nature medicine

LETTERS

https://doi.org/10.1038/s41591-018-0216-2

Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes

Aimee M. Baumann-Dudenhoeffer[1,2*], Alaric W. D'Souza[2], Phillip I. Tarr[3,4], Barbara B. Warner[1] and Gautam Dantas[2,4,5,6*]

Commensal gut bacterial communities (microbiomes) are predicted to influence human health and disease[1,2]. Neonatal gut microbiomes are colonized with maternal and environmental flora and mature toward a stable composition over 2–3 years[3,4]. To study pre- and postnatal determinants of infant microbiome development, we analyzed 402 fecal metagenomes from 60 infants aged 0–8 months, using longitudinal generalized linear mixed models (GLMMs). Distinct microbiome signatures correlated with breastfeeding, formula ingredients, and maternal gestational weight gain (GWG). Amino acid synthesis pathway accretion in breastfed microbiomes complemented normative breastmilk composition. Prebiotic oligosaccharides, designed to promote breastfed-like microflora[5], predicted functional pathways distinct from breastfed infant microbiomes. Soy formula in six infants was positively associated with Lachnospiraceae and pathways suggesting a short-chain fatty acid (SCFA)-rich environment, including glycerol to 1-butanol fermentation, which is potentially dysbiotic. GWG correlated with altered carbohydrate degradation and enriched vitamin synthesis pathways. Maternal and postnatal antibiotics predicted microbiome alterations, while delivery route had no persistent effects. Domestic water source correlates suggest water may be an underappreciated determinant of microbiome acquisition. Clinically important microbial pathways with statistically significant dietary correlates included dysbiotic markers[6,7], core enterotype features[8], and synthesis pathways for enteroprotective[9] and immunomodulatory[10,11] metabolites, epigenetic mediators[1], and developmentally critical vitamins[12], warranting further investigation.

Commensal gut microbes contribute to pathogen exclusion, nutrient acquisition, and immune recognition, thereby preventing or modulating multiple human pathologies[1,2]. Understanding determinants of early microbiome establishment can guide health-promotion and disease-prevention efforts.

Human milk provides optimal infant nutrition[12], and favors gut *Bifidobacterium* and *Lactobacillus* spp[4,13]. While commercial formulas closely approximate breastmilk composition[12,14], and galacto- and fructo-oligosaccharides (GOS and FOS, respectively) are designed to mimic human milk oligosaccharides[5], breastfed and formula-fed infant gut microbiomes remain distinct[4,12]. The impact of specific formula ingredients on gut microbiome acquisition is underdetermined.

To test the hypothesis that specific formula components alter the taxa and gene-encoded functions of developing gut microbiomes, we whole-metagenome shotgun sequenced 402 frozen fecal samples collected monthly from 60 healthy twins (median gestational age, 37 weeks) from birth to 8 months[3,13] (Supplementary Table 1). We constructed longitudinal GLMMs for taxa and genetically encoded functional pathways (for brevity, hereafter referred to as 'pathways') inferred using MetaPhlAn2 and HUMAnN2; all P values are two-tailed, from maximum-likelihood GLMMs Tukey-corrected for multiple comparisons (Methods and Supplementary Tables 2–7). This study, approved by the Human Research Protection Office of Washington University School of Medicine, complied with all ethical regulations. Written informed consent was obtained for all subjects.

We identified multiple known determinants of gut microbiome assembly, confirming the validity of our approach. (Fig. 1 and Supplementary Fig. 1)[4,15–17]. Alpha diversity (Shannon index) correlated positively with time ($n=402$ samples, $P<0.001$) and fruit/vegetable exposure ($n=160$, $P=0.011$), and negatively with maternal intrapartum ampicillin-sulbactam ($n=46$, $P=0.005$) and any postnatal antibiotics ($n=49$, $P=0.043$). Bifidobacteriaceae enrichment correlated with >50% breastfeeding ($n=75$, $P=0.003$) and lifetime GOS exposure ($n=204$, $P=0.005$). Lachnospiraceae increased with time ($n=402$, $P<0.001$) and decreased with any breastfeeding ($n=125$, $P=0.014$), Enterobacteriaceae decreased with time ($n=402$, $P<0.001$) and GOS ($n=204$, $P=0.003$), and Bacteroidaceae decreased with Cesarean delivery ($n=227$, $P=0.003$) and increased with fruit/vegetable exposure ($n=160$, $P=0.004$).

Breastfed infant gut microbiomes accrued amino acid synthesis pathways that complemented breastmilk's changing amino acid content[14], suggesting that parallel milk and microbiome development may reflect physiologic adaptation (Fig. 2). Majority-breastfed (>50%) infant gut microbiomes ($n=75$) had significantly more methionine ($P<0.001$), branched-chain amino acids (BCAA, isoleucine/leucine/valine, $P=0.020$), cysteine/serine ($P=0.012$), threonine ($P=0.004$), and arginine ($P=0.023$) synthesis pathways. All synthesis pathways enriched in breastfed microbiomes except

[1]Division of Newborn Medicine, Department of Pediatrics, Washington University in St. Louis School of Medicine, St. Louis, MO, USA. [2]The Edison Family Center for Genome Sciences and Systems Biology, Washington University in St. Louis School of Medicine, St. Louis, MO, USA. [3]Division of Gastroenterology, Hepatology, and Nutrition, Department of Pediatrics, Washington University in St. Louis School of Medicine, St. Louis, MO, USA. [4]Department of Molecular Microbiology, Washington University in St. Louis School of Medicine, St. Louis, MO, USA. [5]Department of Pathology and Immunology, Washington University in St. Louis School of Medicine, St. Louis, MO, USA. [6]Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, MO, USA. *e-mail: dudenhoeffer@wustl.edu; dantas@wustl.edu

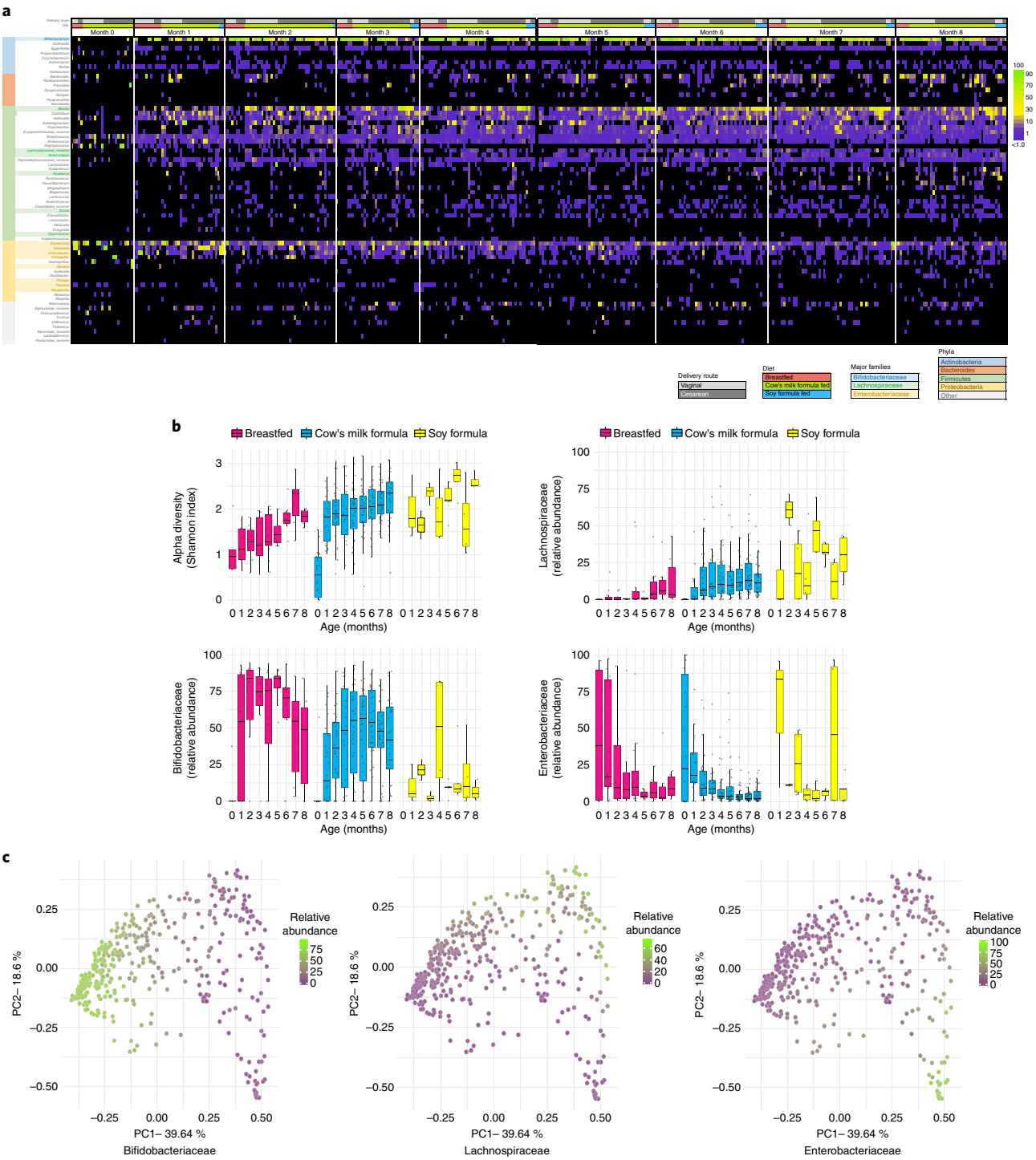

Fig. 1 | Taxonomic composition of infant fecal microbiota. a, Relative abundance of genera, grouped by month, diet, and delivery route. Samples are grouped horizontally by month of life, diet (breastfeeding, cow's milk formula, soy formula), and delivery route. All genera with ≥2% relative abundance in any sample are included, sorted vertically by phylum and relative contribution to the aggregate community of all subjects. **b**, Diversity and major taxa by age and diet. Boxplots (boxes representing IQRs with median shown in black) portray alpha diversity (Shannon index) and relative abundance of Bifidobacteriaceae, Lachnospiraceae, and Enterobacteriaceae over time, separated by diet type: majority breastfeeding ($n=75$ samples), cow's milk formula feeding ($n=295$), and soy formula feeding ($N=32$). Diversity increases with age ($P<0.001$) and soy ($P=0.036$). Bifidobacteriaceae correlated positively with breastfeeding ($P=0.003$) and negatively with soy ($P<0.001$). Lachnospiraceae increased in association with time ($P<0.001$) and soy ($P<0.001$) and decreased with breastfeeding ($P=0.014$). Enterobacteriaceae decreased with time ($P<0.001$) and GOS in cow's milk formula ($P=0.003$). All P values are from multivariate longitudinal maximum-likelihood GLMMs, Tukey-corrected for multiple comparisons (Supplementary Table 3). **c**, Principal Coordinate Analysis (PCoA) plot of taxonomic families, colored by major taxa. PCoA plots of taxonomic families based on the Bray–Curtis dissimilarity index for all samples ($n=402$) are shaded from low (purple) to high (green) relative abundance of Bifidobacteriaceae, Lachnospiraceae, and Enterobacteriaceae, highlighting three distinct clusters. Sequential multivariate analysis of variance (adonis in R, two-tailed) yielded R^2 values of 0.37 for Bifidobacteriaceae ($P=0.001$), 0.13 for Lachnospiraceae ($P=0.001$), and 0.11 for Enterobacteriaceae ($P=0.001$); residual R^2 from a multivariate model including only these three taxa was 0.38. PC1 (principal coordinate axis 1) explains 39.6% of the variance; PC2 (principal coordinate axis 2) explains 18.6% of the variance.

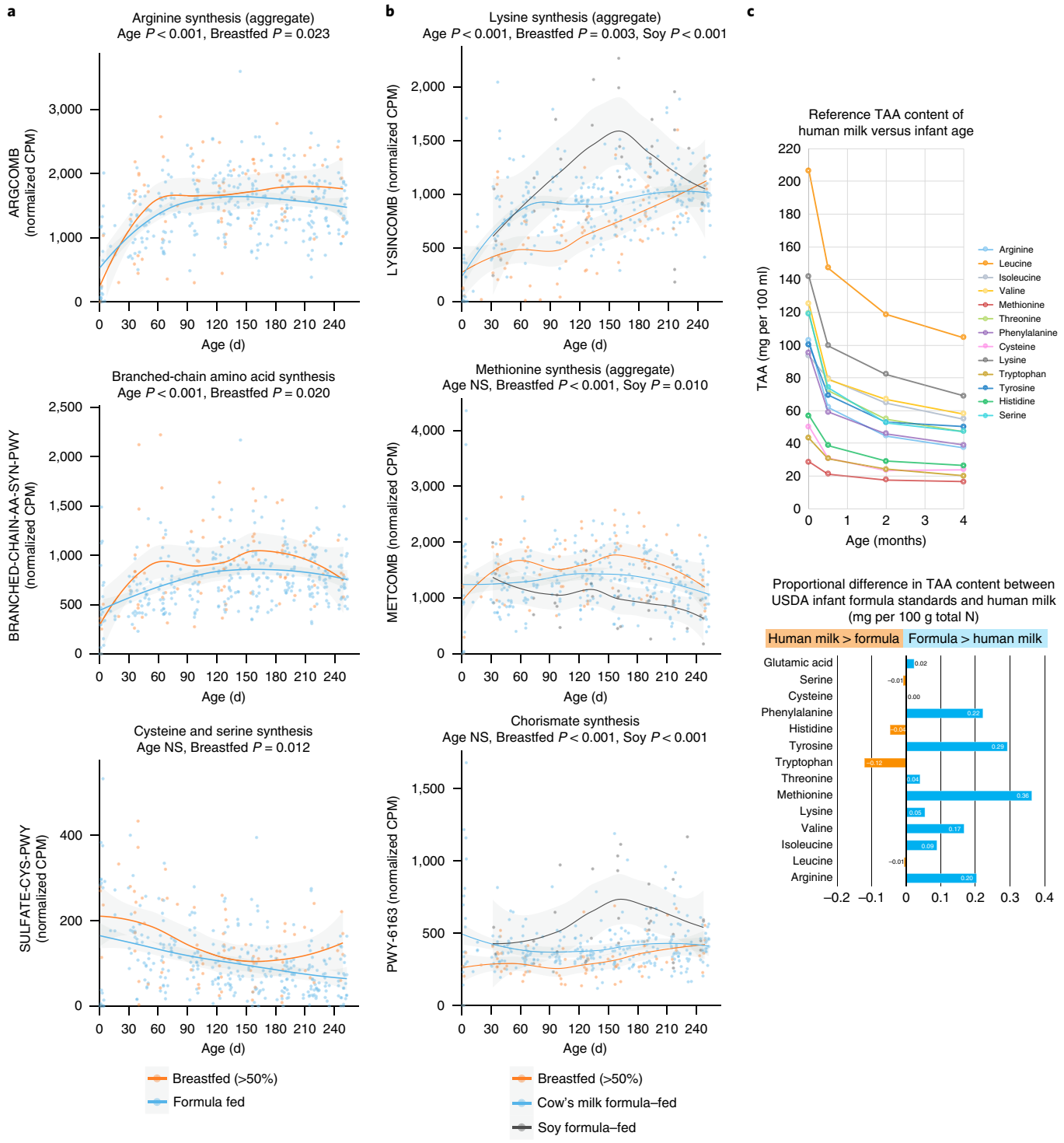

Fig. 2 | Dynamic development of amino acid synthesis pathways. a, Selected amino acid synthesis pathways, by age and diet (breastfeeding versus formula). **b**, Selected amino acid synthesis pathways, by age and diet (breastfed, cow's milk formula fed, soy formula fed). Scatterplots of normalized abundance (CPM) of selected amino acid synthesis pathways versus infant age are shaded according to diet type. In plot **a**, mostly breastfeeding ($n = 75$ samples) is compared with mostly formula feeding ($n = 327$); in plot **b**, current majority breastfeeding ($n = 75$), cow's milk formula feeding ($n = 295$), and soy formula feeding ($n = 32$) are compared. Regression lines with 95% confidence interval shading (light gray) are drawn using the loess method in R. All *P* values are two-tailed, from multivariate longitudinal maximum-likelihood GLMMs Tukey-adjusted for multiple comparisons (Supplementary Table 3). **c**, Known reference ranges for human milk total amino acid (TAA) content. Published TAA reference ranges in term breastmilk are plotted in comparison with US Department of Agriculture standards for infant formula to contextualize panels **a** and **b**. The line graph plots normative human milk TAA content over time for colostrum (origin), transitional milk (0.5 months), and at 2 months and 4 months postdelivery (see Table 4 in ref. [14]). The bar plot shows predicted differences in total amino acid content between US Department of Agriculture 2009 standards for infant formula and mature human milk (see Table 8 in ref. [14]), divided by normative values for human milk; a +0.36 value for methionine indicates that formula has 36% more methionine (mg per total N) than human milk. N, nitrogen; NS, not significant.

cysteine/serine correspond to amino acids less concentrated in breastmilk than in standard infant formula[14,18]. Breastfed microbial arginine and BCAA synthesis pathways increased sharply after birth and plateaued at ~60 d, coinciding precisely with normative declining amino acid content as breastmilk transitions from colostrum to mature milk[14] (Fig. 2). Breastmilk is low in methionine and cysteine in all lactation stages[14]; breastfed microbiomes had more methionine and cysteine synthesis pathways at all time points. Histidine and tryptophan are more abundant in breastmilk than in formula[14], and breastfed microbiomes had significantly fewer histidine-purine-pyrimidine ($P=0.046$) and tryptophan-precursor chorismate ($P<0.001$)[19] synthesis pathways. Glutamate and glutamic acid are abundant in breastmilk[14], and glutamate synthesis pathways (PWY-5505), though too sparse to model, were almost exclusive to formula-fed microbiomes ($n=114$, 90% of total). Lysine was an exception to milk–microbe complementarity. Infant formulas have more lysine than breastmilk, yet formula-fed microbiomes had more lysine synthesis pathways ($P=0.003$). Lysine synthesis pathways mapped to *Bacteroides* and to Firmicutes genera (Supplementary Table 4); formula-associated enrichment probably reflects accelerated microbiome maturation following breastfeeding cessation[4,13].

Milk–microbiome complementarity may be physiologically relevant to neonatal and infant protein balance[12,14]. Although breastmilk's amino acid content declines postpartum[14] and formula composition is static, normative serum arginine, cysteine, and methionine concentrations decline almost identically in breastfed and formula-fed infants[20], suggesting a 'gap' that might be filled by microbially produced amino acids. Breastfeeding-enriched metabolic pathways could mechanistically explain some of its known benefits[11,12]. Arginine and cysteine might prevent serious infections[10,11] and biotin, for which synthesis pathways are enriched in breastfed infants ($P=0.006$), inhibits pathogenic *Escherichia coli* adherence[9]. Many breastfeeding-associated amino acid synthesis pathways mapped to *Bifidobacterium* spp., an exceptionally successful breastfed gut colonizer. Breastfeeding-correlated enrichment of *Bifidobacterium*-identified amino acid synthesis pathways in a pattern contemporaneous and complementary to human milk maturation might reflect ancestral coevolution with commensal microbiota.

GOS and FOS are added to formulas to promote breastfed-like microbial communities[5]. Although lifetime GOS exposure correlated with Bifidobacteriaceae enrichment, prebiotics did not uniformly predict breastfed-like functional pathways, highlighting current technologic limitations of formula design and manufacturing (Supplementary Fig. 2). Concurrent GOS and FOS exposure ($n=26$) predicted increased microbial BCAA ($P<0.001$) and threonine ($P=0.038$) synthesis pathways, mimicking breastfeeding. Lifetime GOS exposure ($n=204$) predicted decreased tyrosine ($P=0.004$), cysteine/serine ($P=0.003$), and arginine-polyamine ($P=0.040$) synthesis pathways, opposing breastfeeding. In all models, prebiotic coefficients approximately equaled or exceeded those for breastfeeding. Pathways depleted with GOS exposure primarily belonged to Enterobacteriaceae (Supplementary Table 4); discordant GOS and breastfeeding correlates might reflect GOS-related decrease in Enterobacteriaceae[5].

Six infants from four families were soy-exposed; sample size ranged from 31–37, depending on soy formula type (+/− FOS) and exposure type (current or lifetime; see Supplementary Table 6). Soy feeding predicted greater alpha diversity (Shannon index, $n=31$, $P=0.036$), low Bifidobacteriaceae ($n=31$, $P<0.001$), and high Lachnospiraceae ($n=32$, $P<0.001$) content; in both taxonomic models, the coefficient for soy was greater than for breastfeeding (Figs. 1b and 3 and Supplementary Fig. 3). Two soy-exposed twin pairs were soy-discordant, permitting comparison with a related control. Twins are expected to have similar microbiomes[3,13], yet soy-discordant twin microbiomes were dissimilar, whereas unrelated soy-exposed microbiomes had strong resemblance. Soy encourages Lachnospiraceae proliferation[16], but has no clear effect on Bifidobacteriaceae[16,21]. Soy formula could disfavor Bifidobacteriaceae via cidal effects of soy isoflavone derivatives[22], by containing prebiotics (FOS) with weak bifidogenic properties[23] or by favoring competing taxa[16]. Pre-soy samples were few ($n=6$), but pre–post soy comparisons did not suggest soy-mediated bifidobacterial suppression: soy-fed microbiomes were low in Bifidobacteriaceae before soy exposure.

Low pre-soy bifidobacterial content suggests that low-Bifidobacteriaceae microbiomes might drive soy formula selection, especially as soy feeding is usually elective[24], rather than required for galactosemia, congenital lactase deficiency, or cow's milk protein allergy[25]. Soy-correlated depletion of Bifidobacteriaceae-identified methionine ($n=31$, $P=0.010$) and *S*-adenosyl methionine ($n=37$, $P=0.019$) synthesis pathways suggests a mechanism for this effect (Supplementary Table 4). Low-Bifidobacteriaceae microbiomes are associated with infant colic, which often prompts formula changes[6]. Methionine is a plausible mediator of enteric symptoms, as it affects both gut epithelia[26] and motility[27]. Indeed, methionine synthesis pathways positively correlated with reported diarrhea in our cohort ($n=16$, $P<0.001$), possibly representing a clinical correlate of methionine's reported prokinetic properties[27]. Soy protein is methionine deficient relative to mammalian casein and whey proteins; soy formula is methionine supplemented with a free methionine content ~125 times that of breastmilk[18,24]. These gut-specific effects of methionine provide a biologically plausible mechanism for symptoms associated with low bifidobacterial and methionine synthesis pathway content to improve after initiation of high-methionine formula.

Several soy-associated pathways—chorismate synthesis ($n=31$, $P<0.001$), lactose/galactose degradation ($n=37$, $P<0.001$), and starch degradation ($n=31$, $P<0.001$)—suggested SCFA-producing Lachnospiraceae proliferation. Soy-correlated enrichment of lysine synthesis ($n=32$, $P<0.001$), riboflavin synthesis ($n=32$, $P<0.001$), and glycerol-to-butanol fermentation ($n=32$, $P<0.001$) pathways suggested adaptation to SCFAs. Lactose/galactose and starch degradation pathways frequently mapped to Lachnospiraceae (Supplementary Table 4), and a greater proportion of chorismate synthesis pathways were *Blautia* identified post-soy exposure (Fig. 3c). Lysine provides an acetate and butyrate synthesis substrate[28], butyrate stress in *Clostridium* spp. induces upregulation of riboflavin and downregulation of methionine synthesis[29], acetate stress promotes glycerol-to-butanol fermentation[29], and many microbes coregulate riboflavin synthesis genes with metabolic stress response modules[30]. Some soy-associated changes are potentially dysbiotic: decreased Bifidobacteriaceae and elevated glycerol-to-1-butanol fermentation combined with high Lachnospiraceae content have been associated with inflammation, allergies, and hepatic steatosis[2,7]. These dysbiotic features raise concerns about the long-term safety and efficacy of elective soy formula feeding.

Maternal GWG has yet-to-be determined effects on infant gut microbiome development[15,31,32]. Here, GWG ($n=402$) predicted persistent enrichment of infants' microbial glucose ($P<0.001$) and glycogen ($P=0.005$) degradation pathways, and phenylalanine ($P=0.011$), cysteine/serine ($P<0.001$), folate ($P=0.015$), thiamine ($P<0.001$), biotin ($P<0.001$), and pyridoxine ($P=0.009$) synthesis pathways, after controlling for gestational age, maternal diabetes, and prepregnancy body mass index (BMI) (Supplementary Table 1). Starch degradation pathways negatively correlated with GWG ($P=0.032$) The GWG distribution in our cohort roughly corresponded with Institute of Medicine guidelines for twin pregnancies (see Methods): women with inadequate and excessive GWG fell into the first and fourth quartiles, respectively. GWG-correlated pathways plotted by age and quartile suggest that GWG-mediated effects persistent at 8 months are most apparent in infants born to mothers who gained the least weight, and low GWG appears more impactful with increasing gestational age (Fig. 4).

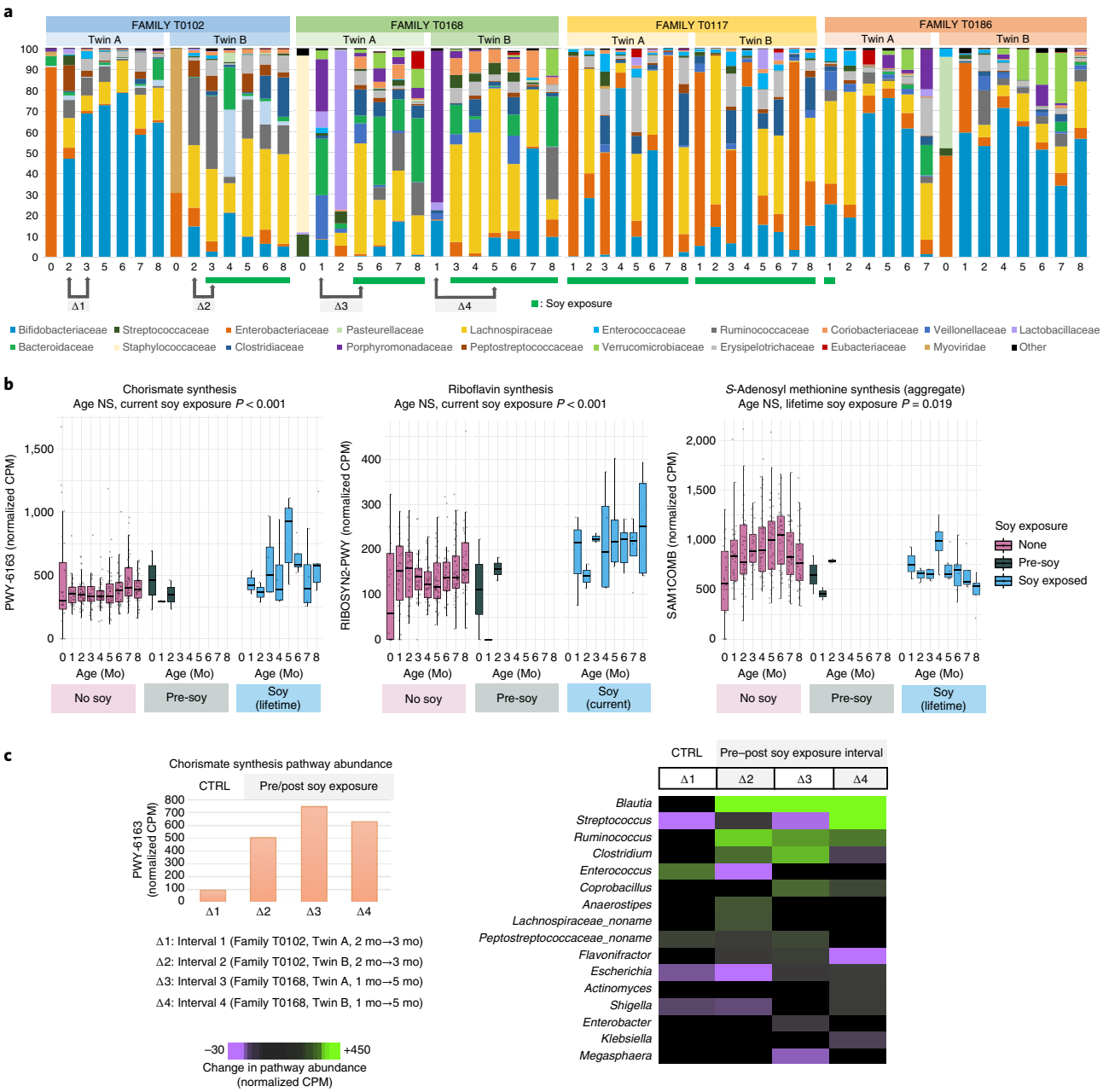

Fig. 3 | Taxonomic and functional changes associated with soy formula. a, Taxonomic structure of soy-exposed infants' gut microbiota. Stacked bar plots show relative abundances of taxonomic families over time from four twin pairs with at least one soy-exposed sibling. All families with ≥3% relative abundance are included (rare taxa aggregated as 'Other'). Pre- and post-soy samples were available for three infants. Bifidobacteriaceae abundance was low pre-soy and throughout the study in all soy-exposed infants, except infant T0186_A, whose Bifidobacteriaceae recovered following soy formula cessation. Soy-discordant twin microbiomes were visibly dissimilar. **b**, Metabolic pathways by age and soy-exposure status. These boxplots (boxes representing IQRs with median shown in black) show normalized abundance (normalized CPM) versus age of three functional pathways: chorismate synthesis (PWY-6163), riboflavin synthesis (RIBOSYN2-PWY), and the aggregate methionine synthesis variable METCOMB. Longitudinal plots are separated into soy-naive ($n = 364$ current, $n = 359$ lifetime), pre-soy ($N = 6$), and soy-exposed samples ($n = 32$ current, $n = 37$ lifetime). Chorismate and riboflavin synthesis pathways increased post-soy, while methionine synthesis pathways decreased. All P values are two-tailed, from multivariate longitudinal maximum-likelihood GLMMs Tukey-adjusted for multiple comparisons (Supplementary Table 3). **c**, Changes in chorismate synthesis pathway homology following soy exposure. The column graph shows numerical differences in total chorismate synthesis pathway (PWY-6163) abundance pre- and post-soy in 3 soy-exposed infants and 1 control ($n = 4$ each time point). Total PWY-6163 abundance qualitatively increased more in soy-exposed infants than in the control. The heatmap plots numerical pre–post soy difference (normalized CPM) in chorismate synthesis pathway-identified genera, which qualitatively shifted towards *Blautia* pathway homology with soy exposure. CTRL, control; mo, months.

Although true malnutrition is unlikely in our cohort, maternal undernutrition increases risk of oxidative injury, glucose dysregulation, adiposity, and cardiovascular disease in offspring[1]. Several GWG-enriched vitamin synthesis pathways (pyridoxine, thiamine, folate) are critical to early infant neurodevelopment[12], and thiamine synthesis pathways are a proposed distinguishing core 'enterotype'

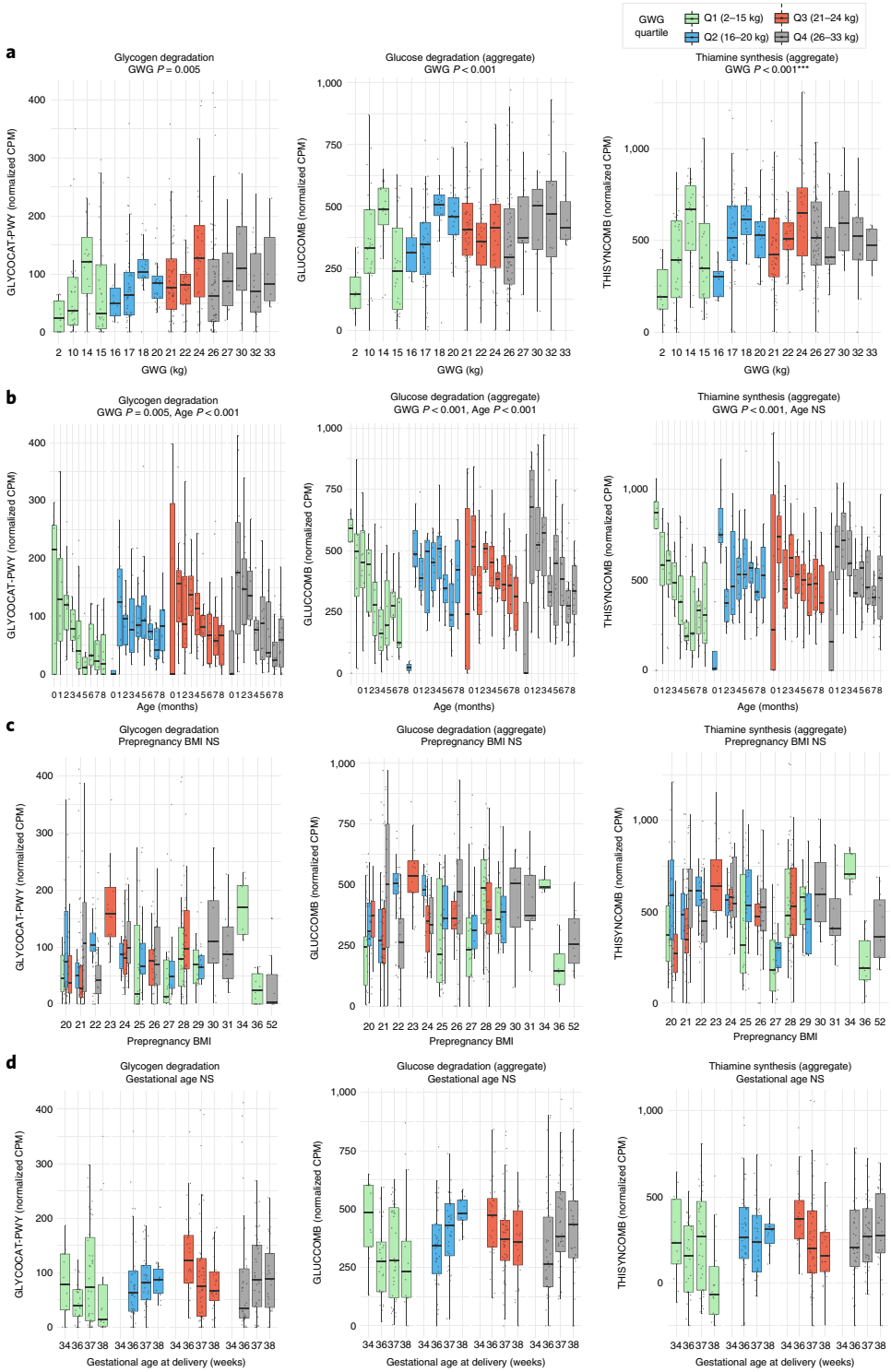

Fig. 4 | Altered development of vitamin synthesis and carbohydrate utilization pathways in association with GWG. Boxplot boxes in all panels represent IQRs with the median line shown in black; total $n = 402$ for all plots. All P values in all panels are two-tailed, from multivariate longitudinal maximum-likelihood GLMMs Tukey-adjusted for multiple comparisons (Supplementary Table 3). **a**, Selected metabolic pathways by GWG. These panels plot normalized abundance (CPM) of the glycogen degradation pathway GLYCOCAT-PWY, the aggregate glucose degradation variable GLUCCOMB, and the aggregate pyridoxine synthesis variable THISYNCOMB versus GWG. The plots are colored according to GWG quartile in this population. All pathways plotted have a significant positive association with GWG. Sample size by quartile: Q1 ($n = 93$ samples), Q2 ($n = 93$), Q3 ($n = 108$), Q4 ($n = 108$). **b**, Selected metabolic pathways by age and GWG. These panels plot normalized abundance (CPM) of the same pathways by month of life, stratified by maternal GWG quartile in this population. The differences between the lowest and highest GWG become more apparent over time. **c**, Selected metabolic pathways by maternal prepregnancy BMI. Normalized abundance (CPM) of the same pathways plotted versus maternal prepregnancy BMI, colored by GWG quartile. **d**, Selected metabolic pathways by gestational age and GWG. These panels plot normalized abundance (CPM) of the same three pathways versus gestational age at delivery, stratified by GWG quartile. In the lowest GWG quartile, the abundance versus gestational age curve slopes in the opposite direction of the abundance versus GWG curve, suggesting that lower GWG is more important with increasing gestational age.

feature[8]. GWG negatively correlates with folate synthesis pathway abundance in the placental microbiome[33]. We observed the inverse relationship in our population (GWG-associated folate pathway enrichment), perhaps representing compensation for the fetal microenvironment. Folic acid is a key epigenetic mediator, and might effectuate enduring host–microbe interactions and mediate fetal origins of disease[1].

GWG-associated microbial metabolic pathway changes persisting eight months postnatally extend current knowledge that GWG influences microbiome development in the first months of human life[15,31,32] and for up to 1 year in nonhuman primates[34]. As maternal dietary records and weight gain by trimester were not collected, we can neither identify trimester-specific modulations nor attribute GWG-associated effects to specific dietary variables (for example, fat content). Enduring GWG-associated changes independent of delivery route or breastfeeding might reflect altered in-utero meconium colonization[35], microbe transfer from caregivers[36], and other genetic or environmental factors (for example, family feeding practices) influencing both GWG and infant microbiome acquisition.

Maternal intrapartum antibiotics predicted postnatal development of taxa and functional pathways, eclipsing the effects of delivery route and postnatal antibiotics (Supplementary Fig. 4). Maternal intrapartum ampicillin-sulbactam exposure ($n=46$) predicted depleted histidine/purine/pyrimidine synthesis ($P=0.012$) and homolactic fermentation ($P<0.001$) pathways in offspring microbiomes. Postnatal amoxicillin exposure ($n=38$), analogous to ampicillin without sulbactam, predicted increased histidine/purine/pyrimidine synthesis pathways ($P=0.011$). Maternal intrapartum clindamycin exposure ($n=25$) positively correlated with Lachnospiraceae ($P=0.008$), Enterobacteriaceae ($P<0.001$), and cysteine/serine ($P<0.001$) and biotin ($P=0.002$) synthesis pathways. When used, clindamycin was given immediately (<30 minutes) before Cesarean delivery in our cohort, but the more frequently administered cefazolin ($n=164$) did not correlate with these pathways. Lack of persistent microbiome effects associated with Cesarean delivery when corrected for confounders is consistent with previous reports[15]. Infant multivitamin with iron exposure ($n=40$, Supplementary Fig. 5) predicted enriched arginine-polyamine ($P=0.018$), folate ($P<0.001$), and heme ($P=0.026$) biosynthesis and homolactic fermentation pathways ($P=0.028$).

Domestic drinking water sources had associated microbiome signatures (Supplementary Fig. 5); sample size depended on exposure type (Supplementary Table 6). Lactose/galactose degradation pathways positively correlated with filtered water exposure ($n=42$, $P=0.004$); enhanced bacterial counts with home water filters might explain this effect[37]. Tap water exposure predicted decreased Enterobacteriaceae ($n=251$, $P=0.016$), glycogen degradation ($n=230$, $P=0.006$), and homolactic fermentation ($n=230$, $P=0.007$) pathways. Bottled water exposure predicted increased homolactic ($n=122$, $P=0.002$) pathways, and boiled/distilled water correlated with increased pyridoxine synthesis pathways ($n=61$, $P=0.003$). Together with animal data[38], these patterns suggest an underappreciated influence of drinking water on microbiome acquisition.

Although this DNA-based study represents genetic potential rather than confirmed functions, our observations are consistent with transcriptomic studies showing enriched arginine biosynthesis transcripts in mother-fed relative to formula-fed piglets[39] and enhanced BCAA synthesis with sialylated oligosaccharide exposure in mice[40]. Further work is required to mechanistically establish a causal relationship between soy exposure and soy-fed microbiome signatures and to definitively show that soy protein per se drives these changes, probably via experimental validation in microbiome-humanized gnotobiotic mice[13].

In summary, our findings suggest host–microbe metabolic mutualism in infancy, whereby gut microbiome gene content expands to counterbalance components relatively lacking in human milk (Supplementary Fig. 6). We propose that this milk–microbiome synergy reflects physiologic coevolution with our earliest commensals, and could play a major teleological role in infant protein nutrition and child growth. The observed discordance between microbial functional correlates of formula components (for example, prebiotics) and breastmilk may warrant revised metrics for evaluating the safety and efficacy of infant formulas. Soy formulas corresponded with profoundly altered taxa and pathways, some of which have pathologic correlates[6,7]. Finally, the association between maternal GWG and altered infant microbiome carbohydrate utilization and vitamin synthesis pathways enduring 8 months postnatally extends current knowledge that maternal GWG influences early microbiome acquisition. These data can inform further ecologic and mechanistic interrogations of gut microbiome development.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability, and associated accession codes are available at https://doi.org/10.1038/s41591-018-0216-2.

Received: 28 February 2018; Accepted: 10 September 2018;
Published online: 29 October 2018

References

1. Indrio, F. et al. Epigenetic matters: the link between early nutrition, microbiome, and long-term health development. *Front. Pediatr.* **5**, 178 (2017).
2. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: an integrative view. *Cell* **148**, 1258–1270 (2012).
3. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
4. Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 852 (2015).
5. Oozeer, R. et al. Intestinal microbiology in early life: specific prebiotics can have similar functionalities as human-milk oligosaccharides. *Am. J. Clin. Nutr.* **98**, 561S–571S (2013).
6. de Weerth, C., Fuentes, S., Puylaert, P. & de Vos, W. M. Intestinal microbiota of infants with colic: development and specific signatures. *Pediatrics* **131**, e550–e558 (2013).
7. Del Chierico, F. et al. Gut microbiota profiling of pediatric nonalcoholic fatty liver disease and obese patients unveiled by an integrated meta-omics-based approach. *Hepatology* **65**, 451–464 (2017).
8. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
9. Yang, B., Feng, L., Wang, F. & Wang, L. Enterohemorrhagic *Escherichia coli* senses low biotin status in the large intestine for colonization and infection. *Nat. Commun.* **6**, 6592 (2015).
10. Badurdeen, S., Mulongo, M. & Berkley, J. A. Arginine depletion increases susceptibility to serious infections in preterm newborns. *Pediatr. Res.* **77**, 290–297 (2015).
11. Zhou, P., Li, Y., Ma, L. Y. & Lin, H. C. The role of immunonutrients in the prevention of necrotizing enterocolitis in preterm very low birth weight infants. *Nutrients* **7**, 7256–7270 (2015).
12. Schwarzenberg, S. J. & Georgieff, M. K. The AAP Committee on Nutrition. Advocacy for improving nutrition in the first 1000 days to support childhood development and adult health. *Pediatrics* **141**, e20173716 (2018).
13. Planer, J. D. et al. Development of the gut microbiota and mucosal IgA responses in twins and gnotobiotic mice. *Nature* **534**, 263–266 (2016).
14. Zhang, Z., Adelman, A. S., Rai, D., Boettcher, J. & Lönnerdal, B. Amino acid profiles in term and preterm human milk through lactation: a systematic review. *Nutrients* **5**, 4800–4821 (2013).
15. Chu, D. M. et al. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat. Med.* **23**, 314–326 (2017).
16. Butteiger, D. N. et al. Soy protein compared with milk protein in a Western diet increases gut microbial diversity and reduces serum lipids in golden Syrian hamsters. *J. Nutr.* **146**, 697–705 (2016).
17. Yassour, M. et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8**, 343ra381 (2016).

Acknowledgements

This work is supported in part by awards to G.D. through the Edward Mallinckrodt, Jr. Foundation (Scholar Award), and the National Institute of General Medical Sciences (http://www.nigms.nih.gov/) of the National Institutes of Health (NIH) under award number R01GM099538. A.M.B.-D. was supported by the National Institutes of Diabetes and Digestive and Kidney Diseases of the NIH under award number K08-DK102673. A.W.D. received support from the Institutional Program Unifying Population and Laboratory-Based Sciences Burroughs Wellcome Fund grant to Washington University. B.B.W. and P.I.T. received support for the cohort and sample collection from the Children's Discovery Institute of Washington University and St. Louis Children's Hospital, and P.I.T. is supported by P30DK052574 (Biobank Core). P.I.T., B.B.W., and G.D. are also supported in part by a grant from the Eunice Kennedy Shriver National Institute of Child Health & Human Development (https://www.nichd.nih.gov/) of the NIH under award number R01HD092414. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. We would like to thank E. Martin, B. Koebbe, and J. Hoisington-López from the Edison Family Center for Genome Sciences & Systems Biology at Washington University School of Medicine for technical support in high-throughput computing and sequencing. We would like to thank A. J. Gasparrini, B. Wang, and B. Berla for technical assistance in experimental and computational protocol optimization for whole-metagenome shotgun sequencing of fecal samples. We would like to thank I. M. Ndao, N. Shaikh, S. Patel, B. Wang, and S. X. Sun for archival and maintenance of frozen fecal sample inventory. We would like to thank F. S. Cole and members of the Dantas lab for general helpful discussions regarding the research presented in this manuscript, and K. Guilonard for helpful comments on the text.

Author contributions

A.M.B.-D., A.W.D., B.B.W., P.I.T., and G.D. conceived of experiments and design of work and analyses. B.B.W. and P.I.T. oversaw collection and stewardship of fecal samples and clinical metadata inventories. A.M.B.-D. performed wet-lab experiments with advice from G.D. A.M.B.-D. performed computational analyses with advice from A.W.D. and G.D. Article drafting was performed by A.M.B.-D. with critical revision performed by A.W.D., B.B.W., P.I.T., and G.D.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-018-0216-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.M.B. or G.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Study population. This study was approved by the Human Research Protection Office of Washington University School of Medicine in St. Louis, and it complied with all ethical regulations. Written informed consent was obtained from all adult participants and from the parents or legal guardians of all minor subjects. We used fecal samples that had been frozen at −80 °C since collection at monthly intervals from a birth cohort of healthy twins in St. Louis, in which the mothers had consented to monthly fecal sample collection from birth until 2 years of age[3,13,41–43]. We selected a time interval of 0–8 months of age to capture transitions from breastfeeding to formula and early introduction of solid food. To minimize potential confounding effects of early illness or antibiotic administration, we excluded any neonates who received antibiotics in hospital following delivery. Because of this predetermined exclusion criterion, we also excluded all infants with a maternal history of chorioamnionitis. A total of 402 samples from 60 infants in 31 families met our predefined coverage threshold of 5,000,000 reads (2,500,000 forward/reverse) before processing[44], for a median of 7 samples per infant (interquartile range (IQR), 6–8). Demographic data are provided in Supplementary Table 1. We excluded neonates treated with antibiotics in the first week of life to avoid potential bias from early illness or antibiotic exposure; there were accordingly no infants with a maternal history of chorioamnionitis. Infant age at stool collection ranged from the day of delivery to 253 d. All infants were exposed to solid food by the end of the study period. The median gestational age was 37 weeks (IQR, 36–38), 43% of infants were delivered vaginally, and 47% of twins were monozygotic, 50% dizygotic, and 3% of unknown zygosity. Four infants' mothers were diabetic (7%), 6 infants' mothers developed preeclampsia (10%), and 2 infants were born to a mother with both conditions.

DNA extraction and sequencing. We extracted fecal metagenomic DNA and a positive control (Zymobiomics microbial community standard D6300), and used a modified Nextera DNA Library Preparation Kit protocol to prepare DNA for Illumina-platform sequencing (NextSeq-High; ~400,000,000 maximum reads, 150 cycles per read). A positive control (Zymobiomics community Standard) and a negative control (nuclease-free water) were included in sequencing runs. Detailed experimental protocols follow.

DNA extraction. We extracted DNA using the MoBio DNEasy PowerSoil Extraction Kit (Qiagen, 12888-100) according to the manufacturer's instructions, with the following modification: in lieu of centrifugation, we used bead beating with a BioSpec Mini-BeadBeater for 4 min. Bead beating consisted of 2 min on the 'homogenize' setting, 2 min on ice, and then 2 min on the 'homogenize' setting. A Zymobiomics microbial community standard (Zymobiomics, D6300) 0.75 ml was also extracted along with fecal DNA samples. DNA was eluted in 100 µl nuclease-free water and quantitated using a Qbit fluorometer and a Qbit dsDNA HS Assay Kit (Invitrogen, Q32854) according to the manufacturer's instructions.

Nextera library preparation. Fecal DNA samples were diluted to a concentration of 0.5 ng µl^{-1}, and 1 µl of each sample (including a nuclease-free water negative control and the Zymo community standard positive control) was added to a 96-well plate. Sequencing libraries were prepared using the Nextera DNA Library Preparation Kit (Illumina, FC-121-1011) protocol according to the manufacturer's instructions, with the following modifications:

A. Tagmentation

1. Tagmentation master mix preparation: Component: 1 reaction (µl), 100 (µl)Tagment DNA buffer: 1.25, 125.0TDE1 enzyme: 0.125, 12.5Nuclease-free water: 0.125, 12.5
2. Tagmentation master mix (1.5 µl) added to 1 µl genomic DNA in each well of the 96-well plate, vortexed, and centrifuged.
3. Plate covered with microseal B and incubated in a Thermocycler at 55 °C for 15 min.

B. Adapter addition

1. KAPA HiFi PCR master mix (KAPA HiFi HotStart 2x ready mix, KK2602/KM2605) used for addition of oligonucleotide index adapters. KAPA PCR MasterMix (11.2 µl) and 8.8 µl of adapters (1 µM) added to each well, vortexed, and centrifuged.
2. PCR done with following Thermocycler protocol:

i. 72 °C, 3 min
ii. 98 °C, 5 min
iii. 98 °C, 10 s
iv. 63 °C, 30 s
v. 72 °C, 30 s
vi. go to iii 13×72 °C, 5 min
vii. 4 °C, forever

C. PCR cleanup

1. Added 22.5 µl AmpPure XP beads to PCR reaction (Agencourt A63881)
2. Incubated 5 min at room temperature
3. Separated beads for 2 min on magnetic stand
4. Removed supernatant
5. Washed beads twice with 200 µl 80% ethanol
6. Air-dried for 15 min
7. Added 30 µl resuspension buffer (10 mM Tris-Cl, 1 mM EDTA, 0.05% Tween-20 (pH 8.0)), pipetted mix
8. Incubated at room temperature for 5 min
9. Separated beads on magnetic plate for 2 min
10. Transferred 27 µl supernatant to new plate
11. Quantitated DNA with Qubit HS dsDNA Assay kit (Invitrogen, Q32854)

Illumina sequencing.

A. Library pooling: after quantitation, sequencing libraries were pooled in triplicate to minimize the effects of pipetting error. Schema for pooling included the following rules:

1. Target of 5 ng DNA per sample, per pool
2. If calculated volume for 5 ng <1 µl, samples were diluted (2×, 5×, 10×, or 20×) so the volume was >1 µl
3. Triplicate pools quantitated with Qubit HS dsDNA Assay kit (Invitrogen, Q32854)
4. Each pool was added to a single pool to make an equimolar solution, and diluted to a concentration of 2 ng µl^{-1}
5. Pool submitted for Illumina platform sequencing (MiSeq flowcell) as a ~500,000-read spike-in sample; reads analyzed to determine evenness of sample distribution. If needed, a fourth corrective pool was pipetted and added to the solution to ensure adequate read coverage (>2,500,000 for both forward and reverse) for all samples.

B. Sequencing

Pooled samples diluted to 2 ng µl^{-1} with nuclease-free water (~10 µM based on expected fragment size) were submitted for Illumina platform sequencing (NextSeq-High; ~400,000,000 maximum reads) with 150 cycles per read. Sequence data were returned as a .fastq file with reads demultiplexed according to oligonucleotide adapter indexes.

Clinical data collection. Clinical data were collected from medical records at the time of delivery, monthly parental surveys at the time of stool sample collection, and outpatient pediatric records, and securely stored on a RedCap database. Parental questionnaires, infant medical records, and formula manufacturers' labels provided a detailed clinical and dietary dataset (including symptoms, medications, and introduction of new foods) associated with each sample. Parental dietary reports included infant formula brands, solid foods, and water sources, as well as fields for reporting daily or weekly frequency of each dietary exposure from the Centers for Disease Control and Prevention Infant Feeding Practices Study II[45]. As exclusive breastfeeding was rare in this twin cohort, infants were classified as breastfed if their parents reported >50% of their feeds as breastmilk in the survey associated with a given stool sample. All breastfed infants received maternal milk; there were no reported exposures to banked or donated human milk. Medication exposures reported on parental surveys were confirmed with medical records from the child's primary care physician. Information from the manufacturer's label for each infant formula was used to generate a suite of variables representing exposure to specific formula ingredients (for example, lactose, sucrose, soy protein, GOS, FOS); full details are below.

Clinical data analytic specifications. Clinical data deidentified of any protected health information was collected from medical records at the time of delivery, monthly parental surveys at the time of stool sample collection, and outpatient pediatric medical records, and was securely stored on the RedCap database. Parental dietary reports included:

A. Binary fields for exposure to human milk, various infant formula brands, foods, medications, experience of symptoms, etc.
B. Fields for frequency of exposure to a food type, expressed either as the number of times an infant received a food per day, or per week
C. Free text options

To transform dietary information into data that were usable in statistical models, the following steps were followed:

A. All frequency information listed as exposures per day was converted into exposures per week for convenience.
B. Percentages of feeds composed of formula were calculated from parental reports of number of formula feeds per week and number of breastfeeds per week. A binary variable for 'Mostly Breastfeeding' was also generated if the percentage of breastfeeds was >50%.
C. Carbohydrate, protein, and prebiotic (GOS, FOS) ingredients were determined for each infant formula according to the manufacturer's label (Supplementary Table 7). Binary variables for exposure to each ingredient at

each time point were generated according to the brand(s) of formula(s) the parents had reported, and the manufacturers' labels. If parents reported using any brand of formula on the survey associated with a stool sample, binary variables for ingredients in that formula were coded as '1', even if the parents otherwise recorded that the infant was exclusively breastfed (that is, if parents reported 100% breastfeeds, but filled in Enfamil Lipil as a formula they selected, the infant was coded as mostly breastfed, but exposed to the ingredients in Enfamil Lipil). If there was ambiguity in the specific brand of formula, then missing values were recorded for binary variables (for example, if it was unclear whether an infant was given Enfamil Lipil or Enfamil Premium, Lactose and Cow's milk Formula, which are present in both, would be coded as '1' but GOS and Polydextrose, which were only present in Premium, were recorded as missing). Twin siblings were assumed to have concordant feeding practices unless the parents specified otherwise.

D. Prebiotic variables were assigned according to exposure to neither, one, or both prebiotics (GOS and FOS):

   a. GOS: exposure to GOS, regardless of concurrent FOS exposure
   b. FOS: exposure to FOS, regardless of concurrent GOS exposure
   c. Only GOS: exposure to GOS with no concurrent FOS exposure
   d. Only FOS: exposure to FOS with no concurrent GOS exposure. Only found in soy formulas.
   e. GOS/FOS: concurrent exposure to GOS and FOS.

E. Solid food binary variables were aggregated as follows:

   a. Fruit or vegetable exposure → Fruit/Veg variable; positive if either Fruit or Vegetables were positive
   b. Meat, fish, or egg exposure → MeatFishEggs variable; positive if any of the components were positive
   c. Juice or sweetened drink exposure → JuiceSweetDrink variable; positive if either component was positive
   d. Cereal or starch exposure → CerealStarch variable; positive if either component was positive
   e. An AnyDairy variable was created for exposure to any dairy product, including cow's milk formula

F. For binary variables reflecting current exposure to a food, medication, or ingredient, a second binary variable was generated reflecting lifetime exposure to that food, medication, or ingredient (exposure at any point previously).

G. Sample size for all binary variables is listed in Supplementary Table 6.

H. Continuous variables (day of life, maternal weight gain, gestational age, weight) were $\log_{10}$-transformed before statistical analysis. Sample size for all continuous variables is 402.

Sequence data processing. A predetermined minimum sequencing depth of 5,000,000 raw reads (2,500,000 forward/reverse) per sample was required for inclusion in the study. Reads were trimmed using Trimmomatic (trimmomatic/0.33, minimum length = 60), and human DNA contamination was removed using Deconseq (Deconseq/0.4.3-chr38). We used MetaPhLan 2 (metaphlan2/2.2.0)[46] to extract taxonomic data, and HUMAnN2 (humann2/0.9.4)[47] to identify microbial functions. All taxonomic data are reported as relative abundance; all functional data were normalized to counts per million (CPM) using the humann2_renorm_table function. Full details are below.

Quality control. Only samples with >2,500,000 raw reads in each direction (>5,000,000 total raw reads) were included in this study. Supplementary Fig. 7 shows reads by subject age (months) and the distribution of samples included in the study by age in months There was no systematic bias in raw reads by age. Neonates and infants <3 months had fewer successful samples than infants 3–8 months of age, with neonates having the lowest number of samples that met our quality threshold. The median number of raw reads per sample was 11,300,000 (IQR, 6,300,000); the median number of reads following trimming and filtering human DNA was 9,200,000 (IQR, 5,600,000).

We trimmed reads using Trimmomatic[48] (trimmomatic/0.33), with the following specifications:

- PE –phred33
- SLIDINGWINDOW:6:10
- LEADING:13
- TRAILING:13
- MINLEN:60

We eliminated human sequences using Deconseq/0.4.3-chr38[49]. All analyses were performed on trimmed and decontaminated samples. Decontaminated sequence data were publicly deposited to protect the privacy of human subjects (Bioproject identity PRJNA473126, accession codes SAMN09259835–SAMN09260236).

Taxonomic data extraction. We used MetaPhlAn2[46] (metaphlan2/2.2.0) to extract taxonomic data from quality-filtered reads, with the following specifications:

- mpa_pkl ${mpa_dir}/db_v20/mpa_v20_m200.pkl
- bowtie2db ${mpa_dir}/db_v20/mpa_v20_m200

Control samples (both a negative control and a positive control from a Zymo community standard) were included in all sequencing runs; the community standard failed in the fifth run. There were no taxa identified from the negative control samples. Although there was some bias in the community standard (Gram-negative organisms overrepresented, Gram-positive underrepresented), probably reflective of bias in DNA extraction, the results were highly reproducible, which is reassuring for analysis of longitudinal trends. There were small proportions (relative abundance <0.1%) of taxa identified in the community standard sample that were not part of the theoretical community composition: *Nauvomozyma* unclassified, *Pantoea* unclassified, and *Eremothecium* unclassified (Supplementary Table 8). *Nauvomozyma* and *Eremothecium* were not identified in any fecal samples, and *Pantoea* unclassified was only found in a relatively small number of fecal samples ($n = 73$ out of 402). There were no taxa identified in the negative control sample. Community standard and negative control results did not suggest any systemic contamination.

Functional data extraction. We used Humann2[47] (humann2/0.9.4) to identify genes and functional pathways from short-read data, with the following specifications:

- input-format fastq
- search-mode uniref50
- bypass-translated-search
- bypass-prescreen
- gap-fill off

We used the function humann2_renorm_table to convert gene and pathway output into normalized CPM.

All models are performed on community-wide counts of MetaCyc-identified functional pathways[47,50]. Individual pathways contributing to aggregate families are detailed in Supplementary Table 2. The proportion of functional pathways identified as homologous to specific genera are summarized in Supplementary Table 4. To model the abundance of pathways related to synthesis of a specific metabolite (for example, clinical predictors of arginine synthetic pathway abundance instead of just the abundance of arginine synthesis I or arginine synthesis IV), pathways that were related to a specific metabolite were aggregated by summing the normalized community-wide abundance.

Statistical analysis. Statistical analysis and generation of figures was performed in R using the vegan, ape, ggplot2, lme4, lmerTest, MuMin, and multcomp packages. Alpha-diversity is reported as the Shannon index, determined from species-level abundance using the vegan diversity() function. PCoA plots were generated from a Bray–Curtis dissimilarity matrix of family-level taxa generated using the vegan vegdist() and ape pcoa() functions. Sequential multivariate analysis of variance was performed using the vegan adonis() function. All GLMMs in this study are maximum-likelihood generalized linear mixed models generated using the lme4 lmer() function, and because the close resemblance between twins' microbial communities represents an important potential confounding factor[3,13,42], we controlled for twin status by including both family and subject (time | subject) as mandatory random effects in all models.

Time, in log(days) was a mandatory fixed effect in all longitudinal GLMMs; all other fixed effects were back-fitted using a stepwise approach, according to the following schema. As the effects of some clinical variables (for example, specific formula ingredients) on the developing gut microbiome are completely unknown, we began the model-fitting process agnostic to which clinical variables would significantly correlate with microbiome features, and screened all variables for inclusion. To broadly screen for covariation between clinical variables and microbiome features, for every taxonomic or pathway variable, we created two arrays of metadata corresponding to the values above and below the median (relative abundance or normalized CPM). We then applied a two-tailed test to compare these two arrays (*t*-test for continuous variables and Fisher's exact test for binary variables) and included all metadata variables with a screening P value <0.05 in a first-approximation GLMM. Because of the potential effects of Cesarean delivery and breastfeeding on the developing microbiome, they were always included in the first-approximation GLMM, even if they did not pass the screening test.

Maternal weight variables represented a special case, with multiple potential confounding variables[15,31,32,35,51–53]. GWG would ideally be classified as normal, inadequate, or excessive according to maternal prepregnancy BMI and estimated gestational age of delivery, according to Institute of Medicine guidelines[54]. However, such calculations are established only for singleton pregnancies, with provisional guidelines available for total weight gain in twin gestation. Thus, in our twin population, we attempted to control for confounding variables such as prepregnancy BMI, gestational age, maternal diabetes, and preeclampsia, by modifying our model selection pathway so that prepregnancy BMI and gestational age at delivery were always included in our first-approximation GLMM, even if they did not pass the initial screening test. The GWG distribution in our cohort roughly corresponded with provisional Institute of Medicine guidelines for GWG in twin pregnancies (16.8–24.5 kg for normal prepregnancy BMI, 14.1–22.7 kg for overweight prepregnancy BMI, 11.3–19.1 kg for obese prepregnancy BMI)[54].

All mothers with inadequate GWG were in the first quartile of our population (2–15 kg), while the fourth quartile from our population (26–33 kg) represented excessive weight gain irrespective of prepregnancy BMI. Additional information can be found in Supplementary Table 1.

First-approximation GLMM was then back-fitted with the lmerTest step() function, and the MuMin rsquaredGLMM() function as a preferred post hoc test for goodness of fit. All *P* values are two-tailed, and are adjusted for multiple comparisons using the multcomp glht() function (tension = Tukey)[55]. Parameters for all GLMMs are in Supplementary Table 1, and statistically significant coefficients are summarized in Supplementary Table 5. Full details are below.

Statistical modeling. All maximum-likelihood longitudinal GLMMs were constructed using the lme4, lmerTest, MuMin, and multcomp packages in R. For all taxonomic and functional pathways, the model formulae took the format of:

lmer(PathwayOrTaxon ~ (1 | Family) + (0 + log(day of life) | Subject) + log(day of life) + x + y + …., REML = FALSE, data = df)

Family and (time | subject) were mandatory random effects and time was a mandatory fixed effect in all models.

Fixed effects were back-fitted according to the following schema:

1. Screening for candidate variables
   a. For each outcome variable of interest (pathway or taxon abundance), the median was determined.
   b. Two arrays of clinical variables were created; one associated with values above the median for the pathway or taxon of interest, and one associated with values below the median.
   c. To screen for covariation of clinical variables with the outcome variable of interest, a two-tailed *t*-test was done for all continuous clinical variables, and a two-tailed Fisher's exact test was done for all binary clinical variables. This screening test was performed to select candidate variables for inclusion in a longitudinal GLMM. No statistical conclusions were made on the basis of this screening test, as this simple screen could not correct for repeated sampling over time, familial effects, and correction for confounding variables.
   d. All clinical variables with $P < 0.05$ on initial *t*-test or Fisher's exact test screening were considered candidate variables for inclusion in the next naive model-fitting set.
   e. Day of life, delivery route, and breastfeeding (>50%) were always included in the set of candidate variables, regardless of significance in the initial variable screening step.
   f. If any maternal weight variable (maternal prepregnancy BMI or maternal GWG) came through the initial screening step, then maternal prepregnancy BMI, maternal GWG, and infant gestational age at delivery were all included in the set of candidate variables, due to the potential for confounding effects.
2. Naive model fitting
   a. As binary variables were in two formats (current exposure to an ingredient versus lifetime exposure to an ingredient), two models were fitted: one for current exposure, one for lifetime exposure. Demographic variables (for example, maternal age, infant birthweight, day of life) were included in all models.
   b. If variables were supersets of other variables (for example, 'Maternal Intrapartum antibiotics' is a superset of 'Maternal Ampicillin' and 'Maternal Cefazolin'), the supersets and subsets were not included in the same model; instead, a specific model (with only subset variables) and a general model (with only superset variables) were created.
   c. Maximum-likelihood GLMMs of all candidate variables identified in step 1 were created using the lmer() function in the lme4 package.
   d. The step() function in the lmerTest package was used to back-fit maximum-likelihood GLMMs for all candidate variables, with a significance cut-off of 0.05 for retaining fixed effects.
   e. Pseudo-R^2 was determined using r.squaredGLMM() in the MuMin package.
3. Testing for contribution of interaction terms
   a. If the correlation matrix of the output model showed any values >0.1 or <−0.1, between infant age and another variable, an interaction term for that variable and infant age ($x \times \log(\text{day of life})$) was added to the set of candidate variables, and back-fitting with the step() function was repeated.
   b. Pseudo-R^2 was determined using r.squaredGLMM() in the MuMin package.
4. Model comparison
   a. The best model was selected from the set of back-fitted models associated with a given outcome variable, which included a current-exposure model and a lifetime-exposure model. If superset/subset variables were part of the candidate set, then the current- and/or lifetime-exposure models were also divided into specific and general models. Pseudo-R^2 was prioritized as a post hoc test to select the best model.
5. Adjustment for multiple comparisons
   a. The glht() function in the multcomp package[55] was used to adjust *P* values in the preferred model for multiple comparison (lincfit = mcp(tension = 'Tukey')).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Example code is available at https://bitbucket.org/alaricwdsouza/twindiet/src/master/TwinDiet_ModelFittingExample.R.

Data availability

Sequence data supporting these findings have been deposited, along with relevant clinical metadata, in the SRA under BioProject ID PRJNA473126, with primary BioSample accession codes SAMN09259835–SAMN09260236 (study SRP148966). Source data for Figs. 1–4 are available online. Any additional data generated and analyzed in this study are available from the corresponding author upon reasonable request.

References

18. Agostoni, C., Carratù, B., Boniglia, C., Riva, E. & Sanzini, E. Free amino acid content in standard infant formulas: comparison with human milk. *J. Am. Coll. Nutr.* **19**, 434–438 (2000).
19. Sharon, G. et al. Specialized metabolites from the microbiome in health and disease. *Cell Metab.* **20**, 719–730 (2014).
20. Haschke-Becher, E., Kainz, A. & Bachmann, C. Reference values of amino acids and of common clinical chemistry in plasma of healthy infants aged 1 and 4 months. *J. Inherit. Metab. Dis.* **39**, 25–37 (2016).
21. Piacentini, G., Peroni, D., Bessi, E. & Morelli, L. Molecular characterization of intestinal microbiota in infants fed with soymilk. *J. Pediatr. Gastroenterol. Nutr.* **51**, 71–76 (2010).
22. Vázquez, L., Flórez, A. B., Guadamuro, L. & Mayo, B. Effect of soy isoflavones on growth of representative bacterial species from the human gut. *Nutrients* **9**, 727 (2017).
23. Li, S. et al. Continuously ingesting fructooligosaccharide can't maintain rats' gut *Bifidobacterium* at a high level. *J. Food Sci.* **80**, M2530–M2534 (2015).
24. Bhatia, J. & Greer, F. The Committee on Nutrition. Use of soy protein-based formulas in infant feeding. *Pediatrics* **121**, 1062–1068 (2008).
25. Vandenplas, Y. Prevention and management of cow's milk allergy in non-exclusively breastfed infants. *Nutrients* **9**, 731 (2017).
26. Bauchart-Thevret, C., Stoll, B., Chacko, S. & Burrin, D. G. Sulfur amino acid deficiency upregulates intestinal methionine cycle activity and suppresses epithelial growth in neonatal pigs. *Am. J. Physiol. Endocrinol. Metab.* **296**, E1239–E1250 (2009).
27. Choe, E. K., Moon, J. S. & Park, K. J. Methionine enhances the contractile activity of human colon circular smooth muscle in vitro. *J. Korean Med. Sci.* **27**, 777–783 (2012).
28. Neis, E. P., Dejong, C. H. & Rensen, S. S. The role of microbial amino acid metabolism in host metabolism. *Nutrients* **7**, 2930–2946 (2015).
29. Alsaker, K. V., Paredes, C. & Papoutsakis, E. T. Metabolite stress and tolerance in the production of biofuels and chemicals: gene-expression-based systems analysis of butanol, butyrate, and acetate stresses in the anaerobe *Clostridium acetobutylicum*. *Biotechnol. Bioeng.* **105**, 1131–1147 (2010).
30. Vitreschak, A. G., Rodionov, D. A., Mironov, A. A. & Gelfand, M. S. Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res.* **30**, 3141–3151 (2002).
31. Stanislawski, M. A. et al. Pre-pregnancy weight, gestational weight gain, and the gut microbiota of mothers and their infants. *Microbiome* **5**, 113 (2017).
32. Collado, M. C., Isolauri, E., Laitinen, K. & Salminen, S. Effect of mother's weight on infant's microbiota acquisition, composition, and activity during early infancy: a prospective follow-up study initiated in early pregnancy. *Am. J. Clin. Nutr.* **92**, 1023–1030 (2010).
33. Antony, K. M. et al. The preterm placental microbiome varies in association with excess maternal gestational weight gain. *Am. J. Obstet. Gynecol.* **212**, 653. e651–616 (2015).
34. Ma, J. et al. High-fat maternal diet during pregnancy persistently alters the offspring microbiome in a primate model. *Nat. Commun.* **5**, 3889 (2014).
35. Hu, J. et al. Diversified microbiota of meconium is affected by maternal diabetes status. *PLoS ONE* **8**, e78257 (2013).
36. Prince, A. L. et al. The perinatal microbiome and pregnancy: moving beyond the vaginal microbiome. *Cold Spring Harb. Perspect. Med.* **5**, a023051 (2015).
37. Sacchetti, R., De Luca, G., Dormi, A., Guberti, E. & Zanetti, F. Microbial quality of drinking water from microfiltered water dispensers. *Int. J. Hyg. Environ. Health* **217**, 255–259 (2014).
38. Dias, M. F. et al. Changes in mouse gut bacterial community in response to different types of drinking water. *Water Res.* **132**, 79–89 (2017).

39. Poroyko, V. et al. Gut microbial gene expression in mother-fed and formula-fed piglets. *PLoS ONE* **5**, e12459 (2010).
40. Charbonneau, M. R. et al. Sialylated milk oligosaccharides promote microbiota-dependent growth in models of infant undernutrition. *Cell* **164**, 859–871 (2016).
41. Lim, E. S. et al. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).
42. Moore, A. M. et al. Gut resistome development in healthy twin pairs in the first year of life. *Microbiome* **3**, 27 (2015).
43. Gurnee, E. A. et al. Gut colonization of healthy children and their mothers with pathogenic ciprofloxacin-resistant *Escherichia coli*. *J. Infect. Dis.* **212**, 1862–1868 (2015).
44. Gibson, M. K. et al. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat. Microbiol.* **1**, 16024 (2016).
45. Fein, S. B. et al. Infant Feeding Practices Study II: study methods. *Pediatrics* **122**(Suppl. 2)), S28–S35 (2008).
46. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
47. McHardy, I. H. et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* **1**, 17 (2013).
48. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
49. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **6**, e17288 (2011).
50. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).
51. Chu, D. M. et al. The early infant gut microbiome varies in association with a maternal high-fat diet. *Genome Med.* **8**, 77 (2016).
52. Robinson, A. et al. Association of maternal gestational weight gain with the infant fecal microbiota. *J. Pediatr. Gastroenterol. Nutr.* **65**, 509–515 (2017).
53. Singh, S., Karagas, M. R. & Mueller, N. T. Charting the maternal and infant microbiome: what is the role of diabetes and obesity in pregnancy? *Curr. Diab. Rep.* **17**, 11 (2017).
54. American College of Obstetricians and Gynecologists. ACOG Committee opinion no. 548: weight gain during pregnancy. *Obstet. Gynecol.* **121**, 210–212 (2013).
55. Joo, J. W., Hormozdiari, F., Han, B. & Eskin, E. Multiple testing correction in linear mixed models. *Genome Biol.* **17**, 62 (2016).

nature research

Corresponding author(s): Aimee Baumann-Dudenhoeffer, Gautam Dantas

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Sequence data processing and extraction of taxonomic data and functional pathways was performed using Trimmomatic (trimmomatic/0.33), Deconseq/0.4.3-chr38, MetaPhlAn2 (metaphlan2/2.2.0), and Humann2 (humann2/0.9.4)
Data analysis	Statistical analysis was performed in R using the vegan, ape, ggplot2, lme4, lmerTest, MuMin, and multcomp packages. R code used in the study is described in the Supplement, and sample code is available at https://bitbucket.org/alaricwdsouza/twindiet/src/master/TwinDiet_ModelFittingExample.R

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data supporting these findings have been deposited, along with relevant clinical metadata, in the SRA under Bioproject ID PRJNA473126 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA473126), with the primary accession codes SAMN09259835-SAMN09260236 (https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP148966). Source data for Figures 1-4 are provided online with the paper. Any additional data generated and analyzed in this study are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Because the expected effect sizes for clinical variables of interest (e.g. formula ingredients) correlating with microbial taxa and functions is unknown, a true power calculation could not be performed. However, our population cohort (N=60), sampling interval (monthly from 0-8 months), and samples successfully whole-genome sequenced (N=402) , is comparable to other published studies of gut microbiome development that were able to detect statistically significant associations of clinical variables with microbiome taxa and functions. For example, Backhed et al, Cell Host and Microbe, 2015 had a larger population (N=98 mothers and infants), but lower resolution longitudinal sampling, with samples collected at birth, 4 months, and 12 months. Similarly, Chu et al, Nature Medicine, 2017 had a larger population (two matched cohorts of N=81), but a smaller number of fecal samples selected for whole-genome sequencing (N=69 meconium samples, and infant and maternal stools). Due to these authors' successful analysis of comparable datasets, it was reasonable to conclude that the sample size of our study was appropriate for investigating the effects of pre- and post-natal clinical factors on early gut microbiome maturation.
Data exclusions	No data that met our predetermined sequencing threshold (>= 5 million total reads prior to processing) were excluded.
Replication	In this longitudinal cohort study, statistical modeling of multiple distinct taxonomic groups and metabolic pathways showed similar patterns and trends, consistently identifying major determinants of functional microbiome maturation (breastfeeding, soy formula exposure, prebiotics, antibiotics, domestic water source, maternal gestational weight gain). Our taxonomic modeling independently confirmed several known determinants of gut microbiome establishment, replicating others' prior work in the field, and confirming the utility of our approach.
Randomization	This was a longitudinal cohort study with no intervention applied; randomization is not applicable
Blinding	This was a longitudinal cohort study with no intervention applied; blinding is not applicable

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	402 samples were included from 60 twin infants in thirty-one families, for a median of 7 samples per infant (IQR 6-8). Infant age at stool collection ranged from the day of delivery to 253 days. The median gestational age at delivery was 37 weeks (IQR 36-38), 43% of infants were delivered vaginally, and 57% were born via Cesarean section. 48% of infants were male and 52% were female. 17% of the infants were Black, 83% were white. 3% of infants were Hispanic, and 97% were non-Hispanic. and 47% of twins were monozygotic, 50% dizygotic, and 3% of unknown zygosity. Four infants' mothers were diabetic (7%), six infants' mothers developed preeclampsia (10%), and two infants were born to a mother with both conditions. Additional detailed population characteristics are included in Supplemental Table S1 and S6
Recruitment	This study was approved by the Human Research Protection Office of Washington University School of Medicine in St. Louis, and complied with all ethical regulations. Written consent was obtained from each adult and a parent or guardian of each minor subject. We used fecal samples that had been frozen at -80 C since collection at monthly intervals from a birth cohort of healthy twins in St. Louis, in which the mothers had consented to monthly fecal sample collection from birth until two years of age. Selecting a population of twins may introduce bias as the mothers are at higher risk for complications of pregnancy (e.g. preeclampsia, diabetes), more likely to deliver via Cesarean section, and more likely to deliver prematurely. To avoid potential bias from early illness and antibiotic administration, we excluded any infants who received antibiotics in the immediate postnatal period.