Title: *Dietary plant diversity predicts early life microbiome maturation*

Teresa McDonald[1], Ammara Aqeel[1], Ben Neubert[1], Anna Bauer[1], Sharon Jiang[1], Olivia Osborne[1], Danting Jiang[10], Filemon Bucardo[6,8], Lester Gutiérrez[23], Luis Zambrana[7], Kirsten Jenkins[10], Jennifer Gilner[11], Javier Rodriguez[12], Amanda Lai[14], Jonathan P. Smith[16,17], Rinn Song[18], Kazi Ahsan[19,20], Sheraz Ahmed[21], Sanam Iram Soomro[21], Fayaz Umrani[21], Michael Barratt[19,20], Jeffrey Gordon[19,20], Asad Ali[21], Najeeha Iqbal[21], Jillian Hurst[10,11], Victoria Martin[2,3,4], Wayne Shreffler[2,3,5], Qian Yuan[2,3,4], Joe Brown[13], Neeraj K. Surana[1,22,15], Samuel Vilchez[6], Sylvia Becker-Dreps[6,7], Lawrence David[1,15]

[1] Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, North Carolina, USA

[2]Harvard Medical School, Boston, Massachusetts, USA.

[3] Food Allergy Center, Massachusetts General Hospital, Boston, Massachusetts, USA

[4] Division of Pediatric Gastroenterology and Nutrition, Massachusetts General Hospital for Children, Boston, Massachusetts, USA.

[5] Division of Pediatric Allergy and Immunology, Massachusetts General Hospital for Children, Boston, Massachusetts, USA.

[6]Department of Family Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

[7]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

[8]Department of Microbiology and Immunology, University of North Carolina Chapel Hill, Chapel Hill, North Carolina, USA.

[9]Universidad Nacional Autónoma de Nicaragua León, León, Nicaragua

[10]Department of Pediatrics, Division of Infectious Diseases, Duke University School of Medicine, Durham, North Carolina, USA

[11]Department of Obstetrics and Gynecology, Division of Maternal-Fetal Medicine, Duke University School of Medicine, Durham, North Carolina, USA

[12]Department of Pediatrics, Children's Clinical Research Unit, Duke University School of Medicine, Durham, North Carolina, USA

[13]Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA.

[14]The Aquaya Institute, San Anselmo, California, USA.

[15]Duke Microbiome Center, Duke University School of Medicine, Durham, North Carolina, USA

[16]Department of Health Policy and Management, Yale School of Public Health, New Haven, Connecticut, USA.

[17]Division of Global HIV and Tuberculosis, U.S. Centers for Disease Control and Prevention, Atlanta, Georgia, USA.

[18]Oxford Vaccine Group, Department of Pediatrics, University of Oxford, Oxford, UK.

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

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

[21] Aga Khan University, Medical College, Pakistan

[22]Department of Pediatrics, Duke University School of Medicine, Durham, North Carolina, USA

[23]Centro de Investigación de Enfermedades Tropicales (CIET). Facultad de Microbiología. Universidad de Costa Rica, San José, Costa Rica.

Abstract

Between birth and adulthood, the human gut is colonized by a complex microbial community. Despite established links between the infant gut microbiome and health, knowledge is limited for how complementary feeding influences colonization. Using FoodSeq, an objective DNA-based dietary assessment technique, we analyzed 1,036 fecal samples from 729 children aged 0-3 years across countries in North America, Central America, Africa, and Asia. We detected a wide diversity of 199 unique plant food sequences, of which only eight staple foods were consistently present across all countries. Despite this variation in global diet, we identified universal trajectories in early life dietary exposure: weaning stage, which tracked with dietary diversity, emerged as the dominant dietary signature across populations. Still, dietary diversity did not correlate with gut microbial diversity. Instead, dietary diversity and weaning stage specifically predicted the abundance of adult-like bacterial taxa, including known fiber-degrading taxa, which colonized after age 1. Our findings support a two-stage model of microbiome maturation: an early phase dominated by milk-adapted taxa independent of complementary feeding, followed by a maturation phase where diet shapes adult-like microbiota colonization. This model suggests that tracking and promoting plant dietary diversity may support the timely emergence of an adult-like microbiome.

Introduction

In the first years of life, the human gut transitions from a sterile state to a diverse microbial ecosystem. During this transition, the gut microbiome transforms from an 'immature' state -- characterized by a limited set of milk-adapted taxa -- to a 'mature' adult-like community harboring fiber-degrading taxa and a diversity of other commensal species[1–3]. Proper microbial succession is fundamental to healthy metabolism, immune development, and disease resistance. Indeed, disruptions in early life colonization and microbiome maturation increase risk of conditions such as obesity[4], allergy[5], and Type I diabetes[6]. Consequently, the first years of life represent a critical window when microbiome development can have lifelong health effects.

Despite the importance of the gut microbiome in healthy infant development, the specific factors guiding early microbiome colonization remain less clear. Well-established influences include antibiotic exposure, which reduces bacterial diversity and stability; caesarean delivery, which alters the initial microbial repertoire; and formula feeding, which lowers *Bifidobacterium* abundance yet increases overall microbial diversity[7]. Another pivotal factor is cessation of breastfeeding[3,2,8], which triggers a decline in the abundance of milk-adapted microbial taxa.

One area of microbiome development, that remains underdeveloped, however, is the precise role that complementary feeding plays in microbial succession[1]. Few studies have found evidence directly linking specific dietary factors to microbiome maturation by concurrently measuring diet and the microbiome in young children. Exceptions have either tested narrow food interventions[9–12] or were conducted within a single cultural and dietary context[13–16]. Other studies have measured microbiome

composition across more diverse dietary traditions[8,17]; still such studies have been limited in sample size or have often lacked detailed dietary data, leading comparisons to be on population-level dietary patterns rather than individual-level diet metrics.

A major barrier to understanding the relationship between diet and microbiome maturation is the challenge of accurately characterizing infant diet. Surveys, the gold standard and traditional approach for collecting dietary data, are labor-intensive[56] require baseline levels of literacy[51], incompletely capture minority or ethnic foods[52-54], , and are vulnerable to social desirability bias[18,19]. Dietary assessment in children poses additional obstacles, as it relies on secondhand reporting by parents and guardians who may not observe all meals when children are under the supervision of other caregivers. In the United States, for instance, a quarter of 2-5 year olds' energy intake is consumed outside the home[20]. Such incomplete observation contributes to reporting bias exemplified by validation studies with doubly-labeled water that show infant dietary intake is overestimated by 7-29%[21–26]. Finally, the technical challenges with infant dietary assessment only multiply in multi-country settings, as assessments must be adapted for each different geographic, cultural, and linguistic setting[27] For example, there are over 100 versions of a standardized survey for comparing population-level global diets[28]. In concert, these barriers have prevented systematic investigation of the relationship between early life microbiome development and infant diet across diverse human populations.

Remarkable diversity in early-life diets worldwide

To perform a global survey of infant microbiome development and diet, we assembled a cohort of 729 children aged 0-3 years with samples from five countries: the

USA (n=144), Nicaragua (n=457), Cambodia (n=178), Kenya (n=228), and Pakistan (n=29). These sampling sites were designed to span multiple continents and encompass variation in socioeconomic status, infectious disease burden, and dietary traditions.

We applied FoodSeq, an objective DNA sequencing-based dietary assessment[29], to 1,036 fecal samples from from our cohort. FoodSeq is an amplicon sequencing-based technique that sequences the leucine gene (*trnL*) from plant plastids or the 12S rRNA gene from animal mitochondria. By mapping *trnL* and 12S sequencing reads to a curated reference database, FoodSeq identifies and quantifies foods of plant and animal origins at a taxonomic level. Unlike traditional dietary surveys, FoodSeq provides objective measurements that are independent of caregiver reporting and remain consistent across varying income levels, health conditions, and infant feeding practices. The diet data generated by FoodSeq also has the advantage of being temporally synchronized with corresponding microbiome data, as both sets of genomic analyses can be generated from the same fecal sample.

FoodSeq revealed extensive variety and breadth in early life diets across our global cohort. With plant FoodSeq, we detected 113 species and 86 assigned sequence variants (ASVs) that matched at a higher level (e.g. genus), encompassing 389 potential dietary plant species. Of the plant ASVs, 42% were detected exclusively in a single country (Fig. S1A), with unique country-specific foods ranging from longan and lotus root in Cambodia, to jute mallow and cassava in Kenya, to raspberry in the USA (Fig. 1A). By contrast, just eight global staples[30] showed consistent prevalence in all five countries (detected in >5% of each country's samples): wheat, corn, rice,

banana/plantain, tomatoes and other nightshades (potato, etc.), mangifera (mangos), and alliums (onion, leek, garlic, chives).

Despite the breadth of plant species in our global survey, principal component analysis (PCA) revealed that the primary axis of dietary variation, PC1, was likely driven by the overall presence of plant matter in the diet rather than specific dietary composition (Fig. 2A-B). Unlike subsequent PCs, which displayed mixed positive and negative loadings indicative of tradeoffs in plant species abundance, PC1 exhibited exclusively positive loadings for common foods (Fig. 2C). This pattern suggests that PC1 captures the extent of plant consumption rather than regional dietary differences. Supporting this interpretation, PC1 was strongly correlated with both infant age (Fig. 2E, Spearman's ρ=0.69, p<0.001) and overall plant FoodSeq richness (pFR, the number of unique sequences assigned to our human foods database in each sample; Fig. 2F, ρ=0.25, p<0.001). These associations are consistent with the expected trajectory of weaning as infants gradually incorporate more variety of solid foods into their diets as they mature. Direct comparisons of dietary diversity and age confirmed our interpretation, showing that pFR consistently increased across the first year of life (Fig. 2G; Spearman's ρ=0.27, p<0.001). Taken together, these findings argue that FoodSeq effectively captures the dietary shifts associated with weaning.

While weaning stage dominated the first PC, the second PC revealed country-specific dietary signatures partitioned by their relative intakes of regional staple plants like rice (enriched in Cambodia), banana/plantain (enriched in Nicaragua), or sorghum and millet (enriched in Kenya). Cross-country differences also emerged in dietary diversity more broadly, with median pFR ranging from 4 in Nicaragua to 10 in Cambodia

(Fig. 1C). Such values aligned with previously reported food variety scores, which found mean food variety ranges from 5.5 to 10 among children from 6-24 months in Africa and Central America[31–33]. The timing and rate of dietary diversification also varied by country. For instance, the diets of Cambodian children diversified early and rapidly until pFR plateaued at 13 months of age, whereas US infants showed a more gradual rise in pFR until 19 months (Welch's t-test, p-value >0.05 at timepoint prior; Fig. 3D).

We note that in contrast to the diversity of plant-associated species and dietary patterns observed in our cohort, we detected a more limited breadth of dietary animals. A total of 28 dietary animal species were detected. Observed animals included widely utilized livestock such as pig, chicken, and cow, as well as more region-specific animals including water buffalo, the primary source of milk in Pakistan, and various fish species in Cambodia. Overall, however, 41% of infant samples had no non-human animal DNA detected, and only 33% had more than one animal species detected. Only the US and Cambodia had a non-zero median count of animal species present (Fig. S1B). The limited detection of dietary animals was consistent with the higher cost of animal-based foods and the majority of human animal-based foods originating from cattle, poultry, and pigs[27] for the locations of infants sampled here. Given the limited detected range of animal species, the distinct nutritional characteristics of different animal products from the same species (beef, dairy), as well as the known importance of nutrients like fiber for microbiome development, we focused our subsequent analyses on plant-based foods and gut microbiome development.

Patterns of microbiome maturation

Prior to analyzing the relationship between early-life dietary and microbiome development, we first confirmed that our cohort followed well-established patterns of early-life gut microbiome maturation[3]. Microbiome alpha-diversity increased steadily from birth (Fig. 3C) through age two (Welch's t-test, p-value < 0.05 at prior timepoint), regardless of country (Fig. 3D). In terms of inter-individual variation (beta-diversity), both age (Fig. 4B) and country of origin (Fig. 4A) were significant factors (age R^2=0.4721, p=0.001; country R^2=0.4675, p=0.001, EnvFit on Unweighted Unifrac. Fig. S3C). In cohorts where data were available on breastfeeding status and birth mode, both of these factors were associated with microbiome composition. (Breastfeeding: USA, Nicaragua, Cambodia, R^2= 0.19, p=0.001; Birth mode: USA, Nicaragua, R^2= 0.0.113, p=0.009, Envfit on unweighted Unifrac, controlling for age and country). Thus, consistent with previous literature, we observed both increased diversification of the microbiome with age, as well as notable roles for breastfeeding status and birth mode in shaping microbiome structure in our cohort[34].

The specific bacterial taxa observed in our infant cohort also aligned with prior studies[1–3,8], as hierarchical clustering revealed a microbial succession pattern (Fig. S5A). We identified an early-life cluster (Cluster 9), enriched in *Bifidobacterium* and *Streptococcus,* and a 'transitional' cluster (Cluster 1) , characterized by plant degraders such as *Blautia* and probiotic *Ligilactobacillus* enriched at ages 12-18 months. Following this transition, a late-microbiome cluster (Cluster 5) emerged between 21-36 months that more closely resembled the adult microbiome, featuring *Bacteroides vulgatus* and *Faecalibacterium prausnitzii,* which ferment plant fiber into butyrate and other short chain fatty acids, and *Prevotella copri,* a core fiber-metabolizing taxon[35]. In parallel, we

constructed a Random Forest model that successfully predicted infant age given input microbiome data. The RF model identified *Faecalibacterium* and *Bifidobacterium* as top predictor variables, mirroring both our hierarchical clustering results and previously published models of infant microbiome development[1] (Fig. S4).

Does dietary diversity predict microbiome diversity?

Having found that infant microbiota in our cohort matured as expected from previous studies, we next compared patterns of dietary maturation to gut microbiome maturation. Surprisingly, we did not see consistent evidence that dietary diversity predicted gut microbial diversity. Microbial alpha diversity continued to increase until at least 14-16 months of age (Fig. 3C) well after dietary diversity (as measured by pFR) had plateaued (Fig. 2G). We did observe associations between dietary diversity and microbiome diversity within several countries but associations were not consistent; for example, pFR correlated with microbiome Shannon diversity in Kenya and the United States, whereas in Nicaragua, the relationship was reversed (Fig. S2E). Further, when we constructed global models controlling for age and country, pFR did not correlate wtih overall microbiome diversity measures (Shannon diversity and 16S richness both non-significant, Fig. 3E, S3D). Thus, although children began eating complex diets soon after introduction of solid foods, the gut microbiome continued to diversify even after dietary complexity had been reached.

Diet is associated with later stage microbes during gut colonization

We next hypothesized that although dietary diversity may not directly correlate with overall microbial diversification, dietary diversity and weaning stage would be associated with microbiome maturity. Indeed, despite substantial geographic variation in diet, we found pFR was positively correlated with the ‘Transitional Microbiome’ cluster and the ‘Late Microbiome’ cluster, comprised of *Bacteroides*, *Faecalibacterium*, and *Prevotella*, even after controlling for age and country (Fig. 4C,F). Likewise, we also found that PC1 was positively correlated with 'Late Microbiome' cluster abundance (Fig. 4G). Because increased dietary diversity (i.e., pFR) and weaning stage (i.e., PC1) both reflect complementary feeding and the introduction of an adult-like diet, these findings argue that introduction of an adult-like diet stimulates the growth of an adult-like microbiome.

Notably, the 'Early Microbiome' cluster showed neither correlations with dietary diversity or within-country dietary PCs (Fig. 4B,4D), consistent with the hypothesis that milk consumption, rather than introduction of solids, shapes gut microbial ecology prior to weaning. These results suggest that the transition to an adult-like microbiome is driven by the increasing introduction of solid foods generally, rather than by specific dietary patterns. More broadly, our findings point to a developmental sequence wherein milk consumption governs early microbiome composition, while the subsequent introduction of solid plant foods, independent of their specific composition, drives the transition toward a mature gut microbial community.

Discussion

In concert, our findings here refine existing models linking weaning to microbiome development in the infant gut. Our data do not reveal simple linear

relationships between dietary and microbiome diversity across early life. This negative finding, however, is consistent with our increasing understanding that microbial ecosystem diversity does not necessarily predict community function[36,37]. Instead, our data support a two-stage developmental model: an early phase governed by milk consumption, followed by a maturation phase shaped by dietary diversity. In this second stage, plant dietary diversity and the child's physiologic age together predict the colonization by select groups of gut bacterial taxa associated with adult-like microbiome function. Notably, despite the broad range of complementary feeding patterns across our geographically diverse cohorts, we observed similar successional trends, underscoring the functional and diet-responsive redundancy[38] that gut microbes can share, regardless of environmental differences.

Early life nutrition remains one of the most critical challenges in global health, with lasting consequences for child development and lifelong health outcomes[39]. Our findings confirm that adequate and diverse plant consumption during the complementary feeding period fosters maturation of the gut microbiome toward an adult-like state rich in fiber-degrading bacteria, adding to the known benefits of dietary diversity for micronutrient adequacy[40,41] and protection against food allergies[42]. Importantly, we show that simply counting the number of distinct plant foods effectively measures this microbial transition, potentially offering a more accessible alternative to complex dietary indices (e.g. minimum dietary diversity score).[43] Implementing such straightforward 'pro-gut' complementary feeding guidance could enable easier monitoring by caregivers and health workers across diverse settings, ultimately supporting appropriate microbiome maturation among infants worldwide.

Materials and Methods

Study populations and stool sampling

Cambodia

Cambodian samples were obtained as extracted DNA collected as part of the Integrated Nutrition, Hygiene and Sanitation Project (NOURISH), which was conducted across three rural provinces[44]. This study involved nutrition and sanitation interventions. The nutrition intervention consisted of nutrition educational direct payments contingent on attending regular health visits, and two $5 food vouchers. Participants were encouraged to provide a diverse diet, breastfeed exclusively until 6 months and continue breastfeeding to age 2, and gather nutritious local foods such as small fish. While the nutrition intervention resulted in a 10% increase in HAZ for children in either nutrition intervention arm, it did not affect breastfeeding or dietary diversity. 101 (56.7%) of samples in our study came from children in either of the nutrition intervention arms. Our study preferentially chose samples from children who had not experienced diarrhea or fever in the week prior to sample collection. N=14 samples came from children who had recently experienced diarrhea, n=156 from children who had not, and n=5 lacked data. N=7 samples came from children who had received RUTF or a local alternative, which was not part of the study intervention. This cohort displayed a high pathogen burden, with most children testing positive for some form of enteric pathogen. Despite this, most children were asymptomatic.

Kenya

Kenyan samples were obtained as extracted DNA, collected in Kisumu, Kenya as part of the TOTO study[45]. These children were generally representative of the overall population in Kenya and were enrolled due to displaying signs of respiratory illness suspected of being tuberculosis infection. 70 of 228 samples came from children who were reported as receiving a nutritional intervention (i.e., RUTF).

Nicaragua

Nicaraguan samples were obtained as stool slurries stored 1:1 in PBS from the *Sapovirus*-Associated Gastro-Enteritis (SAGE) study[46]. This study collected stool samples from children at risk of diarrheal disease in Leon, Nicarauga. In addition to the control and diarrheal samples described in the published study, we obtained n=391 samples from from surveillance stool samples that were collected on a monthly basis from study participants. Non-diarrheal surveillance samples were chosen from children who had not experienced a diarrhea episode in the month prior. 16S data was only obtained from surveillance samples collected from children ages 2-12 months.

Pakistan

The Pakistan samples were obtained as extracted DNA from the SEEM (Study of Environmental Enteropathy) cohort. The SEEM study was a prospective longitudinal study, conducted in Matiari, a rural area in Sindh, Pakistan, between March 2016 and March 2019. The study collected samples from 50 healthy control children and longitudinal samples from malnourished children (WHZ< -2), who were part of a malnourishment intervention arm. Our study used n=29 of stool samples from healthy

control children (WHZ > 0 and not part of a nutritional intervention). Approximately 1g of fecal samples were collected from each participant from community health workers within 10 minutes of stooling. Samples were transported from field sites in 4*C cooler boxes and then stored at -80*C and shipped on dry ice.

United States

US samples were obtained as frozen solid stool or stool slurry from Project HOPE 1000 study[47] (Durham, NC, USA) and Gastrointestinal Microbiome & Allergic Proctocolitis Study (GMAP) study (Boston, MA, USA)[48]. Project HOPE 1000 is a study intended to develop a biorepository of longitudinal samples from healthy participants over the first 1000 years of life in order to identify early life determinants of health. The GMAP study prospectively collected samples from children, some of whom developed allergies and/or proctocolitis. The GMAP samples analyzed in this study came from subjects who did not develop allergic disease. Breastfeeding data at the time of sample collection was collected for Project HOPE 1000. For GMAP, breastfeeding status at 0 months old and 4 months old was collected, but not at the time of sample collection. GMAP samples from children who were exclusively receiving formula at 4 months old were presumed to not be breastfed.

DNA Extraction

Cambodia: Stool samples were preserved 1:1 in Zymo DNA/RNA Shield buffer (Zymo Research, Irvine, CA) and then stored at -20°C. They were then extracted using the QIAamp 96 Virus QIAcube HT (Qiagen, Germany) protocol. 200mg of solid stool or

200uL of liquid stool were combined with 1mL of stool lysis buffer (ASL, Qiagen, Germany) and vortexed in an SK38 soil grinding tube (Bertin Corp., Rockville, MD) for five minutes (Vortex Genie 2, Scientific Industries, Bohemia, NY). Following a ten minute incubation at room temperature, the lysate was centrifuged at 12,000 g for two minutes (Thermo Fisher Scientific, Waltham, MA). 200uL of this lysate was then used for DNA/RNA extraction with the QIAamp 96 Virus QIAcube HT protocol.

Kenya: 1g of stool was diluted in 10ml of sterile water and emulsified with sterile glass beads. The stool slurry was then filtered. DNA was extracted using the Qiagen MagAttract PowerSoil DNA EP according to the manufacturer's instructions.

Pakistan: 50mg of frozen fecal sample was extracted using phenol-chloroform extraction. Briefly, samples were ground in liquid nitrogen using a mortar and pestle. 50mg of this was then transferred to a 2mL vial. 500uL of 0.1mm diameter zirconia/silica beads were added to each sample, along with 500 µL of phenol: chloroform: isoamyl alcohol (25:24:1), 210 µL of 20% SDS and 500 µL of buffer (200 mM NaCl, 200 mM Trizma base, 20 mM EDTA). Samples underwent bead beating (Mini-Beadbeater-8; Biospec) to extract DNA. Extracted DNA was purified, (Qiaquick columns, Qiagen), then eluted in 70 µL of Tris-EDTA (TE) buffer, and then quantified (Quant-iT dsDNA broad range kit; Invitrogen). Each DNA sample was adjusted to a concentration of 1 ng/mL.

Nicaragua and USA:

The Nicaraguan and USA (Durham, Boston) stool samples were diluted 1:10 in filtered phosphate buffered saline (0.2uM filter) and then stored at -80°C. Samples were thawed on ice in an anaerobic chamber. Samples were vortexed 30s and then 250uL aliquots were taken for DNA extraction. Slurry aliquots were then randomized and extracted using the Power Soil Pro MagAttract kit (Qiagen, Germany).

DNA Sequencing

16S rRNA Sequencing

All samples were PCR-amplified using primers targeted to the V4 region of the 16S rRNA gene using primers 515F and 806R as previously described.[49] Kenyan samples were quantified using a Qubit dsDNA HS assay kit (ThermoFisher), pooled in equimolar concentration, and sequenced on a MiSeq sequencer (Illumina; 2x250 base pair). Cambodian, Nicaraguan, Pakistan, and USA samples were processed as previously described[49–51], diluted to a 10nM concentration, and sequenced using an Illumina MiniSeq and a MiniSeq High Output kit (FC420-1004) with paired-end 150-bp reads.

FoodSeq

FoodSeq was carried out as previously described.[52] For Plant FoodSeq, samples were amplified in a 10uL volume with 3 µL of extracted DNA template and either a) 0.5 µL each of 10 µM forward and reverse primers (IDT), 2 µL of 5X KAPA HiFi buffer, 0.3 µL of 10 mM dNTPs, 0.1 µL of 100X SYBR Green I (Life Technologies), 0.1 µL KAPA HiFi polymerase, 3.5 µL nuclease-free water or b) 5uL of KAPA HiFi 2X, 0.3uL each of 10 µM forward and reverse primers (IDT), 0.1uL of 100X SYBR Green I (Life Technologies)

and 1.3uL of nuclease-free water. Method (b) uses the same components as method (a), but in a premixed format with a slightly higher concentration of polymerase. Cycling conditions were 95°C for 3min, followed by 35 cycles of 98°C for 20s, 63°C for 15s, and 72°C for 15s.

For Animal FoodSeq, samples were amplified in a 10uL volume with 1uL of extracted DNA template and 0.5uL of 10uM forward and reverse primer and 1uL of 100uM human blocking primer (IDT), 0.1 uL of 100X SYBR Green, 0.25uL of 20mg/mL bovine serum albumin, and 5uL of 2X Accustart. The amplification primers were 12SV5F and 12SV5R[53] with Illumina overhang adapter sequences added at the 5′ end, and the human blocking primer was DeBarba14 HomoB[54]. Cycling conditions were an initial denaturation at 94°C for 3 minutes, followed by 35 cycles of 94°C for 20 seconds, 57°C for 15 seconds, and 72°C for 1 minute.

For both plant and animal FoodSeq, primary amplification was followed by secondary PCR amplification to add Illumina adapters and dual 8-bp indices for sample multiplexing. Samples were amplified in a 50uL volume using 5uL primary PCR product diluted 1:100 in nuclease free water, 25uL KAPA HiFi 2X Buffer, 0.5uL 100X SYBR Green I, 9.5uL nuclease free water, 5uL each of 2.5uM forward and reverse indexing primers.

Amplicons were cleaned (Ampure XP, Beckman Coulter, Brea, CA), quantified (QuantIT dsDNA assay kit, Invitrogen, Waltham, MA), and pooled in equimolar ratios to create a

sequencing pool. If samples were too low in concentration to be pooled at a equimolar ratio, they were added up to a set volume of 15-20uL. The pooled library was then concentrated, gel purified, quantified by both fluorimeter and either qPCR or BioAnalyzer (Agilent, Santa Clara, CA). If trnL and 12SV5 amplicons were to be sequenced on the same run, they were combined at this point in equimolar amounts. Final sequencing library was spiked with 30% PhiX (Illumina, San Diego, CA) to mitigate low nucleotide diversity. Libraries were sequenced on an Illumina MiniSeq according to manufacturer's instructions using a 150-cycle Mid or High kit, (Illumina, San Diego, CA, USA), depending on the number of samples in each pool.

Data analysis

All packages used are available as a Singularity container hosted by the Duke Gitlabs (https://gitlab.oit.duke.edu/lad-lab).

16S rRNA

All code used in bioinformatic pipeline to process 16S rRNA amplicon data can be found on Github at tjk30/16S-pipeline. To summarize, following conversion to fastq format using Illumina's bcl2fastq package, DNA sequencing results were demultiplexed using qiime2 with the --p-no-golay-error-correction flag. Trimmomatic was used to remove primer sequences. Dada2 was then used to filter, trim, merge, and assign species to reads using the SILVA database v138.1. Samples with <5000 total reads were excluded from analysis.

Richness was calculated as the sum of unique ASVs that had reads>0. Shannon diversity was calculated using estimate_richness() from the phyloseq package.

Unweighted Unifrac was performed using the SILVA v138.1 SSU phylogenetic tree and the Unifrac() function from the phyloseq package. Ordination of sample distances was then performed using cmdscale() from the R package stats with the flag eig=TRUE. Envfit was run using the envfit() function from the R package vegan. For calculation of the variation captured by age and country, both variables were included in the model. For the breastfeeding and birth mode models, each model was run controlling for country and age.

For microbiome clustering analysis, we excluded samples from children known to be experiencing diarrhea at the time of sample collection (n=33/934) or acute malnutrition (WHZ <-2). Diarrhea and malnutrition known to significantly impact gut microbiome composition and we were interested in microbial succession patterns among generally healthy children.[55],[1] However, diarrhea data was only collected in the Cambodian and Nicaraguan cohorts. Following generation of clusters, excluded samples were added back to the dataset and used in subsequent analysis. Clustering was limited to 0-36 month old children to ensure that each age bin consisted of children from multiple countries, as only Kenya had samples from children older than 36 months. Age bins were set 0- 2mo, 3 -5mo, 6-8mo, 8-10mo, 11-13mo, 14-16mo,17-19mo, 20-22mo, 23-25mo, and 26+mo (binned together due to limited sampling >24mo in Cambodia and USA). Rare taxa, defined as having reads>0 in 2 samples or fewer, were removed.

The n=818 samples that passed filtering were then used to calculate the average relative abundance of each ASV within each age bin. Taxa were filtered to those with at least 0.5% relative abundance in at least 1 age bin. In order to normalize across taxa with different relative abundance ranges across age, relative abundances for each taxon were converted to z scores. Clustering analysis was performed on the Euclidean distances of the taxa z scores using UPGMA clustering (hclust function from the R package stats, method=’average’). Clustering analysis was then used to generate an k=10 clusters using the stats function cutree(). We then went back to our full dataset, including the "sick" children we had excluded from cluster generation (n=934), and calculated the cluster abundances for Clusters 1-10 for all samples by merging the taxa within each cluster and then performing centered log ratio (clr) normalization.

FoodSeq

All code and databases used in bioinformatic pipeline to process FoodSeq amplicon data can be found on Github at LAD-lab/mb-pipeline. Bioinformatic processing was conducted as previously described. To summarize, following conversion to fastq format and demultiplexing using Illumina’s bcl2fastq package, primer sequences were trimmed using the qiime2 cutadapt wrapper with the following flags: --p-error-rate 0.15, --p-minimum-length 1 , --p-overlap 5, p-discard-untrimmed. Paired reads were then merged using the qiime2 dada2 wrapper with the following flags: --p-trunc-len-f 0, --p-trunc-len-r 0, --p-max-ee-f 2, --p-max-ee-r 2, --p-trunc-q 2, p-min-overlap 12, --p-pooling-method 'independent'.

As we used a non-error-correcting polymerase for the 12SV5 PCR amplification reactions in order to maximize human blocking primer efficiency, resulting in more error-prone amplification, ASVs mapping to the same species were glommed using tax_glom() in the R package phyloseq prior to analysis. This meant that ASVs were assigned to the most similar species in the database used from Schneider et al.[56]

pFR was calculated as the sum of unique assigned ASVs that had reads>0. Unassigned ASVs that did not match a human food in our database were excluded from pFR. Samples with a pFR >5 standard deviations outside the median were excluded from further analysis.

For food clustering analysis, foods that did not have at least 1 read in 1 sample were removed and a matrix was generated of the percent of samples in each country where each assigned ASV was detected. Clustering was performed on the Euclidean distance between each food based on its prevalence in the 5 countries using the “complete” method in the function hclust from the R package stats. The R package ggtree was then used for dendrogram visualization.

Statistical Analysis

Data Ordinations

To account for the compositional nature of the data, we calculated the centered-log ratio (CLR) transform of DNA sequencing read counts using the transform function of

microbiome v1.26.0. Principal component analysis was then performed on clr transformed read counts using prcomp() in stats v4.4.1.

Linear model design

Generalized linear models were used to model the relationship between dietary factors and microbiome outcomes. We accounted for age due to the strong relationships between our microbiome outcomes of interest and age established in previous literature.[1] We further controlled for country due to differences in age distribution, malnutrition, enteric pathogen exposure, and breastfeeding duration between the countries assessed. Our choice to include interaction terms in our models was determined by choosing the model with the lowest AIC and significant chi-squared p-value in a likelihood ratio test between reduced model and saturated model. The relationship between microbial diversity and dietary diversity was modeled using the call glm(microbiome_diversity ~ pFR + country + age_months + country:age_months + pFR:country, family = 'gaussian'). The relationship between dietary PC1 and microbiome clusters was modeled using the call glm(cluster,~age_months+country+country:age_months+PC1+country:PC1+age_months:PC1, family=’gaussian’). Cluster associations with subsequent PCs were not tested, as country was highly collinear with PC2, accounting for half of its variance (ANOVA $p < 2e–16$; $R^2 \approx 0.50$ for country), which would render PC2 effects indistinguishable from country-level associations.

When fitting within-country models, the following were used:

USA and Kenya (lacked breastfeeding data for most or all samples): glm(microbiome outcome ~ diet_variable + age_months, family = 'gaussian')

Pakistan: (lacked breastfeeding data and all samples came from the same timepoint: glm(microbiome outcome ~ diet_variable, family = 'gaussian')

Cambodia and Nicaragua had breastfeeding data, and therefore had breastfeeding status included in the model:
glm(microbiome_outcome ~ diet_variable + currently_breastfed + age_months + currently_breastfed:age_months + diet_variable:currently_breastfed, family= 'gaussian')

When accounting for the effects of WHZ on microbiome richness, the following models were used:
Kenya: microbiome_diversity ~ WHZ+pFR +age_months+sex

Cambodia: microbiome_diversity ~ WHZ+pFR+age_months+sex+currently_breastfed+pFR:currently_breastfed+currently_breastfed:age_months+WHZ:currently_breastfed

Institutional and Ethical Review

Stool and extracted stool DNA samples were obtained through collaborations with existing studies via an IRB exemption (Pro00100567) from the Instutitional Review Board at Duke University.

Use of Kenyan pediatric stool samples in this analysis was reviewed and approved by the institutional review boards of the US Centers for Disease Control and Prevention, the Kenya Medical Research Institute, and the Jaramogi Oginga Odinga Hospital. Written informed consent for the use of these stool samples was given by parents or legal guardians of participants.

U.S. CDC Disclaimer: This work has been supported in part by the President's Emergency Plan for AIDS Relief through the U.S. Centers for Disease Control and Prevention, The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the funding agencies.

Acknowledgements

Nicaraguan samples: The SAGE study was made possible with the NIH grants R01AI127845 and: K24AI141744.

Kenyan samples: We acknowledge the Duke Microbiome Core Facility for performing the DNA extractions, quality control, and preparing the 16S rRNA gene sequencing library; the Duke Sequencing and Genomic Technologies Shared Resource for sequencing the 16S rRNA gene library. We also thank the Duke Microbiome Core Facility for performing DNA BioAnalyzer quality assessments.

Pakistan samples: SEEM study was a multi-national collaboration between Aga Khan University (Pakistan), Washington University in St. Louis, the University of Virginia, and Cincinnati Children's Hospital Medical Center, funded by the Bill and Melinda Gates Foundation (2016-2019). Marty Meier provided technical assistsance in stool DNA purification.

Figures

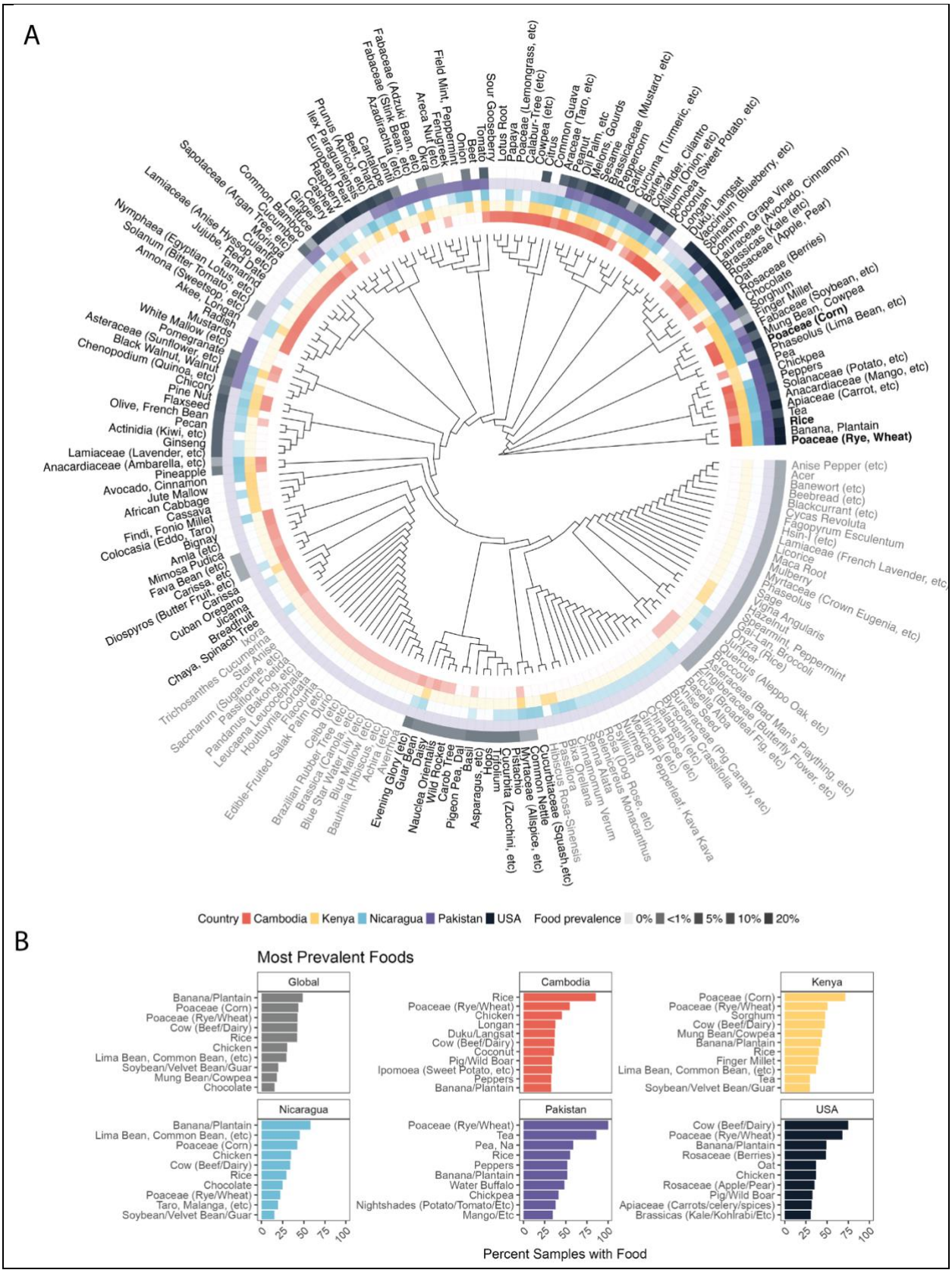
A
Country Cambodia Kenya Nicaragua Pakistan USA
Food prevalence 0% <1% 5% 10% 20%
B
Most Prevalent Foods
Global
Banana/Plantain
Poaceae (Corn)
Poaceae (Rye/Wheat)
Cow (Beef/Dairy)
Rice
Chicken
Lima Bean, Common Bean, (etc)
Soybean/Velvet Bean/Guar
Mung Bean/Cowpea
Chocolate
Cambodia
Rice
Poaceae (Rye/Wheat)
Chicken
Longan
Duku/Langsat
Cow (Beef/Dairy)
Coconut
Pig/Wild Boar
Ipomoea (Sweet Potato, etc)
Peppers
Banana/Plantain
Kenya
Poaceae (Corn)
Poaceae (Rye/Wheat)
Sorghum
Cow (Beef/Dairy)
Mung Bean/Cowpea
Banana/Plantain
Rice
Finger Millet
Lima Bean, Common Bean, (etc)
Tea
Soybean/Velvet Bean/Guar
Nicaragua
Banana/Plantain
Lima Bean, Common Bean, (etc)
Poaceae (Corn)
Chicken
Cow (Beef/Dairy)
Rice
Chocolate
Poaceae (Rye/Wheat)
Taro, Malanga, (etc)
Soybean/Velvet Bean/Guar
Pakistan
Poaceae (Rye/Wheat)
Tea
Pea, Na
Rice
Peppers
Banana/Plantain
Water Buffalo
Chickpea
Nightshades (Potato/Tomato/Etc)
Mango/Etc
USA
Cow (Beef/Dairy)
Poaceae (Rye/Wheat)
Banana/Plantain
Rosaceae (Berries)
Oat
Chicken
Rosaceae (Apple/Pear)
Pig/Wild Boar
Apiaceae (Carrots/celery/spices)
Brassicas (Kale/Kohlrabi/Etc)
0 25 50 75 100
Percent Samples with Food

Fig. 1: Early life diet is diverse and varies globally. (A) Dendrogram of plant prevalences. Rare foods (<1% frequency in at least one country) are labeled with grey text. Sequences that matched to a higher taxonomic level than species are labeled with the lowest taxonomic level assigned, and then a common name of a food that sequence matches. Sequences with multiple foods or food names associated with them are labeled with 'etc'. (B) Top 10 most prevalent foods in each country.

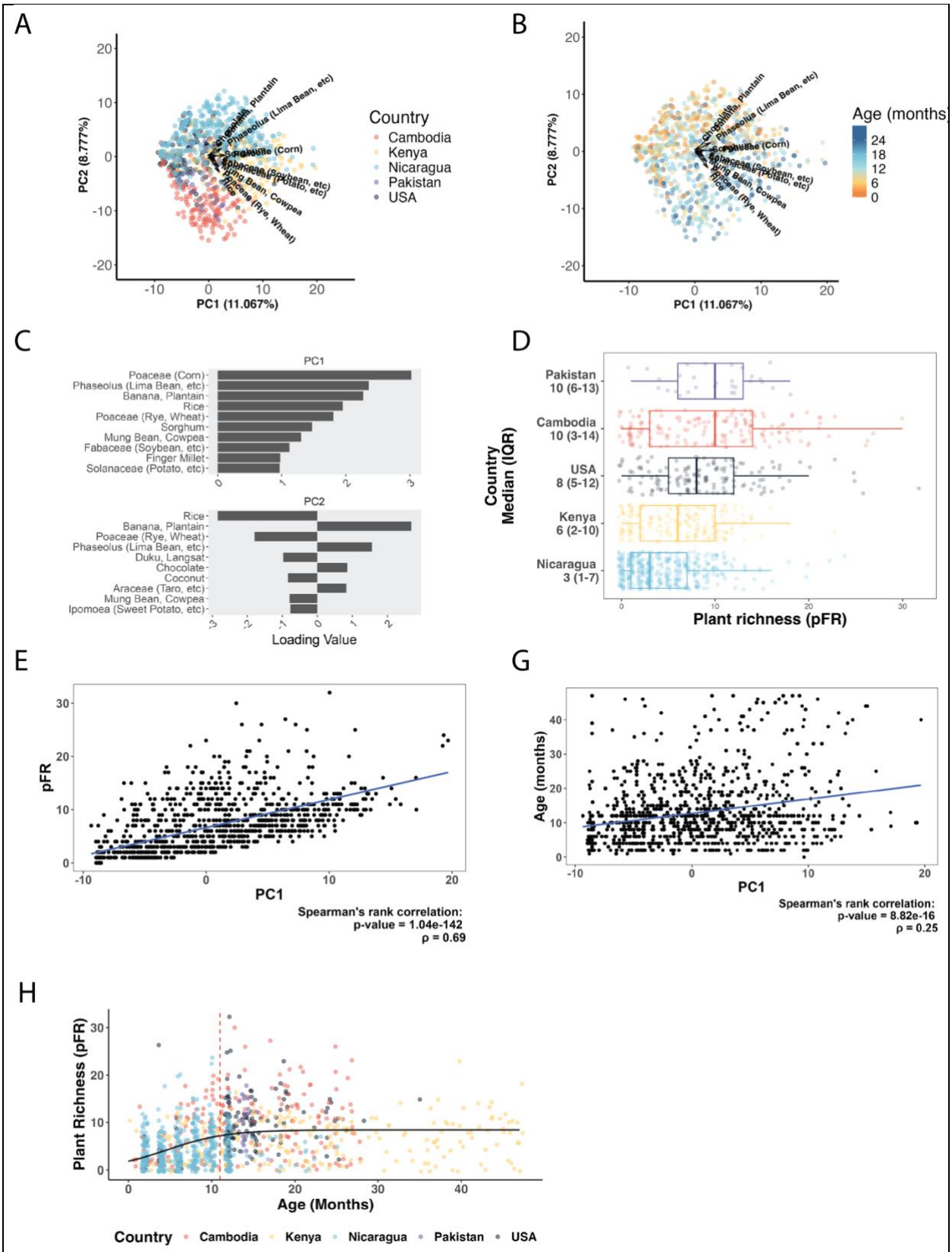

Fig. 2: Dietary variation across age and country. (A) plant FoodSeq PCA colored by country. Country accounted for 15% of variation, age accounted for 0.6% variation, and dietary

diversity accounted for 5% of variation. (PERMANOVA, p-value <0.001 for all). (B) PCA colored by age (C) PCA loadings. (D) Boxplot showing pFR distribution by country. (E) Correlation between age and PC1. (G) Correlation between pFR and PC1. (H) pFR over time. A logistic line (black) was fit to the data. The red dashed line represents the plateau point, calculated as the x value at which the first derivative of the line reaches 50% of its maximum.

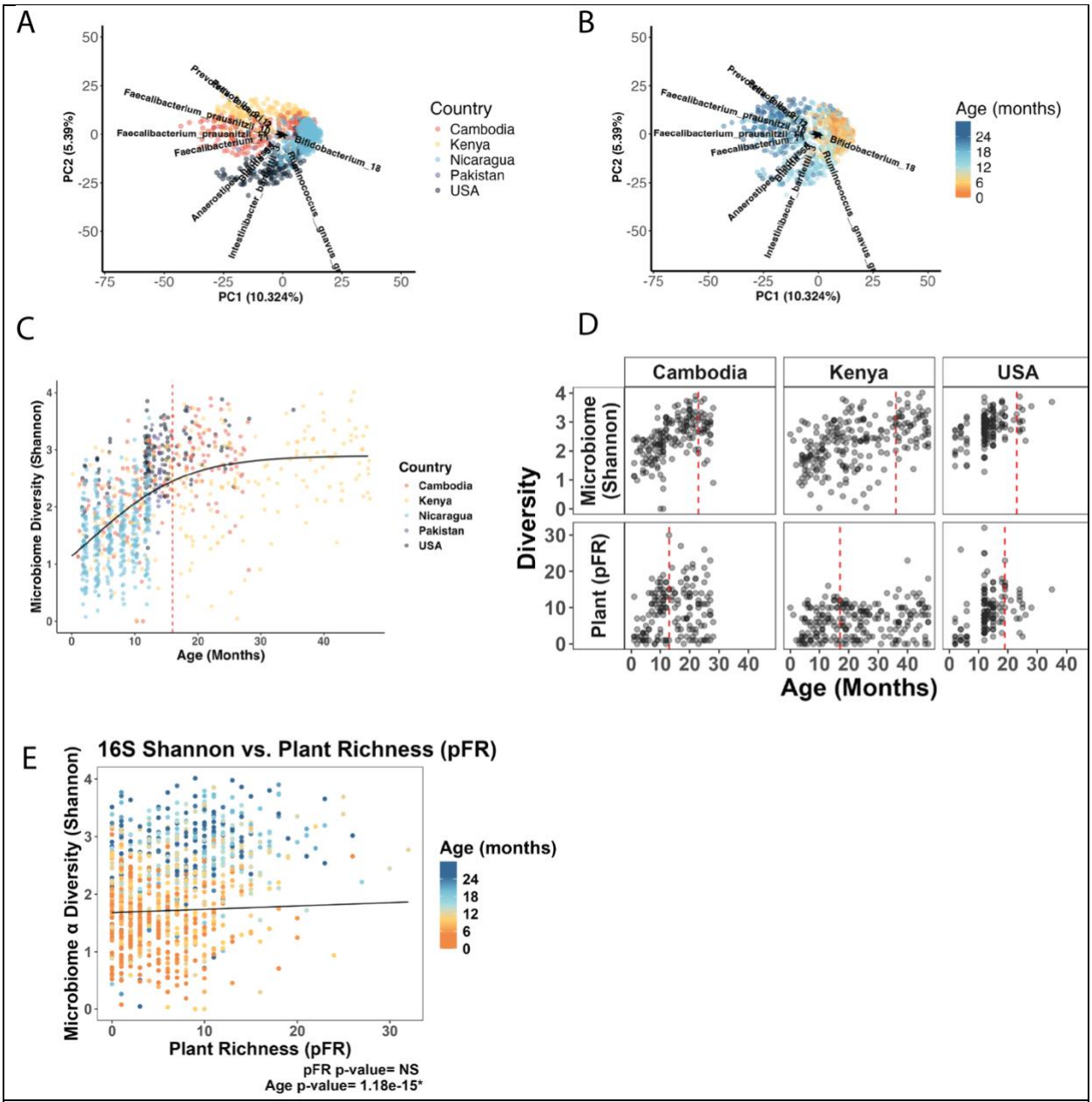

Fig. 3 The early life microbiome varies primarily by geography and age. (A) PCA of 16S rRNA sequencing data colored by country (B) PCA of 16S data colored by age in months (C) Microbiome diversity increases with age. A logistic line (black) was fit to the data. The red dashed line represents the point at which diversity plateaus with respect to age, defined as the point at which the first derivative of the line reached 50% of its maximum. (D) Dietary diversity plateaus prior to microbiome diversity. Scatterplot showing how microbiome diversity and dietary diversity change with age. Red dashed line indicates point at which diversity is no longer distinguishable from diversity at final timepoint (t-test at each timepoint between all samples above vs. below that timepoint). (E) Scatterplot showing relationship between microbial Shannon diversity and plant richness. Lines plot the coefficient derived from a generalized linear model (GLM) predicting microbial diversity from plant dietary diversity and age.

591

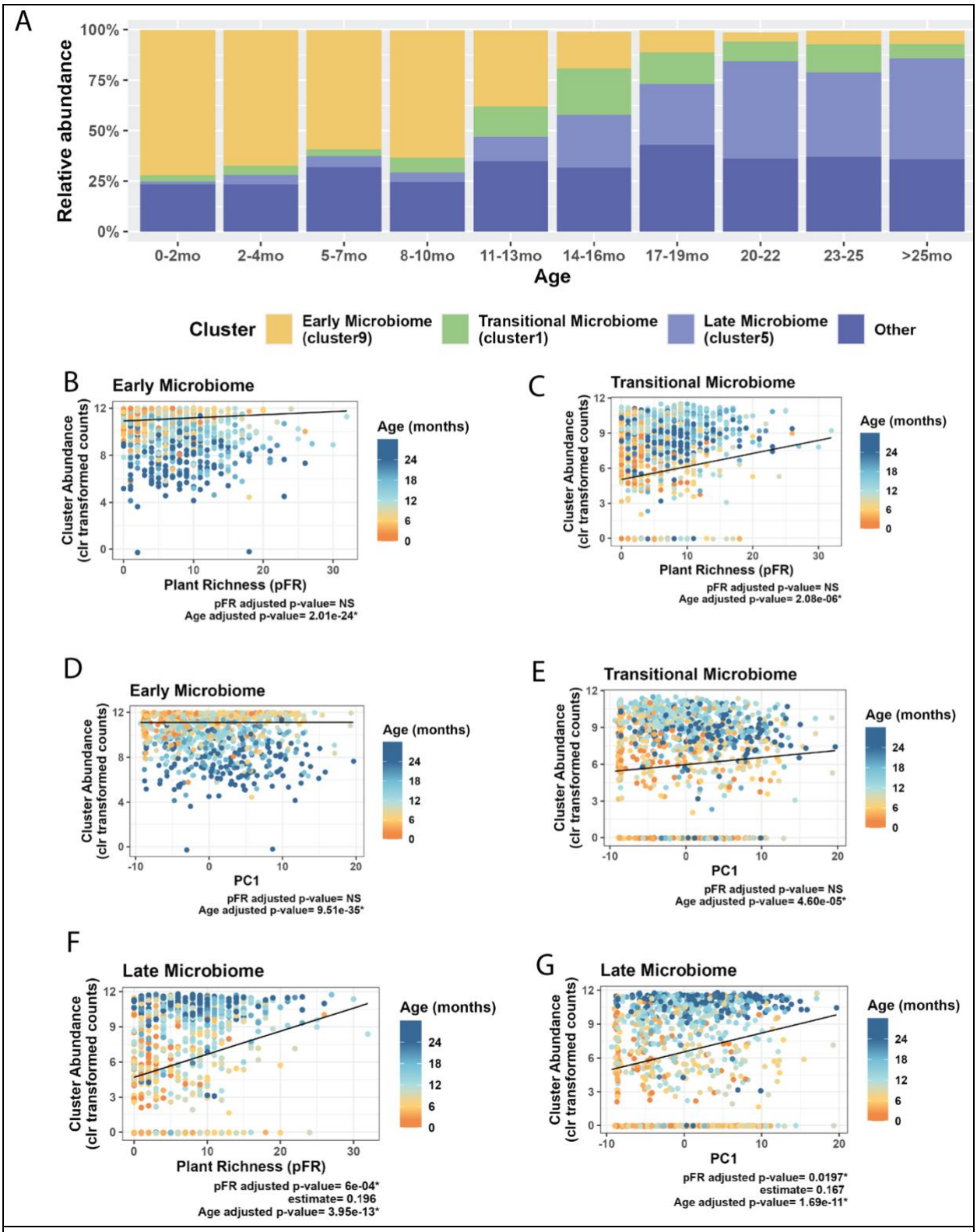

Fig. 4 Plant dietary diversity is correlated with the abundance of mature-microbiome marker bacteria. (A) Cluster abundance over time. (B-C) The abundance of the 'Early Microbiome' cluster, plotted against plant richness, (B) or PC1 (C). (D-E) The

abundance of the 'Transitional Microbiome' cluster, plotted against plant richness (D) and PC1 (E). (F-G) The abundance of the 'Late Microbiome' cluster, plotted against plant richness (F) and PC1 (G). For B-G, lines plot the coefficient derived from a generalized linear model (GLM) controlling for age and country.

Table S1: Cohort Characteristics	Cambodia	Kenya	Nicaragua	Pakistan	USA
Samples	178	228	457	29	144
Subjects	178	228	141	29	132
Age, Months (mean (Range, SD))	14.31 (1-28, 7.47)	22.44 (0-47, 13.28)	7.33 (2-25, 4.09)	13.86 (12-15, 0.58)	13.42 (2-35, 5.84)
Male (%)	95 (53.4)	111 (48.7)	227 (49.7)	17 (58.6)	76 (52.8)
WHZ (mean (SD))	-0.69 (0.98)	-0.92 (1.38)	0.64 (1.19)	-0.15(0.66)	NaN (NA)
HAZ (mean (SD))	-1.09 (2.35)	-0.42(2.41)	0.18 (1.20)	-0.99(1.21)	NaN (NA)
WAZ (mean (SD))	-1.11 (1.47)	-0.86 (1.82)	0.55 (1.02)	-0.59(0.91)	NaN (NA)
Received nutrition (RUTF or local alternative)	7 (4%)	70 (30.8%)	NA	NA	NA
National Human Development Index [57]	Low	Medium	Medium	Low	Very High
Urbanization	Rural	Urban	Urban	Rural	Urban/Suburban
Currently Breastfed (%)					
Don't Know	9 (5.1)	228 (100.0)	0 (0.0)	29 (100.0)	71 (49.3)
No	69 (38.8)	0 (0.0)	133 (29.1)	0 (0.0)	54 (37.5)
Yes	100 (56.2)	0 (0.0)	324 (70.9)	0 (0.0)	19 (13.2)
pFR (mean (SD))	9.37 (6.69)	6.20 (4.30)	4.48 (3.88)	9.90 (4.44)	8.90 (5.71)
Diarrhea (%)					
Don't Know	5 (2.8)	228 (100.0)	392 (85.8)	29 (100.0)	144 (100.0)
No	159 (89.3)	0 (0.0)	33 (7.2)	0 (0.0)	0 (0.0)
Yes	14 (7.9)	0 (0.0)	32 (7.0)	0 (0.0)	0 (0.0)

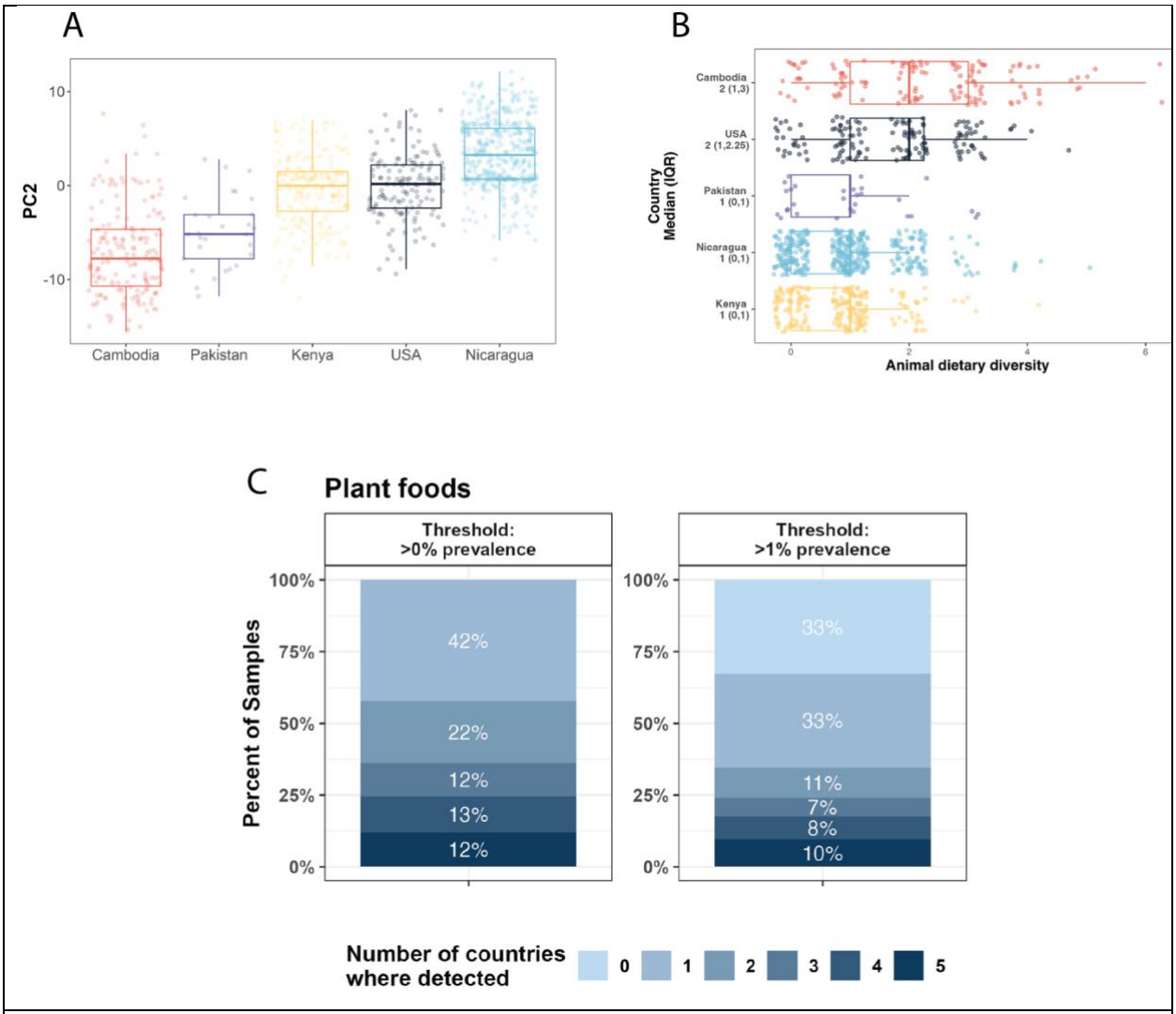

Fig. S1 (A) Boxplot of PC2 distribution by country. PC2 was significantly correlated with country (ANOVA, p-value <0.05). (B) Animal FoodSeq diversity by country. (C) Distribution of foods by how many countries they were each detected in. A food was counted as 'detected' if it was detected in at least 1 sample in that country (>0% threshold), or at least 1% of samples in that country (>1% threshold).

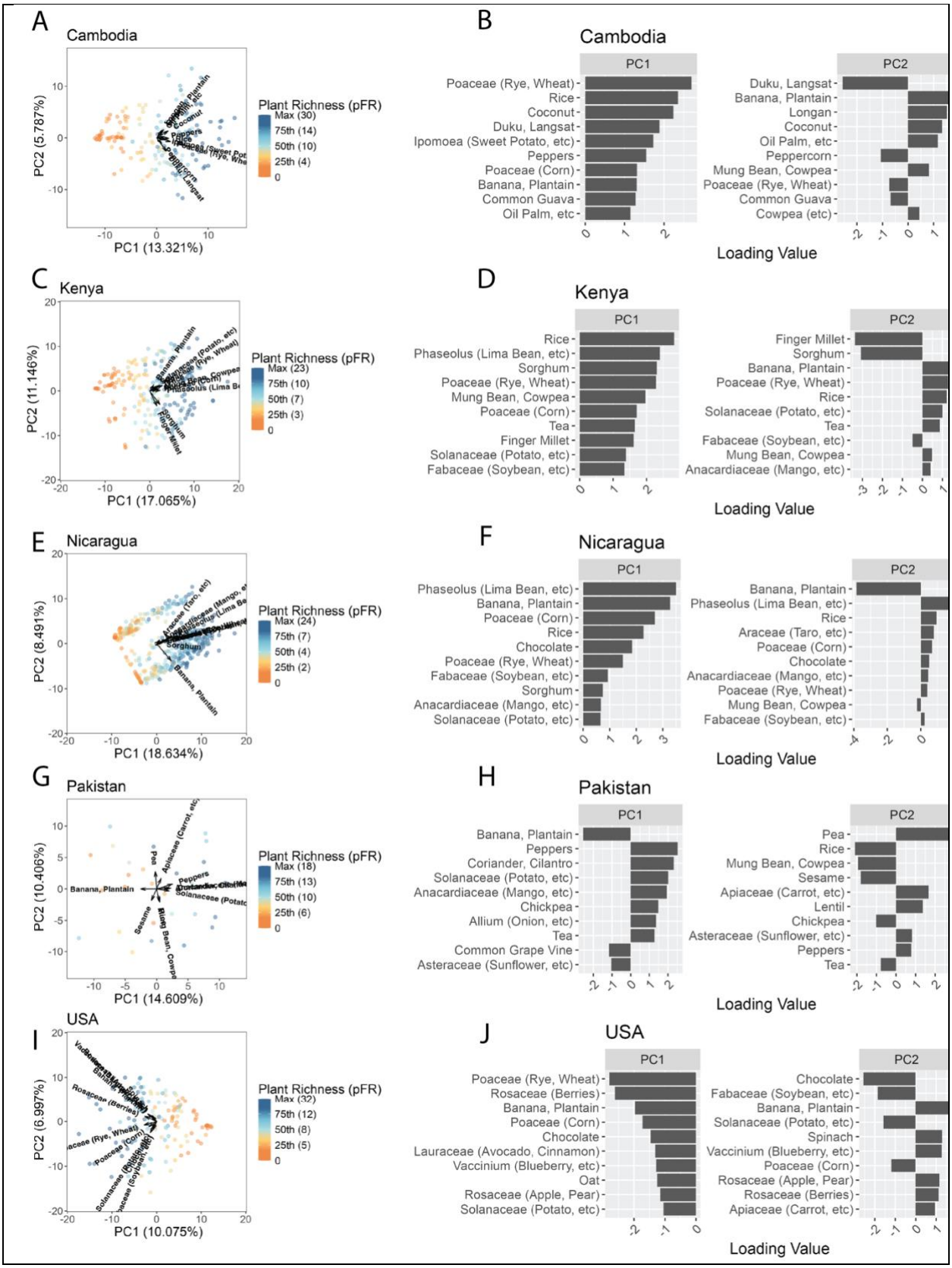
A
Cambodia
PC2 (5.787%)
PC1 (13.321%)
Plant Richness (pFR)
Max (30)
75th (14)
50th (10)
25th (4)
0
B
Cambodia
PC1
Poaceae (Rye, Wheat)
Rice
Coconut
Duku, Langsat
Ipomoea (Sweet Potato, etc)
Peppers
Poaceae (Corn)
Banana, Plantain
Common Guava
Oil Palm, etc
PC2
Duku, Langsat
Banana, Plantain
Longan
Coconut
Oil Palm, etc
Peppercorn
Mung Bean, Cowpea
Poaceae (Rye, Wheat)
Common Guava
Cowpea (etc)
Loading Value
C
Kenya
PC2 (11.146%)
PC1 (17.065%)
Plant Richness (pFR)
Max (23)
75th (10)
50th (7)
25th (3)
0
D
Kenya
PC1
Rice
Phaseolus (Lima Bean, etc)
Sorghum
Poaceae (Rye, Wheat)
Mung Bean, Cowpea
Poaceae (Corn)
Tea
Finger Millet
Solanaceae (Potato, etc)
Fabaceae (Soybean, etc)
PC2
Finger Millet
Sorghum
Banana, Plantain
Poaceae (Rye, Wheat)
Rice
Solanaceae (Potato, etc)
Tea
Fabaceae (Soybean, etc)
Mung Bean, Cowpea
Anacardiaceae (Mango, etc)
Loading Value
E
Nicaragua
PC2 (8.491%)
PC1 (18.634%)
Plant Richness (pFR)
Max (24)
75th (7)
50th (4)
25th (2)
0
F
Nicaragua
PC1
Phaseolus (Lima Bean, etc)
Banana, Plantain
Poaceae (Corn)
Rice
Chocolate
Poaceae (Rye, Wheat)
Fabaceae (Soybean, etc)
Sorghum
Anacardiaceae (Mango, etc)
Solanaceae (Potato, etc)
PC2
Banana, Plantain
Phaseolus (Lima Bean, etc)
Rice
Araceae (Taro, etc)
Poaceae (Corn)
Chocolate
Anacardiaceae (Mango, etc)
Poaceae (Rye, Wheat)
Mung Bean, Cowpea
Fabaceae (Soybean, etc)
Loading Value
G
Pakistan
PC2 (10.406%)
PC1 (14.609%)
Plant Richness (pFR)
Max (18)
75th (13)
50th (10)
25th (6)
0
H
Pakistan
PC1
Banana, Plantain
Peppers
Coriander, Cilantro
Solanaceae (Potato, etc)
Anacardiaceae (Mango, etc)
Chickpea
Allium (Onion, etc)
Tea
Common Grape Vine
Asteraceae (Sunflower, etc)
PC2
Pea
Rice
Mung Bean, Cowpea
Sesame
Apiaceae (Carrot, etc)
Lentil
Chickpea
Asteraceae (Sunflower, etc)
Peppers
Tea
Loading Value
I
USA
PC2 (6.997%)
PC1 (10.075%)
Plant Richness (pFR)
Max (32)
75th (12)
50th (8)
25th (5)
0
J
USA
PC1
Poaceae (Rye, Wheat)
Rosaceae (Berries)
Banana, Plantain
Poaceae (Corn)
Chocolate
Lauraceae (Avocado, Cinnamon)
Vaccinium (Blueberry, etc)
Oat
Rosaceae (Apple, Pear)
Solanaceae (Potato, etc)
PC2
Chocolate
Fabaceae (Soybean, etc)
Banana, Plantain
Solanaceae (Potato, etc)
Spinach
Vaccinium (Blueberry, etc)
Poaceae (Corn)
Rosaceae (Apple, Pear)
Rosaceae (Berries)
Apiaceae (Carrot, etc)
Loading Value

Fig. S2 Principal Component Analysis (PCA) run within each country (A,C,E,G,I), with the loadings for within-country PC1 and PC2 (B,D,F,H,J).

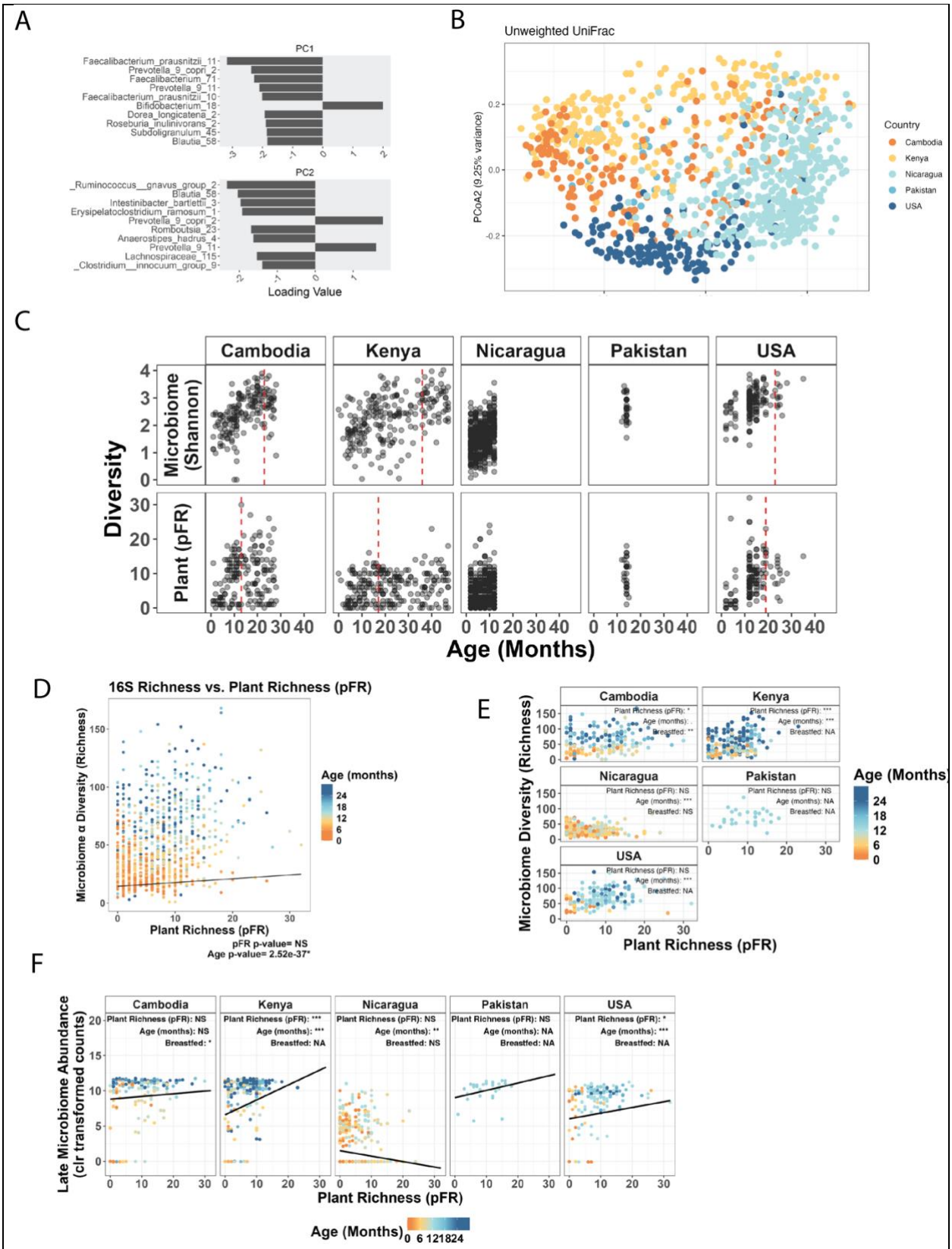

Fig. S3 (A) PCA loadings for 16S PCA (B) Unweighted Unifrac ordination (C) Plant and microbiome diversity versus age, with Pakistan and Nicaragua included. (D) Scatterplot

showing relationship between microbial richness and pFR. (E) Scatterplots showing relationships between microbial richness and pFR within country. (G) Scatterplots showing relationships between microbial Shannon diversity and pFR within country. For D-F, line plots the coefficient of pFR derived from a generalized linear model controlling for age (D-F) and country where applicable (D).

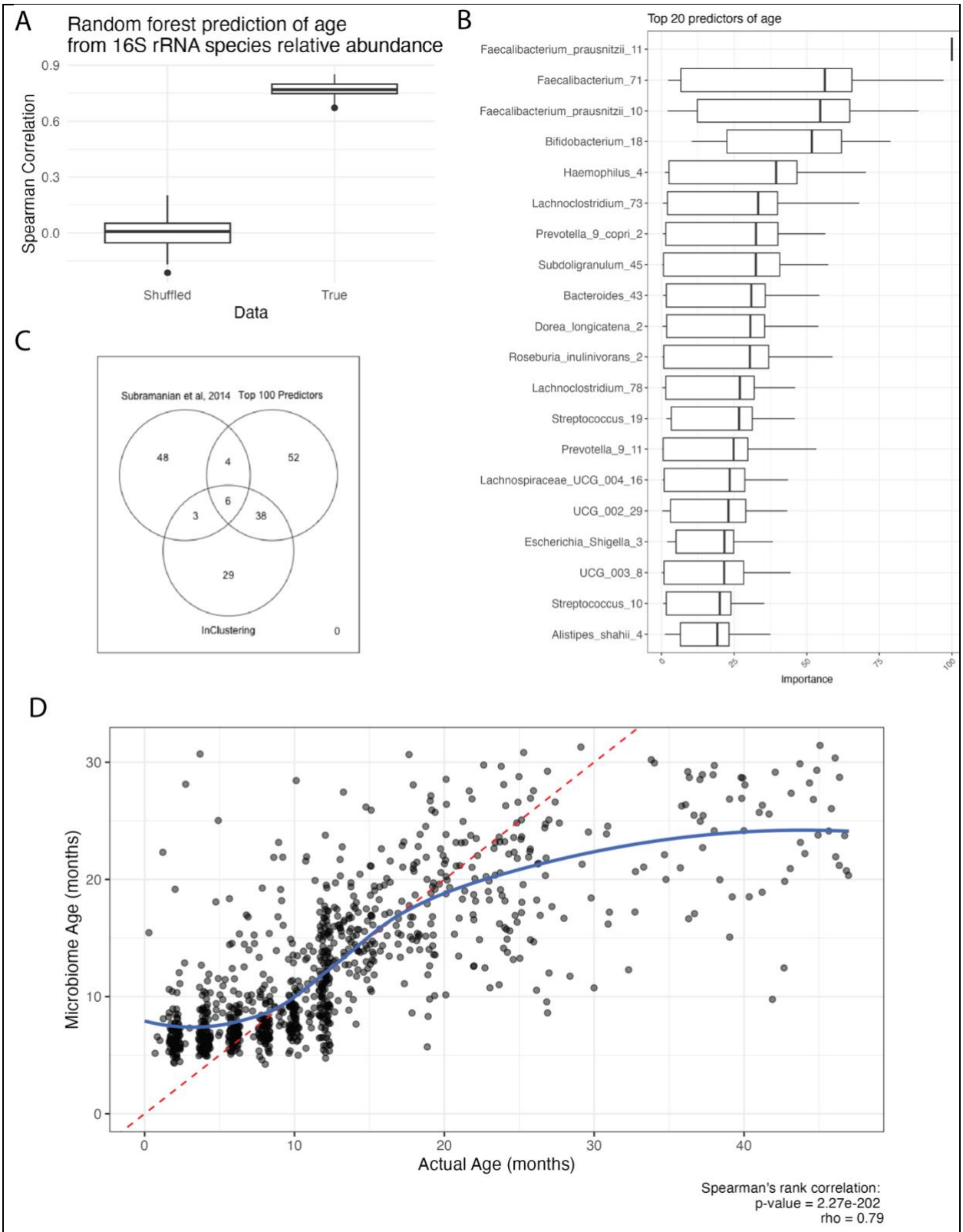

Fig. S4 (A) Model fit for classifier predicting age from shuffled or true 16S rRNA relative abundance data. (B) Top predictors from Random Forest model. (C) Exact sequence matches between the top 100 predictors in the Random Forest model, the taxa identified in the Early,

Transitional, and Late clusters, and a previous study that first employed this analysis. (D) Plot of microbiome age versus actual age. Red dashed line is a theoretical perfect concordance between the two.

1. Subramanian, S. *et al.* Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417–421 (2014).
2. Bäckhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 690–703 (2015).
3. Fahur Bottino, G. *et al.* Early life microbial succession in the gut follows common patterns in humans across the globe. *Nat. Commun.* **16**, 660 (2025).
4. Ajslev, T. A., Andersen, C. S., Gamborg, M., Sørensen, T. I. A. & Jess, T. Childhood overweight after establishment of the gut microbiota: the role of delivery mode, pre-pregnancy weight and early administration of antibiotics. *Int. J. Obes.* **35**, 522–529 (2011).
5. Bager, P., Wohlfahrt, J. & Westergaard, T. Caesarean delivery and risk of atopy and allergic disease: meta-analyses. *Clin. Exp. Allergy J. Br. Soc. Allergy Clin. Immunol.* **38**, 634–642 (2008).
6. Clausen, T. D. *et al.* Broad-Spectrum Antibiotic Treatment and Subsequent Childhood Type 1 Diabetes: A Nationwide Danish Cohort Study. *PloS One* **11**, e0161654 (2016).
7. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).
8. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
9. Krebs, N. F. *et al.* Effects of Different Complementary Feeding Regimens on Iron Status and Enteric Microbiota in Breastfed Infants. *J. Pediatr.* **163**, 416-423.e4 (2013).
10. Qasem, W. *et al.* Assessment of complementary feeding of Canadian infants: effects on microbiome & oxidative stress, a randomized controlled trial. *BMC Pediatr.* **17**, 54 (2017).

11. Tang, M., Frank, D., Hendricks, A., Ir, D. & Krebs, N. Protein Intake During Early Complementary Feeding Affects the Gut Microbiota in U.S. Formula-fed Infants (FS04-03-19). *Nutr. 2019 Abstr.* **3**, nzz048.FS04-03-19 (2019).
12. Plaza-Diaz, J. *et al.* Effects of Whole-Grain and Sugar Content in Infant Cereals on Gut Microbiota at Weaning: A Randomized Trial. *Nutrients* **13**, 1496 (2021).
13. Smith-Brown, P., Morrison, M., Krause, L. & Davies, P. S. W. Dairy and plant based food intakes are associated with altered faecal microbiota in 2 to 3 year old Australian children. *Sci. Rep.* **6**, 32385 (2016).
14. Smith-Brown, P., Morrison, M., Krause, L. & Davies, P. S. W. Microbiota and Body Composition During the Period of Complementary Feeding. *J. Pediatr. Gastroenterol. Nutr.* **69**, 726–732 (2019).
15. Thompson, A. L., Monteagudo-Mera, A., Cadenas, M. B., Lampl, M. L. & Azcarate-Peril, M. A. Milk- and solid-feeding practices and daycare attendance are associated with differences in bacterial diversity, predominant communities, and metabolic and immune function of the infant gut microbiome. *Front. Cell. Infect. Microbiol.* **5**, (2015).
16. Laursen Martin Frederik *et al.* Infant Gut Microbiota Development Is Driven by Transition to Family Foods Independent of Maternal Obesity. *mSphere* **1**, e00069-15 (2016).
17. Olm, M. R. *et al.* Robust variation in infant gut microbiome assembly across a spectrum of lifestyles. *Science* **376**, 1220–1223 (2022).
18. Kipnis, V. *et al.* Structure of Dietary Measurement Error: Results of the OPEN Biomarker Study. *Am. J. Epidemiol.* **158**, 14–21 (2003).
19. Subar, A. F. *et al.* Addressing Current Criticism Regarding the Value of Self-Report Dietary Data12. *J. Nutr.* **145**, 2639–2645 (2015).

20. What We Eat in America, NHANES 2017-March 2020 Prepandemic, individuals 2 years and over (excluding breast-fed children), day 1.
21. Burrows, T. L., Martin, R. J. & Collins, C. E. A Systematic Review of the Validity of Dietary Assessment Methods in Children when Compared with the Method of Doubly Labeled Water. *J. Am. Diet. Assoc.* **110**, 1501–1510 (2010).
22. Davies, P. S. W., Coward, W. A., Gregory, J., White, A. & Mills, A. Total energy expenditure and energy intake in the pre-school child: a comparison. *Br. J. Nutr.* **72**, 13–20 (1994).
23. Fisher, J. O. *et al.* Overestimation of infant and toddler energy intake by 24-h recall compared with weighed food records. *Am. J. Clin. Nutr.* **88**, 407–415 (2008).
24. Lanigan, J. A., Wells, J. C. K., Lawson, M. S. & Lucas, A. Validation of food diary method for assessment of dietary energy and macronutrient intake in infants and children aged 6–24 months. *Eur. J. Clin. Nutr.* **55**, 124–129 (2001).
25. Ravelli, M. N. & Schoeller, D. A. Traditional Self-Reported Dietary Instruments Are Prone to Inaccuracies and New Approaches Are Needed. *Front. Nutr.* **7**, (2020).
26. Wallace, A., Kirkpatrick, S. I., Darlington, G. & Haines, J. Accuracy of Parental Reporting of Preschoolers' Dietary Intake Using an Online Self-Administered 24-h Recall. *Nutrients* **10**, 987 (2018).
27. Data FAO/WHO GIFT | Global Individual Food Consumption Data Tool. *GIFT* https://www.fao.org/gift-individual-food-consumption/data/en.
28. Global Diet Quality Project. *Measuring What the World Eats: Insights from a New Approach.* https://doi.org/10.36072/dqq2022 (2022).

29. Petrone, B. L. *et al.* Diversity of plant DNA in stool is linked to dietary quality, age, and household income. *Proc. Natl. Acad. Sci.* **120**, e2304441120 (2023).

30. Loftus, T. Staple foods: what do people eat? in *Dimensions of Need* (Food and Agriculture organization of the United Nations, 1995).

31. Mank, I. *et al.* Dietary habits associated with growth development of children aged < 5 years in the Nouna Health and Demographic Surveillance System, Burkina Faso. *Nutr. J.* **19**, 81 (2020).

32. Steyn, N. P., Nel, J. H., Nantel, G., Kennedy, G. & Labadarios, D. Food variety and dietary diversity scores in children: are they good indicators of dietary adequacy? *Public Health Nutr.* **9**, 644–650 (2006).

33. Mezgebo, K., Lema, T. B. & Neela, S. Food variety, dietary diversity scores and dietary quality of complementary foods consuming by 6-24 months aged children in Jimma town, southwest Ethiopia. *Nutr. Food Sci.* **51**, 323–344 (2021).

34. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci.* **107**, 11971 (2010).

35. Abdelsalam, N. A., Hegazy, S. M. & Aziz, R. K. The curious case of Prevotella copri. *Gut Microbes* **15**, 2249152.

36. Johnson, K. V.-A. & Burnet, P. W. J. Microbiome: Should we diversify from diversity? *Gut Microbes* **7**, 455–458 (2016).

37. Starke, R., Capek, P., Morais, D., Callister, S. J. & Jehmlich, N. The total microbiome functions in bacteria and fungi. *J. Proteomics* **213**, 103623 (2020).

38. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

39. Branca, F. *et al.* Transforming the food system to fight non-communicable diseases. *BMJ* **364**, l296 (2019).

40. Arimond, M. & Ruel, M. T. Dietary Diversity Is Associated with Child Nutritional Status: Evidence from 11 Demographic and Health Surveys. *J. Nutr.* **134**, 2579–2585 (2004).

41. Kennedy, G. L., Pedro, M. R., Seghieri, C., Nantel, G. & Brouwer, I. Dietary diversity score is a useful indicator of micronutrient intake in non-breast-feeding Filipino children. *J. Nutr.* **137**, 472–477 (2007).

42. Roduit, C. *et al.* Increased food diversity in the first year of life is inversely associated with allergic diseases. *J. Allergy Clin. Immunol.* **133**, 1056-1064.e7 (2014).

43. Infant and young child feeding. https://www.who.int/data/nutrition/nlis/info/infant-and-young-child-feeding.

44. Lai, A. *et al.* Independent and combined effects of nutrition and sanitation interventions on enteric pathogen carriage and child growth in rural Cambodia: a factorial cluster-randomised controlled trial. 2021.05.21.21257546 Preprint at https://doi.org/10.1101/2021.05.21.21257546 (2022).

45. Song, R. *et al.* Sensitive and Feasible Specimen Collection and Testing Strategies for Diagnosing Tuberculosis in Young Children. *JAMA Pediatr.* **175**, e206069 (2021).

46. Becker-Dreps, S. *et al.* Gut Microbiome Composition in Young Nicaraguan Children During Diarrhea Episodes and Recovery. *Am. J. Trop. Med. Hyg.* **93**, 1187–1193 (2015).

47. Study Details | Project HOPE 1000: Health Outcomes Pregnancy Exposures | ClinicalTrials.gov. https://clinicaltrials.gov/study/NCT04417283.

48. Martin, V. M. *et al.* Longitudinal disease-associated gut microbiome differences in infants with food protein-induced allergic proctocolitis. *Microbiome* **10**, 154 (2022).

49. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).

50. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci.* **108**, 4516–4522 (2011).

51. Maurice, C. F., Haiser, H. J. & Turnbaugh, P. J. Xenobiotics Shape the Physiology and Gene Expression of the Active Human Gut Microbiome. *Cell* **152**, 39–50 (2013).

52. Petrone, B. L. *et al.* DNA metabarcoding captures dietary plant diversity in individuals and cohorts. 2022.06.13.22276343 Preprint at https://doi.org/10.1101/2022.06.13.22276343 (2022).

53. Shehzad, W. *et al.* Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (Prionailurus bengalensis) in Pakistan. *Mol. Ecol.* **21**, 1951–1965 (2012).

54. De Barba, M. *et al.* DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Mol. Ecol. Resour.* **14**, 306–323 (2014).

55. David, L. A. *et al.* Gut Microbial Succession Follows Acute Secretory Diarrhea in Humans. *mBio* **6**, 10.1128/mbio.00381-15 (2015).

56. Schneider, J. Comprehensive coverage of human last meal components revealed by a forensic DNA metabarcoding approach | Scientific Reports. *Sci. Rep.* **11**, (2021).

57. Nations, U. *Human Development Index*. *Human Development Reports* https://hdr.undp.org/data-center/human-development-index.