


## ORIGINAL RESEARCH

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

Debmalya Nandy<sup>1</sup> | Sarah J. C. Craig<sup>2,3</sup>  | Jingwei Cai<sup>4</sup> | Yuan Tian<sup>4</sup> | Ian M. Paul<sup>3,5</sup> | Jennifer S. Savage<sup>6,7</sup> | Michele E. Marini<sup>7</sup> | Emily E. Hohman<sup>7</sup> | Matthew L. Reimherr<sup>1,3</sup> | Andrew D. Patterson<sup>4,8</sup> | Kateryna D. Makova<sup>2,3</sup> | Francesca Chiaromonte<sup>1,3,9</sup>

<sup>1</sup>Department of Statistics, Penn State University, University Park, PA, USA

<sup>2</sup>Department of Biology, Penn State University, University Park, PA, USA

<sup>3</sup>Center for Medical Genomics, Penn State University, University Park, PA, USA

<sup>4</sup>Department of Molecular Toxicology, Penn State University, University Park, PA, USA

<sup>5</sup>Department of Pediatrics, Penn State College of Medicine, Hershey, PA, USA

<sup>6</sup>Department of Nutritional Sciences, Penn State University, University Park, PA, USA

<sup>7</sup>Center for Childhood Obesity Research, Penn State University, University Park, PA, USA

<sup>8</sup>Department of Biochemistry & Molecular Biology, Penn State University, University Park, PA, USA

<sup>9</sup>Institute of Economics, EMbeDS, Sant'Anna School of Advanced Studies, Pisa, Italy

## Correspondence

Kateryna D. Makova and Francesca Chiaromonte, Center for Medical Genomics Penn State University University Park, Pennsylvania, USA.  
Email: kdm16@psu.edu (K. D. M.); fxc11@psu.edu (F. C.)

## Present address

Debmalya Nandy, Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA

Jingwei Cai, Department of Drug Metabolism and Pharmacokinetics, Genentech Inc., South San Francisco, California, USA

## Funding information

National Science Foundation, Division of Mathematical Sciences, Grant/Award Number: 1407639; National Institutes of Health, National Center for Advancing Translational Sciences, Grant/Award Number: UL1TR000127; National Center for Research Resources; National Institutes of Health,

## Summary

**Background:** Metabolomic analysis is commonly used to understand the biological underpinning of diseases such as obesity. However, our knowledge of gut metabolites related to weight outcomes in young children is currently limited.

**Objectives:** To (1) explore the relationships between metabolites and child weight outcomes, (2) determine the potential effect of covariates (e.g., child's diet, maternal health/habits during pregnancy, etc.) in the relationship between metabolites and child weight outcomes, and (3) explore the relationship between selected gut metabolites and gut microbiota abundance.

**Methods:** Using <sup>1</sup>H-NMR, we quantified 30 metabolites from stool samples of 170 two-year-old children. To identify metabolites and covariates associated with children's weight outcomes (BMI [weight/height<sup>2</sup>], BMI z-score [BMI adjusted for age and sex], and growth index [weight/height]), we analysed the <sup>1</sup>H-NMR data, along with 20 covariates recorded on children and mothers, using LASSO and best subset selection regression techniques. Previously characterized microbiota community

Debmalya Nandy and Sarah J. C. Craig contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Pediatric Obesity* published by John Wiley & Sons Ltd on behalf of World Obesity Federation.

National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Numbers: R01DK088244, R01DK99364; Penn State Eberly College of Science; Penn State Institute for Computational and Data Sciences; Pennsylvania Department of Health, Grant/Award Numbers: CURE funds, Tobacco Settlement

information from the same stool samples was used to determine associations between selected gut metabolites and gut microbiota.

**Results:** At age 2 years, stool butyrate concentration had a significant positive association with child BMI ( $p$ -value =  $3.58 \times 10^{-4}$ ), BMI z-score ( $p$ -value =  $3.47 \times 10^{-4}$ ), and growth index ( $p$ -value =  $7.73 \times 10^{-4}$ ). Covariates such as maternal smoking during pregnancy are important to consider. Butyrate concentration was positively associated with the abundance of the bacterial genus *Faecalibacterium* ( $p$ -value =  $9.61 \times 10^{-3}$ ).

**Conclusions:** Stool butyrate concentration is positively associated with increased child weight outcomes and should be investigated further as a factor affecting childhood obesity.

#### KEYWORDS

1-HNMR, butyrate, childhood obesity, metabolomics, weight outcomes at 2-years

## 1 | INTRODUCTION

The prevalence of childhood obesity in the United States has increased in the past several decades.<sup>1,2</sup> Based on reports from the Center for Disease Control and Prevention,<sup>3</sup> 13.9% of preschool-aged children (age 2–5 years) were classified as having obesity in 2015–2016. Importantly, childhood obesity is associated with increased risk of adult obesity and its comorbidities.<sup>4</sup> Therefore, identifying early life risk factors, such as variation in microbial communities or their metabolic outputs, to characterize children most at risk of developing obesity is critical.

The gut microbiota has been shown to be an important component of obesity aetiology. It has been over a decade since Turnbaugh's seminal work showing that microbiota from obese mice can stimulate the development of obesity in gnotobiotic mice.<sup>5</sup> In humans, numerous studies have shown that the gut microbiota of adults and children with obesity significantly differ from those without obesity.<sup>6–10</sup> Although there has not been a consensus on what the composition of an “obese microbiota” community looks like, most scientists agree that the microbiota communities of healthy and diseased individuals do differ.

While the role of the gut microbiota in obesity has been broadly investigated, the microorganisms present in the gut explain only part of the picture. The metabolic capacity and outputs of this microbial community, which cannot be characterized through sequence based approaches, is the link between the microbiota and the human host. Metabolomics data generated by proton Nuclear Magnetic Resonance (<sup>1</sup>H-NMR) spectroscopy have popularly been used to study these small molecules to detect important biomarkers for many diseases.<sup>11,12</sup> Faecal metabolome analysis in adults has shown that individuals with obesity have higher levels of short chain fatty acids (SCFAs, e.g., acetate, propionate, and butyrate), branched chain amino acids (e.g., leucine, isoleucine, and valine), and aromatic amino acids (phenylalanine, tryptophan, and tyrosine).<sup>13,14</sup> Higher levels of faecal SCFAs in individuals with obesity, in addition to differences in the gut microbial communities, have also been related to measurements of cardiometabolic disorders (e.g., inflammation, glycemia, and

dyslipidemia) and gut permeability (LPS binding protein).<sup>15</sup> These findings suggest that characterizing the metabolites in the gut is an important step for understanding their roles in the human physiology and risk of diseases.

The characterization of both microbiota and metabolomic environments has mainly been conducted in adults and older children; although a few small studies have been conducted in younger children. Interestingly, the patterns seen in young children do not always match those seen in older children and adults.<sup>16</sup> Furthermore, our knowledge of obesity-related gut metabolites in young children is limited because the majority of studies in children focus on plasma metabolites.<sup>17,18</sup> Our study, which focuses on the gut metabolome of young children and its potential role in their weight outcomes, fills an important gap.

The Intervention Nurses Start Infants Growing on Healthy Trajectories (INSIGHT) study<sup>19</sup> is a clinical trial, in which first-born children were randomized to a responsive parenting behavioural intervention or to a home safety control shortly after birth, and were followed longitudinally. This clinical trial collected data on growth (e.g., weight and height), behaviour (e.g., sleep, activity and temperament), home environment, diet, and medical history. Recently, Craig and colleagues<sup>20</sup> examined the oral and gut microbiota of children from the INSIGHT study at age 2 years, and related these microbial communities to the children's growth trajectories (from birth through age 2). That study, surprisingly, found overall weak links between the gut microbiota and children's growth patterns. Here, we hypothesize that differences in the metabolic output of the gut microbiota community might show stronger associations with children's weight outcomes than the microbiota community itself—specifically, that children with higher weight outcomes at age 2 years will have metabolomic profiles that differ significantly from those of children with lower weight outcomes.

In this article, we report metabolite concentrations from targeted <sup>1</sup>H-NMR<sup>21,22</sup> profiling of stool sampled from 170 children in the INSIGHT cohort at the age of approximately 2 years. Based upon the data from related prior studies,<sup>23,24</sup> we specifically targeted 30 major metabolites commonly found in human stool samples. Our first

objective was to investigate the association of these metabolites with children's BMI (weight/height<sup>2</sup>), BMI z-score (BMI adjusted for age and sex), and growth index (weight/height). Our second objective was to assess the potential effects of 20 other covariates collected from INSIGHT, including children's dietary intake and maternal health and habits during pregnancy, on the relationship between identified metabolites and children's weight outcomes. Our third objective was to investigate whether metabolites significantly associated with weight outcomes have any relationship to the microbial communities we have previously characterized.<sup>20</sup> See Figure 1 for a diagrammatic representation of these objectives.

## 2 | METHODS

### 2.1 | Study population characterization and sample collection

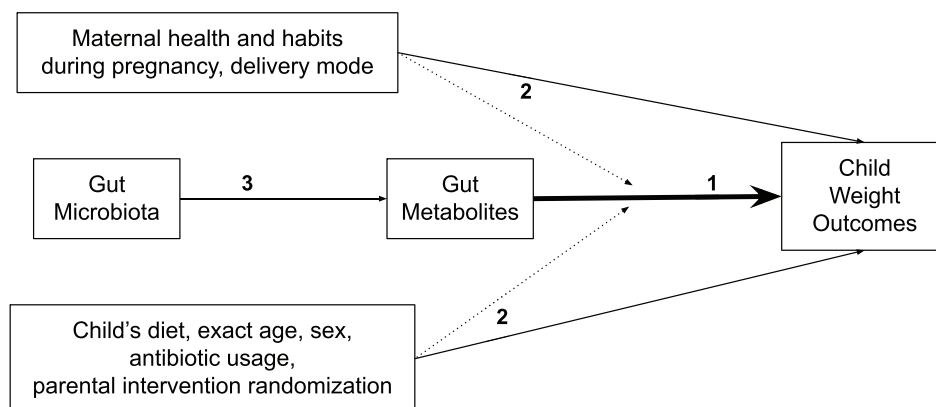
Our sample included 170 of the 279 children enrolled in the INSIGHT<sup>19</sup> study. These children are full-term singletons born to primiparous mothers in Central Pennsylvania and are largely of European descent (Table S1).<sup>25</sup> Stool samples were collected during a study visit occurring around the second birthday of each child. Specifically, the samples were collected by mothers from their child's diaper into sterile containers, placed into the home freezer and transported to the visit (on ice, wrapped in ice packs, in thermo-envelopes). These samples were then stored at  $-80^{\circ}\text{C}$  in the laboratory. During the same visit, weight and standing height of each child were measured by trained research nurses using an electronic scale (Seca 354) and a portable stadiometer (Shorr Productions, Olney, MD), respectively. For each child, we calculated BMI (weight in kilograms divided by squared height in meters), BMI z-score (BMI adjusted for age and sex), and growth index (weight in kilograms divided by height in meters).

Guided by previous literature, we considered a pool of covariates concerning children and mothers that could potentially contribute to

children's weight outcomes. These included the child's sex (female/male; recorded from medical records)<sup>1</sup>; the child's exact age (in months) at the two-year study visit (calculated from date of birth from the medical record and date of visit); the INSIGHT intervention study group (control/responsive parenting)<sup>25</sup>; the child's exposure to antibiotics between birth and 2 years (no/yes, determined from parental survey)<sup>26</sup>; maternal gestational weight gain status (below or at guidelines/exceeded guidelines, determined from medical records)<sup>27</sup>; delivery mode (vaginal/caesarean section; recorded from medical records)<sup>28</sup>; maternal gestational diabetes status (none/controlled with diet and exercise, determined from maternal survey)<sup>29</sup>; maternal smoking status during pregnancy (no/yes, determined from maternal survey)<sup>30</sup>; and the frequencies of 12 food group variables reflecting the qualitative pattern of the child's diet (determined by Infant Food Frequency Questionnaires completed by the parents).<sup>31</sup> More details on these food group variables are provided in Table 1 and Table S2 and in Preprocessing of the Covariates section below.

### 2.2 | Metabolomic profiling

We performed <sup>1</sup>H-NMR profiling following the steps described by Tian and colleagues<sup>23</sup>; the raw <sup>1</sup>H-NMR spectra were generated in two batches. We used the Chenomx NMR Suite of tools (Alberta, Canada) to process the raw spectra (calibrating internal standard, correcting phase and baseline) and fit metabolite profiles following the steps described in Nichols and colleagues.<sup>32</sup> Initially we targeted 35 metabolites, namely: 4-pyridoxate, acetate, alanine, aspartate, butanone, butyrate, creatine, formate, fumarate, galactose, glucose, glutamate, glycine, histamine, histidine, hypoxanthine, isocitrate, isoleucine, lactate, leucine, methionine, niacinamide, O-phosphocholine, phenylalanine, proline, propionate, pyruvate, succinate, trimethylamine N-oxide, tryptophan, tyrosine, uracil, valine, xanthine, and xylose. Because the metabolite assignments of butanone, histamine, isocitrate, niacinamide, and trimethylamine could not be verified with 2D-NMR spectra,<sup>23</sup> we



**FIGURE 1** Schematic of objectives of the current study. The bold solid arrow represents our main objective: to investigate the relationship between gut metabolites and child weight outcomes (1). Lighter arrows represent related objectives: to investigate effects of covariates recorded on children and mothers as direct effects (solid) and through interactions with butyrate (dashed) on child weight outcomes (2). We also investigate the relationship between the gut microbiota and the gut metabolites (3)

**TABLE 1** Summary of the maternal and child-related covariates collected on the 170 INSIGHT children included within this study

Covariates	n = 170
Child's sex (n = female)	84 (49%)
Intervention group (n = intervention)	88 (52%)
Antibiotic usage from birth through age 2 years (n = yes)	110 (65%)
Maternal gestational weight gain status (n = exceeded guidelines)	91 (54%)
Maternal smoking status during pregnancy (n = yes)	13 (8%)
Mode of delivery (n = caesarean section)	46 (27%)
Maternal gestational diabetes (n = controlled with diet and exercise)	9 (5%)
Child's exact age at two-year study visit (months) median (median absolute deviation)/range	24.23 (0.27)/24.01-25.49
Food groups <sup>a</sup> (per day consumption frequency) median (median absolute deviation)/range	
Dairy [8]	3.82 (1.7)/0.50-12.42
Fried foods [3]	0.32 (0.27)/0-2.07
Fruit juice [1]	0.36 (0.53)/0-6
Fruits [18]	2.99 (1.43)/0-18.64
Grains [11]	1.94 (0.74)/0-6.86
Meats [11]	1 (0.56)/0-3.50
Mixed foods [10]	0.85 (0.53)/0-4.36
Non-meat proteins [3]	0.64 (0.42)/0-3
Snacks [5]	1 (0.53)/0-5.07
Sweets [12]	0.60 (0.59)/0-3.29
Vegetables [19]	2.14 (1.19)/0.28-10.83
Water [1]	3 (2.22)/0-6

<sup>a</sup>See Table S1 for the food items considered within each food group. Value in square brackets indicates the number of food items considered within that food group.

removed these five metabolites from further analyses. The 30 verified metabolites were profiled using the “batch fit” tool in Chemomx and then manually calibrated and quantitated for each sample. The profiling procedure described above introduced two sources of variation that we adjusted for in our analyses (see below): (i) batch effect (raw <sup>1</sup>H-NMR spectra were generated in two batches) and (ii) processor effect (metabolites were calibrated and quantitated using Chemomx software by D.N. and S.J.C.C.).

## 2.3 | Statistical analyses

We performed the statistical analyses in two phases: Phase 1 included preprocessing, transforming, cleaning the data and then using the cleaned data to detect statistically significant metabolites associated, separately, with children's BMI, BMI z-score, and growth index

(covariates were not considered in this phase). To validate metabolites' roles after adjusting for the covariates and to investigate the effects of the latter, Phase 2 included analyses of the metabolites selected in Phase 1 along with the covariates and the metabolite-covariate interactions. See Figure S1 for a visual representation of this workflow.

## 2.4 | Data preprocessing

### 2.4.1 | Preprocessing of the metabolite concentrations

First, we normalized the metabolite concentrations in each sample with respect to the total concentration of all the 30 metabolites,<sup>33-35</sup> creating relative concentration, thus accounting also for the dry weight of the samples. We also mitigated the high skewness of the relative concentration values (Figure S2) using a logarithmic transformation. The histograms in Figure S3 indicate that the log-transformed relative concentrations of the metabolites are fairly regularly distributed (they do not deviate strongly from the Gaussian distribution). Outliers were omitted from this analysis (Figure S4).

We then adjusted the log-transformed relative concentrations of the metabolites for potential effects of batch and data processor. For each metabolite, we fitted a two-way ANOVA<sup>36</sup> for log-transformed relative concentrations on “batch” and “processor” (Table S3). The residuals from these ANOVA fits were used as “adjusted metabolite concentrations” in subsequent analyses. We centred and scaled each of these 30 adjusted metabolite concentrations separately to have zero mean and unit variance<sup>33-35</sup>, almost all of the transformed adjusted metabolite concentrations fairly resemble the Gaussian distribution (Figure S5).

### 2.4.2 | Preprocessing of the covariates

The food group variables included as covariates in our Phase 2 analyses (Table 1) are derived from a modified food frequency questionnaire.<sup>37</sup> The raw data from the questionnaire are the same as the data in our previous study<sup>20</sup>; however, we modified the grouping of the food items and created a new food group variable (“Mixed foods”) to more accurately represent the diet. We started with a total of 121 food items divided into 16 food groups (Table S2): beans (4 items), dairy (8 items), fat/oils (4 items), fried foods (3 items), fruit juice (1 item), fruits (18 items), grains (11 items), meats (11 items), mixed foods (10 items), non-meat proteins (3 items), non-sugar sweetened beverages (4 items), snacks (5 items), sugar-sweetened beverages (7 items), sweets (12 items), vegetables (19 items), and water (1 item). For each food item, mothers reported how often their child had consumed it in the past week using response options of 0, 1, 2-3, 4-6 times per week, 1, 2, 3, 4-5 or 6 or more times per day. We converted all records to times per day, using the medians for options that spanned a range of values (see Figure S6). We then omitted four food groups (beans, fats/oils, non-sugar sweetened beverages and sugar-sweetened beverages) for which all items had zero median daily

consumption frequencies. Thus, we had a total of 12 food groups included in the Phase 2 analyses (Table 1). Next, for each of the 12 retained food groups, we calculated the sum total of daily intake frequencies<sup>38</sup> (summing over items within the groups; we replaced missing values with zeros). Finally, after adding 1 to the resulting frequencies, we log-transformed them to mitigate right skewness (Figures S7 and S8), centred to zero mean and scaled to unit variance.

We also log-transformed the child's exact age in months at the time of the two-year visit, centred to zero mean and scaled to unit variance. Seven categorical covariates were dummy-coded as "0/1": intervention group (safety-control/parenting-intervention), child's sex (female/male), child's antibiotic exposure between birth and 2 years (never/yes), maternal gestational weight gain status (below or at guidelines/exceeded guidelines), mode of delivery (caesarean section/vaginal), maternal diabetes mellitus status (none/controlled with diet and exercise), and maternal smoking status during pregnancy (no/yes). For the LASSO regression analysis in Phase 2, we centred and scaled each of the covariates separately, including the dummy-coded ones, to zero mean and unit variance.

## 2.5 | Regression analyses

In Phase 1, our regressions comprised 30 predictor variables (i.e., explanatory variables used in the linear regression models); namely, the 30 adjusted metabolite concentrations. We ran two alternative procedures: (a) LASSO<sup>39</sup> (an  $L_1$ -penalized linear regression method which performs predictor selection) followed by a post-selection least squares fit; and (b) Bayesian information criterion (BIC<sup>40,41</sup>) best subset selection<sup>42</sup> (an exhaustive search over all possible model sizes for linear regression which selects the "best" model as per the BIC criterion), also followed by a post-selection least squares fit. For LASSO, the penalization parameter was tuned with 10-fold cross-validation repeated 100 times on 100 random fold partitions of the data. The selected tuning parameter value, which determines the number of predictors retained by the LASSO, corresponds to the minimum average (over the 100 repetitions) cross-validation mean-squared error. For the BIC best subset selection procedure, we examined all possible models with sizes ranging from 1 up to 30 predictors and selected the model with the minimum BIC. These analyses were completed separately treating each of three weight outcomes (BMI, BMI z-score, and growth index) as the response variable.

In Phase 2, our regressions comprised 41 predictor variables; namely, the butyrate adjusted metabolite concentration (selected in Phase 1; see below), 20 covariates, and the two-way interactions of the butyrate adjusted metabolite concentration with each of the 20 covariates (a total of 20 interaction terms). Just as in Phase 1, we implemented (a) LASSO and (b) BIC best subset selection regression procedures, each followed by a post-selection least squares fit. For LASSO, we tuned the penalization parameter using the same strategy as in Phase 1. For the BIC best subset selection, we again examined all possible models, with sizes ranging from 1 up to 41. Similar to Phase 1, these analyses were completed separately treating each of

the three weight outcomes (BMI, BMI z-score, and growth index) as the response variable.

We performed these statistical analyses using the R software<sup>43</sup> (version 4.0.2). The LASSO procedure with repeated 10-fold cross-validation tuning was implemented using the *ipflasso* CRAN package<sup>44</sup> and the BIC best subset selection procedure was implemented using the *leaps* CRAN package.<sup>45</sup> All codes are publicly available on GitHub: [https://github.com/makovalab-psu/ChildhoodObesity\\_1HNMR-Metabolomics](https://github.com/makovalab-psu/ChildhoodObesity_1HNMR-Metabolomics).

## 2.6 | Microbiota association

The gut microbiota of the INSIGHT children (from the same stool samples used in this study) were characterized by our group in a previous study.<sup>20</sup> The raw data are available on dbGaP (accession number phs001498.v1.p1). From these data, we selected all bacteria identified as a member of the Firmicutes phylum. Next, using information from Vital et al.,<sup>46</sup> we aggregated the abundances of bacteria identified as known or candidate butyrate producers; first as all butyrate producing bacteria, then as butyrate producing Firmicutes, and finally as individual genera within the butyrate producing Firmicutes. See Table S4 for a list of the bacterial groups included in this analysis. We used the statistical software R (version 3.6.1); correlations were interrogated using the *rcorr()* function within the *Hmisc* CRAN package<sup>47</sup> and graphs were generated using the *ggplot2* CRAN package.<sup>48</sup>

## 3 | RESULTS

### 3.1 | Study population and children's weight gain indicators

We considered 170 children enrolled in the INSIGHT<sup>19,25</sup> study who provided stool specimens at the 2-year study visit (see above). A variety of clinical, anthropometric and dietary information was collected on the children along with pregnancy health variables on their mothers (Table 1). We considered the children's height and weight collected during the same visit to compute our main weight indicator—BMI (ratio of weight to squared height, measured in  $\text{kg}/\text{m}^2$ ). BMI ranged from 13.9 to 20.7, with a median of 16.5. Using CDC guidelines,<sup>49</sup> we determined that 4.84% of the children considered were underweight, 79.0% had normal-weight, 11.8% were overweight and 4.30% had obesity (Figure S9). As additional weight outcomes, we computed the BMI z-score (BMI adjusted for age and sex) and the growth index (ratio of weight to height, measured in  $\text{kg}/\text{m}$ ). We used all these three outcomes to investigate the relationships among children's weight, metabolites, and other covariates. Figure S10 demonstrates the significant positive correlations between the pairs from these three outcomes over the range of values covered by our study (A) BMI versus GI:  $R^2 = 0.856$ ;  $p$ -value  $< 2.20 \times 10^{-16}$ ; (B) BMI z-score versus GI:  $R^2 = 0.822$ ;  $p$ -value  $< 2.20 \times 10^{-16}$ ; (C) BMI z-score versus BMI:  $R^2 = 0.983$ ;  $p$ -value  $< 2.20 \times 10^{-16}$ .

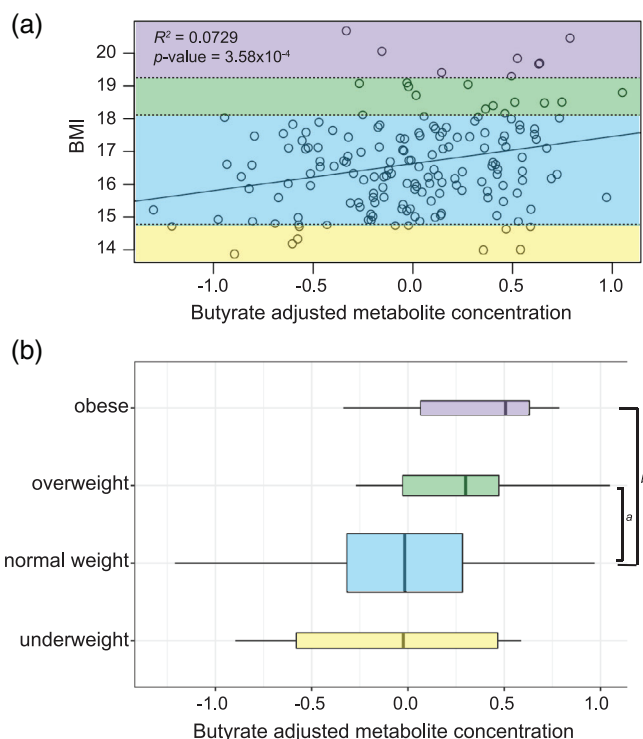


### 3.2 | Analysing metabolites: Butyrate is positively associated with child weight outcomes

The stool samples collected from 170 children were used to generate  $^1\text{H-NMR}$  spectra to determine the concentrations for 30 metabolites (see Methods for details). To identify metabolites associated with weight outcomes, we regressed each weight outcome (response) on 30 adjusted metabolite concentrations (predictors; here defined to mean explanatory variables used in the linear regression models). To conduct each of the regression analyses, we used the LASSO procedure and, separately, the BIC best subset selection procedure, each followed by a post-selection least squares fit (Phase 1 in Figure S1). Considering BMI as the weight outcome, LASSO retained only two metabolites: butyrate and glycine. Of these, only butyrate was significant in the post-selection least squares fit (butyrate:  $p$ -value = 0.0120,  $\hat{\beta}$  = 0.295; glycine:  $p$ -value = 0.147,  $\hat{\beta}$  = -0.169). Similarly, using the BIC best subset procedure, butyrate was the only metabolite selected ( $p$ -value =  $3.58 \times 10^{-4}$ ;  $\hat{\beta}$  = 0.375;  $R^2$  = 0.0729, adjusted  $R^2$  = 0.0677). To assess the out-of-sample predictive performance of butyrate for BMI, we considered 1000 independent 70%:30% random splits of our data into training and test sets ( $n_{\text{total}}$  = 170,  $n_{\text{training}}$  = 120, and  $n_{\text{test}}$  = 50); the median  $R^2$  and the median root mean squared prediction error across the 1000 test sets were 0.0699 and 1.35, respectively. Relatedly, BMI has a significant positive correlation with butyrate adjusted metabolite concentration ( $R^2$  = 0.0729,  $p$ -value =  $3.58 \times 10^{-4}$ , Pearson correlation; Figure 2(A)). Moreover, we found (Figure 2(B)) a significantly lower mean butyrate adjusted metabolite concentration in children with normal weight (mean concentration = -0.0387) than in those with overweight (mean concentration = 0.208,  $p$ -value = 0.0312; one-sided, two-sample  $t$ -test) or obesity (mean concentration = 0.339,  $p$ -value = 0.0176; one-sided, two-sample  $t$ -test).

Considering BMI z-score (BMI adjusted for age and sex) in place of BMI as the weight outcome, LASSO selected butyrate, glycine, and lactate adjusted metabolite concentrations, but only butyrate was significant in the post-selection least squares fit (butyrate:  $p$ -value = 0.0143,  $\hat{\beta}$  = 0.204; glycine:  $p$ -value = 0.532,  $\hat{\beta}$  = -0.0579; lactate:  $p$ -value = 0.333,  $\hat{\beta}$  = -0.0853). Again, the BIC best subset procedure selected only butyrate, which was significant in the post-selection least squares fit ( $p$ -value =  $3.47 \times 10^{-4}$ ,  $\hat{\beta}$  = 0.263,  $R^2$  = 0.0736, adjusted  $R^2$  = 0.0680). The median out-of-sample  $R^2$  and the median root mean squared prediction error of butyrate for BMI z-score were 0.0719 and 0.948, respectively.

As an additional verification of butyrate's association with child's growth, we considered growth index as the third weight outcome (response) and obtained reassuringly similar results compared to those for BMI and BMI z-score. LASSO retained acetate and butyrate adjusted metabolite concentrations, but only the latter was significant in the post-selection least squares fit (butyrate:  $p$ -value = 0.0108,  $\hat{\beta}$  = 0.284; acetate:  $p$ -value = 0.212,  $\hat{\beta}$  = 0.138). The BIC best subset procedure selected again only butyrate, which was significant in the post-selection least squares fit ( $p$ -value =  $7.73 \times 10^{-4}$ ;  $\hat{\beta}$  = 0.342;  $R^2$  = 0.0653, adjusted  $R^2$  = 0.0597). As with the other outcomes, the median out-of-sample  $R^2$  and the median root mean squared



**FIGURE 2** Relationship between butyrate adjusted metabolite concentration and BMI of 170 INSIGHT children. (A) A scatterplot of BMI against butyrate concentration with a simple regression trend line ( $R^2$  = 0.0729,  $p$ -value =  $3.58 \times 10^{-4}$ ). The colours (horizontal bands) correspond to approximate BMI class cut-offs (yellow = underweight, blue = normal weight, green = overweight, purple = obese) and match those in panel (B). (B) Boxplots of butyrate concentration by BMI class. The width of the boxes are proportional to the number of individuals within each class (underweight: 9, normal weight: 137, overweight: 16, obese: 8). a = children with overweight have greater average butyrate adjusted metabolite concentration than children with normal weight,  $p$ -value = 0.0312 (one-sided  $t$ -test). b = children with obesity have greater average adjusted metabolite concentration than children with normal weight,  $p$ -value = 0.0176 (one-sided  $t$  test)

prediction error of butyrate for growth index were 0.0654 and 1.30, respectively.

### 3.3 | Analysing metabolites and covariates

The relationship between metabolites in the stool and a child's growth outcome may need to be considered in conjunction with some of the covariates recorded on INSIGHT children and their mothers. To investigate this possibility, we considered 20 potentially relevant covariates (Table 1, see Methods for details). To incorporate these covariates into our analysis, we considered a regression with the child's growth outcome as the response, and predictor terms, which included the butyrate adjusted metabolite concentration (identified from the Phase 1 analysis), 20 covariates, and the interactions between butyrate adjusted metabolite concentration and each covariate. Using a similar regression approach as in Phase 1, we performed: (a) LASSO and

**TABLE 2** Significant predictors of BMI, BMI z-score, and growth index identified in Phase 2 analyses through two different regression procedures: (1) LASSO and (2) BIC best subset selection, each followed by post-selection least squares fits

	BMI		BMI z-score		Growth index	
	LASSO + post-selection fit <sup>a</sup>	BIC best subset + post-selection fit <sup>b</sup>	LASSO + post-selection fit <sup>c</sup>	BIC best subset + post-selection fit <sup>d</sup>	LASSO + post-selection fit <sup>e</sup>	BIC best subset + post-selection fit <sup>f</sup>
Butyrate adjusted metabolite concentration	<b>0.405 (1.02 × 10<sup>-4</sup>)<sup>g</sup></b>	<b>0.406 (8.02 × 10<sup>-5</sup>)<sup>g</sup></b>	<b>0.263 (3.47 × 10<sup>-4</sup>)<sup>g</sup></b>	<b>0.298 (3.30 × 10<sup>-5</sup>)<sup>g</sup></b>	<b>0.361 (2.33 × 10<sup>-4</sup>)<sup>g</sup></b>	<b>0.340 (6.11 × 10<sup>-4</sup>)<sup>g</sup></b>
Maternal smoking status during pregnancy (yes vs. no)	<b>1.01 (9.01 × 10<sup>-3</sup>)<sup>g</sup></b>	<b>1.04 (6.33 × 10<sup>-3</sup>)<sup>g</sup></b>		<b>0.745 (4.91 × 10<sup>-3</sup>)<sup>g</sup></b>	0.661 (0.0822)	
Child's exact age at a 2-year study visit: Butyrate adjusted metabolite concentration		-0.245 (0.0241)		-0.187 (0.0128)	-0.184 (0.0844)	
Per day meats consumption frequency					0.185 (0.0630)	<b>0.252 (0.0103)<sup>g</sup></b>
Maternal gestational weight gain status (exceeded vs. below/at guidelines)					0.371 (0.0597)	<b>0.497 (0.0115)<sup>g</sup></b>
Child sex (male vs. female)					0.363 (0.0565)	
Per day grains consumption frequency				<b>0.157 (0.0252)</b>		

Note: Each cell reports the estimated coefficient with corresponding p-value in parentheses. Significant terms (p-values <0.05) are boldfaced.

<sup>a</sup>For BMI, post LASSO selection (two terms) least squares fit: R<sup>2</sup> = 0.110, adjusted R<sup>2</sup> = 0.0998, and p-value = 5.71 × 10<sup>-5</sup>.

<sup>b</sup>For BMI, post BIC best subset selection (three terms) least squares fit: R<sup>2</sup> = 0.137, adjusted R<sup>2</sup> = 0.122, and p-value = 1.87 × 10<sup>-5</sup>.

<sup>c</sup>For BMI z-score, post LASSO selection (one term) least squares fit: R<sup>2</sup> = 0.0736, adjusted R<sup>2</sup> = 0.0680, and p-value = 3.47 × 10<sup>-4</sup>.

<sup>d</sup>For BMI z-score, post BIC best subset selection (four terms) least squares fit: R<sup>2</sup> = 0.164, adjusted R<sup>2</sup> = 0.144, and p-value = 5.53 × 10<sup>-6</sup>.

<sup>e</sup>For growth index, post LASSO selection (seven terms) least squares fit (six significant terms shown in the table): R<sup>2</sup> = 0.198, adjusted R<sup>2</sup> = 0.163, and p-value = 6.55 × 10<sup>-6</sup>.

<sup>f</sup>For growth index, post BIC best subset selection (three terms) least squares fit: R<sup>2</sup> = 0.134, adjusted R<sup>2</sup> = 0.118, and p-value = 2.58 × 10<sup>-5</sup>.

<sup>g</sup>The significance of these predictors would remain unaffected by Bonferroni's multiple testing p-value adjustments at α = 0.05 level of significance.

(b) BIC best subset selection, each followed by a post-selection least squares fit on the selected terms (see Phase 2 in Figure S1).

When BMI is used as the outcome, the LASSO procedure retained two terms: butyrate adjusted metabolite concentration and maternal smoking status during pregnancy. The post-selection least squares fit on these two terms produced an  $R^2$  of 0.110, an adjusted  $R^2$  of 0.0998, and a  $p$ -value of  $5.71 \times 10^{-5}$ ; both terms with significant positive coefficient estimates ( $\hat{\beta} = 0.405$  and 1.01, respectively) (Table 2). The BIC best subset procedure selected a model with three terms, all significant in the post-selection least squares fit (Table 2 and Figure S11): butyrate adjusted metabolite concentration and maternal smoking status during pregnancy—both with positive coefficient estimates ( $\hat{\beta} = 0.406$  and 1.04, respectively), and the interaction between butyrate adjusted metabolite concentration and the child's exact age—with a negative coefficient estimate ( $\hat{\beta} = -0.245$ ). The  $R^2$  and the adjusted  $R^2$  for this linear model were, respectively, 0.137 and 0.122, and the  $p$ -value was  $1.87 \times 10^{-5}$ .

Next, we performed the selection procedure using BMI z-score as the weight outcome. The LASSO retained only one term, butyrate adjusted metabolic concentration, which remained significant in the post-selection least squares fit (Table 2;  $\hat{\beta} = 0.263$ ,  $R^2 = 0.0736$ , adjusted  $R^2 = 0.0680$ ,  $p$ -value =  $3.47 \times 10^{-4}$ ). Using the BIC best subset procedure, a model with four terms was selected (Figure S12). These four terms remained significant in the post-selection least squares fit (Table 2;  $R^2 = 0.164$ , adjusted  $R^2 = 0.144$ ,  $p$ -value =  $5.53 \times 10^{-6}$ ). Three terms (butyrate adjusted metabolite concentration,  $\hat{\beta} = 0.298$ ; maternal smoking status during pregnancy,  $\hat{\beta} = 0.745$ ; and child's per day grains consumption frequency,  $\hat{\beta} = 0.157$ ) had positive coefficient estimates, and one term (child's exact age at two-year visit interacting with butyrate adjusted metabolite concentration,  $\hat{\beta} = -0.187$ ) had a negative coefficient estimate.

Finally, we repeated the analysis using the same predictors, but using the growth index as the response. The LASSO procedure retained seven terms. The post-selection least squares fit on these seven terms produced an  $R^2$  of 0.198, an adjusted  $R^2$  of 0.163, and a  $p$ -value of  $6.55 \times 10^{-6}$ . However, only one of the seven terms, butyrate adjusted metabolite concentration, was significant ( $p$ -value =  $2.33 \times 10^{-4}$ ; Table 2) with a positive coefficient estimate ( $\hat{\beta} = 0.361$ ). The BIC best subset procedure selected a model with three terms, all with significant positive coefficient estimates in the post-selection least squares fit—butyrate adjusted metabolite concentration ( $\hat{\beta} = 0.340$ ), child's daily meats consumption frequency ( $\hat{\beta} = 0.252$ ), and maternal gestational weight gain status ( $\hat{\beta} = 0.497$ ) (Table 2 and Figure S13;  $R^2 = 0.134$ , adjusted  $R^2 = 0.118$ ,  $p$ -value =  $2.58 \times 10^{-5}$ ).

### 3.4 | Butyrate concentration is associated with the abundance of butyrate synthesizing Firmicutes

Butyrate is a metabolic product of the microbiota.<sup>46</sup> To connect our results to prior analyses of the gut microbiota that we had published using the same samples from children enrolled in INSIGHT,<sup>20</sup> we tested for associations between bacterial groups and butyrate adjusted

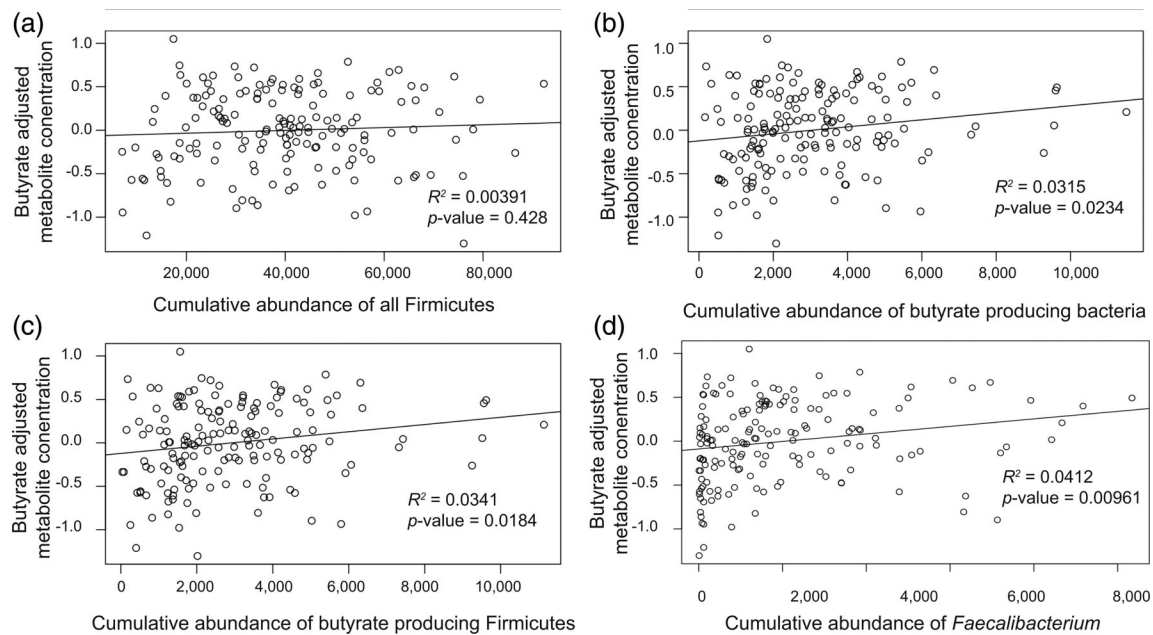
metabolite concentration. First, we considered the cumulative abundance (determined by classified normalized read counts from 16S rDNA gene sequencing) of all bacteria assigned to the Firmicutes phylum, as the majority of butyrate-producing bacteria belong to this phylum.<sup>46</sup> We found no significant association between this abundance and butyrate adjusted metabolite concentration ( $R^2 = 0.00391$ ,  $p$ -value = 0.428; Figure 3(A)). Second, we only considered bacteria that could synthesize butyrate, as identified by Vital and colleagues<sup>46</sup> based on the presence of genes encoding enzymes in the butyrate synthesis pathway, and tested the association of their abundances with butyrate concentration. We found that the abundance of bacteria in this group did correlate positively with butyrate adjusted metabolite concentration ( $R^2 = 0.0315$ ,  $p$ -value = 0.0234; Figure 3(B)). Third, among these butyrate-producing bacteria, we further focused on Firmicutes, which represent the majority of bacteria within this group. This increased the significance of the positive association with butyrate adjusted metabolite concentration ( $R^2 = 0.0341$ ,  $p$ -value = 0.0184; Figure 3(C)). Finally, among butyrate-producing Firmicutes, those belonging to the genus *Faecalibacterium* seem to drive the observed relationship, with an even stronger association between their abundance and butyrate adjusted metabolite concentration ( $R^2 = 0.0412$ ,  $p$ -value =  $9.61 \times 10^{-3}$ ; Figure 3(D)). Thus, butyrate-producing Firmicutes, and those in the *Faecalibacterium* genus in particular, appear to be important contributors to determining the stool butyrate adjusted metabolite concentration in our samples.

## 4 | DISCUSSION

In this study, we found that stool butyrate concentration has a significant positive association with a child's weight outcomes at age 2 years. It is consistently selected as a significant predictor in all regression models for each one of the three child weight outcomes we considered as the response—BMI, BMI z-score and growth index. Butyrate is a SCFA, a group of metabolic products formed by anaerobic microbial fermentation of non-digestible carbohydrates (such as resistant starch, dietary fibre and other polysaccharides) in the host small intestine and colon.<sup>50</sup> The SCFAs, acetate and propionate, are largely produced by the Bacteroidetes phylum, whereas butyrate is mainly the product of the Firmicutes phylum.<sup>50</sup> SCFAs can influence host physiology by signalling through the free fatty acid receptors GPR41 and GPR43. When activated, these receptors lead to an increase in the satiety hormone PYY and in intestinal mobility. GPR41 activation also stimulates the expression of leptin in adipocytes and increases hepatic lipogenesis. Furthermore, butyrate and propionate play a role in decreasing appetite through the formation of the gut hormone GLP-1.<sup>50</sup>

To date, the evidence concerning the association of butyrate with weight gain has been mixed. However, consistent with our findings, some studies have shown higher levels of butyrate in children and adults with obesity.<sup>16,51</sup> This could be due to a higher abundance of gut bacteria that ferment carbohydrates in individuals with obesity, which subsequently results in an increase in production of SCFAs.





**FIGURE 3** Relationship between the abundance of different groups of bacteria and butyrate adjusted metabolite concentration. (A) All bacteria classified as Firmicutes; (B) All bacteria identified in Vital et al. as having butyrate synthesis potential; (C) Firmicute bacteria with butyrate synthesis potential, as identified in Vital et al.; (D) *Faecalibacterium* (one outlier was removed). The black solid lines indicate the fitted linear regression models; the  $R^2$  and the  $p$ -values correspond to Pearson's correlation

This increase in SCFAs can provide extra energy to the host that can be stored as lipids or glucose.<sup>52</sup> Other studies have shown an anti-inflammatory/anti-obesity role for butyrate.<sup>50,53,54</sup> This metabolite is likely to play an important role in gut homeostasis, possibly through the gut-brain axis.<sup>55,56</sup> However, investigating the mechanism by which butyrate influences a child's weight is outside the scope of this study—a larger, longitudinal study collecting data on the host gut cell physiology and the utilization of metabolites (as in Cuesta-Zuluaga et al.<sup>15</sup>) would be necessary to elucidate mechanistically the metabolome's effects on human health. This study provides evidence of an association between butyrate adjusted metabolite concentration and child weight outcomes at the age of 2 years. This brings us closer to more fully understanding factors, which could impact an individual's risk of developing obesity.

As mentioned above, butyrate (and other SCFAs) could act as an intermediary between the gut microbiota and the host physiology. Since we have previously characterized the microbiota of these same samples, it seemed logical to examine the relationship between the microbiota and butyrate. When we combined the abundance of all bacteria categorized as Firmicutes, there was no significant association between bacterial abundance and butyrate concentration. However, although the majority of bacteria that synthesize butyrate belong to the Firmicutes phylum, not all Firmicutes produce butyrate. Instead, we grouped bacteria according to function (rather than according to phylogeny) based on the presence of butyrate synthesis pathway genes.<sup>46</sup> This approach proved successful in identifying *Faecalibacterium* as potentially driving the association between the microbiota and butyrate concentration. These results are consistent with literature on *Faecalibacterium* as a butyrate producer<sup>57,58</sup> being

found in higher concentrations in the gut of children with obesity when compared to children with normal weight.<sup>16,59,60</sup> It is of course important to note that our study established statistical associations among butyrate, microbiota, and BMI—not causative links. We could be observing a compensatory increase in butyrate production to counter the effects of increased BMI or we could be observing a change in butyrate production as a result of a dysbiosis in the gut microbiota. Furthermore, we are also only investigating an association between butyrate concentration and microbiota abundance, which is different from a direct assessment of the microbiota's production of butyrate. There are other potential sources for butyrate in the gut that we cannot account for (e.g., archaea, fungal, dietary sources etc.). In order to properly unravel causality, a large, longitudinal study that records changes in microbiota community composition (including all microorganisms and not just bacteria), anthropometric measurements, physiological measurements of gut cell health, and relevant covariate data (e.g., diet, health history, etc.) is essential.

We used three related outcomes in order to show stability in the results of the statistical analyses. For example, butyrate was selected as a significant variable regardless of the weight outcome considered. The selection of covariates was less consistent, although several of them were selected across multiple outcomes. One such strong finding was maternal smoking status during pregnancy, which had a significant positive relationship with BMI and BMI z-score based on both regression procedures, and a marginally significant positive relationship with growth index based on one of the regression procedures. Many previous studies have documented a role of maternal smoking status during pregnancy, which was found to be positively associated with obesity risk for the child.<sup>30,61–68</sup> However, there is no consensus

on the underlying mechanism linking maternal smoking to a child's weight gain.

Another strong covariate found to have a significant negative association with two weight outcomes (BMI and BMI z-score) and a marginally significant for the third (growth index) was the interaction between a child's exact age (in months) and butyrate adjusted metabolite concentration. This negative statistical effect on weight outcome suggests that the positive statistical effect of butyrate adjusted metabolite concentration in relation to weight outcome decreases in older children (here, the term “statistical effect” refers to the estimated coefficient in the regression model). This could potentially be related to the plasticity of the gut microbiota. For instance, Rivière and colleagues<sup>69</sup> found that the abundance of Clostridial clusters is low immediately after birth, and increases between the ages of 6 and 24 months. Not until around 6 years of age<sup>70</sup> are these bacteria at a higher abundance where they remain through adulthood. Note that the range of children's ages in our study is very narrow (24.01–25.49 months, as recorded at the two-year study visit when stool samples were collected)—yet these ~1.5 months could be sufficient for the gut microbiota to have partially evolved in composition, particularly if there is a concurrent evolution in diet, health, or behaviour. However, to the best of our knowledge, microbiota changes during this small amount of time have not been evaluated previously and should be considered in future studies.

Finally, there are several covariates that are statistically significantly associated with just one weight outcome, such as, excessive maternal gestational weight gain, frequency of a child's per day meats consumption (significantly positive coefficient estimates in regressions with growth index), and frequency of a child's per day grains consumption (significant positive coefficient estimate in regressions with BMI z-score). These variables have also been shown in the literature to be associated with child weight<sup>71–73</sup> but in this study, their association with child's weight seems to be less compelling. However, these data could serve to support future hypotheses regarding the impacts of maternal factors and diet on children's weight trajectories.

#### 4.1 | Statistical considerations

The LASSO and the BIC best subset selection regression procedures produced non-identical but consistent results for the three weight outcomes considered—BMI, BMI z-score and growth index. Many other statistical techniques could be used for analysing the metabolomics data. Among the ones most frequently used in the field are Principal Components Analysis<sup>74,75</sup> and clustering,<sup>76</sup> which are unsupervised techniques, and orthogonal partial least squares—regression/discriminant analysis (OPLS/OPLS-DA<sup>77</sup>), which are supervised techniques.<sup>75,78–80</sup> The regression approach we employed in this study provides a simple and useful alternative. Its main advantage is its interpretability, as we are able to focus on the selection of individual predictor terms (metabolites, covariates, and interactions) instead of linear combinations, as produced by methods such as the OPLS. Additionally,

as previously mentioned, the use of three weight outcomes, and of two regression procedures for each, helps us validate our selection of relevant variables and gauge the robustness of our conclusion. Variables that are selected across multiple outcomes and multiple methods display stable statistical signals. From a different perspective, considering multiple outcomes may also reveal associations that are outcome-specific, and should be considered when conducting multi-study meta-analyses.

Our analyses relied on the tuning of the penalization parameter in the LASSO and on the choice of a criterion for the best subset selection regression procedures. We tuned the LASSO penalization parameter minimizing a 10-fold cross-validation error, averaged over 100 repetitions of the cross-validation random fold splits. For best subset selection, we used the BIC—but other criteria such as the Akaike Information Criterion<sup>81,82</sup> or Mallows' Cp<sup>83,84</sup> could also be used.

#### 4.2 | Limitations and future directions

The individuals enrolled in the INSIGHT study<sup>19</sup> were recruited from Central Pennsylvania and are largely of European descent (see Table S1). Consequently, our results might not generalize to other populations. Furthermore, our sample size ( $n = 170$ ) is rather modest—though similar to those of other metabolomic studies.<sup>85,86</sup> The sample size is one consideration for deciding between a targeted versus an untargeted metabolomics approach, as the dimension of the data increases by orders of magnitude in an untargeted metabolomics approach. In our case, with a modest sample size, a targeted metabolomics approach was more realistic. However, this does limit our ability to detect novel associations. It is possible that metabolites we did not query could impact child weight as well. A significant advantage of our cohort over some other cohorts studied is its deep characterization, that is, the wealth of information on children's and mothers' covariates that were collected and incorporated into our analyses. Nevertheless, for a more comprehensive understanding of the effect of the metabolome on children weight outcomes, our results should be replicated with larger cohorts of children belonging to different ethnicities, ideally also incorporating additional phenotypic and clinical covariates along with longitudinal information. The results that we discuss here are cross-sectional (i.e., the data were collected at a single time point) and as such we can only report on associations between the metabolites and child weight. A longitudinal study would be able to address how changes in the metabolites relate to changes in a child's weight over time.

While the present study focused on metabolites, multiple cross-sectional “omics” data sets are available or are in the process of being generated for the INSIGHT children cohort—including data on microbiota,<sup>20</sup> genetic<sup>87</sup> and epigenetic variants. Moreover, longitudinal metabolomic and microbiomic profiles are being collected for the SIBSIGHT cohort<sup>38,88</sup> (second born siblings of the children within INSIGHT). This will allow us to leverage “multi-omics” approaches to consolidate and expand our findings on childhood obesity. “Multi-

omics” approaches are indeed becoming increasingly popular to gain a comprehensive understanding of diseases—including obesity and other metabolic disorders.<sup>13,89–93</sup>

## ACKNOWLEDGMENTS

We thank the INSIGHT families for their participation in this study and to the study nurses for their assistance in collecting samples and covariate information. We acknowledge help from Jingtao Zhang (2D-NMR sample processing), Imhoi Koo, and Wei Gui (instructions on the use of Chenomx software). We thank Ana Kenney for her input on statistical analyses and Tapas Mal for facilitating our access to the NMR facilities. Finally, we are grateful to all the three anonymous reviewers, the Manuscript Editor (Dr. Wei Perng), and the Editor-in-Chief (Dr. Michael Goran) for their time in providing insightful feedback on our work that has helped us greatly improve this manuscript. Funds provided by the National Science Foundation - Division of Mathematical Sciences (NSF-DMS; grant #1407639) supported Debmalya Nandy and Francesca Chiaromonte. Additionally, funds from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institutes of Health (NIH; grants #R01DK088244 and #R01DK99364); the Penn State Eberly College of Science; the Penn State Institute for Computational and Data Sciences; the National Center for Research Resources; and the National Center for Advancing Translational Sciences, NIH (grant #UL1TR000127), also supported our research. Finally, the Pennsylvania Department of Health provided partial support from the Tobacco Settlement and the CURE funds (the Department specifically disclaims responsibility for any analyses, interpretations or conclusions). Open Access Funding provided by Scuola Superiore Sant’Anna within the CRUI-CARE Agreement.

## CONFLICT OF INTERESTS

The authors declare no financial or other conflicts of interests.

## AUTHOR CONTRIBUTIONS

Study conception and design (Sarah J. C. Craig and Kateryna D. Makova), data generation (Sarah J. C. Craig, Debmalya Nandy, Andrew D. Patterson, Jingwei Cai, and Yuan Tian), statistical analyses (Debmalya Nandy, Francesca Chiaromonte, Matthew L. Reimherr, and Sarah J. C. Craig), resources (Kateryna D. Makova, Francesca Chiaromonte, Ian M. Paul, Emily E. Hohman, Michele E. Marini, and Jennifer S. Savage), manuscript preparation (Sarah J. C. Craig, Debmalya Nandy, Matthew L. Reimherr, Francesca Chiaromonte, Kateryna D. Makova, and comments from all co-authors).

## ORCID

Sarah J. C. Craig  <https://orcid.org/0000-0002-7133-1646>

## REFERENCES

- Ogden CL, Carroll MD, Fryar CD, Flegal KM. Prevalence of obesity among adults and youth: United States, 2011–2014. *NCHS Data Brief*. 2015;(219):1–8.
- Skinner AC, Ravanbakht SN, Skelton JA, Perrin EM, Armstrong SC. Prevalence of obesity and severe obesity in US children, 1999–2016. *Pediatrics*. 2018;141(3):1–9. <https://doi.org/10.1542/peds.2017-3459>
- Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of obesity among adults and youth: United States, 2015–2016. *NCHS Data Brief*. 2017;(288):1–8.
- Zhang T, Whelton PK, Xi B, et al. Rate of change in body mass index at different ages during childhood and adult obesity risk. *Pediatr Obes*. 2019;14(7):e12513.
- Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*. 2006;444(7122):1027–1031.
- Ley RE. Obesity and the human microbiome. *Curr Opin Gastroenterol*. 2010;26(1):5–11.
- Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. *Nature*. 2009;457(7228):480–484.
- Boulangé CL, Neves AL, Chilloux J, Nicholson JK, Dumas M-E. Impact of the gut microbiota on inflammation, obesity, and metabolic disease. *Genome Med*. 2016;8(1):42.
- Mohammadkhalil AI, Simpson EB, Patterson SG, Ferguson JF. Development of the gut microbiome in children, and lifetime implications for obesity and Cardiometabolic disease. *Children*. 2018;5(12):1–20. <https://doi.org/10.3390/children5120160>
- McCann J, Rawls J, Seed P, Armstrong S. The intestinal microbiome and childhood obesity. *Curr Pediatr Rep*. 2017;5(3):150–155.
- Lin Y, Ma C, Bezabeh T, et al. 1 H NMR-based metabolomics reveal overlapping discriminatory metabolites and metabolic pathway disturbances between colorectal tumor tissues and fecal samples. *Int J Cancer*. 2019. 145(6):1679–1689. <https://doi.org/10.1002/ijc.32190>
- Anderson JR, Chokesuwattanasakul S, Phelan MM, et al. 1H NMR metabolomics identifies underlying inflammatory pathology in osteoarthritis and rheumatoid arthritis synovial joints. *J Proteome Res*. 2018;17(11):3780–3790. <https://doi.org/10.1021/acs.jproteome.8b00455>
- Vallianou N, Stratigou T, Christodoulatos GS, Dalamaga M. Understanding the role of the gut microbiome and microbial metabolites in obesity and obesity-associated metabolic disorders: current evidence and perspectives. *Curr Obes Rep*. 2019;8(3):317–332.
- Vernocchi P, Del Chierico F, Putignani L. Gut microbiota profiling: metabolomics based approach to unravel compounds affecting human health. *Front Microbiol*. 2016;7(1144):1–21. <https://doi.org/10.3389/fmicb.2016.01144>
- de la Cuesta-Zuluaga J, de la Cuesta-Zuluaga J, Mueller N, et al. Higher fecal short-chain fatty acid levels are associated with gut microbiome dysbiosis, obesity, hypertension and cardiometabolic disease risk factors. *Nutrients*. 2018;11(1):51. <https://doi.org/10.3390/nu11010051>
- Payne AN, Chassard C, Zimmermann M, Müller P, Stinca S, Lacroix C. The metabolic activity of gut microbiota in obese children is increased compared with normal-weight children and exhibits more exhaustive substrate utilization. *Nutr Diabetes*. 2011;1:e12.
- Saner C, Harcourt BE, Pandey A, et al. Sex and puberty-related differences in metabolomic profiles associated with adiposity measures in youth with obesity. *Metabolomics*. 2019;15(5):75.
- Bervoets L, Massa G, Guedens W, Reekmans G, Noben J-P, Adriaenssens P. Identification of metabolic phenotypes in childhood obesity by H NMR metabolomics of blood plasma. *Future Sci OA*. 2018;4(6):FSO310.
- Paul IM, Williams JS, Anzman-Frasca S, et al. The intervention nurses start infants growing on healthy trajectories (INSIGHT) study. *BMC Pediatr*. 2014;14:184.
- Craig SJC, Blankenberg D, Parodi ACL, et al. Child weight gain trajectories linked to oral microbiota composition. *Sci Rep*. 2018;8(1):14030.
- Barnes S, Benton HP, Casazza K, et al. Training in metabolomics research. I. Designing the experiment, collecting and extracting

- samples and generating metabolomics data. *J Mass Spectrom.* 2016; 51(7):ii-iii. <https://doi.org/10.1002/jms.3672>
22. Emwas A-H, Roy R, McKay RT, et al. NMR spectroscopy for metabolomics research. *Metabolites.* 2019;9(7):123. <https://doi.org/10.3390/metabo9070123>
  23. Tian Y, Zhang L, Wang Y, Tang H. Age-related topographical metabolic signatures for the rat gastrointestinal contents. *J Proteome Res.* 2012;11(2):1397-1411.
  24. Tian Y, Nichols RG, Cai J, Patterson AD, Cantorna MT. Vitamin a deficiency in mice alters host and gut microbial metabolism leading to altered energy homeostasis. *J Nutr Biochem.* 2018;54:28-34.
  25. Paul IM, Savage JS, Anzman-Frasca S, et al. Effect of a responsive parenting educational intervention on childhood weight outcomes at 3 years of age: the INSIGHT randomized clinical trial. *JAMA.* 2018; 320(5):461-468.
  26. Wan S, Guo M, Zhang T, et al. Impact of exposure to antibiotics during pregnancy and infancy on childhood obesity: a systematic review and meta-analysis. *Obesity.* 2020;28(4):793-802.
  27. Voerman E, Santos S, Patro Golab B, et al. Maternal body mass index, gestational weight gain, and the risk of overweight and obesity across childhood: an individual participant data meta-analysis. *PLoS Med.* 2019;16(2):e1002744.
  28. Masukume G, O'Neill SM, Baker PN, Kenny LC, Morton SMB, Khashan AS. The impact of caesarean section on the risk of childhood overweight and obesity: new evidence from a contemporary cohort study. *Sci Rep.* 2018;8(1):15113.
  29. Wang J, Wang L, Liu H, et al. Maternal gestational diabetes and different indicators of childhood obesity: a large study. *Endocrine Connections.* 2018;7(12):1464-1471. <https://doi.org/10.1530/EC-18-0449>
  30. Ino T. Maternal smoking during pregnancy and offspring obesity: meta-analysis. *Pediatr Int.* 2010;52(1):94-99. <https://doi.org/10.1111/j.1442-200x.2009.02883.x>
  31. Rose CM, Birch LL, Savage JS. Dietary patterns in infancy are associated with child diet and weight outcomes at 6 years. *Int J Obes.* 2017; 41(5):783-788.
  32. Nichols RG, Cai J, Murray IA, et al. Structural and functional analysis of the gut microbiome for toxicologists. *Curr Protoc Toxicol.* 2018;78 (1):1-39. <https://doi.org/10.1002/cptx.54>
  33. Euceda LR, Giskeødegård GF, Bathen TF. Preprocessing of NMR metabolomics data. *Scand J Clin Lab Invest.* 2015;75(3):193-203.
  34. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal Chem.* 2006;78(13):4430-4442.
  35. Karaman I. Preprocessing and pretreatment of metabolomics data for statistical analysis. *Adv Exp Med Biol.* 2017;965:145-161.
  36. Draper NR, Smith H. Applied Regression Analysis Wiley Series in Probability and Statistics Published online. 1998 doi:<https://doi.org/10.1002/9781118625590>
  37. Blum RE, Wei EK, Rockett HR, et al. Validation of a food frequency questionnaire in native American and Caucasian children 1 to 5 years of age. *Matern Child Health J.* 1999;3(3):167-172.
  38. Hohman EE, Savage JS, Birch LL, Paul IM. The intervention nurses start infants growing on healthy trajectories (INSIGHT) responsive parenting intervention for firstborns affects dietary intake of Secondborn infants. *J Nutr.* 2020;150(8):2139-2146. <https://doi.org/10.1093/jn/nxaa135>
  39. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc: Ser B (Stat Methodol).* 1996;58(1):267-288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
  40. Schwarz G. Estimating the dimension of a model. *Ann Stat.* 1978;6(2): 461-464. <https://doi.org/10.1214/aos/1176344136>
  41. Murphy KP. *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: The MIT Press; 2012.
  42. Hocking RR, Leslie RN. Selection of the best subset in regression analysis. *Technometrics.* 1967;9(4):531-540. <https://doi.org/10.1080/00401706.1967.10490502>
  43. R Core Team. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. <https://www.R-project.org/>
  44. Boulesteix AL, Fuchs M. ipflasso: Integrative Lasso with penalty factors. R package version 0 1: <https://CRAN.R-project.org/package=ipflasso>. Published online 2015.
  45. Lumley T. leaps: Regression subset selection (using Fortran code by Alan Miller) [Software]. <http://CRAN.R-project.org/package=eaps> (Rpackageversion2.9). Published online 2009.
  46. Vital M, Howe AC, Tiedje JM. Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *MBio.* 2014;5(2): e00889.
  47. Harrell FE Jr, From Charles Dupont WC, Others. M. Hmisc: Harrell Miscellaneous. Published online 2020. <https://CRAN.R-project.org/package=Hmisc>
  48. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Published online 2016. <https://ggplot2.tidyverse.org>
  49. Clinical Growth Charts. Published January 11, 2019. Accessed September 24, 2020. [https://www.cdc.gov/growthcharts/clinical\\_charts.htm](https://www.cdc.gov/growthcharts/clinical_charts.htm)
  50. Chakraborti CK. New-found link between microbiota and obesity. *World J Gastrointest Pathophysiol.* 2015;6(4):110-119.
  51. Liu H, Wang J, He T, et al. Butyrate: a double-edged sword for health? *Adv Nutr.* 2018;9(1):21-29.
  52. Murugesan S, Nirmalkar K, Hoyo-Vadillo C, García-Espitia M, Ramírez-Sánchez D, García-Mena J. Gut microbiome production of short-chain fatty acids and obesity in children. *Eur J Clin Microbiol Infect Dis.* 2018;37(4):621-625.
  53. Chang PV, Hao L, Offermanns S, Medzhitov R. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proc Natl Acad Sci.* 2014;111(6):2247-2252. <https://doi.org/10.1073/pnas.1322269111>
  54. Li X, Xuan LI, Shimizu Y, Kimura I. Gut microbial metabolite short-chain fatty acids and obesity. *Biosci Microbiota Food Health.* 2017;36 (4):135-140. <https://doi.org/10.12938/bmfh.17-010>
  55. Stilling RM, van de Wouw M, Clarke G, Stanton C, Dinan TG, Cryan JF. The neuropharmacology of butyrate: the bread and butter of the microbiota-gut-brain axis? *Neurochem Int.* 2016;99:110-132.
  56. Silva YP, Bernardi A, Frozza RL. The role of short-chain fatty acids from gut microbiota in gut-brain communication. *Front Endocrinol.* 2020;11:25.
  57. Zhou L, Zhang M, Wang Y, et al. Faecalibacterium prausnitzii produces butyrate to maintain Th17/Treg balance and to ameliorate colorectal colitis by inhibiting histone Deacetylase 1. *Inflamm Bowel Dis.* 2018;24(9):1926-1940.
  58. Zhang M, Zhou L, Wang Y, et al. Faecalibacterium prausnitzii produces butyrate to decrease c-Myc-related metabolism and Th17 differentiation by inhibiting histone deacetylase 3. *Int Immunol.* 2019;31 (8):499-514.
  59. Del Chierico F, Abbatini F, Russo A, et al. Gut microbiota markers in obese adolescent and adult patients: age-dependent differential patterns. *Front Microbiol.* 2018;9:1210.
  60. Riva A, Borgo F, Lassandro C, et al. Pediatric obesity is associated with an altered gut microbiota and discordant shifts in Firmicutes populations. *Environ Microbiol.* 2017;19(1):95-105. <https://doi.org/10.1111/1462-2920.13463>
  61. von Kries R, von Kries R, Toschke AM, Koletzko B, Slikker W. Maternal smoking during pregnancy and childhood obesity. *Obstet Gynecol Surv.* 2003;58(5):297-298. <https://doi.org/10.1097/00006254-200305000-00005>
  62. Oken E, Levitan EB, Gillman MW. Maternal smoking during pregnancy and child overweight: systematic review and meta-analysis. *Int J Obes.* 2008;32(2):201-210.



63. Toschke AM, Koletzko B, Slikker W Jr, Hermann M, von Kries R. Childhood obesity is associated with maternal smoking in pregnancy. *Eur J Pediatr*. 2002;161(8):445-448.
64. Widerøe M, Vik T, Jacobsen G, Bakketeig LS. Does maternal smoking during pregnancy cause childhood overweight? *Paediatr Perinat Epidemiol*. 2003;17(2):171-179.
65. Ong KKL. Size at birth and early childhood growth in relation to maternal smoking, parity and infant breast-feeding: longitudinal birth cohort study and analysis. *Pediatr Res*. 2002;52(6):863-867. <https://doi.org/10.1203/01.pdr.0000036602.81878.6d>
66. Ong KKL. Association between postnatal catch-up growth and obesity in childhood: prospective cohort study. *BMJ*. 2000;320(7240):967-971. <https://doi.org/10.1136/bmj.320.7240.967>
67. Salsberry PJ, Reagan PB. Dynamics of early childhood overweight. *Pediatrics*. 2005;116(6):1329-1338.
68. Dubois L, Girard M. Early determinants of overweight at 4.5 years in a population-based longitudinal study. *Int J Obes*. 2006;30(4):610-617.
69. Rivièrè A, Selak M, Lantin D, Leroy F, De Vuyst L. Bifidobacteria and butyrate-producing colon bacteria: importance and strategies for their stimulation in the human gut. *Front Microbiol*. 2016;7:979.
70. Balamurugan R, Janardhan HP, George S, Chittaranjan SP, Ramakrishna BS. Bacterial succession in the colon during childhood and adolescence: molecular studies in a southern Indian village. *Am J Clin Nutr*. 2008;88(6):1643-1647. <https://doi.org/10.3945/ajcn.2008.26511>
71. Barbour LA, Hernandez TL. Maternal non-glycemic contributors to fetal growth in obesity and gestational diabetes: spotlight on lipids. *Curr Diab Rep*. 2018;18(6):37.
72. Vergnaud A-C, Norat T, Romaguera D, et al. Meat consumption and prospective weight change in participants of the EPIC-PANACEA study. *Am J Clin Nutr*. 2010;92(2):398-407. <https://doi.org/10.3945/ajcn.2009.28713>
73. You W, Henneberg M. Meat consumption providing a surplus energy in modern diet contributes to obesity prevalence: an ecological analysis. *BMC Nutr*. 2016;2(22):1-11. <https://doi.org/10.1186/s40795-016-0068-4>
74. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom Intell Lab Syst*. 1987;2(1-3):37-52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9)
75. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Res*. 2012;40:W127-W133.
76. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM Comput Surv*. 1999;31(3):264-323. <https://doi.org/10.1145/331499.331504>
77. Eriksson L, Byrne T, Johansson E, Trygg J, Vikström C. Multi- and Megavariate Data Analysis Basic Principles and Applications. Umetrics Academy. 2013.
78. Liland KH. Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis. *TrAC Trends Anal Chem*. 2011;30(6):827-841. <https://doi.org/10.1016/j.trac.2011.02.007>
79. Worley B, Powers R. Multivariate analysis in metabolomics. *Curr Metabolomics*. 2013;1(1):92-107.
80. Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J*. 2013;4:e201301009.
81. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control*. 1974;19(6):716-723. <https://doi.org/10.1109/tac.1974.1100705>
82. Akaike H. Information theory and an extension of the maximum likelihood principle. In Parzen E., Tanabe K., Kitagawa G. (eds) *Selected Papers of Hirotugu Akaike*; Springer Series in Statistics (Perspectives in Statistics) New York, NY: Springer; 1998:199-213. [https://doi.org/10.1007/978-1-4612-1694-0\\_15](https://doi.org/10.1007/978-1-4612-1694-0_15)
83. Mallows CL. Some comments on Cp. *Technometrics*. 2000;42(1):87-94.
84. Gilmour SG. The interpretation of Mallows's C p -statistic. *The Statistician*. 1996;45(1):49. <https://doi.org/10.2307/2348411>
85. Rauschert S, Uhl O, Koletzko B, Hellmuth C. Metabolomic biomarkers for obesity in humans: a short review. *Ann Nutr Metab*. 2014;64(3-4):314-324.
86. Zhao X, Gang X, Liu Y, Sun C, Han Q, Wang G. Using Metabolomic profiles as biomarkers for insulin resistance in childhood obesity: a systematic review. *J Diabetes Res*. 2016;2016:8160545.
87. Craig SJC, Kenney AM, Lin J, et al. Polygenic risk score based on weight gain trajectories is a strong predictor of childhood obesity. <https://doi.org/10.1101/606277>
88. Ruggiero CF, Hohman EE, Birch LL, Paul IM, Savage JS. The intervention nurses start infants growing on healthy trajectories (INSIGHT) responsive parenting intervention for firstborns impacts feeding of secondborns. *Am J Clin Nutr*. 2020;111(1):21-27.
89. Son MJ, Yang G-J, Jo E-H, et al. Association of atopic dermatitis with obesity via a multi-omics approach. *Medicine*. 2019;98(29):e16527. <https://doi.org/10.1097/md.00000000000016527>
90. Benítez-Páez A, Kjølbæk L, del Pulgar EMG, et al. A multi-omics approach to unraveling the microbiome-mediated effects of Arabinoxylan oligosaccharides in overweight humans. *mSystems*. 2019;4(4):1-16. <https://doi.org/10.1128/msystems.00209-19>
91. Martino D, Ben-Othman R, Harbeson D, Bosco A. Multiomics and systems biology are needed to unravel the complex origins of chronic disease. *Challenges*. 2019;10(1):23. <https://doi.org/10.3390/challe10010023>
92. Virzi GM, Clementi A, Battaglia GG, Ronco C. Multi-Omics approach: new potential key mechanisms implicated in Cardiorenal syndromes. *Cardiorenal Med*. 2019;9(4):201-211.
93. Madrid-Gambin F, Föcking M, Sabherwal S, et al. Integrated Lipidomics and proteomics point to early blood-based changes in childhood preceding later development of psychotic experiences: evidence from the Avon longitudinal study of parents and children. *Biol Psychiatry*. 2019;86(1):25-34.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Nandy D, Craig SJC, Cai J, et al.

Metabolomic profiling of stool of two-year old children from the INSIGHT study reveals links between butyrate and child weight outcomes. *Pediatric Obesity*. 2022;17(1):e12833.

<https://doi.org/10.1111/ijpo.12833>