeds PJ, Davis TA, Gravitt KR.** Porcine colostrum and milk stimulate visceral organ and skeletal muscle protein synthesis in neonatal piglets. *J Nutr* 122: 1205–1213, 1992.
6. **Canani RB, Passareillo A, Buccigrossi V, Terrin G, Guarino A.** The nutritional modulation of the evolving intestine. *J Clin Gastroenterol* 42: S197–S200, 2008.
7. **Catassi C, Bonucci A, Coppa GV, Carlucci A, Giorgi PL.** Intestinal permeability changes during the first month: the effect of natural versus artificial feeding. *J Pediatr Gastroenterol Nutr* 21: 383–386, 1995.
8. **Chapkin RS, Clark AE, Davidson LA, Schroeder F, Zoran DL, Lupton JR.** Dietary fiber differentially alters cellular fatty acid binding protein expression in exfoliated colonocytes during tumor development. *Nutr Cancer* 32: 107–112, 1998.
9. **Chowdhury SR, King DE, Willing BP, Band MR, Beever JE, Lane AB, Loor JJ, Marini JC, Rund LA, Schook LB, Van Kessel AG, Gaskins HR.** Transcriptome profiling of the small intestinal epithelium in germfree versus conventional piglets. *BMC Genomics* 8: 215, 2007.
10. **Conroy ME, Shi HN, Walker WA.** The long-term effects of neonatal microbial flora. *Curr Opin Allergy Clin Immunol* 9: 197–201, 2009.
11. **Coste A, Dubuquoy L, Barnouin R, Annicotte JS, Magnier B, Notti M, Corazza N, Antal MC, Metzger D, Desreumaux P, Brunner T, Auwerx J, Schoonjans K.** LRH-1-mediated glucocorticoid synthesis in enterocytes protects against inflammatory bowel disease. *Proc Natl Acad Sci USA* 104: 13098–13103, 2007.
12. **Cummings AG, Thompson FM.** Postnatal changes in mucosal immune response: a physiological perspective of breast feeding and weaning. *Immunol Cell Biol* 75: 419–429, 1997.
13. **Davidson LA, Aymond CM, Jiang YH, Turner ND, Lupton JR, Chapkin RS.** Non-invasive detection of fecal protein kinase C βII and ζmessenger RNA: putative biomarkers for colon cancer. *Carcinogenesis* 19: 253–257, 1998.
14. **Davidson LA, Jiang YH, Lupton JR, Chapkin RS.** Non-invasive detection of putative biomarkers for colon cancer using fecal mRNA. *Cancer Epidemiol Biomarkers Prev* 4: 643–647, 1995.
15. **Davidson LA, Lupton JR, Miskovsky E, Fields AP, Chapkin RS.** Quantification of human intestinal gene expression profiles using exfoliated colonocytes: a pilot study. *Biomarkers* 8: 51–61, 2003.
16. **Davidson LA, Nguyen DV, Hokanson RM, Callaway ES, Isett RB, Turner ND, Dougherty ER, Wang N, Lupton JR, Carroll RJ, Chapkin RS.** Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. *Cancer Res* 64: 6797–6804, 2004.
17. **Davis CD.** Use of exfoliated cells from target tissues to predict responses to bioactive food components. *J Nutr* 133: 1769–1772, 2003.
18. **De Santa Barbara P, van den Brink GR, Roberts DJ.** Development and differentiation of the intestinal epithelium. *Cell Mol Life Sci* 60: 1322–1332, 2003.
19. **Donovan SM.** Role of human milk components in gastrointestinal development: current knowledge and future needs. *J Pediatr* 149: S49–S61, 2006.
20. **Donovan SM, Odle J.** Growth factors in milk as mediators of infant development. *Annu Rev Nutr* 14: 147–167, 1994.
21. **Dougherty ER, Brun M, Trent JM, Bittner ML.** Conditioning-based modeling of contextual genomic regulation. *IEEE/ACM Trans Comp Biol Bioinf* 6: 310–320, 2009.
22. **Dougherty ER, Kim S, Chen Y.** Coefficient of determination in nonlinear signal processing. *Signal Processing* 80: 2219–2235, 2000.
23. **Dusing MR, Wiginton DA.** Epithelial lineages of the small intestine have unique patterns of GATA expression. *J Mol Histol* 36: 15–24, 2005.
24. **Eisenberg E, Levanon EY.** Human housekeeping genes are compact. *Trends Genet* 19: 362–365, 2003.
25. **Erney RM, Malone WT, Skelding MB, Marcon AA, Kleman-Leyer KM, O'Ryan ML, Ruiz-Palacios G, Hilty MD, Pickering LK, Prieto PA.** Variability of human milk neutral oligosaccharides in a diverse population. *J Pediatr Gastroenterol Nutr* 30: 181–192, 2000.
26. **Favier J, Kempf H, Corvol P, Gasc JM.** Coexpression of endothelial PAS protein 1 with essential angiogenic factors suggests its involvement in human vascular development. *Dev Dyn* 222: 377–388, 2001.
27. **Fornace AJ Jr, Alamo I Jr, Hollander M.** DNA damage-inducible transcripts in mammalian cells. *Proc Natl Acad Sci USA* 85: 8800–8804, 1988.

28. **Gartner H, Graul MC, Oesterreicher TJ, Finegold MJ, Henning SJ.** Development of the fetal intestine in mice lacking the glucocorticoid receptor (GR). *J Cell Physiol* 194: 80–87, 2002.
29. **Jenkins SL, Wang J, Vazir M, Vela J, Sahagun O, Gabbay P, Hoang L, Diaz RL, Aranda R, Martin MG.** Role of passive and adaptive immunity in influencing enterocyte-specific gene expression. *Am J Physiol Gastrointest Liver Physiol* 285: G714–G725, 2003.
30. **Kaeffer B, Des Robert C, Anlexandre-Gouabau MC, Pagniez A, Legrand A, Amarger V, Kuster A, Piloquet H, Champ M, Le Huerou-Luron I, Roze JC.** Recovery of exfoliated cells from gastrointestinal tract of premature infants: a new tool to perform "noninvasive biopsies"? *Pediatr Res* 62: 564–569, 2007.
31. **Kim S, Dougherty ER, Bittner ML, Chen Y, Sivakumar K, Meltzer P, Trent JM.** General nonlinear framework for the analysis of gene interaction via multivariate expression arrays. *J Biomed Optics* 5: 411–424, 2000.
32. **Kim S, Dougherty ER, Chen Y, Sivakumar K, Meltzer P, Trent JM, Bittner M.** Multivariate measurement of gene expression relationships. *Genomics* 67: 201–209, 2000.
33. **Kuntz S, Kunz C, Rudloff S.** Oligosaccharides from human milk induce growth arrest via G_2/M by influencing growth-related cell cycle genes in intestinal epithelial cells. *Br J Nutr* 101: 1306–1315, 2009.
34. **Mace KA, Hansen SL, Myers C, Young DM, Boudreau N.** HOXA3 induces cell migration in endothelial and epithelial cells promoting angiogenesis and wound repair. *J Cell Sci* 118: 2567–2577, 2005.
35. **Martins D, Braga-Neto U, Hashimoto R, Bittner ML, Dougherty ER.** Intrinsically multivariate predictive genes. *IEEE J Selected Topics Signal Processing* 2: 424–439, 2008.
36. **Morikawa J, Li H, Kim S, Nishii K, Ueno S, Suh E, Dougherty ER, Shmulevich I, Shiku H, Zhang W, Kobayashi T.** Identification of signature genes by microarray for acute myeloid leukemia without maturation (FAB-M1) and AML with t(15;17)(q22;q12)(*PML/RARalpha*). *Int J Oncol* 23: 617–625, 2003.
37. **Potten CS, Schofield R, Lajtha LG.** A comparison of cell replacement in bone marrow, testis and three regions of epithelium. *Biochim Biophys Acta* 560: 281–299, 1979.
38. **Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR.** Genomic signatures to guide the use of chemotherapeutics. *Nat Med* 11: 1294–1300, 2006.
39. **Ramsay RG, Gonda TJ.** MYB function in normal and cancer cells. *Nat Rev Cancer* 8: 523–534, 2008.
40. **Scortegagna M, Morris MA, Oktay Y, Bennett M, Garcia JA.** The HIF family member EPAS1/HIF-2a is required for normal hematopoiesis in mice. *Blood* 102: 1634–1640, 2003.
41. **Sima C, Attoor S, Braga-Neto U, Lowey J, Suh E, Dougherty ER.** Impact of error estimation on feature-selection algorithms. *Pattern Recognition* 38: 2472–2482, 2005.
42. **Sima C, Dougherty ER.** What should be expected from feature selection in small-sample settings. *Bioinformatics* 22: 2430–2436, 2006.
43. **Uauy R, Mena P, Rojas C.** Essential fatty acids in early life: structural and functional role. *Proc Nutr Soc* 59: 3–15,2000.
44. **Wagner CL, Taylor SN, Johnson D.** Host factors in amniotic fluid and breast milk that contribute to gut maturation. *Clin Rev Allergy Immunol* 34: 191–204, 2008.
45. **Ward RE, Ninonuevo M, Mills DA, Lebrilla CB, German JB.** In vitro fermentation of breast milk oligosaccharides by *Bifidobacterium infantis* and *Lactobacillus gasseri*. *Appl Environ Microbiol* 72: 4497–4499, 2006.
46. **Waterland RA.** Epigenetic mechanisms and gastrointestinal development. *J Pediatr* 149: S137–S142, 2006.
47. **Waterland RA, Lin JR, Smith CA, Jirtle RL.** Post-weaning diet affects genomic imprinting at the insulin-like growth factor 2 (Igf2) locus. *Hum Mol Genet* 15: 705–716, 2006.
48. **Widdowson EM, Colombo VE, Artavanis CA.** Changes in the organs of pigs in response to feeding for the first 24 hours after birth. II. The digestive tract. *Biol Neonate* 23: 272–281, 1976.
49. **Wiling BP, Van Kessel AG.** Intestinal microbiota differentially affect brush border enzyme activity and gene expression in the neonatal gnotobiotic pig. *J Anim Physiol Anim Nutr* 93: 586–595, 2009.
50. **Zamora R, Vodovotz Y, Betten B, Wong C, Zuckerbraun B, Gibson KF, Ford HR.** Intestinal and hepatic expression of BNIP3 in necrotizing enterocolitis: regulation by nitric oxide peroxynitrite. *Am J Physiol Gastrointest Liver Physiol* 289: G822–G830, 2005.
51. **Zhao C, Ivanov I, Dougherty ER, Hartman TJ, Lanza E, Colburn NH, Lupton JR, Davidson LA, Chapkin RS.** Non-invasive detection of candidate molecular biomarkers in subjects with a history of insulin resistance and colorectal adenomas. *Cancer Prev Res* 2: 590–597, 2009.

Supplemental Table 1. List of 33 housekeeping genes used for normalization of the microarray data.

NCBI Accession #	Gene Name
NM_002635.2	SLC25A3
NM_001863.3	COX6B1
NM_003860.2	BANF1
NM_001675.2	ATF4
NM_001865.2	COX7A2
NM_014730.2	KIAA0152
NM_003314.1	TTC1
NM_007260.2	LYPLA2
NM_006694.2	JTB
NM_000937.2	POLR2A
NM_012412.3	H2AFV
NM_002140.2	HNRPK
NM_002406.2	MGAT1
NM_002794.3	PSMB2
NM_007286.3	SYNPO
NM_021642.2	FCGR2A
NM_004924.3	ACTN4
NM_002686.2	PNMT
NM_021959.1	PPP1R11
NM_002088.3	GRIK5
NM_005175.2	ATP5G1
NM_002636.3	PHF1
NM_004309.3	ARHGDIA
NM_000858.4	GUK1
NM_001619.2	ADRBK1
NM_006612.3	KIF1C
NM_001017.2	RPS13
NM_001694.2	ATP6V0C
NM_019884.2	GSK3A
NM_007245.2	ATXN2L
NM_006389.2	HYOU1
NM_005731.2	ARPC2
NM_004197.1	STK19

SupplementalTable 2. Single gene classification of breastfed versus formula-fed subjects. Single-gene LDA classifiers are shown. $\varepsilon_{bolstered}$ denotes bolstered resubstitution error for the respective classifier.

Gene names	$\varepsilon_{bolstered}$
EPAS1	0.1214
NR5A2	0.1356
NR3C1	0.1364
PCDH7	0.1367
ITGB2	0.1374
FGF5	0.1385
TJP1	0.1400
MYB	0.1456
EPIM	0.1478
BAD	0.1496
PLCB2	0.1507
GATA6	0.1530
PLCD3	0.1544
FOXM1	0.1545
ICAM1	0.1567
REG4	0.1584
TGFB2	0.1601
HOXA3	0.1612
HOXA1	0.1637
MAP17	0.1663
CD9	0.1723
BAX	0.1750
EGFR	0.1788
AXIN2	0.1815
FOXP1	0.1884

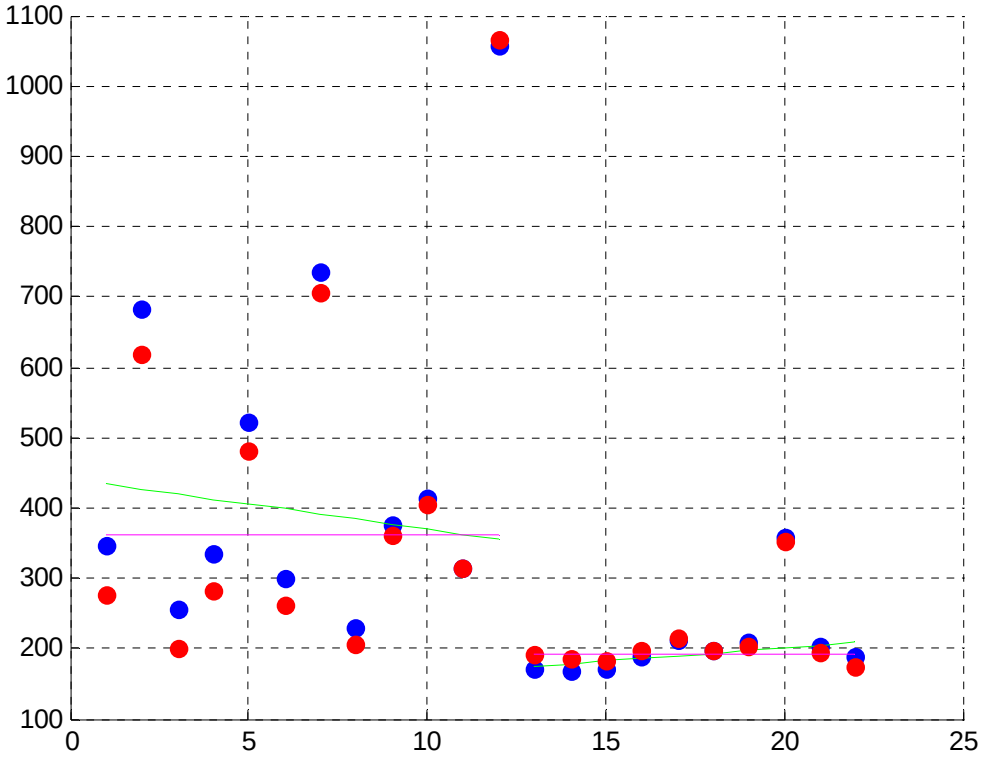
Supplemental Table 3. Pair-wise gene classification of breastfed versus formula-fed subjects. Pair-wise gene LDA classifiers are shown. $\varepsilon_{bolstered}$denotes bolstered resubstitution error for the respective classifier.

Gene names	$\varepsilon_{bolstered}$
EPAS1, UCP20	0.0869
CTDSPL, NR3C1	0.0975
NR3C1, TNFRSF10B	0.0987
FOXP4, NR3C1	0.1026
CDK4, EPAS1	0.1039
EPAS1, SYP	0.1045
NR3C1,SLC26A2	0.1057
GPR41, TJP1	0.1060
FOXP1, NR3C1	0.1064
HSPA1A, NR3C1	0.1064
BAX, EPAS1	0.1083
EPAS1, SCN5A	0.1090
ALOX5, EPAS1	0.1095
EPAS1, PIK3R1	0.1096
FOXE3, NR3C1	0.1096
EPAS1, PHB	0.1103
CAMK2A, EPAS1	0.1106
CACNB2, EPAS1	0.1111
EPAS1, GCM1	0.1112
EPAS1, PTGER4	0.1112
EPAS1, PTK2B	0.1115
EPAS1, FOXE3	0.1116
EPAS1, NR6A1	0.1117
EPAS1, TNFRSF10B	0.1119
BCL2L12, EPAS1	0.1120

Supplemental Table 4. Triplet-wise gene classification of breastfed versus formula-fed subjects. Triplet- wise gene LDA classifiers are shown. $\varepsilon_{bolstered}$denotes bolstered resubstitution error for the respective classifier.

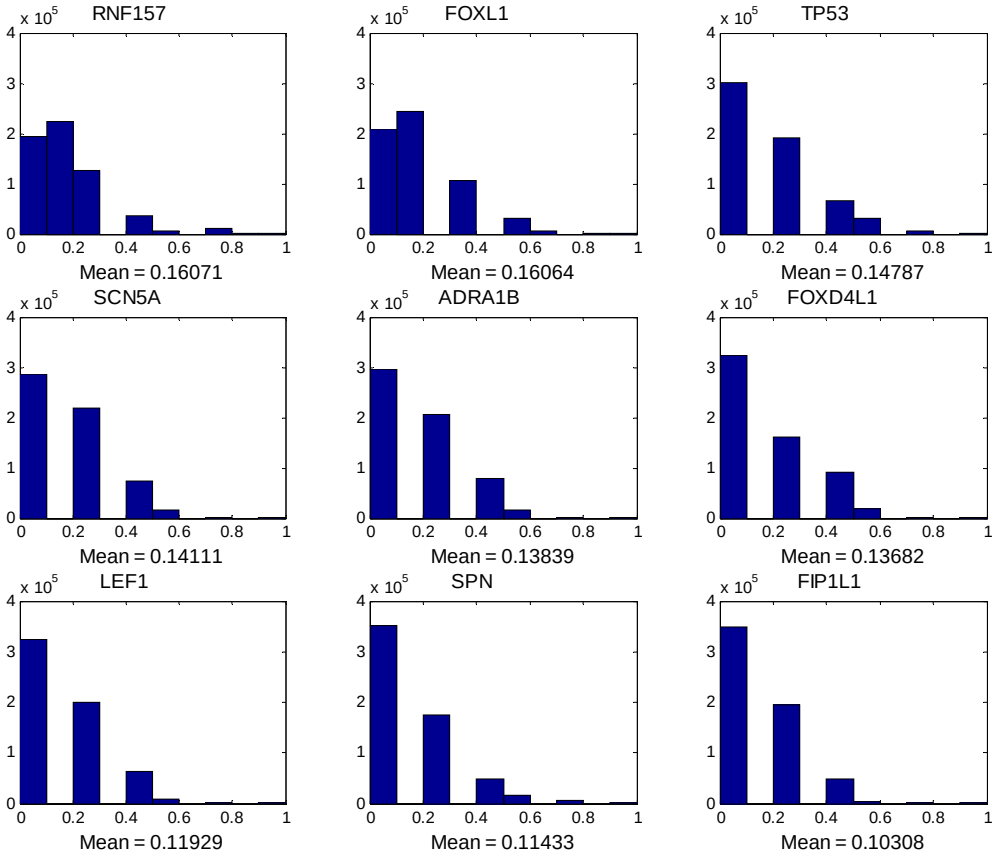
Gene names	$\varepsilon_{bolstered}$
EPAS1, FOXE3, SYP	0.0778
CTDSPL, FOXE3, NR3C1	0.0785
EPAS1, TLR5, UCP2	0.0785
EPAS1, REG4, UCP2	0.0802
EPAS1, LIFR, UCP2	0.0808
EPAS1, NODAL, UCP2	0.0821
EPAS1, HIF3A, UCP2	0.0823
EPAS1, HOXD10, UCP2	0.0823
EPAS1, KIT, UCP2	0.0824
ALOX5, EPAS1, UCP2	0.0827
EPAS1, TGFB3, UCP2	0.0833
EPAS1, HMGCL, UPC2	0.0838
CA12, EPAS1, UPC2	0.0848
CACNB2, EPAS1, UPC2	0.0848
EPAS1, SELL, UPC2	0.0858
EPAS1, PTGER4, UPC2	0.0860
EPAS1, TP53, UPC2	0.0860
BECN1, EPAS1, UPC2	0.0866
FOXE3, FOXP1, NR3C1	0.0876
EPAS1, POU5F1, UCP2	0.0883
FOXE3, NR5A2, SLC26A2	0.0883
CTDSPL, NR3C1,TNFRSF10B	0.0887
EPAS1, MSRB3, UCP2	0.0909
FOXE3, FOXP4, NR3C1	0.0887
CDK4, EPAS1, UCP2	0.0889

Supplemental Figure 1. Housekeeping genes and piecewise median normalization. Blue dots represent average raw intensities (per array) of the 33 housekeeping genes prior to normalization. Red dots represent averagesof the same genes after normalization. The two pink lines represent the median values of the averages across each group prior to normalization. The two green lines represent the robust piecewise linear regression of the sameaverages prior to normalization. The discrepancies between the purpleand the green lines were used to additively shift the values of the 4250common "good" genes on each of the arrays. Y-axis, raw expressionvalues; X-axis, microarray IDs ordered by the grouping of the babies.

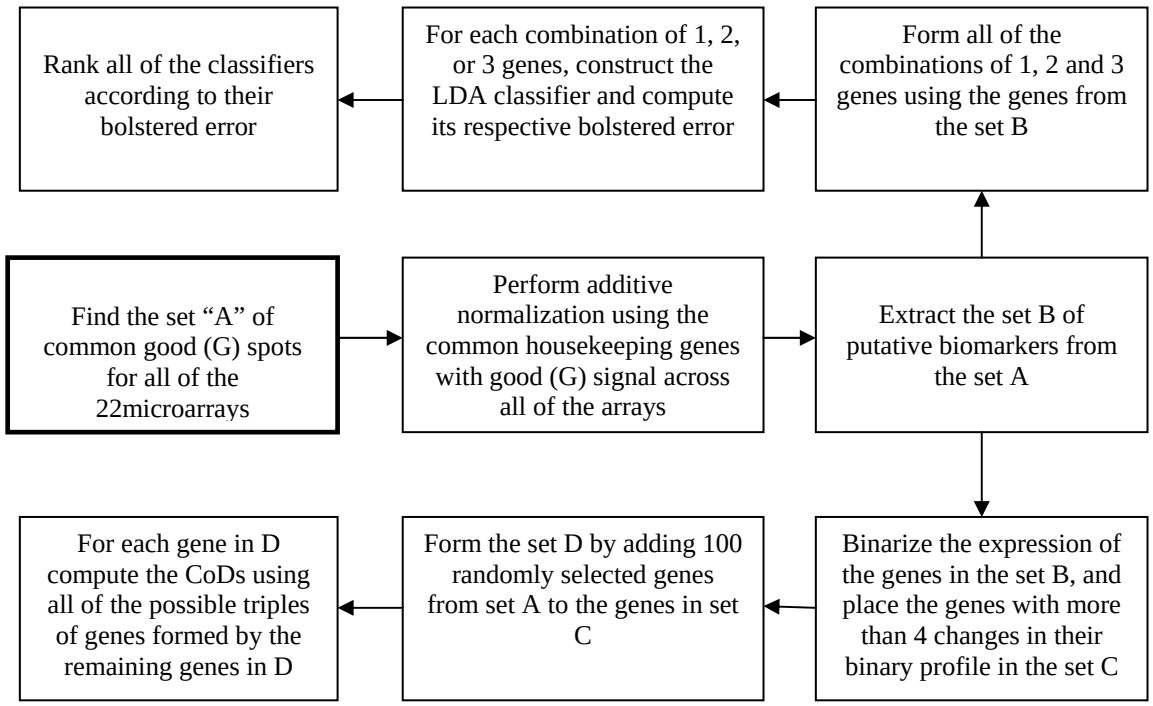


Breastfed babies **Formula-fed babies**

Supplemental Figure 2. Association between the expression of genes (triple predictors) was determined using the coefficient of determination (CoD).This coefficient measures the degree to which the transcriptional levels of an observed gene set can be used to improve the prediction of the transcriptional state of a target gene relative to the best possible prediction in the absence of observations. The CoD histograms in this figure are the ones with the smallest means.



Supplemental Figure 3. Depiction of the "walk-through" steps for data analyses. A detailed description of the procedures is given in the Supplementary Methods.



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.

Bolstered Error Estimation

The simplest approach to estimate the error of a designed classifier ψ is by applying ψ to the sample dataset that is being used for both training and testing. This *resubstitution estimate*, $\hat{\varepsilon}_n^{res}$, is the fraction of errors made by ψ on the sample. The resubstitution estimator is typically (but not always) low-biased, meaning $E[\hat{\varepsilon}_n^{res}] < E[\varepsilon_n]$, and this bias can be severe for small samples, depending on the complexity of the classification rule.

In resubstitution there is no distinction between points near and far from the decision boundary; the *bolstered-resubstitution* estimator is based on the heuristic that, relative to making an error, more confidence should be attributed to points far from the decision boundary than points near it (S1). This is achieved by placing a distribution, called a *bolstering kernel*, at each point. A key issue is the amount of bolstering (spread of the bolstering kernels), and a method has been proposed to compute this spread based on the data (S1).**Supplemental Figure 4** illustrates the error for linear classification when the bolstering kernels are uniform circular distributions. By normalizing their total volume to 1, the collection of bolstering kernels form a probability density

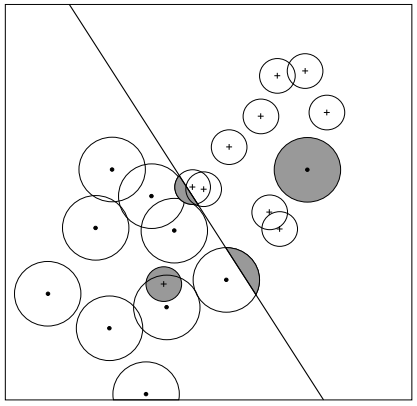
$$\mathbf{f}^{\nabla}(\mathbf{x}, y) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i^{\nabla}(\mathbf{x}-\mathbf{x}_i)I_{y=y_i}$$

where $\mathbf{f}_i^{\nabla}(\mathbf{x}-\mathbf{x}_i)$ is the bolstering kernel at the sample point $\mathbf{x}_i$ and I_A is the indicator function, I_A = 1 if A is true and I_A = 0 otherwise. The bolstered resubstitution error estimate is obtained by

integrating all bolstering kernels over their corresponding error regions (rather than simply counting the erroneously classified points as with resubstitution):

$$\hat{\varepsilon}_n^{bol} = \frac{1}{n}\sum_{i=1}^{n}\left(I_{y_i=0} \int_{\{\mathbf{x}:\psi(\mathbf{x})=1\}} \mathbf{f}_i^{\nabla}(\mathbf{x}-\mathbf{x}_i)\,\mathrm{d}\mathbf{x} + I_{y_i=1} \int_{\{\mathbf{x}:\psi(\mathbf{x})=0\}} \mathbf{f}_i^{\nabla}(\mathbf{x}-\mathbf{x}_i)\,\mathrm{d}\mathbf{x} \right)$$

Bolstered error estimation is illustrated in **Supplemental Figure 4** for circular uniform bolstering kernels and linear classification.



Supplemental Figure 4. Bolstering kernels for linear classification.

Linear Discriminant Analysis

It is common to view classification via *discriminant* functions. For binary classification there are two functions, d_0 and d_1, and $\mathbf{x}$ is classified as $Y = 1$ if $d_1(\mathbf{x}) \geq d_0(\mathbf{x})$and $\mathbf{x}$ is classified as $Y = 0$ if $d_0(\mathbf{x}) \geq d_1(\mathbf{x})$. It has been known for many decades that, if the classes are assumed to possess normally distributed densities, then the optimal classifier is determined by the discriminant

$$d_k(\mathbf{x}) = -(\mathbf{x}-\mathbf{u}_k)'\mathbf{K}_k^{-1}(\mathbf{x}-\mathbf{u}_k) - \log(\det[\mathbf{K}_k]) + 2\log f(k)$$

where $\mathbf{K}_k$ and $\mathbf{u}_k$ are the covariance matrix and mean vector, respectively, and $f(k)$ is the prior probability for class k. The form of this equation shows that the decision boundary $d_0(\mathbf{x}) = d_1(\mathbf{x})$

is quadratic and the equation characterizes *quadratic discriminant analysis* (*QDA*). If the classes possess the same covariance matrix **K**, then the discriminant reduces to

$$d_k(\mathbf{x}) = -(\mathbf{x}-\mathbf{u}_k)'\mathbf{K}^{-1}(\mathbf{x}-\mathbf{u}_k) + 2\log f(k)$$

which is a linear function of **x** and produces hyperplane decision boundaries. This discriminant characterizes *linear discriminant analysis* (*LDA*).

The discriminants for quadratic and linear discriminant analysis are derived under the normality assumption, but in practice can perform well so long as the class conditional densities are not too far from Gaussian and there is sufficient data to obtain good estimates of the relevant covariance matrices, the point being that the QDA and LDA classification rules involve finding the sample covariance matrices and sample means. Owing to the greater number of parameters to be estimated for QDA as opposed to LDA, one can proceed with smaller samples with LDA than with QDA and this is why, with small samples, it is common to use LDA even if it is believed that the covariance matrices are not equal (S2).

Coefficient of Determination

The coefficient of determination(CoD), long used in the context of linear regression, was more recently introduced in the context of nonlinear operators (S3). The CoD measures the degree to which the expression levels of an observed set of “predictor” random variables can be used to improve prediction of the expression of a target variable relative to the best possible prediction in the absence of observations. Given the target variable Y and predictor variables $X_1, X_2, \ldots, X_k$, the CoD for $X_1, X_2, \ldots, X_k$ predicting Y is defined by

$$\text{CoD} = \frac{\varepsilon_0 - \varepsilon_{opt}}{\varepsilon_0}$$

where ε_{opt} is the best estimate, $f_{opt}(X_1, X_2, \ldots, X_k)$, of Y based on the predictor variables and ε_0 is the error of the best constant estimate of Y in the absence of any observations. It is easily seen

that the CoD must be between 0 and 1 and measures the relative decrease in error by estimating *Y* via the predictor variables rather than by just the best constant estimate. In practice, the CoD must be estimated from training data with designed approximations being used in place of f_{opt}.

Identifying Multivariate Discriminators (Feature Gene Sets) for Diet Classification

Because we only have data and not the underlyingfeature-label distributions, the errors have been estimated from thedata. This approach takes into account that, in small-sample settings,we do not have much confidence in any single feature set and that it ismuch more likely, if there is an adequate sized collection of good performingfeature sets, that some will performwell on the overall population (S4).

Identifying Potential "Master" and "Slave" Genes

If "master" gene activities completely determine the activities of all genes in a given pathway, then the "master" and the genes in the governed pathway will exhibit closelyrelated expression profiles. Hence, their CoD histograms will be very close. In effect, one can think of the pathway as a "master pathway," in the sense that its collective activity mirrors that of the true "master" gene, which is itself controlling many genes, including the given pathway. Put simply, the CoD cannot distinguish between genes possessing the same expression profiles.

Gene Binarization

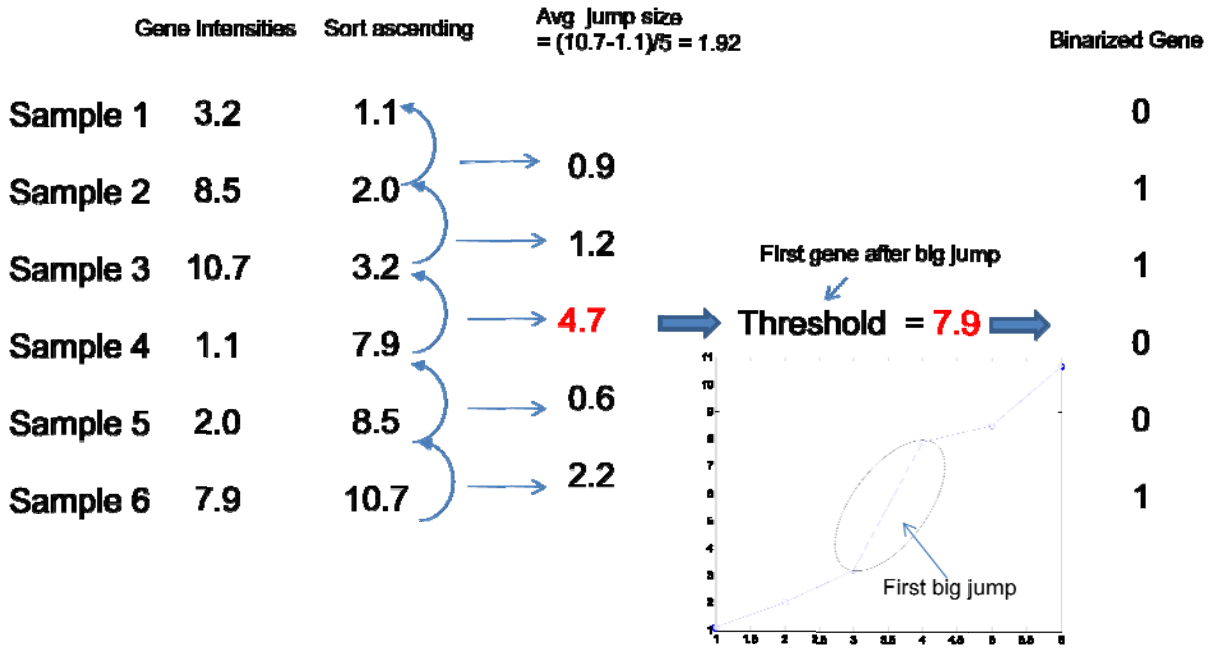
Procedure:

1. Sort gene intensities in ascending order and compute the average increase in those intensities (jump size).

2. Find the first “big jump” which is larger than the average jump size and determine the threshold to be the first value after the big jump.

3. For each gene expression value, if it is greater than the threshold found in step 2, set the binarized gene expression to be 1 and if ithe gene expression value is less than the threshold, then set the binarized expression to be 0.

This procedure is illustrated by the example given in the **Supplemental Figure 5**.



Supplemental Figure 5. An example of the gene expression binarization procedure.

Supplemental References

S1. **Braga-Neto UM, Dougherty ER.** Is Cross-Validation Valid for Small-Sample Microarray Classification. *Bioinformatics* 20:374-380, 2004.

S2. **Shmulevich I, Dougherty ER.***Genomic Signal Processing*, Princeton University Press, Princeton, 2007.

S3. **Dougherty ER, Kim S, Chen Y.** Coefficient of Determination in Nonlinear Signal Processing. *EURASIP JSignal Proc* 80(10): 2219-2235, 2000.

S4. **Kim S, Dougherty ER, Shmulevich I, Hess KR, Hamilton SR, Trent JM, Fuller GN, Zhang W.** Identification of combination gene sets for Glioma classification. *Molec Cancer Ther***1**(13):1229-1236, 2002.