SUPPLEMENT: Metabolomic profiling of stool of two-year old children from the INSIGHT Study reveals links between butyrate and child weight outcomes

Debmalya Nandy1#$, Sarah J. C. Craig2,3#, Jingwei Cai4$$, Yuan Tian4, Ian M. Paul3,5, Jennifer S. Savage6,7, Michele E. Marini7, Emily E. Hohman7, Matthew L. Reimherr1,3, Andrew D. Patterson4,8, Kateryna D. Makova2,3*, and Francesca Chiaromonte1,3,9*

1) Department of Statistics, Penn State University, University Park, PA, USA
2) Department of Biology, Penn State University, University Park, PA, USA
3) Center for Medical Genomics, Penn State University, University Park, PA, USA
4) Department of Molecular Toxicology, Penn State University, University Park, PA, USA
5) Department of Pediatrics, Penn State College of Medicine, Hershey, PA, USA
6) Department of Nutritional Sciences, Penn State University, University Park, PA, USA
7) Center for Childhood Obesity Research, Penn State University, University Park, PA, USA
8) Department of Biochemistry & Molecular Biology, Penn State University, University Park, PA, USA
9) Institute of Economics, EMbeDS, Sant’Anna School of Advanced Studies, Piazza Martiri della Libertà, Pisa, Italy

#Authors contributing equally
*Co-corresponding authors
Kateryna Makova (kdm16@psu.edu)
Francesca Chiaromonte (fxc11@psu.edu)

$Present address:
- Debmalya Nandy: Department of Biostatistics & Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
- Jingwei Cai: Department of Drug Metabolism and Pharmacokinetics, Genentech Inc., South San Francisco, CA, USA
Contents:

Table S1: Racial breakdown of the 170 INSIGHT children included in this study. 3
Table S2. Food items considered within each of the 16 food-groups. 4
Table S3. p-values from the 2-way ANOVA on batch and processor. 5
Table S4. Bacteria groupings. 6
Figure S1. Flowchart for data preprocessing and statistical analyses. 8
Figure S2. Boxplots of relative concentrations (RCs) of the 30 metabolites used in our analyses. 9
Figure S3. Histograms of logarithmic values of relative concentrations (log(RC)’s) of the 30 metabolites used in our analyses. 10
Figure S4. Boxplots of logarithmic relative concentrations (log(RC)’s) of the 30 metabolites used in our analyses. 11
Figure S5. Histograms of adjusted metabolite concentrations (log(RC)’s adjusted for batch and processor effects) for the 30 metabolites used in our analyses. 12
Figure S6. Boxplots of consumption frequencies of food items within the 16 initial food groups. 13
Figure S7. Histograms of the children’s per day consumption frequencies for the 12 food groups considered in Phase 2 analyses. 21
Figure S8. Histograms of the children’s per day consumption frequencies (adjusted and log-transformed) for the 12 food groups considered in Phase 2 analyses. 22
Figure S9. Distribution of BMI among study participants at age two years. 23
Figure S10. Scatterplots between pairs from the three child weight outcomes considered: body mass index (BMI), BMI z-score, and growth index (GI). 24
Figure S11. Plot of Bayesian Information Criterion (BIC) values versus the model size (ranging from 1 to 41), using BMI as the response in Phase 2 (Figure S1) BIC best subset selection regression procedure. 27
Figure S12. Plot of Bayesian Information Criterion (BIC) values versus the model size (ranging from 1 to 41), using BMI z-score as the response in Phase 2 (Figure S1) BIC best subset selection regression procedure. 28
Figure S13. Plot of Bayesian Information Criterion (BIC) values versus the model size (ranging from 1 to 41), using GI as the response in Phase 2 (Figure S1) BIC best subset selection regression procedure. 29
References 30
Table S1: Racial breakdown of the 170 INSIGHT children included in this study.

| Race                                | n=170 |
|-------------------------------------|-------|
| Black                               | 6     |
| White                               | 155   |
| Indian or Alaskan Native            | 0     |
| Native Hawaiian or Pacific Islander | 0     |
| Asian                               | 5     |
| Other - specified                   | 1*    |
| Other - not specified               | 0     |
| Not Answered                        | 3     |

*Specified as Dominican Republic
Table S2. Food items considered within each of the 16 food-groups.

Number in brackets indicates the number of food items considered within the associated food-group. Four food-groups were omitted from the Phase 2 analyses after preprocessing: Beans, Fat/oils, Non-sugar sweetened beverages, and Sugar sweetened beverages.

| Food-groups [# of food items] | List of food items within the groups |
|-------------------------------|-------------------------------------|
| Beans [4]                     | Chick-peas, beans, cooked beans, hummus |
| Dairy [8]                     | Whole milk, 2% milk, 1% milk, skim milk, cheese, cottage cheese, yogurt, yogurt melts |
| Fat/Oils [4]                  | Dressing, mayonnaise, butter, margarine |
| Fried foods [3]               | Fried potatoes, fried fish, fried chicken |
| Fruit juice [1]               | Fruit juice |
| Fruits [18]                   | Banana, peaches, fruit cocktail, oranges, grapefruit, apples, pears, blueberries, grapes, strawberries, applesauce, melon, watermelon, pineapple, mangoes, kiwi, raisins/prunes, raspberries/blackberries |
| Grains [11]                   | Infant cereal, hot cereal, cold cereal (sweetened), unsweetened cereal, rice, pasta, bread, cornbread/tortilla, bagel, biscuit, pancakes/etc. |
| Meats [11]                    | Hot dogs, sausage, hamburgers, canned tuna, other fish, cold cuts, pork/ham, beef, liver, non-fried chicken/turkey, bacon |
| Mixed foods [10]              | Granola bars, fruit bars, pizza, tacos, mac-n-cheese, fruit snacks, veggie soup, other soup, hot pockets, PB & J sandwich |
| Non-meat proteins [3]         | Eggs, nuts, peanut butter |
| Non-sugar sweetened beverages [4] | Non-dairy milk, coffee, diet soda, unsweetened tea |
| Snacks [5]                    | Chips, popcorn, pretzels, goldfish crackers, other crackers |
| Sugar sweetened beverages [7] | Flavored water, sugar-water, fruit-drink, flavored milk, regular soda, sweetened tea, hot chocolate |
| Sweets [12]                   | Cookies/brownies, cake, pie, pudding, jello, chocolate candy, other candy, ice-cream, milkshakes, frozen soy desserts, donut, sweet rolls |
| Vegetables [19]               | Corn, peas, tomatoes, peppers, carrots, avocado, asparagus, broccoli, cauliflower, green beans, spinach, greens, mixed veggies, squash, zucchini, sweet potatoes, other potatoes, cabbage/coleslaw, salad |
| Water [1]                     | Water |
Table S3. p-values from the 2-way ANOVA on batch and processor.

Analysis performed separately for the log(relative concentrations) of the 30 metabolites used in our analysis to control for batch and processor effects.

| Metabolites      | Batch   | Processor |
|------------------|---------|-----------|
| 4-pyridoxate     | 0.4292  | 0.0749    |
| Acetate          | 0       | 0.3363    |
| Alanine          | 2.00E-04| 0.4162    |
| Aspartate        | 0.0099  | 0.1287    |
| Butyrate         | 0.4213  | 0.1433    |
| Creatine         | 0       | 0         |
| Formate          | 0       | 0.0163    |
| Fumarate         | 0.038   | 0.0541    |
| Galactose        | 0.5483  | 0.0107    |
| Glucose          | 0.7377  | 0.8824    |
| Glutamate        | 0.0171  | 0.0443    |
| Glycine          | 0.6923  | 0.6102    |
| Histidine        | 0.2538  | 0.9994    |
| Hypoxanthine     | 0.3196  | 0.0132    |
| Isoleucine       | 0.0145  | 0.0189    |
| Lactate          | 0.002   | 0.1658    |
| Leucine          | 0       | 0.4932    |
| Methionine       | 0       | 0.2944    |
| O-phosphocholine | 0.3684  | 0.4923    |
| Phenylalanine    | 0.0545  | 0.1849    |
| Proline          | 0       | 0.1423    |
| Propionate       | 0.6019  | 0.5984    |
| Pyruvate         | 0       | 0         |
|          | Value 1 | Value 2 |
|----------|---------|---------|
| Succinate| 0.0115  | 0.0226  |
| Tryptophan| 1.00E-04 | 0.7602  |
| Tyrosine | 0.0029  | 0.0984  |
| Uracil   | 6.00E-04 | 0.0127  |
| Valine   | 0.0608  | 0.1199  |
| Xanthine | 0.2269  | 0.0078  |
| Xylose   | 0.5285  | 0.0099  |
## Table S4. Bacteria groupings.

List of bacteria included as they were identified by Vital et al.\(^1\) as potentially butyrate producing.

| Phylum          | Genus                                                                 |
|-----------------|-----------------------------------------------------------------------|
| Firmicute       | Anaerococcus, Carnobacterium, Clostridium, Thermoanaerobacterium, Coprococcus, Roseburia, Shuttleworthia, Acidaminococcus, Alkaliphilus, Anaerofustis, Anaerostipes, Butyrivibrio, Lachnospira, Ruminococcus, Desulfotobacter, Anaerotruncus, Faecalibacterium, Symbiobacterium, Syntrophomonas, Megasphaera, Thermosinus, Dethiobacter, Halanaerobium, Natranaerobius, Peptoniphilus |
| Actinobacteria  | Janibacter, Kribbella, Nocardioides, Verrucosispora, Salinispora       |
| Fusobacteria    | Fusobacterium                                                          |
| Spirochaetes    | Treponema, Candidatus Cloacamonas                                       |
| Thermotogae     | Fervidobacterium, Kosmotoga, Petrotoga, Thermosipho, Thermostoga       |
| Bacteroidetes   | Propionibacterium, Bacteroides, Porphyromonas, Odoribacter             |
| Deferribacter   | Deferribacter                                                          |
| Proteobacteria  | Rhodoferax, Anaeromyxobacter, Desulfarculus, Desulfobulbus, Geobacter, Halangiun, Myxococcus, Sorangium, Syntrophobacter |
Figure S1. Flowchart for data preprocessing and statistical analyses.

Phase 1 includes data preprocessing and selection of potentially influential metabolites. Phase 2 includes selection of potentially influential covariates.
Figure S2. Boxplots of relative concentrations (RCs) of the 30 metabolites used in our analyses.

The distributions of relative concentration values show remarkable positive skewness for certain metabolites, such as succinate.
Figure S3. Histograms of logarithmic values of relative concentrations (log(RC)’s) of the 30 metabolites used in our analyses.
Figure S4. Boxplots of logarithmic relative concentrations (log(RC)’s) of the 30 metabolites used in our analyses.

The five outliers marked in red correspond to four unique subjects.
Figure S5. Histograms of adjusted metabolite concentrations (log(RC)’s adjusted for batch and processor effects) for the 30 metabolites used in our analyses.

Means are centered at 0 and standard deviations are set to 1.
Figure S6. Boxplots of consumption frequencies of food items within the 16 initial food groups.

The dummy codes for frequency indicate the following: 0=Never or 0 times per week; 1=1 time per week; 2=2-3 times per week; 3=4-6 times per week; 4=1 time per day; 5=2 times per day; 6=3 times per day; 7=4-5 times per day; 8=6 or more times per day. Four food groups (Beans, Non-sugar sweetened beverages, Sugar sweetened beverages, and Fats/Oils) were omitted from the Phase 2 analyses (see Figure S1) because each of the food items within these four groups has zero median consumption frequency.

A) Dairy [8 food items].
B) Vegetables [19 food items]

C) Beans [4 food items]. Note: All food items within this food group have zero medians, and therefore, this food group was removed from the Phase 2 analysis.
D) Meats [11 items]

E) Non-meat proteins [3 food items]
F) Grains [11 food items]

G) Sugar sweetened beverages [7 food items]. All food items within this food group have zero medians, and therefore, this food group was removed from the Phase 2 analysis.
H) Non-sugar sweetened beverages [4 food items]. Note: All food items within this food group have zero medians, and therefore, this food group was removed from the Phase 2 analysis.

I) Fruit-juice [1 item] and Water [1 item]
J) Fruits [18 food items]

K) Snacks [5 food items]
L) Sweets [12 food items]

M) Fried foods [3 food items]
N) Fats/Oils [4 food items]. Note: All food items within this food group have zero medians, and therefore, this food group was removed from the Phase 2 analysis.

O) Mixed foods [10 food items]
Figure S7. Histograms of the children’s per day consumption frequencies for the 12 food groups considered in Phase 2 analyses.
Figure S8. Histograms of the children’s per day consumption frequencies (adjusted and log-transformed) for the 12 food groups considered in Phase 2 analyses.
Figure S9. Distribution of BMI among study participants at age two years.

Number within a bar indicates the number of individuals in that category. The BMI percentiles and the associated raw BMI cutoffs for boys and girls are indicated as per CDC guidelines\(^2\).
Figure S10. Scatterplots between pairs from the three child weight outcomes considered: body mass index (BMI), BMI z-score, and growth index (GI).

The blue solid line is the linear regression fit, and the gray shaded area around it denotes the 95% confidence band.

A) BMI vs. GI
B) BMI z-score vs. GI

Scatterplot of BMI z-score and GI (R^2 = 0.82, p-value < 2.2 E-16)
C) BMI z-score vs. BMI

Scatterplot of BMI z-score and BMI ( $R^2 = 0.98$, p-value $< 2.2 \times 10^{-16}$)
Figure S11. Plot of Bayesian Information Criterion (BIC) values versus the model size (ranging from 1 to 41), using BMI as the response in Phase 2 (Figure S1) BIC best subset selection regression procedure.

The red dot indicates that the minimum BIC value corresponds to model size 3.
Figure S12. Plot of Bayesian Information Criterion (BIC) values versus the model size (ranging from 1 to 41), using BMI z-score as the response in Phase 2 (Figure S1) BIC best subset selection regression procedure.

The red dot indicates that the minimum BIC value corresponds to model size 4.
Figure S13. Plot of Bayesian Information Criterion (BIC) values versus the model size (ranging from 1 to 41), using GI as the response in Phase 2 (Figure S1) BIC best subset selection regression procedure.

The red dot indicates that the minimum BIC value corresponds to model size 3.
References

1. Vital, M., Howe, A. C. & Tiedje, J. M. Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *MBio* 5, e00889 (2014).

2. Clinical Growth Charts. https://www.cdc.gov/growthcharts/clinical_charts.htm (2019).