***Prevotella copri*-linked effects of a therapeutic food for malnutrition**

Yi Wang[1,2,3†], Hao-Wei Chang[1,2,3†], Evan M. Lee[1,2†], Cyrus Zhou[1,2], Kali M. Pruss[1,2,3], Suzanne Henrissat[1,2,4], Robert Y. Chen[1,2], Clara Kao[1,2], Matthew C. Hibberd[1,2,3], Hannah M. Lynn[1,2], Daniel M. Webber[1,2,3], Marie Crane[1,2], Jiye Cheng[1,2], Dmitry A. Rodionov[5], Aleksandr A. Arzamasov[5], Juan J. Castillo[6], Garret Couture[6], Ye Chen[1,2,6], Nikita P. Balcazo, Jr. [6], Carlito B. Lebrilla[6], Nicolas Terrapon[4], Bernard Henrissat[7,8], Olga Ilkayeva[9,10,11], Michael J. Muehlbauer[9,10], Christopher B. Newgard[9,10,11,12], Ishita Mostafa[13], Subhasish Das[13], Mustafa Mahfuz[13], Andrei L. Osterman[5], Michael J. Barratt[1,2,3], Tahmeed Ahmed[13] and Jeffrey I. Gordon[1,2,3,*]

[†]Contributed equally

[1]Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63110 USA

[2]Center for Gut Microbiome and Nutrition Research, Washington University School of Medicine, St. Louis, MO 63110 USA

[3]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110 USA

[4]Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille University, F-13288, Marseille, France

[5]Infectious and Inflammatory Disease Center, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA 92037 USA

[6]Department of Chemistry, University of California, Davis, CA 95616 USA

[7]Department of Biotechnology and Biomedicine (DTU Bioengineering), Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

[8]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

[9]Sarah W. Stedman Nutrition and Metabolism Center, Duke University Medical Center, Durham, NC 27710 USA

[10]Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC 27710 USA

[11]Department of Medicine, Duke University Medical Center, Durham, NC, 27710 USA

[12]Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC 27710 USA

[13]International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka 1212, Bangladesh

[*]Address correspondence to: jgordon@wustl.edu

Key words: childhood malnutrition; gut microbiome development/repair; therapeutic foods; polysaccharide structure/metabolism; intestinal gene expression and metabolism; single nucleus RNA-Seq

Short title: Role of *P. copri* in treating malnutrition

Preclinical and clinical studies are providing evidence that the healthy growth of infants and children reflects, in part, healthy development of their gut microbiomes[1-5]. This process of microbial community assembly and functional maturation is perturbed in children with acute malnutrition. Gnotobiotic animals, colonized with microbial communities from children with severe and moderate acute malnutrition, have been used to develop microbiome-directed complementary food (MDCF) formulations for repairing the microbiomes of these children during the weaning period[5]. In the accompanying paper[1], we analyze bacterial genomes assembled from sequencing the fecal microbiomes of Bangladeshi children with moderate acute malnutrition (MAM) who had participated in a previously reported 3-month randomized controlled clinical study of an MDCF[6]. This formulation, MDCF-2, produced significantly improved weight gain compared to a commonly used nutritional intervention despite the lower caloric density of the MDCF. Characterizing the metagenome assembled genomes (MAGs) of bacterial strains present in the microbiomes of study participants revealed a significant correlation between accelerated ponderal growth and the expression by two *Prevotella copri* MAGs of metabolic pathways involved in processing of MDCF-2 glycans[1]. To provide a direct test of these relationships, in this report we perform 'reverse translation' experiments. These experiments center on a gnotobiotic mouse model of mother-to-offspring microbiome transmission involving sequential introduction of defined consortia of age- and ponderal growth-associated gut bacterial strains cultured from Bangladeshi infants/children in the study population, with or without *P. copri* isolates resembling the MAGs. By combining analyses of microbial community assembly, gene expression and processing of glycan constituents of MDCF-2 with single nucleus RNA-Seq and mass spectrometric analyses of the intestine, we establish a key role for the *P. copri* strains in mediating metabolism of MDCF-2 glycans, as well as the activities of metabolic pathways involved in lipid, amino acid, carbohydrate plus other facets of energy metabolism within epithelial cells positioned at different locations in intestinal crypts and villi. Together, the results described in these two papers provide insights into structure/function relationships between MDCF-2 and members of the gut communities of malnourished children; they also have implications for developing future prebiotic, probiotic and/or synbiotic therapeutics for microbiome restoration in children with already manifest malnutrition, or who are at risk for this pervasive health challenge.

In a companion study[1], we describe (i) sequencing DNA from fecal samples that were serially collected from participants during the course of this clinical trial, (ii) assembly of the resulting sequencing reads into MAGs, (iii) identification of MAGs whose abundances are associated with anthropometric measures of ponderal growth (changes in weight-for-length (height) expressed as WLZ scores), (iv) characterization of how these MAGs change their patterns of gene expression in response to treatment, and (v) analysis of the metabolism of carbohydrate components of MDCF-2. A notable finding was that among the 75 MAGs found to be positively correlated with WLZ, two belonging to *Prevotella copri* were a principal source of expression of carbohydrate utilization pathways that target MDCF-2 polysaccharides. Moreover, expression of these pathways and the levels of products of glycan metabolism were significantly positively correlated with the magnitude of the children's improvement in WLZ. Here we describe 'reverse translation' experiments, conducted in gnotobiotic mice, that directly examine the role of *P. copri* in mediating MDCF-2 metabolism and the effects of this therapeutic food on aspects of intestinal function.

The small intestine is lined with a layer of continuously renewing epithelial cells. Renewal is fueled by $Lgr5^+$ stem cells located at the base of mucosal invaginations (crypts of Lieberkühn). Stem cell progenies adopt four principal fates. Members of the predominant enterocytic lineage execute their differentiation program as they migrate from crypts up adjacent small intestinal villi. Goblet cells, enteroendocrine cells, tuft cells, and Paneth cells represent other lineal descents of the multipotential crypt stem cell and its oligo-potential daughters. Substantial metabolic investments are needed to support the normal daily replacement of large numbers of gut epithelial cells[7] as well as the functions they normally express in their differentiating and differentiated states.

To test the hypothesis that epithelial cell gene expression and metabolism would be sensitive reporters of functional changes associated with *P. copri* strains representing the MAGs, we have performed dam-to-pup microbial community transmission studies. This involved sequential introduction, during the suckling and weaning periods, of defined collections of genome-sequenced bacterial strains cultured from the study population. The strains represented organisms whose prominence normally changes at different stages of postnatal gut microbial community assembly in healthy Bangladeshi children ('age-discriminatory' taxa)[4,5,8,9]; they also include WLZ-correlated bacterial strains selected based on their shared features with MAGs identified in the clinical study[1,6]. Two groups of mice were compared; those that did or did not harbor *P. copri* in their defined consortia of cultured organisms. All pups were subjected to the same dietary sequence of exclusive milk feeding (from the dam) followed by weaning onto an MDCF-2 supplemented diet. The results highlight the central role played by *P. copri* strains closely resembling the identified WLZ-associated MAGs in metabolizing glycans present in MDCF-2 plus their capacity, in the context of this consortium of strains and MDCF-2, to promote weight gain and influence expression of metabolic functions and other activities in enterocytes as they migrate up the villus.

Results

A manipulatable model of mother-pup transmission of cultured age- and WLZ-associated bacterial taxa

Selection of bacterial strains – To test the role of *P. copri* in the context of a defined human gut microbial community that captured features of the developing communities of children who had been enrolled in the clinical study of MDCF-2[1], we selected 20 bacterial strains, 16 of which were cultured from the fecal microbiota of 6- to 24-month-old Bangladeshi children living in Mirpur (**Supplementary Table 1a**). They included strains initially identified by the close correspondence of their 16S rRNA gene sequences to (i) a group of taxa that describe a normal program of development of the microbiota in healthy Bangladeshi children[8,9] and (ii) taxa whose abundances had statistically significant associations (positive or negative) with the rate of weight gain (β-WLZ) and statistically significant correlations with plasma levels of WLZ-associated proteins in children from the previously reported MDCF-2 clinical study[5,6]. The relatedness of these strains to the 1,000 MAGs assembled from fecal samples obtained from all participants in the clinical study[1] was determined by average nucleotide sequence identity (ANI) scores, alignment coverage parameters[10,11] and their encoded metabolic pathways. [See the accompanying paper for how *in silico* reconstructions were performed to predict which of 106 different pathways reflecting the major metabolic capabilities of the human gut microbiome were present or absent, and how the results are expressed as 'binary phenotype' scores for each pathway ('1' or '0' denoting present or absent)]. We deemed a cultured, bacterial strain as representing a specific MAG if the whole genome alignment coverage was >10%, ANI was >94%, and the binary phenotype concordance score was >90% (see *Methods*). Based on these criteria, we classified four of the 20 strains as corresponding to MAGs positively associated with WLZ, including *P. copri,* and eight strains as corresponding to MAGs negatively associated with WLZ (**Supplementary Table 1b-d**).

As noted in the accompanying paper[1], liquid chromatography-mass spectrometry (LC-MS) analysis of glycosidic linkages and polysaccharides in MDCF-2 and RUSF disclosed that cellulose, galactan, arabinan, xylan, and mannan represent the principal non-starch polysaccharides in MDCF-2. Gene set enrichment analysis (GSEA) of fecal microbial RNA-Seq datasets generated from children in the MDCF-2 and RUSF arms of the clinical trial disclosed that MDCF-2 produced a meta-transcriptome that was enriched for components of metabolic pathways involved in the utilization of arabinose, α-arabinooligosaccharides (aAOS), and fucose. One-third of the 'leading-edge' transcripts associated with these pathways (*i.e.*, transcripts most discriminatory for the pathway response to MDCF-2[1]) were derived from the two *P. copri* MAGs whose abundances were positively correlated with WLZ (MAGs Bg0018 and Bg0019[1]); These leading-edge transcripts include 11 of the 14 related to aAOS utilization. Moreover, a comparison of the fecal meta-transcriptomes of children in the MDCF-2 arm of the clinical study

revealed that those in the upper quartile of WLZ responses (*i.e.*, had the greatest rate of weight gain) exhibited significant enrichment in the expression of metabolic pathways for utilization of MDF2-glycans. A majority of the leading-edge transcripts in these pathways were also derived from the two *P. copri* MAGs. Another feature that distinguished these two MAGs from the other nine *P. copri* MAGs present in the microbiomes of study participants is that they share 10 functionally conserved PULs, including seven that are completely conserved and three that are partially conserved, albeit structurally distinct (see *Methods* and ref. 1 for the criteria used to classify the degree of PUL conservation). These 10 PULs encode a diverse set of glycoside hydrolases (**Supplementary Table 1e**) with broad substrate specificities for glycans present in MDCF-2 (range of substrates: arabinogalactan, pectin, pectic galactan, β-glucan, β-mannan, glucomannan, xylan, arabinoxylan, and xyloglucan)[1]. Notably, the degree of representation of the seven completely conserved PULs among the 11 *P. copri* MAGs identified in study participants was highly predictive of each MAG's association with β-WLZ, suggesting a link between metabolism of carbohydrates by *P. copri* and ponderal growth responses among the malnourished children[1].

The Bangladeshi *P. copri* strain PS131.S11 (abbreviated *P. copri* Bg131) was the only *P. copri* strain in the 20-member collection. There were several reasons why we chose Bg131 over four other cultured *P. copri* strains initially obtained from Bangladeshi children. *First*, based on phylogenetic distance, *P. copri* Bg131 was most similar to MAGs Bg0018 and Bg0019 (**Extended Data Fig. 1a**). *Second*, the results of our *in silico* metabolic reconstructions were expressed as a binary phenotype matrix (BPM), where for (i) amino acids and B vitamins, "1" denotes a predicted prototroph and "0" an auxotroph, (ii) for carbohydrates, "1" and "0" refer to a strain's predicted ability or inability, respectively, to utilize the indicated mono-, di- or oligosaccharide, and (iii) for fermentation end products, a "1" and "0" indicate a strain's predicted ability/inability to produce the indicated compound (see **Supplementary Table 1c**). Based on this classification, strain Bg131 has an overall binary phenotype concordance score of 97% and 96% when compared to Bg0018 and Bg0019, respectively (**Supplementary Table 1d**). Among 55 carbohydrate utilization pathways analyzed, Bg131 shares 53 (96%) with MAG Bg0018 and 54 (98%) with MAG Bg0019 (**Supplementary Table 1d**). *Third*, *P. copri* Bg131 contains 32 PULs including five of the 10 highly and partially conserved PULs shared by Bg0018 and Bg0019. These five PULs in *P. copri* Bg131 are predicted to be involved in utilizing β-glucan (PUL8 and PUL30), pectin (PUL3), pectic galactan (PUL14), starch (PUL27a), and xylan (PUL8) (**Supplementary Table 1e,f**). Although it did not meet our strict criteria for conservation with these Bg0018/Bg0019 PULs, an additional arabinogalactan-targeted PUL (PUL27b) immediately adjacent to the conserved PUL27a was also identified.

P. stercorea was the only other *Prevotella* species present in the 20-member collection. Although none of the WLZ positively (or negatively) associated MAGs identified in the clinical study belonged to *P. stercorea*, this isolate was included in the collection to assess the specificity of the responses of *P. copri* to MDCF-2. The *P. stercorea* isolate did not possess any of the PULs present in *P. copri* Bg131 or Bg0018/Bg0019, even after relaxing the criteria for sequence conservation to account for the taxonomic divergence between the two species. The cultured *P. stercorea* strain has 10 PULs, only five of which encode known carbohydrate utilization enzymes. The glycoside hydrolases in these five PULs are predicted to have very different carbohydrate specificities from those found in the *P. copri* strain and two *P. copri* MAGs (the *P. stercorea* PULs mainly target non-plant glycans) (**Supplementary Table 1g**).

Bifidobacterium longum subsp. *infantis* (*B. infantis*) is a prominent early colonizer of the infant gut. Therefore, we wanted to ensure that it was well represented at the earliest stages of assembly of the defined community so that later colonizers such as *P. copri* could establish themselves. The collection of cultured isolates included two strains of *B. infantis* recovered from Bangladeshi children – *B. infantis* Bg463 and *B. infantis* Bg2D9. The Bg463 strain had been used in our earlier preclinical studies that led to development of MDCF-2[5,9]. In a recent study, simultaneous introduction of five different Bangladeshi *B. infantis* strains, without any other organisms, into just-weaned germ-free mice fed a diet representative of that consumed by 6-month-old children living in Mirpur[12] revealed that *B. infantis* Bg2D9 exhibited greater fitness (absolute abundance) than any of the other strains, including Bg463. Based on comparative

genomic and microbial RNA-Seq analyses, this improved fitness was attributed to additional carbohydrate utilization pathways that the Bg2D9 strain possesses[12].

Initial colonization and phenotyping experiment

Design - We used the 20-strain collection to perform a 3-arm, fixed diet study that involved 'successive' waves of maternal colonization with four different bacterial consortia (**Fig. 1a-c**). The sequence of introduction of taxa into dams was designed to emulate temporal features of the normal postnatal development of the human gut community; *e.g.*, consortia 1 and 2 were comprised of strains that are prominent colonizers of healthy infants/children in the first postnatal year (including the *B. infantis* isolates) while those in consortium 3 are prominent during weaning in the second postnatal year[4,5,8,9]. This dam-to-pup colonization strategy also helped overcome the technical challenge of reliable delivery of bacterial consortia to newborn pups via oral gavage.

Dually-housed germ-free dams were switched from a standard breeder chow to a 'weaning-diet' supplemented with MDCF-2 on postpartum day 2, two days before initiation of the colonization sequence. This weaning diet was formulated to emulate the diets consumed by children in the clinical trial during MDCF-2 treatment (See *Methods*; **Fig 1a**; **Supplementary Table 2**). It contained (i) powdered human infant formula, (ii) complementary foods consumed by 18-month-old children living in Mirpur, Bangladesh where the study took place, and (iii) MDCF-2. The contributions of the milk, complementary food and MDCF-2 'modules' to total caloric content (53%, 17%, and 30%, respectively) were based on published studies of the diets of cohorts of healthy and undernourished 12- to 23-month-old children from several low- and middle-income countries, including Bangladesh, as well as the amount of MDCF-2 given to the 12-18-month-old children with MAM in the clinical study[1,5,6,13-16].

In Arm 1, dams received the following series of oral gavages: (i) on postpartum day 4, a consortium of five 'early' infant gut community colonizers with *B. infantis* represented by Bg2D9; (ii) on postpartum day 7, *P. copri and P. stercorea*; (iii) on postpartum day 10, additional age-discriminatory and WLZ-associated taxa, and (iv) on postpartum day 21, *P. copri, P. stercorea,* and *Faecalibacterium prausnitzii* (**Fig. 1b,c**). At this last time point, the three strains were given by oral gavage to both the dams and their offspring to help promote successful colonization. In Arm 2, pups were subjected to the same sequence of microbial exposures and the same diet switches as in Arm 1, except that *B. infantis* Bg463 rather than *B. infantis* Bg2D9 was included in the first gavage mixture. Arm 3 was a replicate of Arm 2 but without the *Prevotella* gavages. Pups in all three arms were subjected to a diet sequence that began with exclusive milk feeding (from the nursing dam) followed by a weaning period where pups had access to the weaning phase diet supplemented with MDCF-2. Pups were weaned at postnatal day 24 (P24), after which time they received MDCF-2 alone *ad libitum* until P53 when they were euthanized. Our rationale for the timing of the first three gavages was based on the diet sequence [gavage 1 of early colonizers at a time (P4) when mice were exclusively consuming the dam's milk, gavage 2 as the pups were just beginning to consume the human weaning (complementary food) diet, gavage 3 somewhat later during this period of 'complementary feeding and the fourth gavage to help to ensure a consistent level of *P. copri* colonization at the end of weaning (and subsequently through the post-weaning period)]. The relative abundances of these strains in fecal samples collected from dams on days postpartum days 21, 24, and 35, as well as the absolute abundances of these strains in fecal samples collected from their offspring on P21, P24, P35, and P53 were quantified by shotgun sequencing of community DNA (n=2 dams and 5-8 pups analyzed/arm). Colonization of each consortium member was highly consistent among all animals in each treatment group (**Supplementary Table 3).** [Note that control dam-to-pup transmission experiments where we attempted to mono-colonize mice with *P. copri* showed that *P. copri* is a poor and unreliable colonizer by itself (**Extended Data Fig. 1**), with absolute abundance levels in the cecum up to several orders of magnitude lower than that achieved in Arm 1. Moreover, the offspring of dams that had just been gavaged with *P. copri* had massive, fluid-filled cecums similar to those typically seen in germ-free mice (n=9 mice per experiment; 2 independent control experiments)].

A relationship between B. infantis and P. copri colonization – *B. infantis* Bg2D9 successfully colonized pups at P21 in Arm 1 [8.4±0.5 $\log_{10}$ (genome equivalents/mg feces) (mean±SD); relative

abundance, 9.0±3.9% (mean±SD)]. In contrast, *B. infantis* Bg463 colonized at 5-8 orders-of-magnitude lower absolute abundance levels in Arms 2 and 3 [3.22±1.9 and 0.6±1.5 $\log_{10}$ (genome equivalents/mg feces) (mean±SD), respectively]. These differences were sustained through P53 (**Fig. 1d**). Moreover, exposure to *B. infantis* Bg2D9 in Arm 1 was associated with 3-orders of magnitude greater levels of *P. copri* colonization in the pre-weaning period (P21) than in Arm 2 mice exposed to *B. infantis* Bg463 [$P<0.005$, Mann-Whitney U test] (**Fig. 1e; Supplementary Table 3c**). Administering the fourth gavage on P21 elevated the absolute abundance of fecal *P. copri* in Arm 2 to a level comparable to Arm 1; this level was sustained throughout the post-weaning period (P24 to P53) [**Fig. 1e;** $P>0.05$; linear mixed effects model (*Methods*); also see **Supplementary Table 3c**]. This effect of the fourth gavage was also evident in the ileal and cecal microbiota at the time of euthanasia (**Supplementary Table 3d**).

The effects of *B. infantis* on *P. copri* did not generalize to *P. stercorea.* Unlike *P. copri*, the absolute abundance of *P. stercorea* in feces sampled on P21 and P24 was not significantly different in mice belonging to Arms 1 and 2 ($P>0.05$, Mann-Whitney U test). Prior to weaning at P24, the absolute abundance of *P. stercorea* was 5 orders of magnitude lower than that of *P. copri*. Throughout the post-weaning period, the absolute abundance of *P. stercorea* remained similar in members of both treatment arms ($P>0.05$, Mann-Whitney U test) but 2 orders of magnitude below that of *P. copri* (**Supplementary Table 3c**).

Based on these results, we directly tested the colonization dependency of *P. copri* on *B. infantis* in two independent experiments whose designs are outlined in **Extended Data Fig. 1b**. Dually-housed germ-free dams were switched from standard breeder chow to the weaning Bangladeshi diet supplemented with MDCF-2 on postpartum day 2. On postpartum day 4, one group of dams was colonized with *B. infantis* Bg2D9 while the other group received a sham gavage. On postpartum days 7 and 10, both groups of gnotobiotic mice were gavaged with a consortium containing five *P. copri* strains. These five *P. copri* strains (1A8, 2C6, 2D7, G8, and Bg131) were all isolated from fecal samples obtained from Bangladeshi children (**Supplementary Table 4**). Pups were separated from their dams at the completion of weaning and their diet was switched to MDCF-2 until euthanasia on P42. At this time point, the absolute abundance of *P. copri* in feces collected from mice that had received *B. infantis* Bg2D9 was 3-orders of magnitude higher than in animals never exposed to *B. infantis* (**Extended Data Fig. 1c; Supplementary Table 5**) – a finding that confirmed what we had observed between Arms 1 and 2 of the initial colonization experiment (see **Fig. 1e)**. There was no statistically significant difference in weight gain from P23 to P42 between the mono- and bi-colonization groups. However, interpretation of this result is confounded by the fact that compared to the bi-colonized animals with significantly higher levels of *P. copri*, mono-colonized mice with low levels of *P. copri* had massive, fluid-filled cecums, similar to those commonly seen in germ-free mice. This pronounced cecal enlargement adds substantially to body weight and in a comparison of the two treatment groups obscures the ability to discern whether increased levels of *P. copri* has ponderal growth-promoting effects.

Effects on weight gain and metabolism of MDCF-2 glycans – Gnotobiotic mice in Arm 1 exhibited a significantly greater increase in weight gain between P23 (the first time point measured, 2 days after the final gavage) and P53 compared to mice in the two other experimental arms [$P < 0.05$ for Arm 1 versus Arm 2; $P<0.01$ for Arm 1 versus Arm 3; linear mixed-effects model (see *Methods*)] (**Fig. 1f**). Unlike the mono- and bi-colonization experiments described above, cecal sizes were comparable across the three treatment groups. It is important to note that in this experiment isolating the effects of *P. copri* on weight gain requires a comparison of Arms 2 and 3 which both contain the same *B. infantis* 463 strain rather than Arms 1 and 3 which also differ by *B. infantis* strain. While there was not a significant difference in the weights between mice belonging to Arms 2 and 3, the absolute abundance of *P. copri* was significantly lower in mice containing *B. infantis* Bg463 in Arm 2 compared to those harboring *B. infantis* Bg2D9 in Arm 1 during the period before completion of weaning (**Fig. 1e; Supplementary Table 3c**). This observation raised the possibility that the levels of *P. copri* (and presumably, its expressed functions) attained *early* in the weaning period was a contributing factor to the accelerated weight gain observed in animals in Arm 1. Given the experiment was ended on P53 we cannot exclude

the possibility that delayed attainment of these levels of *P. copri* in Arm 2 might also result in delayed (catch-up) weight gain had the experiment been continued.

Given these results, we advanced samples collected from mice in Arms 1 and 3 for additional analyses of the metabolism of MDCF-2 glycans. Because (i) *P. copri* was the only *Prevotella* sp. in our defined community that encodes and expresses CAZymes capable of degrading linkages in MDCF-2 glycans (**Supplementary Table 1f**), (ii) *P. copri* achieved markedly higher absolute abundance than *P. stercorea* in this community (**Supplementary Table 3c**,**d**), (iii) analyses of MAGs in the fecal microbiomes of clinical study participants had linked the abundance of *P. copri* but not *P. stercorea* MAGs to ponderal growth (WLZ)[1] and (iv) two different *B. infantis* strains were used, we refer to Arm 1 of this experiment as 'w/ *P. copri* & w/ *B. infantis* Bg2D9' and Arm 3 as 'w/o *P. copri* & w/ *B. infantis* Bg463'. In our companion study[1], we report increased levels of enzyme-resistant arabinose linkages, such as 5-Ara*f*, 2-Ara*f*, and 2,3-Ara*f*, in the feces of MDCF-2 treated children in the upper- compared to lower-quartile of WLZ responses. The results suggested increased consumption and preferential/partial degradation of MDCF-2 and arabinose-containing complementary foods among these MDCF-2 responders. Compared to the human study, our reverse translation experiments conducted in gnotobiotic mice provided a more controlled and sensitive environment for assaying the effects of *P. copri* on MDCF-2 glycan metabolism. We had access to and focused on the mouse cecum because we wanted to compare microbial gene expression with polysaccharide degradative capacity in a large gut habitat specialized for microbial fermentation[17]. Moreover, a uniform diet comprised only of MDCF-2 could be monotonously administered to mice, eliminating potential confounding effects of variations in background diet encountered in humans.

Integrating results from mass spectrometric and microbial RNA-Seq data generated from cecal contents harvested from mice at the time of euthanasia (P53) provided several lines of evidence for the important role played by *P. copri* in metabolizing the principal polysaccharide components of MDCF-2. *First,* unlike *P. stercorea*, *P. copri* Bg131 contains PULs involved in processing MDCF-2 glycans: *i.e.*, PUL27a and PUL27b specify CAZymes known or predicted to digest starch and arabinogalactan, while PUL2 possesses a fucosidase that could target the terminal residues found in arabinogalactan II (**Supplementary Table 1f**). *Second*, ultra-high performance liquid chromatography-triple quadrupole mass spectrometric (UHPLC-QqQ-MS)-based measurements of 49 glycosidic linkages in cecal contents disclosed that animals in Arm 1 harboring *P. copri* had (i) significantly lower levels of t-Ara*p*, t-Ara*f*, 2-Ara*f*, 2,3-Ara*f*, and 3,4-Xyl*p*/3,5-Ara*f* ($P<0.05$; Mann-Whitney U Test; **Fig. 2a**; **Supplementary Table 6a)** and (ii) significantly lower amounts of arabinose in cecal glycans ($P<0.05$; Mann-Whitney U test; **Fig. 2b**; **Supplementary Table 6b**). *Third,* GC-MS-based measurements of cecal short-chain fatty acids showed significantly higher levels of acetate, indicating increased fermentation by the *P. copri* -containing microbial community ($P<0.01$; Mann-Whitney U test) (**Fig. 2c**; **Supplementary Table 7**). Together, our findings indicate that mice with the *P. copri*-containing community exhibit more complete degradation of branched arabinans and a greater degree of liberation of arabinose from MDCF-2 glycans.

Effects of P. copri colonization on metabolic functions expressed by other community members – To investigate the transcripts driving the observed differences in microbial glycan processing in the 'w/ *P. copri* & w/ *B. infantis* Bg2D9' and 'w/o *P. copri* & w/ *B. infantis* Bg463' arms, we performed microbial RNA-Seq on cecal contents collected at the time of euthanasia (**Supplementary Table 8**). Transcript abundance tables were filtered and counts were aggregated based on mcSEED reconstructions of metabolic pathways to provide an average expression value across the genes in a given metabolic pathway in a given organism (see *Methods*). We then performed principal component analysis (PCA) on these aggregated tables and compared the clustering of samples from each experimental arm in a space determined by PC1 and PC2 and the contribution of each expressed metabolic pathway to each principal component (see **Fig. 2d** for details). The results revealed significant separation of meta-transcriptomes aggregated by metabolic pathway between cecal samples obtained from animals in the two experimental arms ($P<0.001$; PERMANOVA) (**Fig. 2e)**.

To identify metabolic pathways that drive separation of samples along PC1, we used the contribution of each pathway in each community member to PC1 to rank the pathways. Notably,

arabinose utilization was consistently among the pathways most highly expressed with *P. copri* colonization (**Fig. 2f**). Moreover, three of the four bacteria capable of arabinose utilization (*Blautia obeum*, *Bifidobacterium catenulatum*, and *Mitsuokella multacida*) were significantly more abundant in the cecums of mice colonized with *P. copri* (**Supplementary Table 3d**).

We subsequently performed differential expression analysis (using edgeR[18]; see *Methods*) to further assess the effects of *P. copri*-colonization on the transcriptional profiles of community members at gene-level resolution. Differentially expressed transcripts associated with complete metabolic pathways are provided in **Supplementary Table 8c** and summarized in **Fig. 2g**. Among the four arabinose-utilizing strains, *B. obeum* and *M. multacida* demonstrated significantly higher expression of all of their genes involved in arabinose utilization with *P. copri* colonization, as well as statistically significant upregulation of all or most of their genes involved in biosynthesis of glutamine and glutamate plus branched-chain amino acids (isoleucine, leucine, and valine). *B. obeum* also displayed elevated expression of genes involved in acetate production in the *P. copri*-colonized mice. Integrating the mass spectrometric and microbial RNA-Seq results generated from this defined consortium indicates that *P. copri* colonization leads to the liberation of arabinose from MDCF-2 glycans, which in turn becomes available to other community members, resulting in their increased fitness and altered expressed metabolic functions.

snRNA-Seq of intestinal gene expression

Histomorphometric analysis of villus height and crypt depth in jejunums harvested from mice harboring communities with and without *P. copri* (n=8 and 7, respectively) disclosed no statistically significant architectural differences between the two treatment groups ($P>0.05$; Mann-Whitney U test; **Supplementary Table 9**). We subsequently turned to snRNA-Seq to investigate whether these two colonization states produced differences in expressed functions along the crypt-villus axis in jejunal tissue collected from P53 animals (n=4/treatment arm; **Fig 3a-c**; **Extended Data Fig. 2**; **Supplementary Table 10**). A total of 30,717 nuclei passed our quality metrics (see *Methods*). Marker gene-based annotation disclosed cell clusters that were assigned to the four principal intestinal epithelial cell lineages (enterocytic, goblet, enteroendocrine, and Paneth cell) as well as to vascular endothelial cells, lymphatic endothelial cells, smooth muscle cells and enteric neurons (**Extended Data Fig. 2a,b**). Marker gene analysis allowed us to further subdivide the enterocytic lineage into three clusters: 'villus-base', 'mid-villus' and 'villus-tip'. Pseudobulk snRNA-seq analysis, which aggregates transcripts for each cell cluster and then uses edgeR to identify differentially expressed genes in each cluster[18,19], disclosed that a majority of all statistically significant differentially expressed genes (3,651 of 5,765; 63.3%) were assigned to the three enterocyte clusters (**Extended Data Fig. 2c; Supplementary Table 10b**).

We initially used *NicheNet*[20] to evaluate the effects of the *P. copri*-containing community on intercellular communications. *NicheNet* integrates information on signaling and gene regulation from publicly available databases to build a "prior model of ligand-target regulatory potential" and then predicts potential communications between user-defined "sender" and "receiver" cell clusters [20]. After incorporating snRNA-Seq-based expression data from both sender and receiver cells, *NicheNet* computes a list of potential ligand-receptor interactions between senders and receivers. The ligand-receptor interactions in the resulting list are then ranked based on their effect on expression of downstream genes in their signaling pathway (*i.e.*, more downstream genes are expressed in a 'high-ranking' interaction). After this ranking step, an additional filter is applied, with ligand-receptor interactions having firm experimental validation in the literature designated as "bona fide" interactions. Finally, *NicheNet* uses information generated by Seurat[21] from a snRNA-Seq dataset to identify altered "bona fide" ligand-receptor interactions.

We designated six of the epithelial cell clusters (crypt stem cells, proliferating transit amplifying (TA)/stem cells, villus base, mid-villus, and villus tip enterocytes and goblet cells) as "receiver cells" while all clusters (both epithelial and mesenchymal) were designated "sender cells". We then conducted *NicheNet* analysis for each sender-receiver pair. **Extended Data Fig. 3** shows bona fide ligand-receptor interactions that are altered between the two colonization conditions for each receiver cell cluster.

Ligands identified include those known to affect cell proliferation (*igf-1*), cell adhesion (*cadm1, cadm3*, *cdh3*, *lama2*, *npnt*), zonation of epithelial cell function/differentiation along the length of the villus (*bmp4*, *bmp5*), as well as immune responses (*cadm1*, *il15*, *tgfb1*, *tnc*) (**Extended Data Fig. 3**). Among all receiver cell clusters, crypt stem cells exhibited the highest number of altered bona fide ligand-receptor interactions. For example, Igf-1 is known to enhance intestinal epithelial regeneration[22]. We found that colonization with the *P. copri*-containing consortium was associated with markedly elevated expression of *igf-1* in goblet and lymphatic endothelial sender cells that signals to crypt stem cell receivers.

We subsequently applied the *Compass* algorithm[23] to our snRNA-Seq datasets to generate *in silico* predictions of the effects of the consortia containing and lacking *P. copri* on the metabolic states of (i) stem cell and proliferating TA cell clusters positioned in crypts of Lieberkühn, (ii) the three villus-associated enterocyte clusters, and (iii) the goblet cell cluster. *Compass* combines snRNA-seq data with the Recon2 database [24]. This database describes 7,440 metabolic reactions grouped into 99 Recon2 subsystems[24], plus information about reaction stoichiometry, reaction reversibility, and associated enzyme(s). Using snRNA-seq data, *Compass* computes a score for each metabolic reaction. If the metabolic reaction is reversible, then one score is calculated for the "forward" reaction and another score is calculated for the "reverse" reaction[24]. We calculated a 'metabolic flux difference' (see *Methods*) to quantify the difference in net flux for a given reaction (*i.e.*, the forward and reverse activities) between the two treatment groups.

Fig. 3 shows the predicted metabolic flux differences for Recon2 reactions in enterocytes distributed along the length of the villus and in goblet cells. In clusters belonging to the enterocyte lineage, the number of statistically significant differences is greatest in villus base enterocytes and decreases towards the villus tip (**Fig. 3a**). Mice in the 'w/ *P. copri* & w/ *B. infantis* Bg2D9' treatment group, relative to their 'w/o *P. copri* & w/ *B. infantis* Bg463' counterparts, had the greatest predicted increases in the activities of subsystems related to energy metabolism, the metabolism of carbohydrates, amino acids and fatty acids, as well as various transporters, in villus base and mid-villus enterocytes (**Fig. 3b**, **Extended Data Fig. 4**).

While enterocytes prioritize glutamine as their primary energy source, they are also able to utilize fatty acids and glucose. The *Compass*-defined increase in reactions related to fatty acid oxidation that occur in the villus enterocytes of mice in the 'w/ *P. copri* & w/ *B. infantis* Bg2D9' group extended to their crypts of Lieberkühn (**Fig. 3b**). Fatty acid oxidation has been linked to intestinal stem cell maintenance and regeneration[25]. Mice colonized with *P. copri* exhibited 'statistically credible increases' in the proportional representation of crypt stem cells and proliferating TA/stem cells but not in their villus-associated enterocytic clusters (**Fig. 3c**; also see **Supplementary Table 10c** for results regarding all identified epithelial and mesenchymal cell clusters). [The term 'statistically credible difference' was defined by *scCODA*[26] (see *Methods*)]. Compared to mice lacking *P. copri*, those colonized with this organism also had predicted increases in energy metabolism in their goblet cells, as judged by the activities of subsystems involved in glutamate (Glu) metabolism, the urea cycle, fatty acid oxidation and glycolysis (**Fig. 3b**).

Citrulline is generally poorly represented in human diets; it is predominantly synthesized via metabolism of glutamine in small intestinal enterocytes and transported into the circulation[27]. Both glutamate and arginine are important for citrulline production in enterocytes[28]. Glutaminase (Gls) and glutamate dehydrogenase (GluD) in the glutamine pathway provide ammonia for generating carbamoyl phosphate (**Fig. 3d**), while arginine is a primary precursor for ornithine synthesis[29]. Ornithine transcarbamylase (Ots) produces citrulline from carbamoyl phosphate and ornithine. *Compass* predicted that mice harboring *P. copri* exhibit statistically significant increases in these reactions in their villus base and mid-villus enterocyte clusters [$q<0.05$ (adjusted *P*-value); Wilcoxon Ranked Sum test; **Fig. 3d**]. Follow-up targeted mass spectrometric analysis confirmed that citrulline was significantly increased in jejunal, ileal and colonic tissue segments, as well as in the plasma of mice harboring *P. copri* ($P<0.05$; Mann-Whitney U test; **Fig. 3e**). Notably, studies of various enteropathies and short bowel syndrome have demonstrated that citrulline is a quantitative biomarker of metabolically active enterocyte mass and its levels in plasma are indicative of the absorptive capacity of the small intestine[27]. Citrulline has also been

reported to be markedly lower in blood from children with severe acute malnutrition compared to healthy children[30]. Low plasma citrulline levels have also been described in cohorts of undernourished children with environmental enteric dysfunction and higher levels were predictive of future weight gain[31,32].

The presence of *P. copri* was also associated with significantly greater predicted activities in Recon2 subsystems involved in transport of nine amino acids (including the essential amino acids leucine, isoleucine, valine, and phenylalanine), dipeptides and monosaccharides (glucose and galactose) in villus base and mid-villus enterocytes (**Fig. 3f**). These predictions suggest a greater absorptive capacity in *P. copri*-colonized animals for these important growth-promoting nutrients, which are known to be transported within the jejunum at the base and middle regions of villi[33].

Further assessment of the effects of *P. copri* on host metabolism and weight gain

To validate some of these *Compass*-based predictions, we repeated the experiment described above with a larger number of animals (4 dually housed germ-free dams yielding 18-19 viable pups per arm). To test whether the weight gain phenotype and metabolic alterations observed in the initial colonization and phenotyping experiment could be attributed to the presence or absence of *P. copri* in the microbial community, we administered *B. infantis* Bg2D9 in both arms of this experiment. Outside of this change, the same cultured strains, the same sequence of their introduction and the same sequence of diet switches were applied (**Extended Data Fig. 5a**). Reproducible colonization of consortium members within each arm was confirmed by quantifying their absolute abundances in cecal samples collected at the time of euthanasia (P53; see **Supplementary Table 11**). As in the previous experiment, animals in the arm containing *P. copri* exhibited significantly greater weight gain between P23 and P53 than those in the no *P. copri* arm [$P<0.05$; linear mixed-effects model (see *Methods*)] (**Extended Data Fig. 5b**).

Mass spectrometric analysis of host metabolism – We used targeted mass spectrometry to quantify levels of 20 amino acids, 19 biogenic amines, and 66 acylcarnitines in the jejunum, colon, gastrocnemius, quadriceps, heart muscle, and liver of the two groups of mice. Additionally, we quantified the 66 acylcarnitines in their plasma. The results are described in **Supplementary Table 12** and **Extended Data Fig. 5c-e**. Consistent with the previous experiment, citrulline, the biomarker for metabolically active enterocyte biomass, was significantly elevated in the jejunums of mice belonging to the w/ *P. copri* group ($P<0.05$; Mann-Whitney U test) (**Extended Data Fig. 5c**; **Supplementary Table 12a**).

We observed significant elevations of acylcarnitines derived from palmitic acid (C16:0), stearic acid (C18:0), oleic acid (C18:1), linoleic acid (C18:2), and linolenic acid (C18:3) in the jejunums of *P. copri*-colonized animals ($P<0.01$; Mann-Whitney U test) (**Extended Data Fig. 5d**); these are the major fatty acids found in soybean oil[34], which is the principal source of lipids in MDCF-2. These acylcarnitine chain lengths were found at higher abundance than all other medium or long-chain acylcarnitine species in our jejunal samples, indicating their role as primary dietary lipid energy sources (**Supplementary Table 12b)**. Elevation of these species suggests increased transport and ß-oxidation of long-chain dietary lipids in the jejunums of the *P. copri*-colonized animals.

Analysis of colonic tissue showed significant elevation of C16:0, C18:1, and C18:2 acylcarnitines in *P. copri*-colonized animals, suggesting that β-oxidation is also elevated in tissue compartments not directly involved in lipid absorption ($P<0.01$; Mann-Whitney U test) (**Extended Data Fig. 5e**; **Supplementary Table 12b**). This finding was matched by a significant elevation in plasma levels of non-esterified fatty acids in w/ *P. copri* animals, suggesting higher circulation of dietary lipids which would support fatty acid β-oxidation in peripheral tissues ($P<0.05$; Mann-Whitney U test) (**Extended Data Fig. 5f**; **Supplementary Table 12c**). We extended our targeted LC-MS to liver, gastrocnemius muscle, quadriceps and heart. The only statistically significant difference in levels of acylcarnitines whose chain length corresponded to components of soybean oil was an increase in 18:2 and 18:3 species in the myocardium of mice with *P. copri* compared to those lacking this organism (**Supplementary Table 12b**). Additionally, jejunal levels of C3 and C4 acylcarnitines as well as colonic levels of C4 and C5 acylcarnitines known to be derived from branched-chain amino acid catabolism, were significantly

elevated in the *P. copri*-colonized animals ($P<0.05$; Mann-Whitney U test; **Extended Data Fig. 5d,e** and **Supplementary Table 12b**).

P. copri colonization and weight gain in another diet context

The focus of these 'reverse translation' experiments in gnotobiotic mice was to establish whether *P. copri* MAGs identified in the human study play a key role in the metabolism of MDCF-2 glycans and host responses. As such, our mouse experiments examined the effects of *P. copri* in the context of the MDCF-2 diet. The experiment shown in **Extended Data Fig. 6** illustrates the challenge posed in testing the effects of *P. copri* colonization in the context of a 'control' diet representative of that typically consumed by 18-month-old Bangladeshi children living in Mirpur ('Mirpur-18 diet')[5]. The design was similar to that employed for the experiments described in **Fig. 1** and **Extended Data Fig. 5** with two exceptions; (i) *B. infantis* Bg2D9 was used in both groups and (ii) on P24 pups from different litters were mixed and randomly assigned to two diet treatment groups, MDCF-2 and Mirpur-18 (n=2 dams and 12 pups per group). The absolute abundances of community members were quantified in cecal contents harvested at the time of euthanasia on P53. While the absolute abundance of *P. copri* Bg131 was not significantly different between the two diet groups ($P>0.05$; Mann-Whitney), there were statistically significant differences between the abundances of 11 of the 19 community members (**Extended Data Fig. 6b,c**). Nonetheless, we proceeded to test whether the increased weight gain phenotype associated with the presence of *P. copri* in the community was evident in the Mirpur-18 diet context. To do so, we repeated the dam-to-pup microbial transmission experiment where all animals were weaned onto the Mirpur-18 diet but where one group had received *P. copri* Bg131 (n=8 animals) and the other had not (n=9). All animals were euthanized on P53. *P. copri* successfully colonized mice and was maintained throughout the experiment at levels comparable to previous experiments (10.4 ± 0.1 $\log_{10}$ genome equivalents per gram cecal contents at P53). Importantly, there was no statistically significant difference in weight gain between the two groups [$P=0.297$; linear mixed-effects model (see *Methods*)]. These findings provide evidence that the effect of *P. copri* on weight gain in this preclinical model is diet dependent.

Functional evaluation of cultured *P. copri* strains more closely resembling MAGs Bg0018 and Bg0019

To further test the causal relationship between *P. copri* MAGs Bg0018/Bg0019, glycan and host metabolism, and weight gain in the context of MDCF-2, we turned to cultured *P. copri* strains whose genomic features were closer to the MAGs than those represented in strain Bg131. *P. copri* Bg131 shared notable but incomplete conservation of PULs and other genomic features represented in MAGs Bg0018 and Bg0019. In our companion study, we characterize five additional fecal *P. copri* strains that we cultured from Bangladeshi children living in Mirpur[1]. Two of these strains (BgD5_2 and BgF_2), had greater genomic similarity to MAGs Bg0018 and Bg0019 than the other isolates, including *P. copri* Bg131, as quantified by phylogenetic distance and PUL content. Nine of the 10 functionally conserved PULs in Bg0018/Bg0019 were present in *P. copri* BgD5_2 and BgF5_2 compared to five in the case of *P. copri* Bg131 (**Supplementary Table 13a,b**; **Fig. 4a**). All 103 genes in these nine conserved PUL homologs shared 100% nucleotide sequence identity between the two isolates. Moreover, of the 198 genes comprising the 22 PULs contained in each isolate's genome, 195 shared 100% nucleotide sequence identity (**Supplementary Table 13c**). *In silico* metabolic reconstructions disclosed that both the BgD5_2 and BgF5_2 strains shared 103/106 (97%) binary phenotypes with MAG Bg0018 and 102/106 (96%) with MAG Bg0019, including 53/55 (96%) and 54/55 (98%) of their carbohydrate utilization pathways with MAGs Bg0018 and Bg0019, respectively (**Supplementary Table 13d**). In addition, results of *in vitro* growth assays conducted in defined medium supplemented with different glycans represented in MDCF-2 disclosed that strain BgF5_2 exhibited stronger preference than Bg131 for glycans enriched in and/or unique to MDCF-2 compared to RUSF (i.e., arabinan, arabinoxylan, galactan, galactomannan)[1].

To directly determine whether pre-weaning colonization with *P. copri* strains resembling MAGs Bg0018 and Bg0019 is sufficient to promote growth and produce the metabolic effects described above, we performed an experiment whose design (**Fig. 4b**) was similar to our previous experiments (**Fig. 1b** and

Extended Data Fig. 5a), but with several modifications to address limitations of the initial colonization and phenotyping experiment. *First*, because of their greater genomic similarity to WLZ-associated MAGs Bg0018 and Bg0019, we used *P. copri* BgD5_2 and BgF5_2 in place of *P. copri* Bg131. *Second*, to control for differences in *B. infantis* strain used in the initial experiment (**Fig. 1**), both arms received *B. infantis* Bg2D9 in this 'third' experiment (as was the case in the second experiment described in **Extended Data Fig. 5**). *Third*, because *P. stercorea* had colonized at a lower abundance than *P. copri* and did not express CAZymes related to MDCF-2 glycans, it was not included in the second gavage mixture, which now only contained *P. copri*. *Fourth*, given that *B. infantis* Bg2D9 promoted pre-weaning colonization of *P. copri* in the initial experiment, we omitted the fourth gavage, previously administered at the end of the weaning period, that had included *P. copri* and *P. stercorea*. As before, the control group of animals were those that did not receive *P. copri* (n=2 dams and 13 pups/treatment group).

Shotgun sequencing of DNA isolated from cecal contents collected at the time of euthanasia (P53) confirmed that animals in the experimental group had been colonized with both *P. copri* isolates as well as all other members of the defined consortia (**Supplementary Table 14**). Even though strain BgD5_2 grew much more poorly than BgF5_2 when cultured in defined medium[1], in animals colonized with both isolates, the BgD5_2 strain was present at higher absolute abundance than the BgF5_2 strain (**Fig. 4c**); their relative abundances were 37.8±4.4% and 15.5±1.0%, respectively, compared with 31±6.6% and 24±8.0% for *P. copri* Bg131 in the experiments described in **Fig. 1b** and **Extended Data Fig. 5a**. Comparing the two groups disclosed that colonization with BgD5_2 and BgF5_2 augmented community biomass without displacing other bacteria (**Fig. 4c; Supplementary Table 14b**).

We observed a significantly greater increase in body weight between P23 and P53 in mice colonized with *P. copri* BgD5_2 and BgF5_2 compared to those without *P. copri* [$P<0.0001$; linear mixed-effects model] (**Fig. 4d)**. The difference in the mean percent increase in postweaning weight between the experimental and control groups (24%) was comparable to that documented in the two previous experiments shown in **Fig. 1b** and **Extended Data Fig. 5a** (25% and 13% respectively). As in these previous experiments, the weight difference was not attributable to differences in cecal size.

Mass spectrometry confirmed that preweaning colonization with *P copri* affected intestinal lipid metabolism and was a major determinant of MDCF-2 glycan degradation. Targeted LC-MS of ileal and colonic tissue revealed a significant elevation of long-chain acylcarnitines corresponding to soybean oil lipids (**Extended Data Fig. 7; Supplementary Table 15**), consistent with changes observed in the experiment described in **Extended Data Fig. 5**. Microbial RNA-Seq of cecal contents revealed that among all PUL genes, those present in the three conserved PULs with predicted arabinan/starch, pectin, and pectic galactan substrates were significantly enriched for higher levels of expression (See *Methods*) (**Fig. 4e; Supplementary Table 16**). UHPLC-QqQ-MS of monosaccharides in glycans present in cecal contents indicated that the presence of *P. copri* BgD5_2 and F5_2 resulted in significantly lower levels of arabinose, consistent with our previous observations using *P. copri* Bg131, as well as galactose (a finding specific to this experiment) (**Fig. 4f; Supplementary Table 17**). Colonization with *P. copri* BgD5_2 and BgF5_2 also significantly lowered levels of all arabinose-containing glycosidic linkages measured, as well as three galactose-containing linkages (**Fig. 4g,h; Supplementary Table 17**). Together, these data indicate that the PUL content of these two isolates is associated with enhanced degradation of MDCF-2 glycans compared to the *P. copri* Bg131-containing microbial community. Targeted UPHLC-QqQ-MS measurements of all 20 amino acids and seven B vitamins also revealed that compared to the control group, colonization with these strains was associated with significantly higher cecal levels of two essential amino acids (tryptophan, lysine), seven non-essential amino acids (glutamate, glutamine, aspartate, asparagine, arginine, proline, glycine) and pantothenic acid (vitamin B5) (**Supplementary Table 18**).

Based on all of these experiments, we concluded that (i) pre-weaning colonization with *P. copri* augments weight gain in the context of the MDCF-2 diet, (ii) the presence of specific strains of this species is a major determinant/effector of MDCF-2 glycan degradation, and (iii) incorporating these strains into the gut community changes metabolism in intestinal cell lineages.

Discussion

Accessing tissue from different regions of the human intestine, as well as from extra-intestinal sites, represents a major challenge when trying to characterize the mechanisms by which microbiome-targeted nutritional interventions impact the operations of microbial community members, and how these changes alter human physiology at a molecular, cellular and systems level. In this study, we illustrate a 'reverse translation' strategy that can be used to address this challenge. Gnotobiotic mice were colonized with defined consortia of age- and WLZ-associated bacterial strains cultured from fecal samples collected from children living in a Bangladeshi community where the prevalence of malnutrition is high. *Prevotella copri* was represented by cultured isolates whose genomic features, including PULs and metabolic pathways involved in carbohydrate utilization, are highly similar to MAGs associated with improved weight gain in the clinical trial. Dam-to-pup transmission of these communities occurred in the context of a sequence of diets that re-enacted those consumed by children enrolled in a clinical study of a microbiome-directed complementary food (MCDF-2). Microbial RNA-Seq and targeted mass spectrometry of glycosidic linkages present in intestinal contents provided evidence that *P. copri* plays a key role in the metabolism of polysaccharides contained in MDCF-2. Consistent with this, *P. copri* associated weight gain in the preclinical model was dependent on consumption of MDCF-2; this phenotype was not observed when a diet commonly consumed by Bangladeshi children was administered instead, despite comparable levels of *P. copri* colonization in the two diet contexts. Single nucleus RNA-Seq and targeted mass spectrometry of the intestine indicated that *P. copri* increases the uptake and metabolism of lipids (including those fatty acids that are most prominently represented in the soybean oil that comprises the principal lipid component of MDCF-2). Additional effects on uptake and metabolism of amino acids (including essential amino acids) and monosaccharides were predicted and in select cases validated by mass spectrometric assays. These effects on nutrient processing and energy metabolism involve proliferating epithelial progenitors in the crypts as well as their descendant lineages distributed along the villus. snRNA-Seq revealed discrete spatial features of these effects, with populations of enterocytes positioned at the base-, mid- and tip regions of villi manifesting distinct patterns of differential expression of a number of metabolic functions.

Inspired by the results of the clinical trial, the goal of our reverse translation experiments was to ascertain the impact of the presence or absence of *P. copri* in a model that emulated postnatal gut microbial community assembly and exposure to MDCF-2. The current study raises several questions that have both mechanistic and therapeutic implications. We were not able to successfully mono-colonize mice with our cultured Bangladeshi *P. copri* strains. Therefore, we could not directly test the effects of these strains *in vivo* on MDCF-2 glycan metabolism, weight gain and/or gut epithelial biology in the absence of other potential microbial interactions. Additional work, involving systematic manipulation of the composition of the bacterial consortia introduced into dams (and subsequently transmitted to their offspring) will be required to ascertain the extent to which *P. copri* has (i) direct effects on intestinal epithelial gene expression and host metabolism versus (ii) effects of other community members that are dependent upon its presence or absence. If *P. copri* has direct effects on the host, it remains to be determined whether the mediators of these effects are the direct products of MDCF-2 glycan metabolism, or the products of other metabolic pathways in *P. copri* whose activities are regulated by biotransformation of these glycans, or other MDCF-2 components. Future studies are also needed to disambiguate the extent to which the observed effects of *P. copri* on gut epithelial carbohydrate, lipid and amino acid metabolism contribute to weight gain. The spatial features of metabolic pathway expression documented by snRNA-seq must be characterized further. This effort will be technically challenging; e.g., it could require (i) documenting the distribution of *P. copri* and other community members along the crypt-villus axis, (ii) advancing methods for spatial transcriptomics[33,35] so that they can be (simultaneously) applied to both epithelial cell lineages and microbial community members, and (iii) using *in situ* mass spectrometry to directly characterize the metabolic profiles of discrete gut cell populations.

The relationship between prominent initial colonization by *B. infantis* Bg2D9 and the capacity of *P. copri* to subsequently colonize also needs further investigation. As noted above, *B. infantis* Bg2D9 contains several genomic loci not represented in most other cultured *B. infantis* strains, which could enhance its ability to utilize a variety of dietary carbohydrates[12]. In principle, these loci could support increased fitness of *B. infantis* Bg2D9 in malnourished children whose consumption of breast milk is low. Bangladeshi infants and young children with severe acute malnutrition (SAM) have markedly lower or absent levels of *B. infantis* compared to their healthy counterparts[12]. The *B. infantis*-*P. copri* interaction documented in this preclinical study provides a rationale for testing the effects of first administering *B. infantis* Bg2D9 to individuals with SAM and subsequently MDCF-2 to restore age-appropriate microbiome configurations and promote healthy growth.

In summary, this and our companion study[1] illustrate one approach for identifying members of a gut microbial community that function as principal metabolizers of dietary components as well as key effectors of host biological responses. The results can provide a starting point for developing microbiome-based diagnostics for stratification of populations of undernourished children who are candidates for treatment with a given MDCF, and for monitoring their treatment responses including in adaptive clinical trial designs. Another potential return on investment for this approach is a knowledge base needed for (i) creating 'next generation' MDCFs composed of (already) identified bioactive components, but from alternative food sources that may be more readily available, affordable and culturally acceptable for populations living in different geographic locales; (ii) making more informed decisions about dosing of an MDCF for undernourished children as a function of their stage of development and state of disease severity and (iii) evolving policies about complementary feeding practices based on insights about how food components impact the fitness and expressed beneficial functions of growth-promoting elements of a child's microbiome. Finally, this approach and resulting knowledge base can also serve as a guide for culture initiatives designed to recover candidate probiotic strains, and/or components of synbiotics for repairing gut microbial communities that cannot be resuscitated with food-based interventions alone.

Methods

Bacterial genome sequencing and annotation

Monocultures of each isolate were grown overnight at 37 ºC in Wilkins-Chalgren Anaerobe Broth (Oxoid Ltd.; catalog number: CM0643) in a Coy Chamber under anaerobic conditions (atmosphere; 75% N_2, 20% CO_2 and 5% H_2) without shaking. Cells were recovered by centrifugation (5000 × g for 10 minutes at 4 ºC). High molecular weight genomic DNA was purified (MagAttract® HMW DNA kit, Qiagen) following the manufacturer's protocol and the amount was quantified (Qubit fluorometer). The sample was passed up and down through a 29-gauge needle 6-8 times and the fragment size distribution was determined (~30 kbp; TapeStation, Agilent).

Fragmented genomic DNA (400-1000 ng) was prepared for long-read sequencing using a SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences) adapted to a deep 96-well plate (Fisher Scientific) format. All DNA handling and transfer steps were performed with wide-bore, genomic DNA pipette tips (ART). Barcoded adapters were ligated to A-tailed fragments (overnight incubation at 20 °C) and damaged or partial SMRTbell templates were subsequently removed (SMRTbell Enzyme Cleanup Kit). High molecular weight templates were purified (volume of added undiluted AMPure beads = 0.45 times the volume of the DNA solution). Libraries prepared from different strains were pooled (3-6 libraries/pool). A second round of size selection was then performed; AMPure beads were diluted to a final concentration of 40% (v/v) with SMRTbell elution buffer with the resulting mixture added at 2.2 times the volume of the pooled libraries. DNA was eluted from the AMPure beads with 12 μL of SMRTbell elution buffer. Pooled libraries were quantified (Qubit), their size distribution was assessed (TapeStation) and sequenced [Sequel System, Sequel Binding Kit 3.0 and Sequencing Primer v4 (Pacific Biosystems)]. The resulting reads were demultiplexed and Q20 circular consensus sequencing (CCS) reads were generated (Cromwell workflow configured in SMRT Link software). Genomes were assembled using Flye[36] (v2.8.1) with HiFi-error set to 0.003, min-overlap set at 2000, and other options

set to default. Genome quality was evaluated using CheckM[37] (v1.1.3) (**Supplementary Table 1a, 4, and 13a**).

We applied Prokka[38] (v1.14) to identify potential open reading frames (ORF) in each assembled genome. Additional functional annotation of these ORFs using a 'subsystems' approach adapted from the SEED genome annotation platform[39] was performed as described in the accompanying paper[1]. We assigned functions to 9,820 ORFs in 20 isolate genomes using a collection of mcSEED metabolic subsystems that capture the core metabolism of 98 nutrients/metabolites in four major categories (amino acids, vitamins, carbohydrates, and fermentation products) projected over 2,856 annotated human gut bacterial genomes[40-42]. *In silico* reconstructions of selected mcSEED metabolic pathways were based on functional gene annotation and prediction using homology-based methods and genome context analysis. Reconstructions were represented as a binary phenotype matrix (BPM) where, as noted in the main text, assignment of a "1" for amino acids and B vitamins denotes a predicted prototroph and a "0" denotes an auxotroph. For carbohydrates, "1" and "0" refer to a strain's predicted ability or inability, respectively, to utilize the indicated mono-, di- or oligosaccharide while for fermentation end products, a "1" and "0" indicate a strain's predicted ability/inability to produce the indicated compound, respectively (see **Supplementary Table 1c, 13d**).

To calculate phylogenetic relationships between five *P. copri* isolates and MAGs Bg0018 and Bg0019, we first used CheckM[37] (v1.1.3) to extract and align the amino acid sequences of 43 single copy marker genes in each isolate or each of the two MAGs, plus an isolate genome sequence of *Bacteroides thetaiotaomicron* VPI-5482 (PATRIC accession number: 226186.12). Concatenated marker gene sequences were analyzed using fasttree[43] (v2.1.10) to construct a phylogenetic tree using the Jones-Taylor-Thornton model and CAT evolution rate approximation, followed by tree rescaling using the Gamma20 optimization. The tree was subsequently processed in R using ape[44] (v5.6-2) to root the tree with the *B. thetaiotaomicron* genome and extract phylogenetic distances between genomes, followed by ggtree[45] (v3.2.1) for tree plotting.

The similarity between the genomes of these strains and MAGs was quantified by calculating the ANI score with pyani[46] (ANIm implementation of ANI, v0.2.10). We first calculated ANIm scores for all possible combinations between MAGs and the genomes of cultured bacterial strains and subsequently removed any MAG-strain genome combination with <10% alignment coverage[11]. For the remaining MAGs, a "highly similar" genome in the collection of cultured bacterial strains was defined as having > 94% ANIm score (**Supplementary Table 1b**). We then determined the degree of binary phenotype concordance between each genome in the collection of cultured bacterial strains and its "highly similar" MAG. A binary phenotype concordance score was calculated by dividing the number of binary phenotypes[1] shared between a cultured strain's genome and a MAG by the total number of binary phenotypes annotated in the strain and MAG. A 'Representative MAG' for each genome was defined as having a binary phenotype concordance score >90% (**Supplementary Table 1c**).

PULs were predicted based on the method described in ref. 47 and displayed with the PULDB interface[48]. PULs were placed into three categories: (i) 'functionally conserved' (PULs containing shared ORFs encoding the same CAZymes and SusC/SusD proteins in the same organization in their respective genomes with ≥90% amino acid identity between proteins); (ii) 'structurally distinct' (PULs present in respective genomes but where one or more CAZymes or one or both SusC/SusD proteins are missing or fragmented in a way likely to impact function, or where extra PUL elements are present), and (iii) 'not conserved' (PULs present in respective genomes but with mutations likely to completely compromise function, or no PUL identified).

Colonization and husbandry

Gnotobiotic mouse experiments were performed using protocols approved by the Washington University Animal Studies Committee. Germ-free C57BL/6J mice were maintained in plastic flexible film isolators (Class Biologically Clean Ltd) at 23 ºC under a strict 12-hour light cycle (lights on a 0600h). Autoclaved paper 'shepherd shacks' were kept in each cage to facilitate natural nesting behaviors and provide environmental enrichment.

A weaning diet containing MDCF-2 was formulated as described in the main text. Ingredients represented in the different diet modules were combined and the mixture was dried, pelleted, and sterilized by gamma irradiation (30-50 Kgy). Sterility was confirmed by culturing the pellets in LYBHI medium[49] and in Wilkins-Chalgren Anaerobe Broth under aerobic and anaerobic conditions for 7 days at 37 ºC followed by plating on LYBHI- and blood-agar plates. Nutritional analysis of each irradiated diet was performed by Nestlé Purina Analytical Laboratories (St. Louis, MO) (**Supplementary Table 2c**).

Pregnant C57BL/6J mice originating from trio matings were given *ad libitum* access to an autoclaved breeder chow (Purina Mills; Lab Diet 5021) throughout their pregnancy and until postpartum day 2. Key points about the experimental design of the gnotobiotic mouse experiments described in **Fig. 1b, Extended Data Fig. 5a, Extended Data Fig. 6a,** and **Fig. 4a** are; (i) all bacterial strains were cultured in Wilkins-Chalgren Anaerobe Broth (except for *F. prausnitzii* which was cultured in LYBHI medium) and were harvested after overnight growth at 37 ºC (**Supplementary Table 1a**), (ii) all gavage mixtures contained equivalent amounts (by OD_{600}) of their constituent bacterial strains except for *F. prausnitzii* which was concentrated 100-fold before preparing the gavage mixture, (iii) each bacterial consortium was administered to the postpartum dams in a volume of 200 µL using an oral gavage needle (Cadence Science; catalog number: 7901), (iv) the number of dams and pups per treatment group (two dams and 7-8 pups/treatment group for the experiment described in **Fig. 1b**; four dams and 18-19 pups/treatment group for the experiment outlined in **Extended Data Fig. 5a;** two dams and 13 pups/treatment group for the experiment illustrated in **Fig. 4a;** two dams and 12 pups/treatment group for the experiment shown in **Extended Data Fig. 6a**), (v) half of the bedding was replaced with fresh bedding in each cage each day from postpartum day 1 to 14, after which time bedding was changed every 7 days, (vi) diets were provided to mothers as well as to their weaning and post-weaning pups *ad libitum*, and (vii) biospecimens collected from mice when they were euthanized (without prior fasting) were snap frozen in liquid nitrogen and stored at -80 °C before use.

Pups were weighed on P23, P35, and P53, and normalized to the weight on P23. A linear mixed-effects model was used to evaluate the effect of different microbial communities on normalized mouse weight gain:

$$normalized\ weight \sim \beta_1(Arm) + \beta_2(postnatal\ day) + (1 \mid mouse)$$

Defining the absolute abundances of bacterial strains in ileal, cecal and fecal communities

The absolute abundances of bacterial strains were determined in the microbiota using previously described methods with minor modifications[50-52]. In brief, 3.3×10^6 cells of *Alicyclobacillus acidiphilus* DSM 14558 and 1.49×10^7 cells of *Agrobacterium radiobacter* DSM 30147 (ref. 50) were added to each weighed frozen sample prior to DNA isolation and preparation of barcoded libraries for shotgun sequencing. Sequencing was performed using an Illumina NextSeq instrument. Bacterial abundances were determined by assigning reads to each bacterial genome, followed by a normalization for genome uniqueness in the context of a given community[51]. The resulting count table was imported into R[52] (v4.0.4). We calculated the absolute abundance of a given strain *i* in sample *j* in reference to the spike-in *A. acidiphilus* (*Aa*) and *A. radiobacter* (*Ar*) genomes using the following equation:

$$strain_{i,j} = \left(\frac{counts_{i,j} \times Aa\ cells\ added_j}{Aa\ counts_j \times sample\ weight_j} + \frac{counts_{i,j} \times Ar\ cells\ added_j}{Ar\ counts_j \times sample\ weight_j} \right) \times 0.5$$

The statistical significance of observed differences in the abundance of a given strain between experimental groups in ileal, cecal, and fecal samples was determined by using the Kruskall-Wallis test followed by Dunn's test for each pairwise comparison among the three arms in the experiment described in **Fig. 1** or the Mann-Whitney U test between the two arms in experiments described in **Extended Data Fig. 1b**, **Extended Data Fig. 5**, and **Fig. 4**. For the experiment described in **Fig. 1**, 96 fecal samples were sequenced [$2.2 \times 10^6 \pm 1.2 \times 10^5$ unidirectional 75 nt reads/sample (mean ± SD)], along with 20 cecal samples

[$1.5x10^6 \pm 4.2x10^5$ unidirectional 75 nt reads/sample] and 20 ileal samples [$1.5x10^6 \pm 9.1x10^4$ unidirectional 75 nt reads/sample] (**Supplementary Table 3a**). For the experiment described in **Extended Data Fig. 1b**, 37 fecal samples were sequenced [$5.8x10^6 \pm 1.6x10^6$ unidirectional 75 nt reads/sample] (**Supplementary Table 5a)**], while for the experiment described in **Extended Data Fig. 5a**, 37 cecal samples were sequenced [$1.3x10^6 \pm 1.3x10^5$ unidirectional 75 nt reads/sample (**Supplementary Table 11a**)]. For the experiment described in **Fig. 4b**, 26 cecal samples were sequenced [$2.9x10^6 \pm 5.6x10^6$ 150nt paired-end reads/samples (**Supplementary Table 14a**)].

The statistical significance of observed differences in the abundance of a given strain across different treatment groups and time was tested using a linear mixed effects model within the R packages lme4[53] (v1.1-27) and lmerTest[54] (v3.1-3). The change in *P. copri* absolute abundance in fecal samples during the course of the experiment was determined by a linear mixed-effects model:

$$P.copri\ absolute\ abundance \sim \beta_1(Arm) + \beta_2(postnatal\ day) + (1 \mid mouse)$$

Microbial RNA-Seq

RNA was isolated[5] from cecal contents collected at the end of the experiment. cDNA libraries were generated from isolated RNA samples using the 'Total RNA Prep with Ribo-Zero Plus' kit (Illumina). Barcoded libraries were sequenced [Illumina NovaSeq instrument]. For the experiment described in **Fig. 1b**, DNA recovered from 15 different samples of cecal contents were each sequenced to a depth of $7.9x10^7 \pm 1.1x10^7$ bidirectional 150 nt reads (mean±SD) (**Supplementary Table 8a**). For the experiment shown in **Fig. 4b**, DNA from 26 different samples of cecal contents were each sequenced to a depth of $6.5x10^7 \pm 2.1x10^7$ bidirectional 150 nt reads (**Supplementary Table 16a**). Raw reads were trimmed by using TrimGalore[55] (v0.6.4). Trimmed reads longer than 100 bp were mapped to reference genomes with kallisto[56] (v0.43.0).

For the analysis of metabolic pathway expression described in **Fig. 2d-f**, pseudocounts for each transcript in a given cecal sample were first normalized by the absolute abundance of the corresponding strain (**Supplementary Table 3d**) in order to minimize the confounding effects of differences in strain abundance[57]. The rlog transformation in the R DESeq2 package (v1.34.0) was then applied to this dataset for variance stabilization[58]. To obtain relative expression of metabolic pathways, only transcripts corresponding to complete mcSEED metabolic pathways (*i.e.*, pathways with binary phenotype scores of 1; see above) were retained and transformed expression values were averaged across genes within a given pathway in each strain. The resulting aggregated pathway expression dataset was then centered prior to singular value decomposition with NumPy[59] (v1.21.5). Principal component analysis of samples was calculated by projecting the centered pathway dataset onto right singular vectors. Sample projections along the first two principal components were converted to a Euclidian distance matrix, and PERMANOVA was used to test for the significance of separation of samples by experimental group[60]. Pathway PCA loadings were calculated by projecting the transpose of the centered dataset onto left singular vectors. Metabolic pathways were subsequently ranked within each strain by their corresponding loadings, and directionality was resolved such that more negative left singular vector 1 projections corresponded to higher pathway expression in the *P. copri* colonization context (**Fig. 2f**). The PC1 contributions for each pathway were represented by a dot plot, while the box plots were generated using the boxplot function of the seaborn[61] (v0.12.1) with default settings such that the box represents the median and quartiles and the whiskers represent the range excluding outliers falling more than 1.5x the interquartile range from the box edges.

For differential expression analysis of microbial transcripts described in **Fig. 2g**, the filtered version of kallisto pseudocounts was imported into edgeR[18]. For each community member, 'within-taxon sum-scaling' was applied by calculating the trimmed mean of M-value library size corrections based on the total pool of RNA reads from that member. This organism-scaled transcript set was then used for dispersion estimation and fitting of a generalized linear model (GLM). In addition to 'experiment arm', we included the absolute abundance of the organism in each cecal sample as a covariate in the GLM to

reduce false discoveries due to differences in the abundances of community members. A likelihood ratio test (edgeR[18]) was then used to detect differential expression between samples obtained from members of the w/ *P. copri* versus w/o *P. copri* treatment groups. Transcripts with statistically significant differences in their expression were identified [q-value < 0.05 (adjusted *P* value < 0.05)] after multiple hypothesis correction was applied to the entire set of transcripts from a given organism via the Benjamini-Hochberg method.

To analyze PUL expression discussed in the experiment described in **Fig. 4e** involving *P. copri* strains BgD5_2 and BgF5_2, kallisto TPMs were generated by mapping reads to the entire collection of bacteria present in the cecums of *P. copri*-colonized animals. Due to the high fraction of genes with 100% nucleotide identity between *P. copri* BgD5_2 and BgF5_2, the kallisto index was generated to represent the set of unique genes between the two isolates by including the entire *P. copri* BgF5_2 gene set and only the subset of *P. copri* BgD5_2 genes that did not share identical sequences in *P. copri* BgF5_2 (61 out of 3,066 genes). TPMs were filtered down to the set of unique genes in the *P. copri* BgD5_2 and BgF5_2 PULs [n=201 unique genes, consisting of 195 genes with identical nucleotide sequences between the two isolates and 3 unique genes from each isolate (**Supplementary Table 13c**)]. Log_2 TPM with a pseudocount of 1 were visualized for all PUL genes using the seaborn[61] (v0.12.1) Python package, splitting the set to show homologs of specific consensus PULs in color on the right side of each violin plot in **Fig. 4e** against the remainder of the PUL genes shown in gray on the left side of each violin plot. Adjusted GSEA *P*-values were calculated with fgsea[62] (v1.20.0), ranking genes by their mean log_2 TPM across the *P. copri* colonized samples, with each PUL comprising a gene set against the background of all PUL genes with minimum and maximum gene set sizes of 5 and 50 genes, respectively (**Supplementary Table 16b**).

Histomorphometric analysis of villus height and crypt depth

Jejunal and ileal segments were fixed in formalin, embedded vertically in paraffin, 5 µm-thick sections were prepared and the sections were stained with hematoxylin and eosin. Slides were scanned (NanoZoomer instrument, Hamamatsu). For each animal, 10 well-oriented crypt-villus units were selected from each intestinal segment for measurement of villus height and crypt depth using QuPath[63] (v0.3.2). Measurements were performed with the investigator blinded with respect to colonization group. A two-tailed Mann-Whitney U test was applied to the resulting datasets.

Single nucleus (sn) RNA-seq

Jejunal segments of 1.5 cm in length were collected from mice and snap frozen in liquid nitrogen [n=4 animals/treatment group (2 males and 2 females); 2 treatment groups in total]. The method for extracting nuclei was adapted from a previously described protocol for the pancreas[64]. Briefly, tissues were thawed and minced in lysis buffer [25mM citric acid, 0.25M sucrose, 0.1% NP-40, and 1× protease inhibitor (Roche)]. Nuclei were released from cells using a pestle douncer (Wheaton), washed 3 times with buffer [25mM citric acid, 0.25M sucrose, and 1× protease inhibitor], and filtered successively through 100µm, 70µm, 40µm, 20µm and finally 5µm diameter strainers (pluriSelect) to obtain single nuclei in resuspension buffer [25mM KCl, 3mM $MgCl_2$, 50mM Tris, 1mM DTT, 0.4U/µL rNase inhibitor (Sigma) and 0.4U/µL Superase inhibitor (ThermoFisher)]. Approximately 10,000 nuclei per sample were subjected to gel bead-in-emulsion (GEM) generation, reverse transcription and construction of libraries for sequencing according to the protocol provided in the 3' gene expression v3.1 kit manual (10X Genomics). Libraries were balanced, pooled and sequenced [Illumina NovaSeq S4; $3.23\times10^8 \pm 1.39\times10^7$ paired-end 150 nt reads/nucleus (mean±SD) from jejunal samples]. Read alignment, feature-barcode matrices and quality controls were processed by using the 10X Genomics CellRanger[65] 5.0 pipeline with the flag '-include-introns' to ensure that reads would be allowed to map to intronic regions of the mouse reference genome (GRCm38/mm10). Nuclei with over 2.5% reads from mitochondria-encoded genes or ribosomal protein genes were filtered out.

Analysis of snRNA-seq datasets-- Sample integration, count normalization, cell clustering and marker gene identification was performed using Seurat 4.0[21]. Briefly, filtered feature-barcode matrices outputted from CellRanger were imported as a Seurat object using *CreateSeuratObject* (min.cells = 5, min.features = 200). Each sample was normalized using *SCTransform*[66,67] and integrated using *SelectIntegrationFeatures*, *PrepSCTIntegration*, *FindIntegrationAnchors*, and *IntegrateData* from the Seurat software package. The integrated dataset, incorporating nuclei from all samples, was subjected to unsupervised clustering using *FindNeighbors* (dimensions = 1:30) and *FindClusters* (resolution = 1) from the Seurat package, which executes a shared nearest-neighbor graph clustering algorithm to identify putative cell clusters. Cell type assignment was performed manually based on expression of reported markers[33,68,69].

Cross-condition differential gene expression analysis was performed based on a "pseudobulk" strategy: for each cell cluster, gene counts were aggregated to obtain sample-level counts; each pseudo-bulked sample served as an input for edgeR-based differential gene expression analysis[18,19].

For *NicheNet*-based analysis[20] (v1.1.0), all clusters in our snRNA-seq dataset were used as senders for crypt stem cells, proliferating TA/stem cells, villus base enterocytes, mid-villus enterocytes and villus tip enterocytes, plus goblet cells. We used the *nichenet_seuratobj_aggregate* (assay_oi = "RNA") function with its default settings to incorporate differential gene expression information from Seurat into our *NicheNet* analysis and to select bona fide ligand-receptor interactions.

Compass-based *in silico* metabolic flux analysis[23] (v0.9.10.2) was performed using transcripts from each of six epithelial cell clusters (crypt stem cells, proliferating TA cells, villus-base, mid-villus and villus tip enterocytes and goblet cells). The reaction scores calculated by *Compass* were filtered based on (i) the confidence levels of the Recon2 reactions and (ii) the completeness of information for Recon2 reaction annotations. Only Recon2 reactions that are supported by biochemical evidence (defined by Recon2 as having a confidence level of 4; ref. 24) and that have complete enzymatic information for the reaction were advanced to the follow-on analysis (yield: 2,075 pass filter reactions in 83 Recon2 subsystems).

We subsequently calculated a "metabolic flux difference" to determine whether the presence or absence of *P. copri* affected *Compass*-based predictions of metabolic activities at the Recon2 reaction level in the six cell clusters. The "net reaction score" was calculated as follows

$$c = c_f - c_r$$

where c_f denotes the *Compass* score for a given reaction in the "forward" direction, and, if the biochemical reaction is reversible, c_r denotes the score for the "reverse" reaction.

A Wilcoxon Rank Sum test was used to test significance of the net reaction score between the two treatment groups. *P* values from the Wilcoxon Rank Sum tests were adjusted for multiple comparisons with the Benjamini-Hochberg method.

Cohen's d can be used to show the effect size of c_f or c_r for each reaction between two groups[23] (in our case mice harboring communities with and without *P. copri*). Briefly, Cohen's d of two groups, j and k, was calculated based on the following two equations[70] where n, s, and a represent the number, the variance, and the mean of the observations (in our case, the net reaction scores).

$$s_{pool} = \sqrt{\frac{(n_j - 1)s_j^2 - (n_k - 1)s_k^2}{n_j + n_k - 2}}$$

$$d = \frac{a_j - a_k}{s_{pool}}$$

If both a_j and a_k are non-negative numbers, a positive Cohen's d indicates that the mean of group j is greater than that of group k, whereas a negative Cohen's d means the mean of group j is smaller in that comparison. The magnitude of Cohen's d represents the effect size and is correlated with the difference between the means of the two groups. Because the mean of the net subsystem scores as well as the net reaction scores could be negative, we made the following adjustments to Cohen's d in order to preserve the concordance of sign and the order of group means. The adjusted Cohen's d represents the metabolic flux difference m, and is defined as:

$$\begin{cases} a_j > 0; a_k < 0; |a_j| < |a_k|: m = -d \\ a_j < 0; a_k > 0; |a_j| > |a_k|: m = -d \\ a_j < 0;\ a_k < 0: m = \dfrac{|a_j| - |a_k|}{s_{pool}} \\ \text{other}: m = d \end{cases}$$

scCODA[26] (v0.1.8) is a Bayesian probabilistic model for detecting 'statistically credible differences' in the proportional representation of cell clusters, identified from snRNA-seq datasets, between different treatment conditions[26]. This method accounts for two main challenges when analyzing snRNA-seq data: (i) low sample number and (ii) the compositionality of the dataset (an increase in the proportional representation of a specific cell cluster will inevitably lead to decreases in the proportional representation of all other cell clusters). Therefore, applying univariate statistical tests, such as a t-test, without accounting for this inherent negative correlation bias will result in reported false positives. *scCODA* uses a Bayesian generalized linear multivariate regression model to describe the 'effect' of treatment groups on the proportional representation of each cell cluster; Hamiltonian Monte Carlo sampling is employed to calculate the posterior inclusion probability of including the effect of treatment in the model. The type I error (false discovery) is derived from the posterior inclusion probability for each effect. The set of "statistically credible effects" is the largest set of effects that can be chosen without exceeding a user-defined false discovery threshold α (α=0.05 by default). We applied *scCODA* using default parameters, including choice of prior probability in the Bayesian model and the setting for Hamiltonian Monte Carlo sampling. The enteroendocrine cell cluster was used as the reference cluster in accordance with recommendations from the creators of *scCODA* to choose a cell cluster that has consistent proportional representation across samples.

Mass spectrometry

UHPLC-QqQ-MS of cecal glycosidic linkages and GC-MS of short-chain fatty acids – Ultra-high performance liquid chromatography-triple quadrupole mass spectrometric (UHPLC-QqQ-MS) quantification of glycosidic linkages and monosaccharides present in cecal glycans was performed using methods described in the accompanying study[1]. Levels of short-chain fatty acid levels in cecal contents were measured by GC-MS using a procedure outlined in ref. 5.

LC-MS of acylcarnitines, amino acids, and biogenic amines in host tissues – Acylcarnitines were measured in jejunum, colon, liver, gastrocnemius, quadriceps, and heart muscle, plus plasma according to ref. 71, while 20 amino acids plus 19 biogenic amines were quantified in jejunum, liver, and muscle according to methods detailed in ref. 72. Plasma levels of non-esterified fatty acids were measured using a UniCel DxC600 clinical analyzer (Beckman Coulter).

Targeted mass spectrometry of cecal amino acids and B vitamins – Methods for targeted LC-QqQ-MS of amino acids and B vitamins were adapted from a previous publication[73]. Cecal samples were extracted with ice-cold methanol, and a 200µL aliquot was dried (vacuum centrifugation; LabConco CentriVap) and reconstituted with 200µL of a solution containing 80% methanol in water. A 2µL aliquot of extracted metabolites was then injected into an Agilent 1290 Infinity II UHPLC system coupled with an Agilent 6470 QqQ-MS operated in positive ion dynamic multiple reaction monitor mode (dMRM). The native metabolites were separated on HILIC column (ACQUITY BEH Amide, 2.1 x 150 mm, 1.7 µm

particle size, Waters) using a 20 minute binary gradient with constant flow rate of 0.4 mL/minute. The mobile phases were composed of 10mM ammonium formate buffer in water with 0.125% formic acid (Phase A) and 10mM ammonium formate in 95% acetonitrile/H_2O (v/v) with 0.125% formic acid (Phase B). The binary gradient was as follows: 0-8 minutes, 91-90% B; 8-14 minutes, 90-70% B; 15-15.1 minutes, 70-91% B; 15.1-20 minutes, 91% B. A pool of 20 amino acids and 7 B vitamins standards with known concentrations (amino acid pool: 0.1ng/mL-100µg/mL; B vitamin pool: 0.01ng/mL-10µg/mL) was injected along with the samples as an external calibration curve for absolute quantification.

References

1. Hibberd, M. C. et al. Bioactive glycans in a microbiome-directed food for malnourished children. *medRxiv* doi:10.1101/2023.08.14.23293998 (2023).

2. Smith, M. I. et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* **339**, 548-554 (2013).

3. Kau, A. L. et al. Functional characterization of IgA-targeted bacterial taxa from undernourished Malawian children that produce diet-dependent enteropathy. *Sci. Transl. Med.* **7**, 276ra24 (2015).

4. Blanton, L. V. et al. Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children, *Science* **351**, aad3311 (2016).

5. Gehrig, J. L. et al. Effects of microbiota-directed foods in gnotobiotic animals and undernourished children, *Science* **365**, eaau4732 (2019).

6. Chen, R. Y. et al. A microbiota-directed food intervention for undernourished children. *N. Engl. J. Med.* **384**, 1517-1528 (2021).

7. Sender, R., Milo, R. The distribution of cellular turnover in the human body. *Nat. Med.* **27**, 45–48 (2021).

8. Subramanian, S. et al. Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417-421 (2014).

9. Raman, A. S. et al. A sparse covarying unit that describes healthy and impaired human gut microbiota development. *Science* **365**, eaau4735 (2019).

10. Richter, M. & Rosselló-Móra, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19126-19131 (2009).

11. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864-2868 (2017).

12. Barratt, M. J. et al. *Bifidobacterium longum* subsp. *infantis* strains for treating severe acute malnutrition in Bangladeshi infants. *Science Trans. Med.*, **14**, eabk1107 (2022).

13. Prentice A. M. et al. Energy requirements of pregnant and lactating women. *European J Clin. Nutr.* **50**, S82-S110 (1996).

14. World Health Organization Department of Nutrition for Health and Development. Complementary feeding of young children in developing countries: a review of current scientific knowledge. (1998).

15. Dewey, K. G. & Brown, K. H. Update on technical issues concerning complementary feeding of young children in developing countries and implications for intervention programs. *Food Nutr. Bull.* **24**, 5-28 (2003).

16. Islam, M. M. et al. Risk factors of stunting among children living in an urban slum of Bangladesh: findings of a prospective cohort study. *BMC Public Health* **18**, 197 (2018).

17. Karasov, W. H. & Douglas, A. E. Comparative digestive physiology. *Comprehensive Physiology*, **3**, 741 (2013).

18. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

19. Squair, J. W. et al. Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).

20. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159-162 (2020).

21. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 (2021).

22. Zheng, Y., Song, Y., Han, Q., Liu, W., Xu, J. & Yu., Z. et al. Intestinal epithelial cell-specific IGF1 promotes the expansion of intestinal stem cells during epithelial regeneration and functions on the intestinal immune homeostasis. *Am. J. Physiol. Endocrinol. Metab*. **315**, E638-E649 (2018).

23. Wagner, A. et al. Metabolic modeling of single Th17 cells reveals regulators of autoimmunity. *Cell* **184**, 4168-4185 (2021).

24. Thiele I. et al. A community-driven global reconstruction of human metabolism. *Nat. Biotechnol*. **31**, 419-425 (2013).

25. Mihaylova, M. M. Fasting activates fatty acid oxidation to enhance intestinal stem cell function during homeostasis and aging. *Cell Stem Cell* **22**, 769-778 (2018).

26. Büttner, M., Ostner, J., Müller, C. L., Theis, F. J. & Schubert, B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nat. Commun.* **12**, 6876 (2021).

27. Crenn, P., Messing, B. & Cynober, L. Citrulline as a biomarker of intestinal failure due to enterocyte mass reduction. *Clin. Nutr.* **27**, 328-339 (2008).

28. Guihot, G. et al. Effect of an elemental vs a complex diet on L-citrulline production from L-arginine in rat isolated enterocytes. *J. Parenter. Enteral. Nutr*. **21**, 316-323 (1997).

29. Blachier, F. et al., Arginine metabolism in rat enterocytes. *Biochim. Biophys. Acta* **1092**, 304-310 (1991).

30. Lanyero, B. Correlates of gut function in children hospitalized for severe acute malnutrition, a cross-sectional study in Uganda. *J. Pediatr. Gastroenterol. Nutr.* **69**, 292-298 (2019).

31. Semba, R. D. et al. Child stunting is associated with low circulating essential amino acids. *eBioMedicine*. **6**, 246-252 (2016).

32. Guerrant, R. L. et al. Biomarkers of environmental enteropathy, inflammation, stunting, and impaired growth in children in northeast Brazil. *PLoS One* **11**, e0158772 (2016).

33. Moor, A. E. et al. Spatial reconstruction of single enterocytes uncovers broad zonation along the intestinal villus axis. *Cell* **175**, 1156-1167 (2018).

34. Clemente, T. E. & Cahoon, E. B. Soybean oil: genetic approaches for modification of functionality and total content. *Plant Physiol.* **151**, 1030-1040 (2009).

35. Fawkner-Corbett, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810-826 (2021).

36. Kolmgorov, M., Yuan, J. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol*. **37**, 540-546 (2019).

37. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes *Genome Res*. **25**, 1043-1055 (2015).

38. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).

39. Aziz, R. K. et al. SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *pLoS One*. **7**, e48053 (2012).

40. Rodionov, D. A. et al. Micronutrient requirements and sharing capabilities of the human gut microbiome. *Front. Microbiol*. **10**:1316 (2019).

41. Frolova, M. S., Suvorova, I. A., Iablokov, S. N., Petrov, S. N. & Rodionov, D. A. Genomic reconstruction of short-chain fatty acid production by the human gut microbiota. *Front. Mol. Biosci.* **9**:949563 (2022).

42. Ashniev, G. A., Petrov, S. N., Iablokov, S. N. & Rodionov, D. A. Genomics-based reconstruction and predictive profiling of amino acid biosynthesis in the human gut microbiome. *Microorganisms* **10**, 740 (2022).

43. Price, M. N. et al. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *pLoS One* **5**, e9490 (2010).

44. Paradis, E. and Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2018).

45. Yu, G. Using ggtree to visualize data on tree-like structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).

46. Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G. & Toth, I. K. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods* **8**, 12-24 (2016).

47. Terrapon, N., Lombard, V., Gilbert, H. J. & Henrissat, B. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics* **31**, 647-655 (2015).

48. Terrapon, N. et al. PULDB: the expanded database of Polysaccharide Utilization Loci. *Nucleic Acids Res*. ***46***, D677-D683 (2018).

49. Sokol, H., Pigneur, B., Watterlot, L. & Langella, P. *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16731-16736 (2008)

50. Stammler, F. et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **21**, 28 (2016).

51. McNulty, N. P. et al. Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus* WH2, a symbiont with an extensive glycobiome. *pLoS Biol.* **11**, e1001637 (2013).

52. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2020).

53. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J Stat. Softw.* **67**, 1-48 https://doi.org/10.18637/jss.v067.i01 (2015).

54. Kuznetsova, A., Brockhoff, P. B. & Bojesen-Christensen, R. H. lmerTest: Tests in linear mixed effects models. *J. Stat. Soft.* **82**, 1-12. https://doi.org/10.18637/jss.v082.i13 (2017).

55. Krueger F, Trimgalore. Online deposit. https://doi.org/10.5281/zenodo.5127899 (2021).

56. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

57. Zhang, Y. et al. Statistical approaches for differential expression analysis in metatranscriptomics. *Bioinformatics* **37**, i34-i41 (2021).

58. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

59. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357-362 (2020).

60. Anderson, M. J. Permutational Multivariate Analysis of Variance (PERMANOVA). https://doi.org/10.1002/9781118445112.stat07841 (2017)

61. Waskom, M. L. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021. https://doi.org/10.21105/joss.03021 (2021).

62. Korotkevich, G. et al., Fast gene set enrichment analysis. *bioRxiv.* 060012 (2021).

63. Bankhead, P. et al. QuPath: Open source software for digital pathology image analysis. *Scientific Reports* **7**, 16878 (2017).

64. Tosti, L. et al. Single-nucleus and *in situ* RNA–sequencing reveal cell topographies in the human pancreas. *Gastroenterology* **160**, 1330-1344 (2021).

65. Zheng, G., Terry, J., Belgrader, P. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun.* **8**, 14049 (2017).

66. Hafemeister, C., Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

67. Choudhary, S., Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol.* **23**, 20 (2022).

68. Capdevila, C., Calderon, R. I., Bush, E. C., Sheldon-Collins, K., Sims, P. A. & Yan, K. S. Single-cell transcriptional profiling of the intestinal epithelium. *Methods Mol. Biol*. **2171**, 129-153 (2020).

69. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium, *Nature* **551**, 333-339 (2017).

70. Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.)

71. An, J. et al. Hepatic expression of malonyl-CoA decarboxylase reverses muscle, liver and whole-animal insulin resistance. *Nature Med*. **10**, 268-274 (2004).

72. Gray, N. et al. High-speed quantitative UPLC-MS analysis of multiple amines in human plasma and serum via precolumn derivatization with 6-aminoquinolyl-N-hydroxysuccinimidyl carbamate: application to acetaminophen-induced liver failure. *Anal. Chem*. **89**, 2478-2487 (2017).

73. Konieczna, L. et al. Bioanalysis of underivatized amino acids in non-invasive exhaled breath condensate samples using liquid chromatography coupled with tandem mass spectrometry. *J. Chromatogr. A*. **1542**, 72–81 (2018).

Acknowledgments: We thank David O'Donnell and Maria Karlsson for their invaluable assistance with mouse husbandry, Martin Meier for his essential role in generating shotgun sequencing and microbial RNA-Seq datasets, Janaki Guruge for her help with culturing and maintaining bacterial strains, plus Su Deng and Justin Serugo for oversight of biospecimen archives. This work was supported by grants from the NIH (DK30292). the Centene-Washington University Personalized Medicine Initiative and the Bill & Melinda Gates Foundation. Y.W. is the recipient of a career development award from the NIH (K01 DK134840).

Author Contributions: Y.W., H.-W.C., E.M.L., C.Z. and J.I.G. designed the gnotobiotic mouse experiments, which were performed by Y.W., H.-W.C., E.M.L. and C.Z.. Mouse diets were formulated by H.-W.C. with assistance from M.J.B. and were based on the results of human diet surveys provided by I.M, S.D., M.M. and T.A.. Bacterial genomes were sequenced and assembled by D.M.W., H.-W.C. and M.C. with S.H., D.A.R., A.A.A., N.T., B.H. and A.L.O. contributing to genome annotation. Shotgun sequencing and microbial RNA-Seq datasets were generated by H.-W.C., E.M.L. and Y.W., and analyzed together with M.C.H., H.M.L., and D.M.W.. Mass spectrometry-based metabolomic studies were conducted by J.C., O.I., M.J.M. and C.B.N.. Glycan analyses were performed by J.J.C., G.C., Y.C., N.P.B. and C.B.L.. Morphometric analysis was conducted by Y.W.. snRNA-Seq datasets were generated and analyzed by Y.W., H.-W.C., K.M.P., R.Y.C., C.K. and M.C.. Y.W., H.-W.C., E.M.L. and J.I.G. wrote this manuscript with invaluable assistance from co-authors.

Competing interests: A.O. and D.R. are co-founders of Phenobiome Inc., a company pursuing development of computational tools for predictive phenotype profiling of microbial communities. C.B.L is a co-founder of Infinant Health, interVenn Bio, and BCD Bioscience, companies involved in the characterization of glycans and developing carbohydrate applications for human health.

Materials & Correspondence

Correspondence and material requests should be sent to Jeffrey Gordon (jgordon@wustl.edu)

Figures

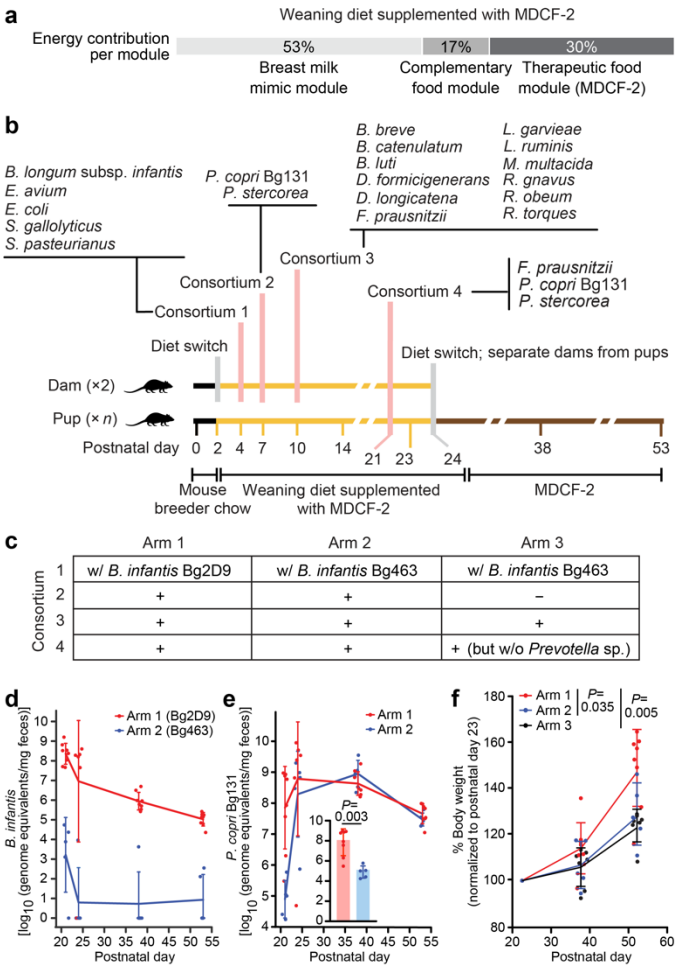


Fig. 1. Identifying factors that affect the efficiency of colonization of gnotobiotic dam-pup dyads with *P. copri* in the presence of other cultured age-discriminatory and WLZ-associated bacterial strains and the effects of colonization on pup weight gain. (a) Energy contribution from different modules of the 'weaning diet supplemented with MDCF-2'. (**b,c**) Study design (n=2 dams and 5-8 offspring/treatment arm). Panel b outlines the timing of bacterial colonization of dams and diet switches. Panel c describes the gavages administered to members of each treatment arm. (**d**) Absolute abundance of *B. infantis* Bg2D9 (Arm 1) and *B. infantis* Bg463 (Arm 2) in fecal samples obtained from pups. (**e**) Absolute abundance of *P. copri* in fecal samples collected from pups in the indicated treatment arms at the indicated postnatal time points. Inset: the absolute abundance of *P. copri* in fecal samples collected from pups at P21. (**f**) Body weights of the offspring of dams, normalized to postnatal day 23. [linear mixed effects model (see *Methods*)]. Mean values ± SD are shown. Each dot in panels d-f represents an individual animal. *P*-values were calculated using a Mann-Whitney U test (panel e inset) or a linear mixed effect model (panel f).

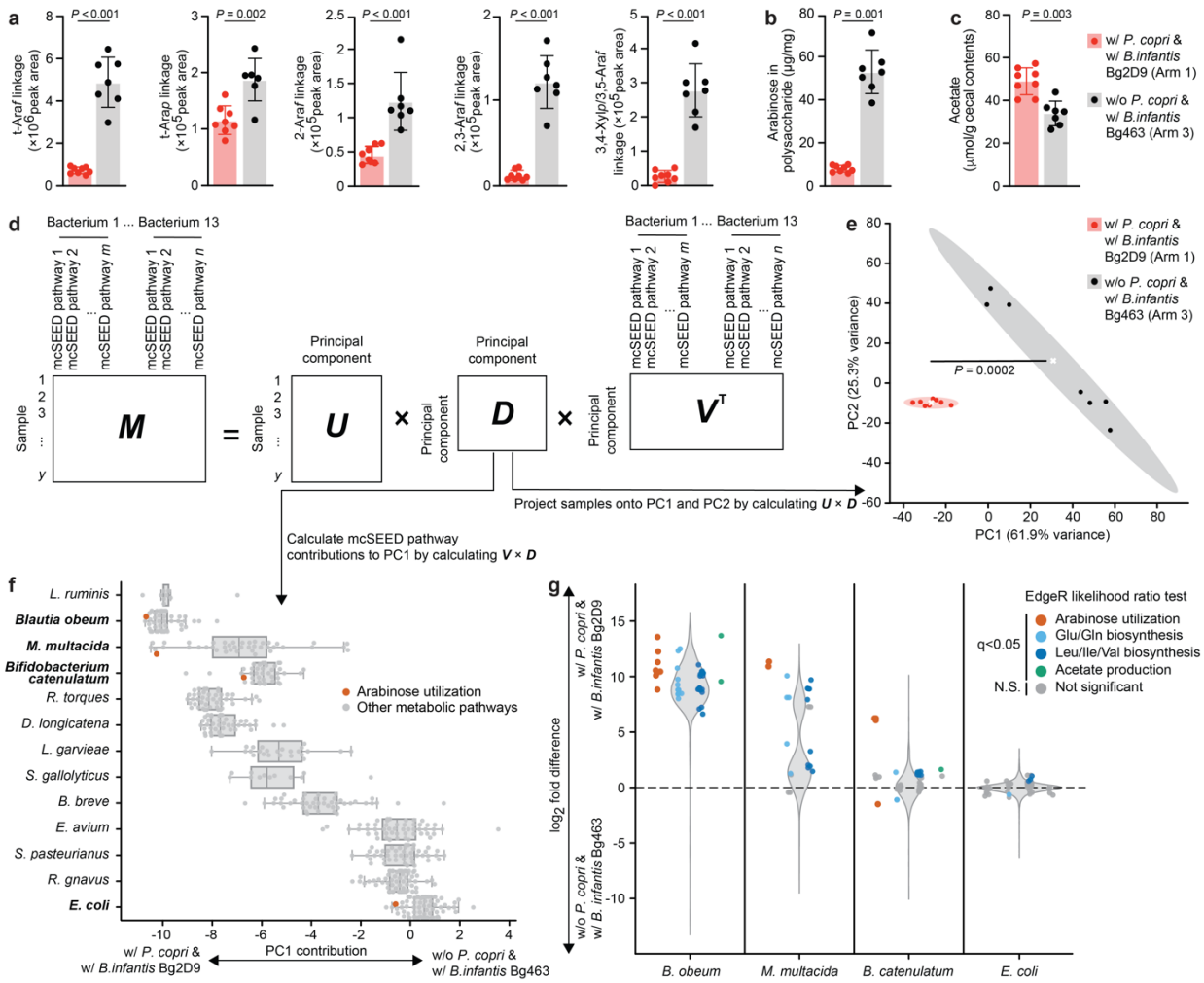


Fig. 2. Targeted mass spectrometric and microbial RNA-Seq analyses of consortia of cultured age-discriminatory and WLZ-associated bacteria strains that colonized gnotobiotic mice. Cecal contents collected from Arm 1 (w/ *P. copri* & w/ *B. infantis* Bg2D9) and Arm 3 (w/o *P. copri* & w/ *B. infantis* Bg463) at the end of the experiment described in **Fig. 1a** were analyzed. Each dot in panels a-e represents an individual animal (n=7-8 mice/treatment arm). Mean values ± SD are shown. *P*-values were calculated using a Mann-Whitney U test for panels a-c. **(a,b)** UHPLC-QqQ-MS-based quantitation of levels of arabinose-containing glycosidic linkages (panel a) and total arabinose (panel b) in cecal glycans. Abbreviations: Ara*f*, arabinofuranose; Ara*p,* arabinopyranose; Xyl*p*, xylopyranose. (**c**) GC-MS quantitation of cecal acetate levels. (**d**) Illustration of singular value decomposition and its application to microbial RNA-Seq analysis. Matrix ***M*** stores the aggregated expression for each metabolic pathway in each bacterium in each sample (see *Methods*). (**e**) Projection of samples onto a space determined by PC1 and PC2. Centroids are denoted by a white "X". Shaded ellipses represent the 95% confidence interval of the sample distribution. (**f**) Contributions of the transcriptional responses of reconstructed metabolic pathways in ***M*** for each species listed to the same PC1 depicted in panel e. Bacteria that can utilize arabinose, based on mcSEED metabolic reconstruction, are highlighted using bold font. Each box plot represents median and quartile PC1 contributions within each species, while the whiskers represent 1.5x the interquartile range (See *Methods*). **(g)** Differential expression analysis of genes involved in carbohydrate utilization, amino acid biosynthesis, and fermentation in arabinose-utilizing bacteria. Violin plots show the distribution of $\log_2$ fold-differences for all expressed genes in the indicated strain. Dots

represent individual genes involved in the corresponding pathway. Abbreviations: Glu, glutamate; Gln, glutamine; Leu, leucine; Ile, isoleucine; Val, valine.

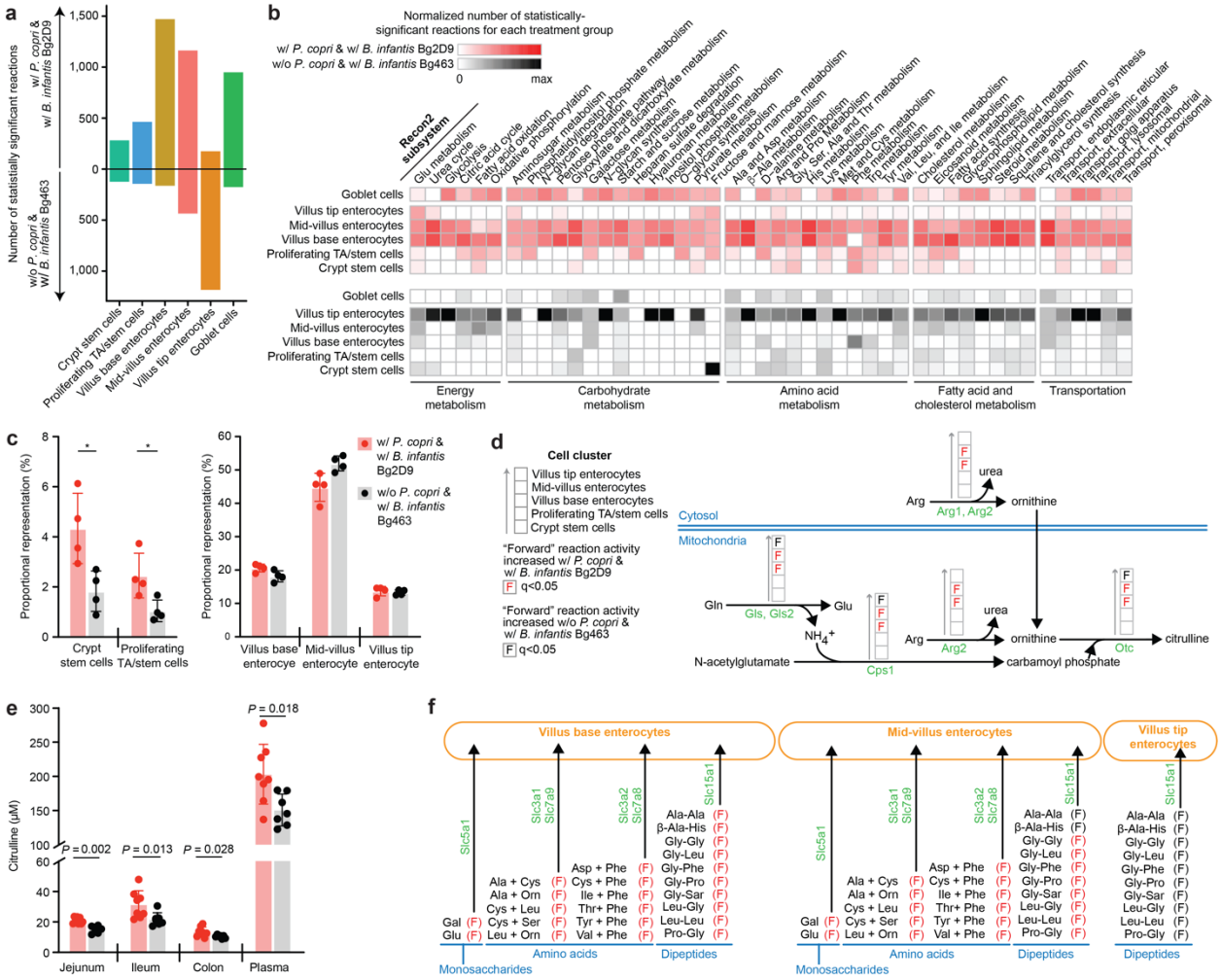


Fig. 3. snRNA-Seq analysis and targeted mass spectrometric analysis of intestinal tissue and plasma harvested from mice containing bacterial communities with or without *P. copri*. Jejunal tissue samples collected from Arm 1 (w/ *P. copri* & w/ *B. infantis* Bg2D9) and Arm 3 (w/o *P. copri* & w/ *B. infantis* Bg463) at the end of the experiment (P53) described in **Fig. 1a** were analyzed (n=4 samples/treatment arm). (**a**) The number of Recon2 reactions with statistically significant differences in their predicted flux between mice in Arm 1 and Arm 3. (**b**) The number of Recon2 reactions in each Recon2 subsystem that are predicted to have statistically significant differences in their activities between the two treatment groups. Colors denote values normalized to the sum of all statistically significantly different Recon2 reactions found in all selected cell clusters for a given Recon2 subsystem in each treatment group. (**c**) Proportional representation of cell clusters identified by snRNA-Seq. Asterisks denote 'statistically credible differences' as defined by *scCODA* (see **Supplementary Table 10c** and *Methods*). (**d**) Selected Recon2 reactions in enterocyte clusters distributed along the villus involved in the urea cycle and glutamine metabolism. (**e**) Targeted mass spectrometric quantifications of citrulline levels along the length of the gut and in plasma. Mean values ± SD and *P*-values from the Mann-Whitney U test are shown. Each dot represents an individual animal (n=7-8). (**f**) Effect of colonization with bacterial consortia containing or lacking *P. copri* on extracellular transporters for monosaccharides, amino acids and dipeptides. Sar: sarcosine. These transporters were selected and the spatial information of their expressed region along the length of the villus was assigned based on published experimental evidence[34]. Arrows in panels b and e indicate the "forward" direction of each Recon2 reaction. The Wilcoxon Rank Sum test was used to evaluate the statistical significance of the net reaction scores (panel a, b, d and e)

between the two treatment groups. *P*-values were calculated from Wilcoxon Rank Sum tests and adjusted for multiple comparisons (Benjamini-Hochberg method); a q-value < 0.05 was used as the cut-off for statistical significance.

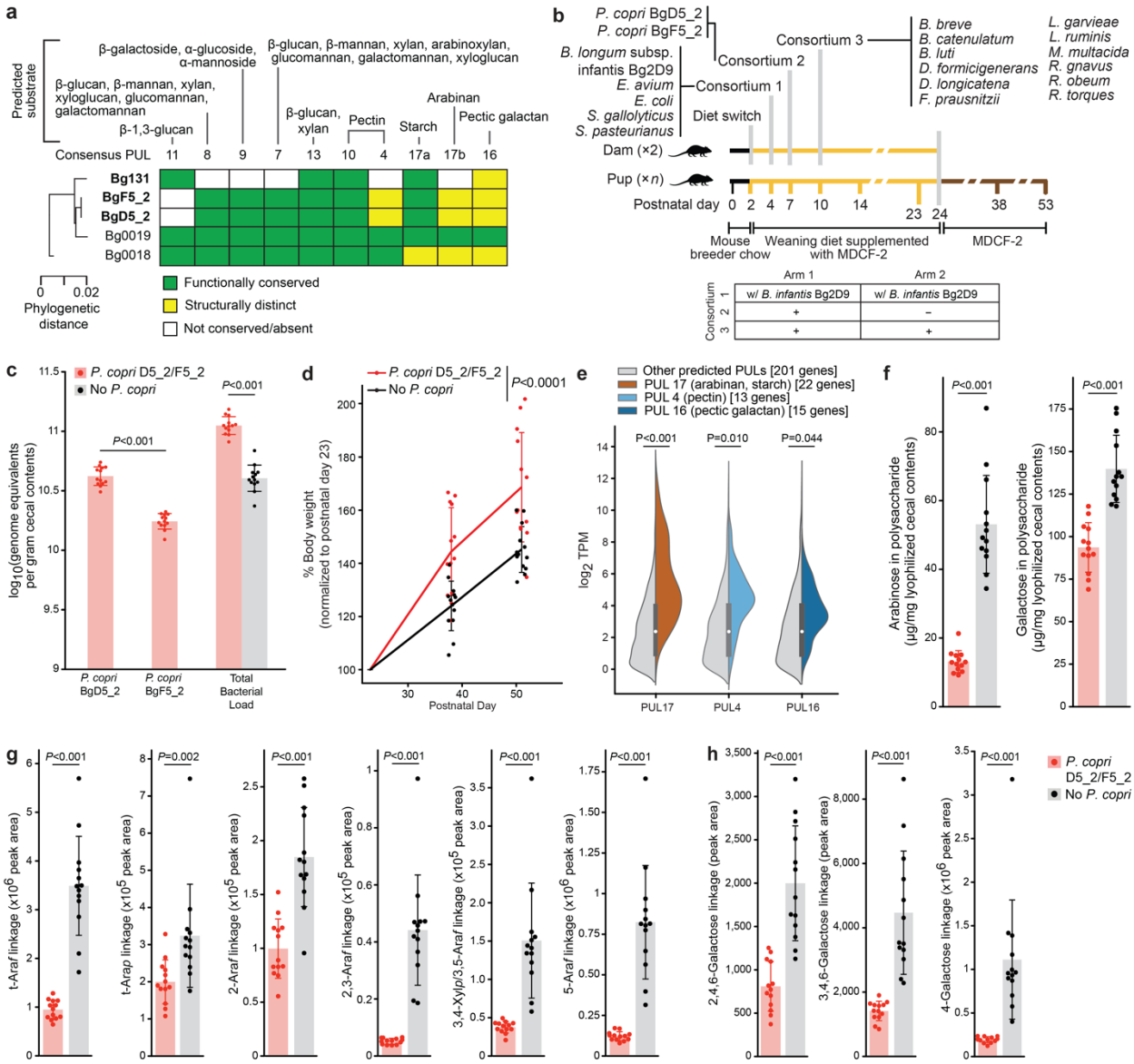
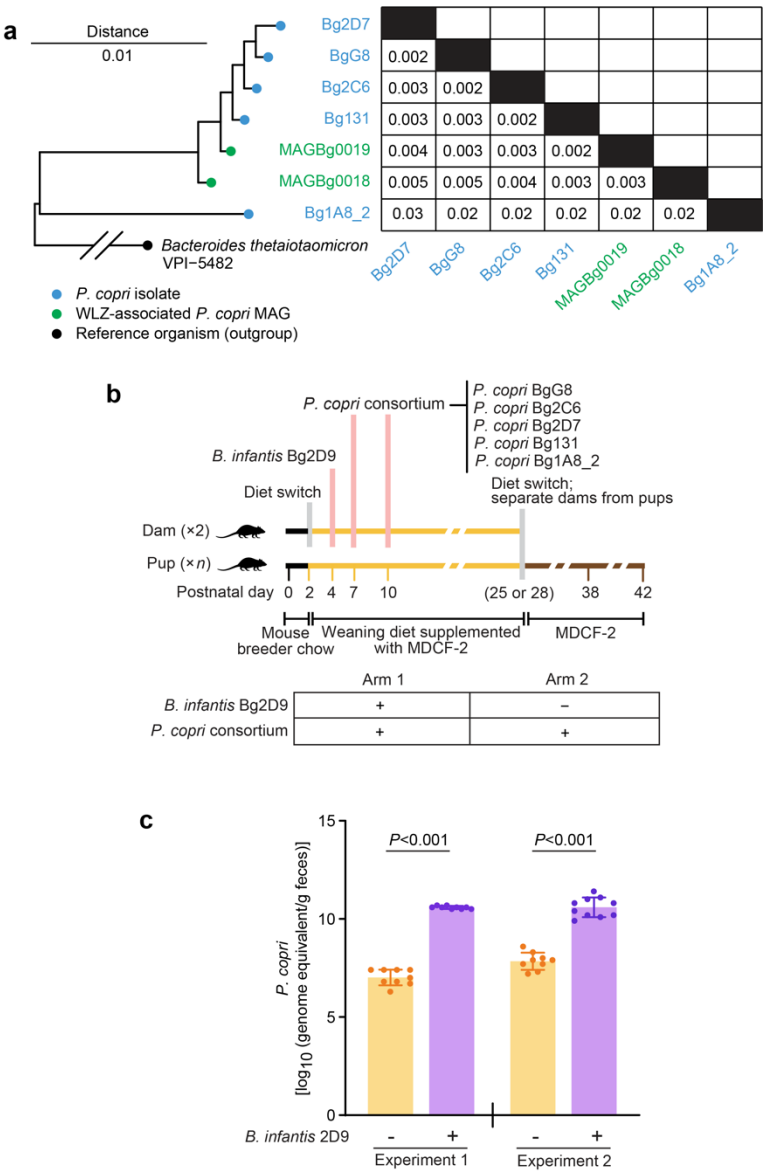


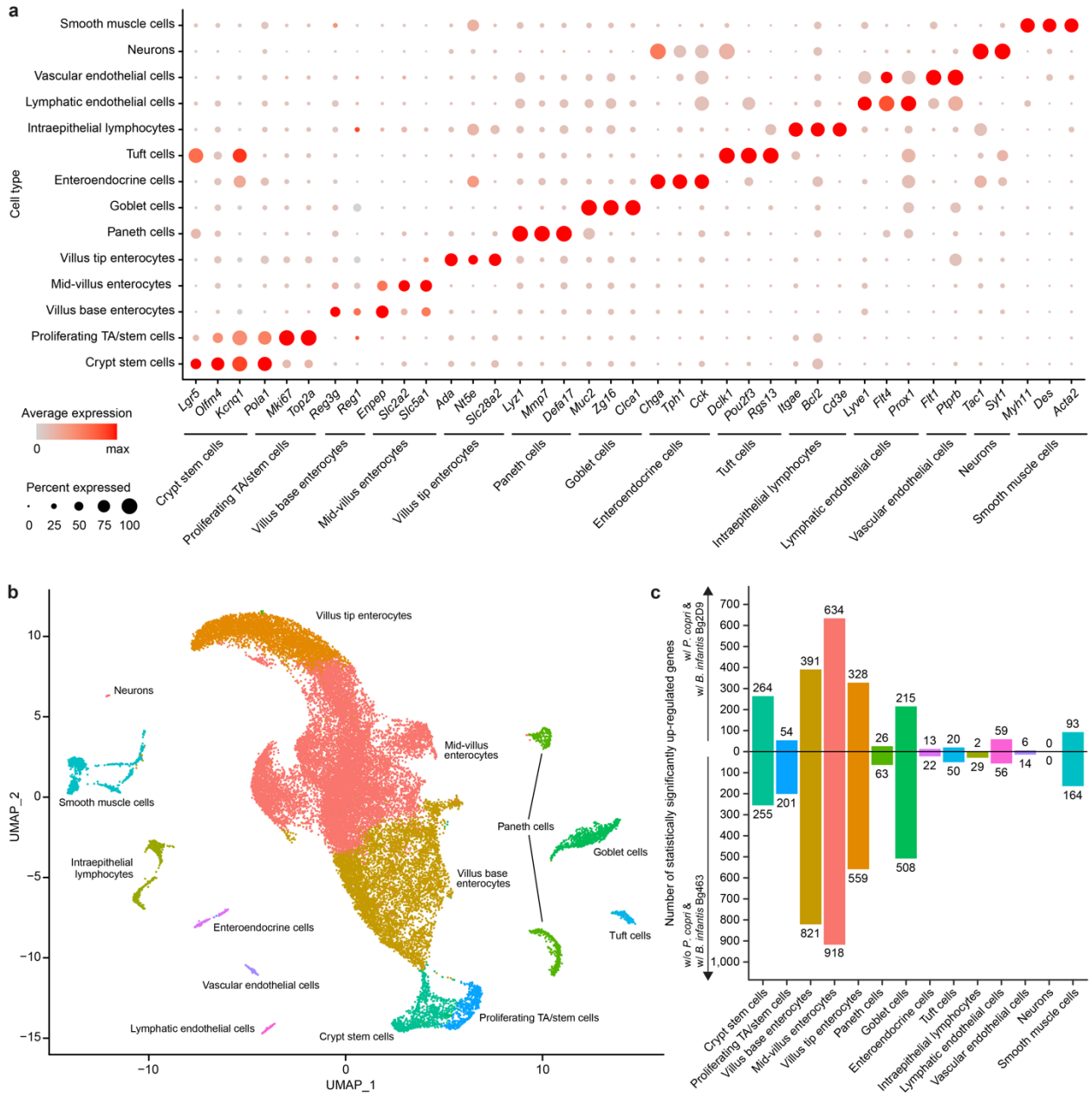
Fig. 4. Testing the effects of pre-weaning colonization with two *P. copri* strains closely related to MAGs Bg0018 and Bg0019 on host weight gain and MDCF-2 glycan degradation. (**a**) Comparison of polysaccharide utilization loci (PULs) highly conserved in the two *P. copri* MAGs with their representation in the three cultured *P. copri* strains. (**b**) Study design (n=2 dams and 13 offspring/ treatment arm). (**c**) Absolute abundance of *P. copri* strains and total bacterial load in cecal contents collected at the end of the experiment (P53). (**d**) Body weights of the offspring of dams, normalized to postnatal day 23 [linear mixed effects model (see *Methods*)]. (**e**) GSEA of expression of PULs shared by *P. copri* BgD5_2 and BgF5_2 in the cecal contents of animals. Adjusted *P*-values were calculated using GSEA ranking genes by their mean $\log_2$ TPM across the *P. copri* colonized samples, with each PUL comprising a gene set against the background of all predicted PUL genes. Violin plots show the $\log_2$ TPM of all genes assigned to any of the 22 predicted PULs in each isolate (n=201 genes) in each of the samples, split to show homologs of consensus PUL 17 (arabinan, starch; n=22 genes), PUL 4 (pectin; n=13 genes), and PUL 16 (pectic galactan; n=15 genes) in color compared to the remainder of all PUL genes in gray. Internal box plots show the median (circle) and quartiles (thick line) for all genes assigned to PULs. (**f**) UHPLC-QqQ-MS analysis of total arabinose and galactose in glycans present in cecal

contents collected at P53. (**g-h**) UHPLC-QqQ-MS of glycosidic linkages containing arabinose (panel g) and galactose (panel h) in cecal contents. Mean values ± SD are shown. *P*-values were calculated using a Mann-Whitney U test. Each dot in panels b-h represents an individual animal.

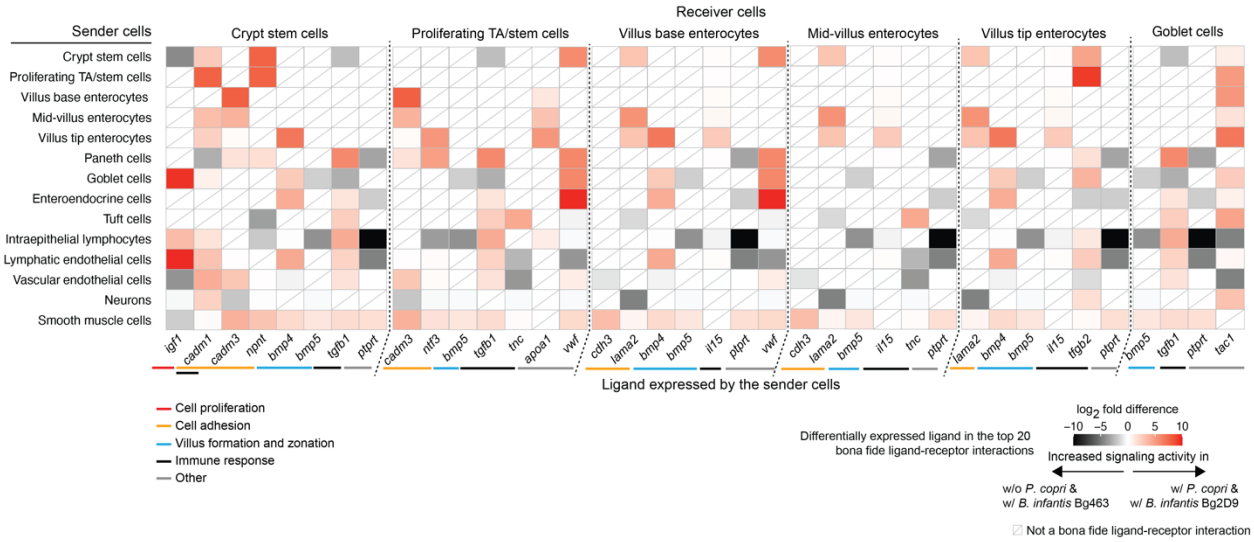
Extended Data Figures



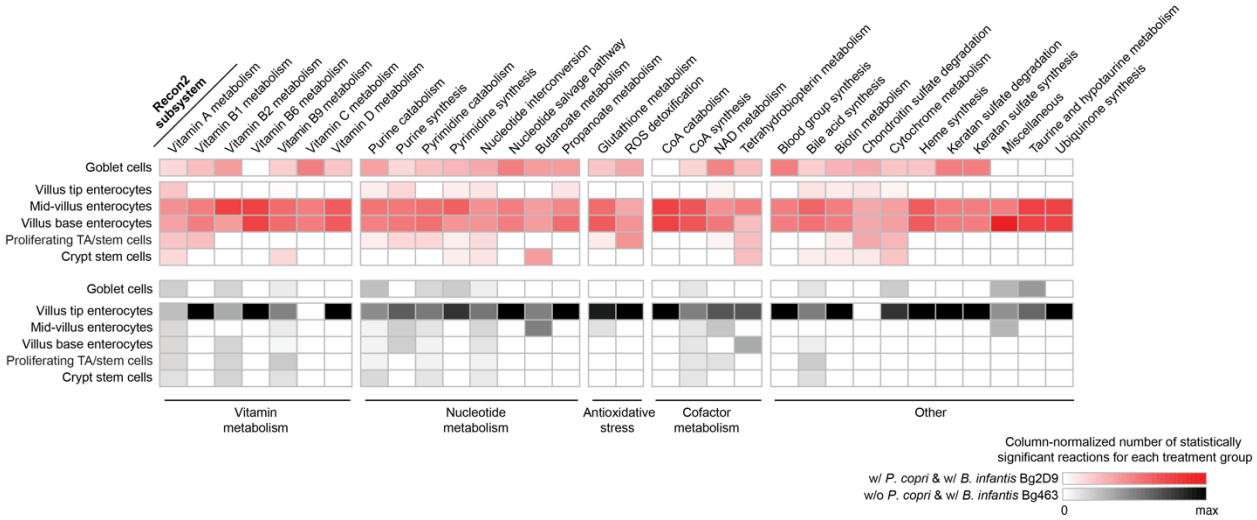
Extended Data Fig. 1. Determining the relationship between *P. copri* colonization efficiency and pre-colonization with *B. longum* subsp. *infantis*. (a) Phylogenetic tree of cultured *P. copri* isolates used in the mouse studies described and the two MAGs positively associated with WLZ in the randomized controlled human study. The phylogenetic distance between each pair of comparisons is shown in the matrix. **(b)** Experimental design. Mice were weaned at P28 and P25 for experiments 1 and 2, respectively. **(c)** Total absolute abundances of *P. copri* strains in fecal samples collected from mice at P42. Mean values ± SD are shown. Each dot represents a separate animal. *P*-values were calculated using a Mann-Whitney U test.



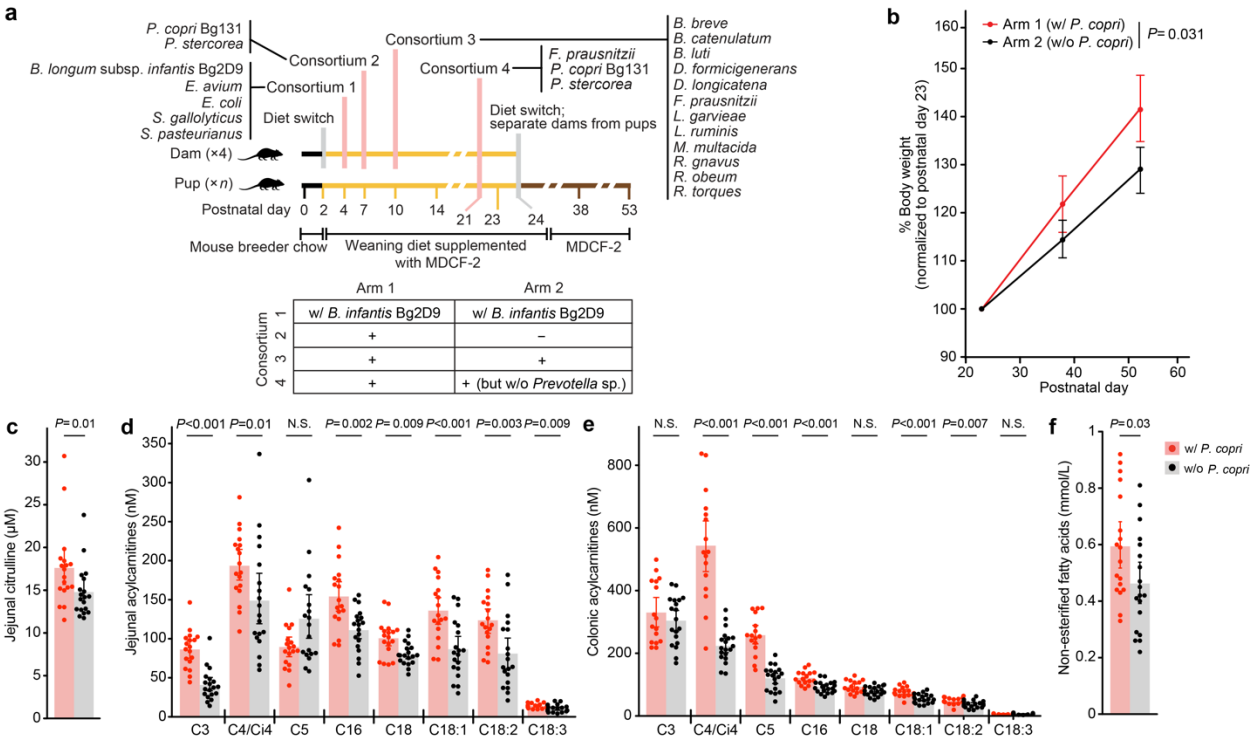
Extended Data Fig. 2. snRNA-Seq analysis of differential intestinal gene expression in mice colonized with bacterial consortia with or without *P. copri*. Jejunal tissue samples collected from 'w/ *P. copri* & w/ *B. infantis* Bg2D9' and 'w/o *P. copri* & w/ *B. infantis* Bg463' groups at the end of the experiment described in **Fig. 1** were analyzed. (**a**) Dot plot of marker gene expression across epithelial cell types. The average expression level and percentage of nuclei that express a given gene within a cell type are indicated by dot color and size, respectively. (**b**) Integrated UMAP plot for all jejunal nuclei isolated from animals in both arms (n=4 mice/arm). (**c**) The number and directionality of statistically significant differentially expressed genes in each cell cluster.



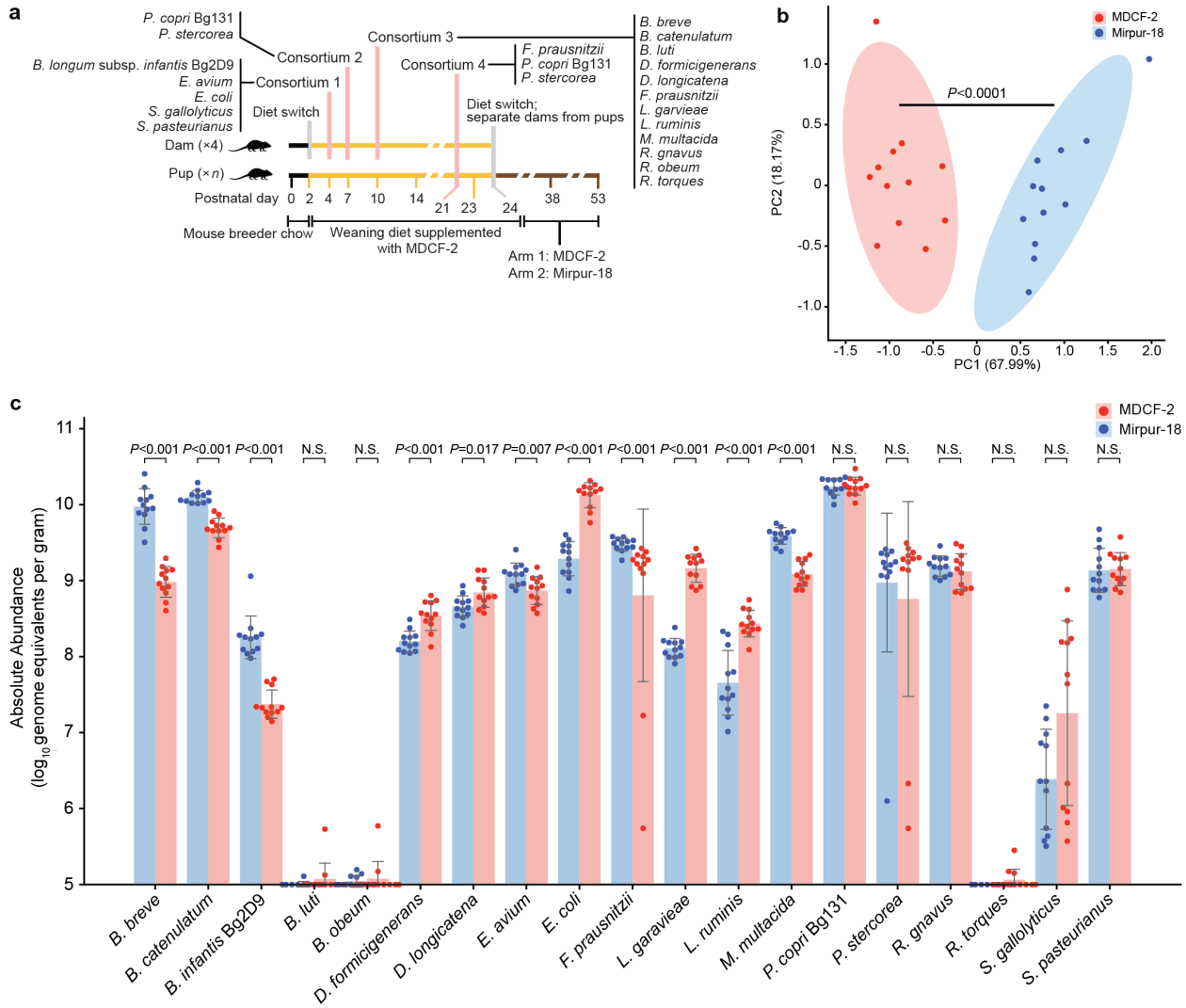
Extended Data Fig. 3. *NicheNet*-based analysis of the effects of *P. copri* colonization on cell-cell signaling activities. Each row represents different sender cell clusters. Each column represents ligands expressed by these sender cells. Cells are colored based on the $\log_2$-fold difference in expression of ligands in the sender cell clusters between mice in 'w/ *P. copri* & w/ *B. infantis* Bg2D9' and 'w/o *P. copri* & w/ *B. infantis* Bg463' groups from the experiment described in **Fig. 1**. Ligands (columns) are grouped based on receiver cell clusters and the indicated functions of downstream signaling pathways in these receiver cells.



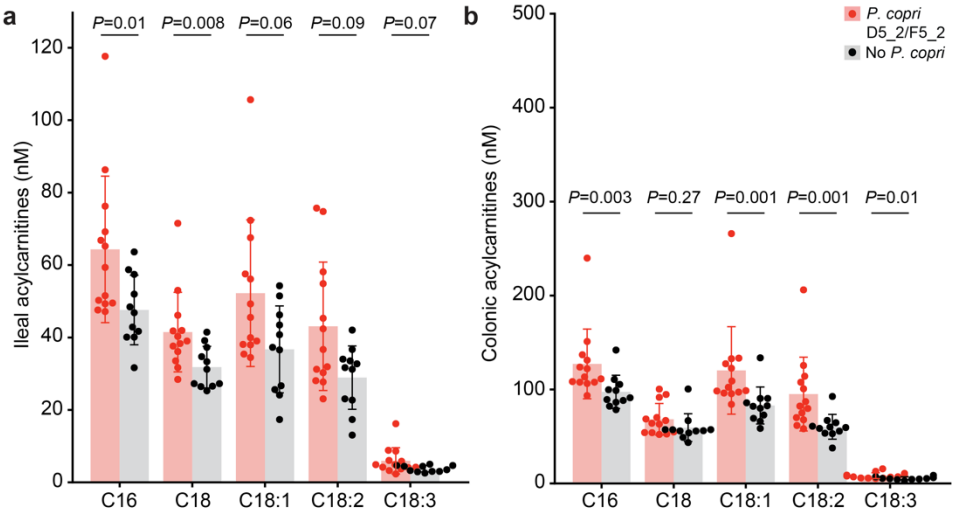
Extended Data Fig. 4. Normalized number of Recon2 reactions in Recon2 subsystems predicted to have statistically significant differences in their activities between mice in the 'w/ *P. copri* & w/ *B. infantis* Bg2D9' and 'w/o *P. copri* & w/ *B. infantis* Bg463' groups. See legend to **Fig. 3b**, which shows other affected subsystems, for details.



Extended Data Fig. 5. Validating the effects of *P. copri* colonization on postnatal weight gain and host metabolism in gnotobiotic mother-pup dyads. (**a**) Study design (n=4 dams and 18-19 offspring/ treatment arm). (**b**) Body weights of the offspring of dams, normalized to postnatal day 23 [linear mixed effects model (see *Methods*)]. (**c-e**) Targeted mass spectrometric analysis of jejunal citrulline (panel c) and acylcarnitine levels (panel d), plus colonic acylcarnitine levels (panel e). (**f**) Plasma levels of non-esterified fatty acids. Each dot represents a single animal. Mean values ± SD are shown for panels b-f. *P*-values were calculated from the linear mixed effect model (panel b) or Mann-Whitney U test (panels c-f). N.S., *P*-value > 0.05.



Extended Data Fig. 6 Evaluating the effect of diet on the defined community in gnotobiotic mother-pup dyads. (a) Experimental design (n=2 dams and 12 offspring/diet treatment). **(b)** Principal component analysis showing the significant differences in community structure in the cecums of mice euthanized on P53 (P<0.0001; PERMANOVA). Ellipses represent 95% confidence intervals. **(c)** Absolute abundances of the defined community members in cecal contents at P53; P-values were calculated by the Mann-Whitney U test. N.S., P-value > 0.05. Mean values ± SD are shown. Each dot represents an individual animal.



Extended Data Fig. 7. LC-MS of ileal and colonic acylcarnitines in gnotobiotic mice colonized with *P. copri* BgD5_2 and BgF5_2. (a) LC-MS of ileal acylcarnitines corresponding to soybean oil lipids. (b) LC-MS of colonic acylcarnitines corresponding to soybean oil lipids. Each dot represents an individual animal. Mean values ± SD are shown. *P*-values were calculated using a Mann-Whitney U test.

Supplementary Information

Data Availability: Microbial community and bacterial strain genome shotgun sequencing datasets plus microbial RNA-Seq and snRNA-Seq datasets have been deposited in the European Nucleotide Archive under study accession PRJEB47807. UHPLC-QqQ-MS datasets are available in Glycopost (ID GPST000245).

Biological Materials: Human specimens and bacterial strains cultured from fecal samples collected from Bangladeshi children are the property of icddr,b. Material Transfer Agreements exist between icddr,b, and Washington University for the use of these samples. Requests for materials should be made to J.I.G.

Computer Code: All software used was from publicly available sources.

Supplementary Tables

Supplementary Table 1. Bacterial strains used in the defined community gnotobiotic mouse experiments. (**a**) Quality metrics for assembled bacterial genomes. (**b**) Comparison between the genomes of cultured bacterial strains and MAGs. (**c**) Binary phenotype predictions for cultured bacterial strains in the 20-member consortium. (**d**) Comparison of the representation of metabolic pathways in the genomes of cultured bacterial strains and MAGs (ANIm > 94%; alignment coverage > 10%). (**e**) Comparison of PULs conserved between MAG Bg0018 and Bg0019 with PULs in *P. copri* Bg131 and other *P. copri* MAGs. (**f**) *P. copri* Bg131 PULs. (**g**) Global distribution of *P. copri* Bg131, *P. copri* MAGs, and *P. stercorea* CAZymes in different GH, PL, CE, and CBM families.

Supplementary Table 2. Diets used in gnotobiotic mouse studies. (**a**) Ingredients in each diet module. (**b**) Representation of modules in the weaning diet supplemented with MDCF-2. (**c**) Nutritional analysis of diets.

Supplementary Table 3. Absolute abundances of bacterial strains in dam-pup dyads colonized with cultured bacterial consortia in the experiment described in Fig. 1. (**a**) Sample metadata. (**b**) Temporal changes in the absolute abundances of strains in fecal samples collected from dams [$\log_{10}$(genome equivalents/milligram feces)]. (**c**) Temporal changes in the absolute abundances of bacterial strains in fecal samples collected from progeny of dams [$\log_{10}$(genome equivalents/milligram feces)]. (**d**) Absolute abundance of strains [$\log_{10}$(genome equivalents/milligram sample)] in ileal and cecal contents collected from P53 progeny of dams.

Supplementary Table 4. Bacterial strains used in the gnotobiotic mouse experiments shown in Extended Data Fig. 1.

Supplementary Table 5. Absolute abundances of bacterial strains in dam-pup dyads colonized with cultured bacterial consortia in the experiment described in Extended Data Fig. 1. (a) Sample metadata. (**b**) Absolute abundance of strains [$\log_{10}$(genome equivalents per gram sample)] in cecal contents collected from progeny of dams at euthanasia.

Supplementary Table 6. UHPLC-QqQ-MS analysis of the levels of glycosidic linkages and monosaccharides in cecal glycans present in mice from the various treatment arms of the experiment described in Fig. 1. (**a**) Glycosidic linkage composition (peak area in arbitrary units). (**b**) Monosaccharides (μg/mg lyophilized cecal sample).

Supplementary Table 7. Targeted mass spectrometric analysis of short-chain fatty acids in the cecal contents of gnotobiotic mice from the various treatment arms of the experiment described in Fig. 1.

Supplementary Table 8. Cecal microbial RNA-Seq dataset generated from the gnotobiotic mouse experiment described in Fig. 1. (**a**) Sample metadata. (**b**) Level of gene expression in *P. copri* Bg131 PULs (TPM normalized). (**c**) Differentially expressed genes (q-value < 0.05) with annotations in mcSEED pathways for carbohydrate utilization, amino acid, and vitamin biosynthesis, or with annotations in the Carbohydrate Active Enzyme (CAZyme) database.

Supplementary Table 9. Quantification of jejunal villus height and crypt depth in gnotobiotic mice in the experiment described in Fig. 1.

Supplementary Table 10. snRNA-Seq dataset generated from the jejunums of gnotobiotic mice in the experiment described in Fig. 1. (**a**) Sample metadata. (**b**) Differentially expressed genes from

pseudobulk analysis (q-value < 0.05). (**c**) Proportional representation of cell clusters identified from snRNA-seq dataset.

Supplementary Table 11. Absolute abundances of bacterial strains in dam-pup dyads colonized with cultured bacterial consortia in the experiment described in Extended Data Fig. 5. (**a**) Sample metadata. (**b**) Absolute abundance of strains [$\log_{10}$(genome equivalents per gram)] in cecal contents collected from P53 offspring of the dams.

Supplementary Table 12. Targeted LC-MS analysis of levels of acylcarnitines, amino acids and biogenic amines in gnotobiotic mice in the experiment described in Extended Data Fig. 5. (**a**) Amino acids/other biogenic amines (μM). (**b**) Acylcarnitines (nM). (**c**) Plasma non-esterified fatty acids (NEFA; mmol/L).

Supplementary Table 13. Comparisons of genomes of *P. copri* isolates BgD5_2 and BgF5_2 to *P. copri* MAGs Bg0018 and Bg0019. (**a**) Quality metrics for genomes assembled from *P. copri* isolates BgD5_2 and BgF5_2. (**b**) Comparison of mcSEED metabolic pathway annotations between the genomes of P. copri isolates and MAGs Bg0018 and Bg0019. (**c**) Comparison of PUL content between the genomes of *P. copri* isolates and MAGs Bg0018 and Bg0019. (**d**) Comparison of *P. copri* BgD5_2 and BgF5_2 genome identity and uniqueness.

Supplementary Table 14. Absolute abundances of bacterial strains in gnotobiotic mice colonized with cultured bacterial consortia in the experiment described in Fig. 4. (**a**) Sample metadata. (**b**) Absolute abundance of bacterial strains [$\log_{10}$(genome equivalents per gram)] in cecal contents collected from P53 offspring of the dams.

Supplementary Table 15. Targeted LC-MS analysis of levels of acylcarnitines, amino acids and biogenic amines in gnotobiotic mice in the experiment described in Fig. 4. (**a**) Amino acid and biogenic amine results for jejunum (μM). (**b**) Acylcarnitine results for jejunum, ileum, colon, and liver (nM).

Supplementary Table 16. Cecal microbial RNA-Seq dataset generated from gnotobiotic mice in the experiment described in Fig. 4. (**a**) Sample metadata. (**b**) Gene Set Enrichment Analysis (GSEA) of predicted PULs from BgD5_2 and BgF5_2.

Supplementary Table 17. UHPLC-QqQ-MS analysis of levels of glycosidic linkages and monosaccharides in cecal glycans recovered from gnotobiotic mice in the experiment described in Fig. 4. (**a**) Glycosidic linkage composition (peak area in arbitrary units). (**b**) Monosaccharides (μg/mg lyophilized cecal contents).

Supplementary Table 18. Targeted UHPLC-QqQ-MS analysis of amino acids and B vitamins in cecal contents collected from gnotobiotic mice in the experiment described in Fig. 4. (a) Amino acids (μg/g of cecal contents). (**b**) B vitamins (ng/g of cecal contents).