Early life microbial succession in the gut follows common patterns in humans across the globe

Guilherme Fahur Bottino[1], Kevin S. Bonham[1], Fadheela Patel[2], Shelley McCann[1], Michal Zieff[2], Nathalia Naspolini[3], Daniel Ho[4], Theo Portlock[4], Raphaela Joos[5,6], Firas S. Midani[7,8], Paulo Schüroff[3], Anubhav Das[5,6], Inoli Shennon[4], Brooke C. Wilson[4], Justin M. O'Sullivan[4], Robert A. Britton[7,9], Deirdre M. Murray[9], Mairead E. Kiely[9], Carla R. Taddei[10], Patrícia C. B. Beltrão-Braga[10], Alline C. Campos[11], Guilherme V. Polanczyk[12], Curtis Huttenhower[13], Kirsten A. Donald[2], Vanja Klepac-Ceraj[1#]

[#] Corresponding Author

[1]Department of Biological Sciences, Wellesley College, Wellesley, MA, USA
[2]University of Cape Town, Cape Town, Western Cape, ZAF
[3]School of Arts, Sciences and Humanity, University of São Paulo, São Paulo, SP, BRA
[4]The Liggins Institute, The University of Auckland, Auckland, NZL
[5]APC Microbiome Ireland, Cork, IRL
[6]School of Microbiology, University College Cork, Cork, IRL
[7]Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA
[8]Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine, Houston, TX, USA.
[9]INFANT Maternal and Child Health Centre, Dept of Paediatrics and Child Health, University College Cork, IRL.
[10]Microbiology Department, Institute of Biomedical Sciences (ICB-II), University of São Paulo, São Paulo, SP, BRA.
[11]Pharmacology of Neuroplasticity Lab- Department of Pharmacology, Ribeirão Preto Medical School- University of São Paulo, São Paulo, SP, BRA.
[12]Division of Child & Adolescent Psychiatry, Department & Institute of Psychiatry, Faculdade de Medicina FMUSP, Universidade de São Paulo, São Paulo, SP, BRA.
[13]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Running Title: *Global infant gut microbiome age model*

ORCIDs:

Guilherme Fahur Bottino:	0000-0003-1953-1576
Kevin S. Bonham:	0000-0003-3200-7533
Fadheela Patel:	0000-0001-5177-7416
Shelley McCann:	0000-0002-9753-7968
Michal Zieff:	0000-0001-9352-9947
Nathalia Naspolini:	0000-0002-0768-3284
Paulo Schuroff:	0000-0002-7808-7818
Daniel Ho	0000-0002-3699-0419
Inoli Shennon	0009-0004-3952-7894
Brooke C. Wilson	0000-0002-2213-542X
Theo Portlock	0000-0001-5971-3847
Anubhav Das	0009-0008-5264-8737
Raphaela Joos:	0000-0002-8124-7747
Firas S. Midani	0000-0002-2473-7758
Robert A. Britton:	0000-0001-8983-9539
Deirdre M. Murray:	0000-0002-2201-9912

Mairead E. Kiely:	0000-0003-0973-808X
Justin M. O'Sullivan:	0000-0003-2927-450X
Carla R. Taddei:	0000-0003-0869-0651
Patrícia C. B. Beltrão-Braga:	0000-0003-0004-1091
Aline C. Campos:	0000-0003-4258-3198
Guilherme V. Polanczyk:	0000-0003-2311-3289
Curtis Huttenhower:	0000-0002-1110-0096
Kirsten A. Donald:	0000-0002-0276-9660
Vanja Klepac-Ceraj:	0000-0001-5387-5706

Keywords:

microbiome age, infant gut microbiome, metagenomics, microbiome clock, succession, microbiome development

Abstract

Characterizing the dynamics of microbial community succession in the infant gut microbiome is crucial for understanding child health and development, but no normative model currently exists. Here, we estimate child age using gut microbial taxonomic relative abundances from metagenomes, with high temporal resolution (±3 months) for the first 1.5 years of life. Using 3,154 samples from 1,827 infants across 12 countries, we trained a random forest model, achieving a root mean square error of 2.61 months. We identified key taxonomic predictors of age, including declines in *Bifidobacterium* spp. and increases in *Faecalibacterium prausnitzii* and Lachnospiraceae. Microbial succession patterns are conserved across infants from diverse human populations, suggesting universal developmental trajectories. Functional analysis confirmed trends in key microbial genes involved in feeding transitions and dietary exposures. This model provides a normative benchmark of “microbiome age” for assessing early gut maturation that can be used alongside other measures of child development.

Introduction

The human gut microbiome is a complex ecosystem consisting of diverse microorganisms that interact with each other and form tight partnerships with their host. These are crucial for several physiological processes, including digestion, metabolism, and immune function[1]. The first major colonization event of an infant's gastrointestinal tract happens at birth, and microbial succession continues over the first few years of life[2,3]. Age-dependent aspects of this succession are shaped by a combination of natural history and environmental exposures, such as breastfeeding behavior and the introduction of solid food[4,5]. Altered colonization events, especially in early life, may have significant implications on a child's health, including the development of inflammatory disorders (e.g., allergies and asthma), metabolic disease (e.g., diabetes), neurocognitive outcomes, and other chronic conditions[6,7].

Specific microbial taxa tend to proliferate at different stages during early infancy[8]. Initial gastrointestinal tract colonizers include microorganisms capable of metabolizing human milk oligosaccharides or scavenging simple molecules[9]. The later introduction of a solid, complex and diverse diet brings opportunity for more fastidious colonizers and a more diverse community[10]. Recurring patterns of colonization and microbial succession across different life stages, from birth to late life and death[11–15], have shown consistent links between chronology and microbiome development.

These chronology-based approaches have been used to describe the phenotypic implications of an underdeveloped gut microbiome. Studies suggest that when the gut microbial community does not match the expected stage for a child's age, there can be significant health associations, particularly with growth and immune function[16,17]. This underdevelopment may respond to and contribute to a cycle of poor health and malnutrition, potentially affecting various aspects of the child's physiology and behavior[18,19]. To measure this temporal mismatch, two things are necessary: a reference developmental trajectory of the gut microbiome in early life and a way to measure a subject's deviation from such trajectory. One possible solution is to develop age estimation models using gut microbial communities sequenced across large and diverse cohorts. Those models can be trained to accurately produce an estimate of host age that can then be compared with the age at sample collection[17]. Following this approach, links between model outputs and health outcomes in childhood have been reported in multiple areas[20,21].

Despite showing promise, existing age models face several challenges to be applied in early childhood. Most existing models in this age range utilize data from 16S rRNA gene amplicon sequencing to estimate gut microbiome maturation[17,22] but this provides only a limited taxonomic resolution as closely related taxa are often binned together[23–25]. Most quantitative age models focus on aging[26–29] and span large age ranges that either exclude early childhood, or lack the necessary temporal resolution to produce meaningful predictions within the first year of life. Many models that account for early microbiome development with age do not produce a numeric age estimate, instead relying on unsupervised learning and qualitative predictions or associations[30]. Models also tend to be trained on individual cohorts and not validated on external populations, and cross-geographic analyses[31,32] have been lacking. In recent years, shotgun metagenomic sequencing data has become available from appropriately powered and diverse populations[3], but these datasets have not yet been incorporated into multi-site age models. Therefore, there is an opportunity and need to develop a comprehensive, global-scale quantitative age model focused on early childhood.

Here, we present such a model for age estimation developed using gut microbial taxonomic relative abundances, with high temporal resolution for the first 1.5 years of life. This model incorporates a large and geographically diverse population, comprising 3,154 shotgun-sequenced samples from 12 countries spanning Africa, Europe, Asia and America.

Results

Global metagenomes enable large-scale meta-analysis

We investigated developmental trajectories of the infant gut microbiome using a pooled dataset combining 3,154 stool samples sequenced with shotgun metagenomic sequencing from 1,827 healthy individuals obtained from 12 studies. The metagenomes spanned 12 countries from 4 continents (**Table 1, Fig. 1A**). All samples that matched inclusion criteria (see **Methods**) collected between ages 2-18 months (mean = 7.90 mo, SD = 3.99 mo) were incorporated into the model, resulting in a slight overrepresentation of younger samples (ages 2-4 months, **Fig. 1B**, Supplementary Fig. 1). Building the analysis dataset from a wide array of global sources enabled us to include a significant portion of data from low- and middle-income countries (LMICs), representing approximately 46 % of our total sample pool. The 1kD Wellcome LEAP effort contributed a total of 1,817 samples that have not been used previously in age-related studies. 427 of those samples were collected by the Khula study in South Africa[33] and have not been published before. These 1kD-LEAP samples are slightly younger (mean = 6.86 mo, SD = 3.55 mo), and the majority (80.57 %) are from LMICs.

Harmonized computational processing provides a continuous diversity landscape

After processing all sequence data using the same bioinformatics pipeline (BioBakery V3, **Fig. 1C**), we pooled all community profiles for the downstream analyses. To quantify the variation in gut microbial taxa associated with both age and data source, we used permutational analysis of variance (PERMANOVA) accounting for those factors (**Fig. 1D-E**). Sample group (source) and age explained 5.03% ($p = 0.001$) and 3.38% ($p = 0.001$) of the variance, respectively. In a multivariable analysis combining both factors, age still explained 2.28% ($p = 0.001$) of the variance after accounting for the data source contribution.

Pooled metagenomes predict age with high resolution

To assess the predictive potential of gut taxonomic profiles for the chronology of gut development, we trained a 5-fold cross-validated (CV) random forest (RF) model on features derived exclusively from the community composition obtained from shotgun metagenomic sequencing. Our inputs were the relative abundances of species present in at least 5% of samples, alongside the α-diversity estimated as the Shannon index. After removing samples with no reads assigned to at least one of the prevalence-filtered species, our analysis comprised 3,153 samples (~630 per fold) and 149 species. Our model targeted continuous age as a univariate regression output and generated validation-set predictions that reach a root mean square error of cross-validation (RMSECV) of 2.56 months (16.0% of the effective dynamic range, 64.1% of output SD) and a Pearson correlation of 0.803 with the ground truth values, on a 100x repeated 5-fold CV setting (**Fig. 2A**).

Changing taxa show feeding transitions and dietary exposures

To derive biological insight from the trained models, we analyzed the fitness-weighted variable importances on the cross-validated models, producing a list of top predictive features (**Fig. 2B**). The 35 highest ranking predictors (23.3% of inputs) were responsible for 70% of the cumulative weighted variable importance. Among those, 25 (71.4%) were positively correlated with age (mean $R_{(age)}$ = 0.18, SD = 0.12), with the remaining 10 (28.6)% negatively correlated to age (mean $R_{(age)}$ = -0.11, SD = 0.07, Supplementary Fig. 2). α-diversity measured as the Shannon index was the third most important predictor (4.86% of total importance, $R_{(age)}$ = +0.52, **Fig. 2C**). All but one of the top predictive taxa (97%) were present in every major cohort (*see* **Methods**), with only *Roseburia intestinalis* remaining undetected in the 1kDLEAP-M4EFaD samples. Additionally, there were several examples of site-biased or site-specific importances. For instance, *Dorea longicatena* and *Dorea formicigenerans* (**Fig. 2D**) were elevated in the South African cohort, and *Escherichia coli* (**Fig. 2E**) was elevated in the Brazilian cohort. Most of the top predictive taxa are species consistently prevalent across all cohorts, indicating that the relevant predictors are robust indicators of age across diverse populations, overcoming population-specific effects.

Across all cohorts, *Faecalibacterium prausnitzii* (**Fig. 2F**) and *Anaerostipes hadrus* were the taxa with the greatest importance scores for age prediction, accounting for 17.3% of the total weighted variable importance, together. Individually, those species positively correlate with age in our dataset (respectively, +0.41 and +0.32). The opposite trend is observed in another key group of predictors that include *Bifidobacterium longum* and *Bifidobacterium breve* (**Fig. 2G**), with 2.2% combined importance, exhibiting negative prior correlations with age (respectively, -0.14 and -0.14). The presence of certain species in the family Lachnospiraceae previously tied to developmental outcomes, such as *Ruminococcus gnavus* and *Blautia wexlerae*[34] is also noteworthy as a cluster of high-importance predictors of age. The former follows the same trend as the *Bifidobacterium* spp. (2.5% of total importance, $R_{(age)}$ = -0.063, p = 0.001), in agreement with previous studies[35].

Learned gut microbial patterns generalize across different sites

To evaluate the generalizability of our model across different data sources and test the predictive ability of each data source toward age, we performed a leave-one-datasource-out cross-validation (LOOCV) experiment. LOOCV yielded an average RMSE of leave-one-out cross-validation of 3.03 +- 0.63 months (Supplementary Table 1, Supplementary Fig. 3). We hypothesized that this generalizability resulted from combined effects from abundance trends and underlying prevalence trends (**Fig. 2D-G**, Supplementary Fig. 2). This would mean that predictors would be important, in part, because they would appear and disappear from the infant gut following similar trends, regardless of geographical origin.

By grouping a subset of our samples by location - Baltic countries, United States and South Africa - and binning them by age (in months), we computed monthly prevalences for the 34 top taxonomic predictors of gut chronology. Strikingly similar patterns of succession emerged between all tested locations (**Fig. 3**), evidenced by whole-matrix mean prevalence correlations: Baltic/USA = 0.799; USA/SA = 0.750; Baltic/SA = 0.749. This consistency suggests that many of the succession patterns identified by our model are likely universal, transcending local environmental influences.

Hierarchical cluster analysis of the binned prevalence time series revealed one large universal cluster of species and succession patterns containing 18 (53%) of the top 34 taxonomic predictors, which correlated highly between sites, along with smaller clusters of decentralized patterns. Representatives of the larger, common core are species as mentioned above, such as *F. prausznitzii,* positively correlated with the outcome on all three cases, alongside early colonizers such as *E. coli* (1.3% of total importance, R = -0.25 with age), that follow the opposite pattern consistently on the three sites. Among the divergent cluster, besides the aforementioned *Dorea* genus (*D. longicatena* and *D. formigicerans,* 2.8% combined importance) in South Africa, we identified taxa such as *Prevotella copri* (0.9% of total importance, R = +0.22 with age), which exhibit distinct abundance and prevalence patterns between westernized and non-westernized populations[36].

Enzyme changes in the first year corroborate prior studies

We hypothesized that, as was the case with taxonomic composition, the functional composition in terms of microbial metabolic enzymes would change similarly between sites. Utilizing longitudinal samples in the South African cohort, we measured the consistency of the direction of EC abundance transitions between earlier and later samples from the same subject using a Transition Score (TS, *see* ***Methods***). We then selected the top hits in both directions - later enrichment (highest scores) and later depletion (lowest scores), and stratified their abundances into the corresponding top predictive taxa (**Fig. 4**).

The lowest-scored EC (decreasing on most subjects) was transaldolase (2.2.1.2), with a TS of -0.84 and a variation of -86.74 ± 11.46 counts per million reads (CPM). It is followed by succinate-CoA ligase (ADP-forming, 6.2.1.5) and pyridoxal kinase (2.7.1.35), both with a TS of -0.81 and variations of -119.89 ± 20.15 CPM and -67.40 ± 11.44, respectively. The expanded list of stratified ECs decreasing in abundance with age was dominated by functions associated with *B. longum, B. breve, R. gnavus* and *E. coli*, consistent with the aforementioned depletion of those species along the first year of life. That group of species and the highlighted functions account for a consistent average fold change of -0.46 ± 0.01 $\log_{10}$ CPM between younger and older samples.

The highest-scored ECs (increasing on most subjects) were [ribosomal protein S12] (aspartate(89)-C(3))-methylthiotransferase (2.8.4.4, TS = +0.84, Δ = +53.89 ± 9.49 CPM), and coproporphyrinogen dehydrogenase (1.3.99.22, TS = +0.79, Δ = +31.54 ± 5.18 CPM). Stratification of the ECs that increase in abundance with age is more diverse, and contains ECs assigned to a wider array of fastidious anaerobes: *F. prausnitzii*, *A. hadrus*, *B. wexlerae*, *Blautia obeum, D. longicatena* and *P. copri*. Combined, highlighted functions assigned to those species exhibit an average fold change of +0.99 ± 0.10 $\log_{10}$ CPM between younger and older samples.

When compared to the results published by Vatanen and colleagues[37], our list of the top 1.5% increasing or decreasing ECs (**Fig. 4**) contains 11 (27.5%) of the previously-reported transitioning ECs. This overlap between the results happened on both major trend clusters, as exemplified by the previously reported decreases in ribokinase (2.7.1.15, TS = -0.73, Δ = -155.44 ± 22.25 CPM) and transaldolase or the increase in 6-phosphofructokinase (2.7.1.11, TS = +0.59, Δ = +102.66 ± 25.10 CPM). Furthermore, we identified transitioning ECs not previously reported. In this group of novel ECs, notable variations were the decrease in pyridoxal kinase and the increase in malate dehydrogenase (1.1.1.40, TS = 0.66, Δ = +39.62 ± 8.50 CPM).

Discussion

In this study, we show that the succession of a small number of key taxa in the early-life gut microbiome follows common patterns, even across various geographical and socioeconomic settings. These patterns are strong and consistent enough to be learned by our microbiome age model, allowing it to generalize beyond individual cohort boundaries. One of the main reasons why we were able to build such a robust model was our large-scale pooling strategy, which enabled us to sample diverse backgrounds in, for example, dietary practices and diet composition, an exposure strongly reflected on the learned patterns. As a result, we captured a broad and representative spectrum of microbial profiles, enhancing the robustness of our model towards regional variations, considered a key obstacle to the generalization of microbiome-based models for a variety of phenotypes[38].

Most studies to date characterized microbiome age using taxonomic classifications from amplicon sequencing of the 16S rRNA gene. Some of the limitations associated with this sequencing technique are the biases introduced by the choice of primers and target region for the experiment, and substantially reduced taxonomic resolution[23–25]. In our work, by building a model using well-defined species identified by metagenomic sequencing, rather than solely relying on 16S rRNA sequencing, we leveraged the ability of the metagenomic approach to sample all genes in a complex sample. The bacterial genes themselves are too highly dimensional and sparse to act as raw simultaneous inputs to multivariable predictive models, but, when processed, allow for the identification of a broader array of taxa at a higher resolution when compared with the depth of information offered by 16S rRNA gene sequencing[23]. Additionally, through the identification of the functional pathways to which those genes belong, we can get a better understanding of how the functional repertoire of the microbial communities evolved with age.

Importance analysis of the fitted random forest models revealed that the main age predictors were the taxa involved in the microbiome's natural succession influenced by key events such as changes in diet. For example, *F. prausnitzii* and *A. hadrus are* important age predictors in the first two years of life. Those taxa are butyric acid producers[39] that usually appear with the cessation of breastfeeding, which marks the transition to a Firmicutes-dominated gut characterized with increased production of short-chain fatty acids (SCFA)[40,41]. The same phenomenon explains the learned importance of known metabolizers of human milk oligosaccharides, namely *Bifidobacterium* spp.[42], characteristic of the early stages of infancy, especially in locations where exclusive breastfeeding is prevalent. Alongside these taxa, the Shannon index (alpha diversity) also emerged as an important predictor. This was expected, as microbial diversity in the gut increases with age in early infancy[25]. Many of the top predictive taxa showed similar succession patterns during the first 13 months of life (**Fig. 3**) across all tested geographical sites (USA, Europe, South Africa), despite significant socioeconomic differences. This suggests that there is a strong, consistent, and machine-learnable pattern for determining age based on microbial succession, regardless of metadata variations, among the geographical sites tested in this work.

Our study corroborates a significant portion of the results from a previous study[37] that also examined temporal transitions in ECs in early life. This implies that age-determining taxa and their functions are consistent across different microbial communities, even with the diverse lifestyles and ethnic backgrounds of the several cohorts sampled[32]. The ECs that showed most change were primarily involved in central carbohydrate metabolisms, many of which are associated with bifidobacteria. For example, *B. breve* utilizes ribokinase (2.7.1.15) to harvest ribose as a carbon source in the early gut[43], and several *Bifidobacterium* spp. have transaldolase (2.2.1.2)[44,45]. The presence of glycolytic and pentose-phosphate cycle enzymes supports the idea that diet-related transitions, particularly those tied to the intake of complex carbohydrates, are major drivers of age-determining patterns. In this context, one enzyme of particular interest is pyridoxal kinase (2.7.1.35), which plays a role in the GABA synthesis pathway typical of bifidobacteria[46]. Notably, GABA concentrations in infant stool have been associated with behavioral traits in early infancy[47,48]. Our findings suggest a specific functional link of this association between GABA and *Bifidobacterium* spp. that is also related to age, highlighting a pathway that can be a strong candidate for studying behavioral outcomes in the first year of life.

Despite the strong benchmarks reported by our models, there are several limitations that future studies need to address. For example, one key decision in our model development was to exclude all additional participant and biospecimen metadata, using only participant age and microbial data. This decision was made due to the lack of uniformity in metadata collection and annotation across studies. However, previous studies have shown that metadata such as feeding practices[14], socioeconomic status[49], delivery mode and gestational age[50] can enhance the predictive power in microbial-based models. Notably, in our case, including these covariables would have resulted in a significant loss of samples due to missing metadata, which would have compromised the model's generalizability and made comparative benchmarks unfeasible. Another area of improvement would be to incorporate season as an external effect to model the time-serial succession patterns, accounting[51] for different hemispheres. It is also worth mentioning that, even though there are many reference genomes for the early-life gut microorganisms, detailed information on their functions and biochemical characteristics is still biased toward a few well-characterized microorganisms[52]. While we were able to corroborate findings from Vatanen et al. (2018) despite the time gap between the studies, this may partly be due to the limited characterization of the annotated functional space.

Studying developmental changes associated with dynamic processes can be challenging without benchmarks or standards that provide expected ranges of values. Given the high dimensional and highly dynamic nature of microbial composition, simple standards such as those used in anthropometrics (e.g., age-standardized Z-scores for length or weight in infants) are not feasible, and studying microbial associations with child development has been challenging without such an agreed upon normative developmental trajectory. The microbiome-age model provided here, built from a diverse and global population of human children provides a model of development that may be deployed to advance our understanding of the gut microbiome in child growth and flourishing.

Methods

Sample collection and processing for the Khula cohort

Participants and study design

Infants were recruited from local community clinics in Gugulethu, an informal settlement in Cape Town, South Africa as part of an ongoing longitudinal study (most of the enrollment happened prenatally with 38.82% of infants enrolled shortly after birth[33]). The first language of the majority of residents in this area is Xhosa. Study procedures were offered in English or Xhosa depending on the language preference of the mother. This study was approved by the Health Research Ethics Committees (study number: 666/2021). Informed consent was collected from mothers on behalf of themself and their infants. Demographic information, including maternal place of birth, primary spoken language, maternal age at enrollment, maternal educational attainment, and maternal income, was collected at enrollment (**Table 2**).

Families were invited to participate in three in-lab study visits over their infant’s first 18 months of life. At the first in-lab study visit (hereafter Visit 1), which took place when the infants were between approximately 2 and 6 months of age (M=3.63, SD=0.78, range=2.13-5.34), the following data were collected: the infants' age (in months), sex, and infant stool samples. At the second study visit (hereafter Visit 2), occurring when infants were between approximately 6 months and 12 months of age (age in months: M=8.77, SD=1.39, range=5.38-11.90) and at the third study visit (hereafter Visit 3), occurring when infants were between approximately 12 months and 17 months of age (age in months: M=14.01, SD=1.31, range=11.63-17.97), infant stool samples were collected again. At visits where infants could not donate stool samples on the same day, samples were collected on different days close to the visit date.

Sample collection

Stool samples (n=427) were collected in the clinic by the research assistant directly from the diaper and transferred to the Zymo DNA/RNA Shield^{TM} Fecal collection Tube (#R1101, Zymo Research Corp., Irvine, USA) and immediately frozen at -80 ˚C. Stool samples were not collected if the subject had taken antibiotics within the two weeks prior to sampling.

DNA extraction

DNA was extracted at the Medical Microbiology Department, University of Cape Town, South Africa from stool samples collected in DNA/RNA Shield™ Fecal collection tube using the Zymo Research Fecal DNA MiniPrep kit (# D4300, Zymo Research Corp., Irvine, USA) following manufacturer’s protocol. To assess the extraction process's quality, ZymoBIOMICS® Microbial Community Standards (#D6300 and #D6310, Zymo Research Corp., Irvine, USA) were incorporated and subjected to the identical process as the stool samples. The DNA yield and purity were determined using the NanoDrop® ND -1000 (Nanodrop Technologies Inc. Wilmington, USA).

Sequencing

Shotgun metagenomic sequencing was performed on all samples at the Integrated Microbiome Research Resource (IMR, Dalhousie University, NS, Canada). A pooled library (max 96 samples per run) was prepared using the Illumina Nextera Flex Kit for MiSeq and NextSeq from 1 ng of each sample. Samples were then pooled onto a plate and sequenced on the Illumina NextSeq 2000 platform using 150+150 bp paired-end P3 cells, generating on average 24M million raw reads and 3.6 Gb of sequence per sample[53].

Public metagenomic data acquisition

Publicly available metagenome metadata was obtained from the CuratedMetagenomicsData database[54]. Database entries considered for inclusion were those annotated as stool samples on the “body_site” property, pertaining to subjects identified as either “newborn” or “child” on the “age_category” property and containing a valid numeric “infant_age” annotation in days. From that set, samples identified as belonging to premature-born children were excluded. We also excluded samples belonging to children suffering from acute infectious conditions - including sepsis - at the time of sample collection. Future T1D-annotated samples, however, (3.9% of the CMD-DIABIMMUNE samples) were not excluded. For the three DIABIMMUNE cohorts, complementary metadata containing harmonized annotation was gathered from the DIABIMMUNE study website and merged with the original set. Sequence data was then downloaded from originally referenced data repositories (**Table 1**).

Computational processing, analyses and statistics

Metagenome processing

For the 1kDLEAP-Khula cohort samples, raw metagenomic sequence reads (Mean = 20.19M, SD = 6.75M reads per sample) were processed using tools from the bioBakery suite, following already-established protocols [55]. Initially, KneadData v0.10.0 was employed with default settings to trim low-quality reads and eliminate human sequences, using the hg37 reference database. Subsequently, MetaPhlAn v3.1.0, utilizing the mpa_v31_CHOCOPhlAn_201901 database, was applied with default parameters to map microbial marker genes and generate taxonomic profiles. The taxonomic profiles, along with the same reads obtained in the initial step, were then processed with HUMAnN v3.7 to produce stratified functional profiles. Utilizing this pipeline, the 1kDLEAP-Khula, the ECHO-Resonance[34] (Mean = 9.34M, SD = 6.75M reads per sample) and the CMD sequence reads (Mean = 15.35M, SD = 13.72M reads per sample) were processed at Wellesley College, USA; the 1kDLEAP-Germina (Mean = 8.32M, SD = 6.48M reads per sample) sequences were processed at the University of Sao Paulo, Brazil; the 1kDLEAP-Combine (Mean = 8.32M, SD = 6.48M reads per sample) sequences were processed at the APC Microbiome Ireland, Ireland; and the 1kDLEAP-M4EFaD (Mean = 41.45M, SD = 6.63M reads per sample) sequences were processed at the Liggins Institute, New Zealand.

Sample pooling

Samples were pooled into the same collective dataset and were annotated to differentiate their original data source. For the 4 Wellcome LEAP 1kD studies, every individual study became one separate annotated data source. ECHO-Resonance samples were also annotated as their individual data source. For simplification purposes in downstream analysis, all the CMD-derived samples were annotated as belonging to the same meta-datasource, “CMD.” In analyses that warranted a higher degree of discrimination, we divided this meta-group into two meta-subgroups, “CMD-DIABIMMUNE” (containing 642 samples from Vatanen et al.[56], Kostic et al.[57]

and Yassour et al.[58]) and “CMD-OTHER” (containing 471 samples from Asnicar et al.[59], Bäckhed et al.[3], Pehrsson et al.[60], Shao et al.[61]).

Microbial community analysis

Computational analysis was conducted using the Julia programming language[62]. Microbial community profiles (taxonomic and functional) were parsed and processed using the BiobakeryUtils.jl and Microbiome.jl packages[63]. Principal coordinates analysis (PCoA) with the Bray-Curtis dissimilarity was calculated for all pairs of samples, focusing on species-level classifications, using Distances.jl. Classical multidimensional scaling (MDS) was then performed on the dissimilarity matrix with MultivariateStats.jl. Additionally, permutational analysis of variance (PERMANOVA) was conducted using PERMANOVA.jl.

Machine Learning

Machine learning analysis was performed using the MLJ.jl package[64] and the associated framework. Random forest regression utilized the backend from the DecisionTree.jl package[65]. Linear bias correction was applied to forest outputs when necessary[66] using GLM.jl[67]. Data visualization was built using the Makie.jl package[68].

Functional Analysis

EC abundance profiles were obtained for each subject of the 1kDLEAP-Khula cohort that had longitudinal samples collected on the 3 month and 12 month timepoints, for a total of 73 sample pairs. Only ECs that could be assigned to at least one detected species were analyzed. ECs were then assigned a transition score (TS) to represent the directionality and consistency of the change in its abundance between the timepoints. For each EC, the TS score was calculated according to the following expression:

$$TS = \frac{\sum_{i=1}^{n} \square \left(p_i \ \ sgn(a_i^{12mo} - a_i^{3mo})\right)}{n}$$

where *n* is the total number of samples; $sgn(a_i^{12mo} - a_i^{3mo})$ is the sign of the difference in community-wide enzyme abundance for the *i*th sample pair between the 12mo and 3mo timepoints; and p_i is a factor that controls for the significance of the EC abundance in either timepoint, according to the expression:

$$p_i := \ (1\ if\ ((a_i^{3mo}\ \ >=\ 10\ CPM)\ or\ (a_i^{12mo}\ \ >=\ 10\ CPM));\ 0\ otherwise)$$

A score close to +1.0 means that the enzyme is consistently increasing from 3 to 12 months, and a score close to -1.0 means that the enzyme is consistently decreasing from 3 to 12 months. After scoring and ranking the ECs, we selected 1.5% of the total scored functions (48 ECs) equally distributed between the highest and lowest-scoring enzymes (24 in each major trend cluster) for stratified functional analysis and visualization.

Data availability

The processed datasets generated and/or analyzed during the current study have been deposited in Data Dryad under DOI: https://doi.org/10.5061/dryad.dbrv15f9z. The raw sequencing data for the Khula study have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA1128723. All other relevant data supporting the key findings of this study and instruction on how to obtain it are available within the article and its Supplementary Information files, or are available from the corresponding author upon reasonable request.

Code availability

Information for replicating the package environment and code for data analysis and figure generation, as well as scripts for automated download of input files, are available on GitHub at https://github.com/Klepac-Ceraj-Lab/MicrobiomeAgeModel2024 and archived on Zenodo under DOI: https://zenodo.org/doi/10.5281/zenodo.12822332.

References

1. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res.* **30**, 492–506 (2020).

2. Milani, C. *et al.* The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health Implications of the Infant Gut Microbiota. *Microbiol. Mol. Biol. Rev.* (2017) doi:10.1128/mmbr.00036-17.

3. 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).

4. Bolte, E. E., Moorshead, D. & Aagaard, K. M. Maternal and early life exposures and their potential to influence development of the microbiome. *Genome Med.* **14**, 4 (2022).

5. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).

6. Durack, J. *et al.* Delayed gut microbiota development in high-risk for asthma infants is temporarily modifiable by Lactobacillus supplementation. *Nat. Commun.* **9**, 707 (2018).

7. Stokholm, J. *et al.* Publisher Correction: Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 704 (2018).

8. Beller, L. *et al.* Successional Stages in Infant Gut Microbiota Maturation. *MBio* **12**, e0185721 (2021).

9. Dawod, B., Marshall, J. S. & Azad, M. B. Breastfeeding and the developmental origins of mucosal immunity: how human milk shapes the innate and adaptive mucosal immune systems. *Curr. Opin. Gastroenterol.* **37**, 547–556 (2021).

10. McKeen, S. *et al.* Adaptation of the infant gut microbiome during the complementary feeding transition. *PLoS One* **17**, e0270213 (2022).

11. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4578–4585 (2011).

12. Feng, L. *et al.* Identifying determinants of bacterial fitness in a model of human gut microbial succession. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 2622–2633 (2020).

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

14. Stewart, C. J. *et al.* Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).

15. Odamaki, T. *et al.* Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol.* **16**, 90 (2016).

16. Hoskinson, C. *et al.* Delayed gut microbiota maturation in the first year of life is a hallmark of pediatric allergic disease. *Nat. Commun.* **14**, 1–14 (2023).

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

18. Robertson, R. C. *et al.* The gut microbiome and early-life growth in a population with high prevalence of stunting. *Nat. Commun.* **14**, 1–15 (2023).

19. Fontaine, F., Turjeman, S., Callens, K. & Koren, O. The intersection of undernutrition, microbiome, and child development in the first years of life. *Nat. Commun.* **14**, 1–9 (2023).

20. Schoch, S. F. *et al.* From Alpha Diversity to Zzz: Interactions among sleep, the brain, and gut microbiota in the first year of life. *Prog. Neurobiol.* **209**, 102208 (2022).

21. Shen, W. *et al.* Postnatal age is strongly correlated with the early development of the gut microbiome in preterm infants. *Transl Pediatr* **10**, 2313–2324 (2021).

22. Wernroth, M.-L. *et al.* Development of gut microbiota during the first 2 years of life. *Sci. Rep.* **12**, 9080 (2022).

23. Durazzi, F. *et al.* Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Sci. Rep.* **11**, 3030 (2021).

24. Brumfield, K. D., Huq, A., Colwell, R. R., Olds, J. L. & Leddy, M. B. Microbial resolution of whole genome shotgun and 16S amplicon metagenomic sequencing using publicly

available NEON data. *PLoS One* **15**, e0228899 (2020).

25. Peterson, D. *et al.* Comparative Analysis of 16S rRNA Gene and Metagenome Sequencing in Pediatric Gut Microbiomes. *Front. Microbiol.* **12**, 670336 (2021).

26. Wang, H. *et al.* A gut aging clock using microbiome multi-view profiles is associated with health and frail risk. *Gut Microbes* **16**, 2297852 (2024).

27. Chen, Y. *et al.* Human gut microbiome aging clocks based on taxonomic and functional signatures through multi-view learning. *Gut Microbes* **14**, 2025016 (2022).

28. Galkin, F. *et al.* Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. *iScience* **23**, 101199 (2020).

29. Huang, S. *et al.* Human Skin, Oral, and Gut Microbiomes Predict Chronological Age. *mSystems* **5**, (2020).

30. Eggesbø, M. *et al.* Development of gut microbiota in infants not exposed to medical interventions. *APMIS* **119**, 17–35 (2011).

31. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).

32. Olm, M. R. *et al.* Robust variation in infant gut microbiome assembly across a spectrum of lifestyles. *Science* **376**, 1220–1223 (2022).

33. Zieff, M. R. *et al.* Characterizing developing executive functions in the first 1000 days in South Africa and Malawi: The Khula Study [version 1; peer review: 2 approved with reservations]. *Wellcome Open Research* **9**, (2024).

34. Bonham, K. S. *et al.* Gut-resident microorganisms and their genes are associated with cognition and neuroanatomy in children. *Sci Adv* **9**, eadi0497 (2023).

35. Sagheddu, V., Patrone, V., Miragoli, F., Puglisi, E. & Morelli, L. Infant Early Gut Colonization by Lachnospiraceae: High Frequency of Ruminococcus gnavus. *Front Pediatr* **4**, 57 (2016).

36. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative

study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14691–14696 (2010).

37. Vatanen, T. *et al.* The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).

38. He, Y. *et al.* Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).

39. Singh, V. *et al.* Butyrate producers, 'The Sentinel of Gut': Their intestinal significance with and beyond butyrate, and prospective use as microbial therapeutics. *Front. Microbiol.* **13**, 1103836 (2022).

40. Tsukuda, N. *et al.* Key bacterial taxa and metabolic pathways affecting gut short-chain fatty acid profiles in early life. *ISME J.* **15**, 2574–2590 (2021).

41. Walker, A. W., Duncan, S. H., McWilliam Leitch, E. C., Child, M. W. & Flint, H. J. pH and peptide supply can radically alter bacterial populations and short-chain fatty acid ratios within microbial communities from the human colon. *Appl. Environ. Microbiol.* **71**, 3692–3700 (2005).

42. Kitaoka, M. Bifidobacterial enzymes involved in the metabolism of human milk oligosaccharides. *Adv. Nutr.* **3**, 422S–9S (2012).

43. Pokusaeva, K. *et al.* Ribose utilization by the human commensal Bifidobacterium breve UCC2003. *Microb. Biotechnol.* **3**, 311–323 (2010).

44. Klaassens, E. S., de Vos, W. M. & Vaughan, E. E. Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Appl. Environ. Microbiol.* **73**, 1388–1392 (2007).

45. Requena, T. *et al.* Identification, detection, and enumeration of human bifidobacterium species by PCR targeting the transaldolase gene. *Appl. Environ. Microbiol.* **68**, 2420–2427 (2002).

46. Ham, S. *et al.* Gamma aminobutyric acid (GABA) production in Escherichia coli with

pyridoxal kinase (pdxY) based regeneration system. *Enzyme Microb. Technol.* **155**, 109994 (2022).

47. Zuffa, S. *et al.* Early-life differences in the gut microbiota composition and functionality of infants at elevated likelihood of developing autism spectrum disorder. *Transl. Psychiatry* **13**, 257 (2023).

48. Laue, H. E., Korrick, S. A., Baker, E. R., Karagas, M. R. & Madan, J. C. Prospective associations of the infant gut microbiome and microbial function with social behaviors related to autism at age 3 years. *Sci. Rep.* **10**, 15515 (2020).

49. Lewis, C. R. *et al.* Family SES Is Associated with the Gut Microbiome in Infants and Children. *Microorganisms* **9**, (2021).

50. Toubon, G. *et al.* Early Life Factors Influencing Children Gut Microbiota at 3.5 Years from Two French Birth Cohorts. *Microorganisms* **11**, (2023).

51. de Goffau, M. C. *et al.* Gut microbiomes from Gambian infants reveal the development of a non-industrialized Prevotella-based trophic network. *Nat Microbiol* **7**, 132–144 (2022).

52. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).

53. Comeau, A. M. & Filloramo, G. V. Preparing multiplexed WGS/MetaG libraries with the Illumina DNA Prep kit for the Illumina NextSeq or MiSeq. (2023).

54. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).

55. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, (2021).

56. Vatanen, T. *et al.* Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **165**, 842–853 (2016).

57. Kostic, A. D. *et al.* The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).

58. Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8**, 343ra81 (2016).
59. Asnicar, F. *et al.* Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. *mSystems* **2**, (2017).
60. Pehrsson, E. C. *et al.* Interconnected microbiomes and resistomes in low-income human habitats. *Nature* **533**, 212–216 (2016).
61. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
62. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.* (2017) doi:10.1137/141000671.
63. Bonham, K. S., Kayisire, A. A., Luo, A. S. & Klepac-Ceraj, V. Microbiome.jl and BiobakeryUtils.jl - Julia packages for working with microbial community data. *Journal of Open Source Software* **6**, 3876 (2021).
64. Blaom, A. D. *et al.* MLJ: A Julia package for composable machine learning. *Journal of Open Source Software* **5**, 2704 (2020).
65. Sadeghi, B. *et al. DecisionTree.jl - A Julia implementation of the CART Decision Tree and Random Forest algorithms*. (Zenodo, 2022). doi:10.5281/zenodo.7359268.
66. Chen, L., Gamage, P. W. & Ryan, J. Debias random forest regression predictors. *J. Statist. Res.* **56**, 115–131 (2022).
67. Bates, D. *et al. JuliaStats/GLM.jl: v1.9.0*. (Zenodo, 2023). doi:10.5281/zenodo.8345558.
68. Danisch, S. & Krumbiegel, J. Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software* **6**, 3349 (2021).
69. Fatori, D. *et al.* Identifying biomarkers and trajectories of executive functions and language development in the first 3 years of life: design, methods and initial findings of the Germina cohort study. Preprint at https://doi.org/10.31219/osf.io/ed4fb (2024).
70. Hemmingway, A. *et al.* A detailed exploration of early infant milk feeding in a prospective

birth cohort study in Ireland: combination feeding of breast milk and infant formula and early breast-feeding cessation. *Br. J. Nutr.* **124**, 440–449 (2020).

71. O'Sullivan, J. *et al.* Alterations in gut microbiota composition, plasma lipids, and brain activity, suggest inter-connected pathways influencing malnutrition-associated cognitive and neurodevelopmental changes. Preprint at https://doi.org/10.21203/rs.3.rs-4115616/v1 (2024).

Acknowledgments

We would like to extend our thanks to all the mothers and their infants who generously provided their time and samples for this study. We are also grateful to the dedicated nurses and researchers who recruited participants and collected data, ensuring the success of this project. We would like to express our appreciation to the members of the Huttenhower laboratory for their insightful comments and edits. This research was supported by the Wellcome LEAP 1kD program.

Author Contributions

Conceptualization - GFB, KSB, VKC; Data curation - GFB, KSB, ID, BCW, DMM; Formal Analysis - GFB, KSB, BCW, SM, FP, NN, PS, DH, AD, RJ, FSM; Funding acquisition - VKC, JMO, KAD, DMM, MEK, ACC, GVP; Investigation - GFB, Methodology - GFB, KSB, CH, FP, SM, VKC; Project administration - VKC; Resources - VKC, KD; Software - GFB, KSB; Supervision - VKC, KSB, FSM, RAB, JMO, CRT, PCBBB, CH, KAD; Validation - GFB; Visualization - GFB; Writing/original draft - GFB; Writing/review & editing - All authors.

Competing interests

GVP has served as a speaker and/or consultant to Abbott, Ache, Adium, Apsen, EMS, Libbs, Medice, Takeda, developed CME material for Mantecorp and receives authorship royalties from Manole Editors.

Author correspondence

Vanja Klepac-Ceraj, vklepacc@wellesley.edu, 781-283-3541, 106 Central St, Wellesley, MA 02481

Figure Legends

Figure 1. A continuous diversity landscape arises from pooling a large number of globally sampled, uniformly (computationally wise) processed early-life metagenomes. (A) Geographical distribution of sample sources (total n=3,154), color-coded by major data source. (B). Distribution of age at sample collection, binned by months since birth, in the dynamic range of the age model, color-coded by major data source. Donut plot details the total sample contribution by major data source. (C) Overview of methodology, from data acquisition (via sampling, sourcing on public repositories or data collaboration), through the same processing pipeline and downstream statistical analysis. (D-E) NMDS ordination of Bray-Curtis β diversity colored by categorical data source (D) and by continuous age in months (E). Axis percentages denote variance explained by principal coordinates.

Figure 2. Gut microbial taxon abundances from shotgun metagenomics predict host age with high accuracy in early infancy. (A) Validation-set predicted ages versus ground-truth ages for all samples, color-coded by major data source. (B) Directional importances of top predictive features measured as mean decrease in impurity (MDI) for the trained RF models, multiplied by sign of correlation between predictor and outcome. Absolute values in the x-axis represent a proportion of the total fitness-weighted importance assigned to features. (C) Shannon index with respect to host age, color-coded by major data source. (D-G) Relative abundances color-coded by major data source and average month-by-month prevalences of the indicated important species, *D. formicigenerans* (D)*, E. coli* (E)*, F. prausnitzii* (F), and *B. breve* (G), with respect to host age.

Figure 3. Temporal succession patterns for a common core of age-predictive taxa generalize beyond geographical boundaries. Heatmaps of average taxon prevalence for each of the top 30 predictive species highlighted in **Fig. 2**. Species are ordered vertically by minimal-distance hierarchical clustering. Samples are binned horizontally from 2 to 13 months. Each cell represents the mean prevalence of that species in the samples collected on that specific month. Panels represent samples belonging to (A) Baltic countries (FIN, EST, RUS, SWE); (B) the United States (USA) and (C) South Africa (ZAF).

Figure 4. Functional changes are driven by taxonomic changes and centered around diet-associated pathways. Top 24 increasing and top 24 decreasing ECs (in community-wide abundance), stratified in a selected subset of the top taxonomic predictors of age. Cell colors

reflect taxon-stratified EC abundance on younger (A) and older (B) samples, measured in $\log_{10}$ CPM (counts per million reads). Blue and red triangles indicate species that increase and decrease in abundance in the first year of life, respectively.

Tables

Table 1. Sources of data for the pooled analysis

Study #	Reference	Repository	Repository ID	Number of samples	Mean age in months (SD)	Country(ies)
1	Asnicar, F. et al. (2017)[59]	SRA	PRJNA339914	3	2.95 (0.0)	ITA
2	Backhed, F. et al. (2015)[3]	SRA	PRJEB6456	180	8.03 (4.04)	SWE
3	Kostic, A. D. et al. (2015)[57]	SRA*	PRJNA231909	59	12.28 (3.88)	EST, FIN
4	Pehrsson, E. et al. (2016)[60]	SRA	PRJNA300541	3	10.44 (6.91)	SLV
5	Shao, Y. et al. (2019)[61]	ENA	PRJEB32631	285	8.65 (1.95)	GBR
6	Vatanen, T. et al. (2016)[56]	SRA**	PRJNA290380	479	10.90 (4.28)	FIN, EST, RUS
7	Yassour, M. et al. (2018)[58]	SRA***	PRJNA290381	104	7.4 (4.25)	FIN
8	Bonham, K. et al. (2023)[34]	SRA	PRJNA695570	224	7.50 (4.37)	USA
9	This work	SRA	PRJNA1128723	427	9.36 (4.46)	ZAF
10	Fatori, D. et al. (2024)[69]	SRA	PRJNA1072081	963	5.41 (2.13)	BRA
11	Hemmingway, A. et al. (2020)[70]	ENA	PRJEB77202	353	6.75 (3.11)	IRL
12	O'Sullivan, J. et al. (2024)[71]	SRA	PRJNA1087376	74	11.88 (0.51)	BGD

** - This is the NCBI BioProject ID for the DIABIMMUNE T1D cohort, but the data was instead obtained from the Broad Institute mirror (https://diabimmune.broadinstitute.org/diabimmune/t1d-cohort)

** - This is the NCBI BioProject ID for the DIABIMMUNE Three Country cohort, but the data was instead obtained from the Broad Institute mirror (https://diabimmune.broadinstitute.org/diabimmune/t1d-cohort)

*** - This is the NCBI BioProject ID for the DIABIMMUNE Antibiotics cohort, but the data was instead obtained from the Broad Institute mirror (https://diabimmune.broadinstitute.org/diabimmune/antibiotics-cohort)

Table 2. Summary demographics of Khula study participants (mothers)

	Overall (N=252[a])
Maternal Place of Birth	
South Africa	249 (98.8%)
In the African Continent (not South Africa)	3 (1.2%)
Primary Spoken Language	
Xhosa Language	245 (97.2%)
Sotho Language	2 (0.8%)
English Language	2 (0.8%)
Zulu Language	1 (0.4%)
Ndebele Language	1 (0.4%)
Afrikaans Language	1 (0.4%)
Maternal Educational Attainment[b]	
Completed Grade 6 (Standard 4) to Grade 7 (Standard 5)	5 (2.0%)
Completed Grade 8 (Standard 6) to Grade 11 (Standard 9) i.e., high school without matriculating	116 (46.0%)
Completed Grade 12 (Standard 10) i.e., high school	102 (40.5%)
Part of university/ college/ post-matric education	15 (5.9%)
Completed university/ college/ post-matric education	14 (5.6%)
Maternal Monthly Income[c] (South African Rand/ZAR)	
Unknown	22 (8.7%)
Less than R1000 per month	44 (17.5%)
R1000 - R5000 per month	121 (48.0%)
R5000 - R10,000 per month	57 (22.6%)
More than R10 000 per month	8 (3.2%)
Depression Score[d]	

	Overall (N=252[a])
Mean (SD)	12.9 (8.79)
Median [Min, Max]	12.0 [0, 42.0]
Infant Biological Sex	
Female	119 (47.2%)
Male	133 (52.8%)

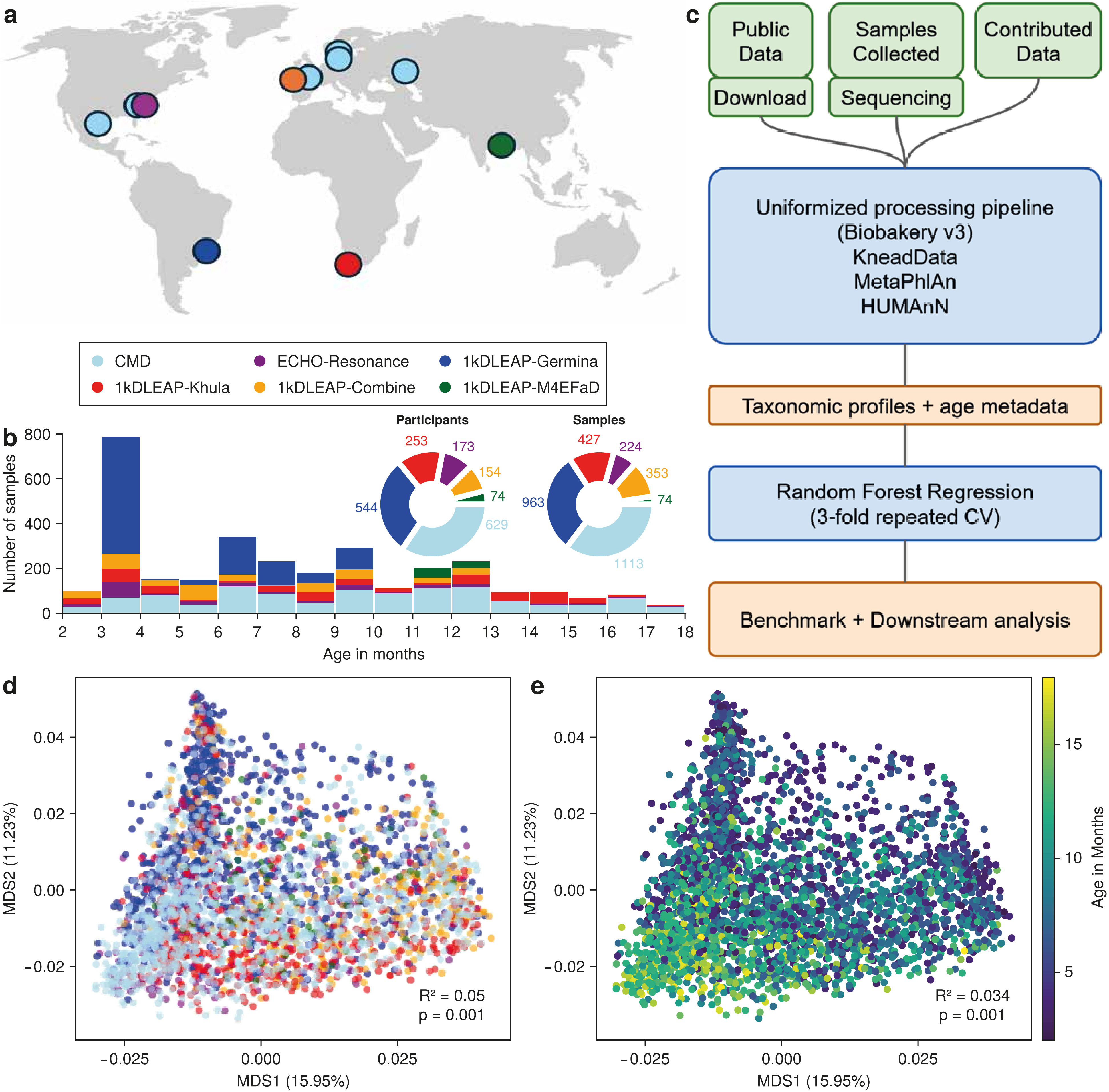
[a] Table lists only Khula study participants that had at least one sample included in this work. For the full cohort demographics, see

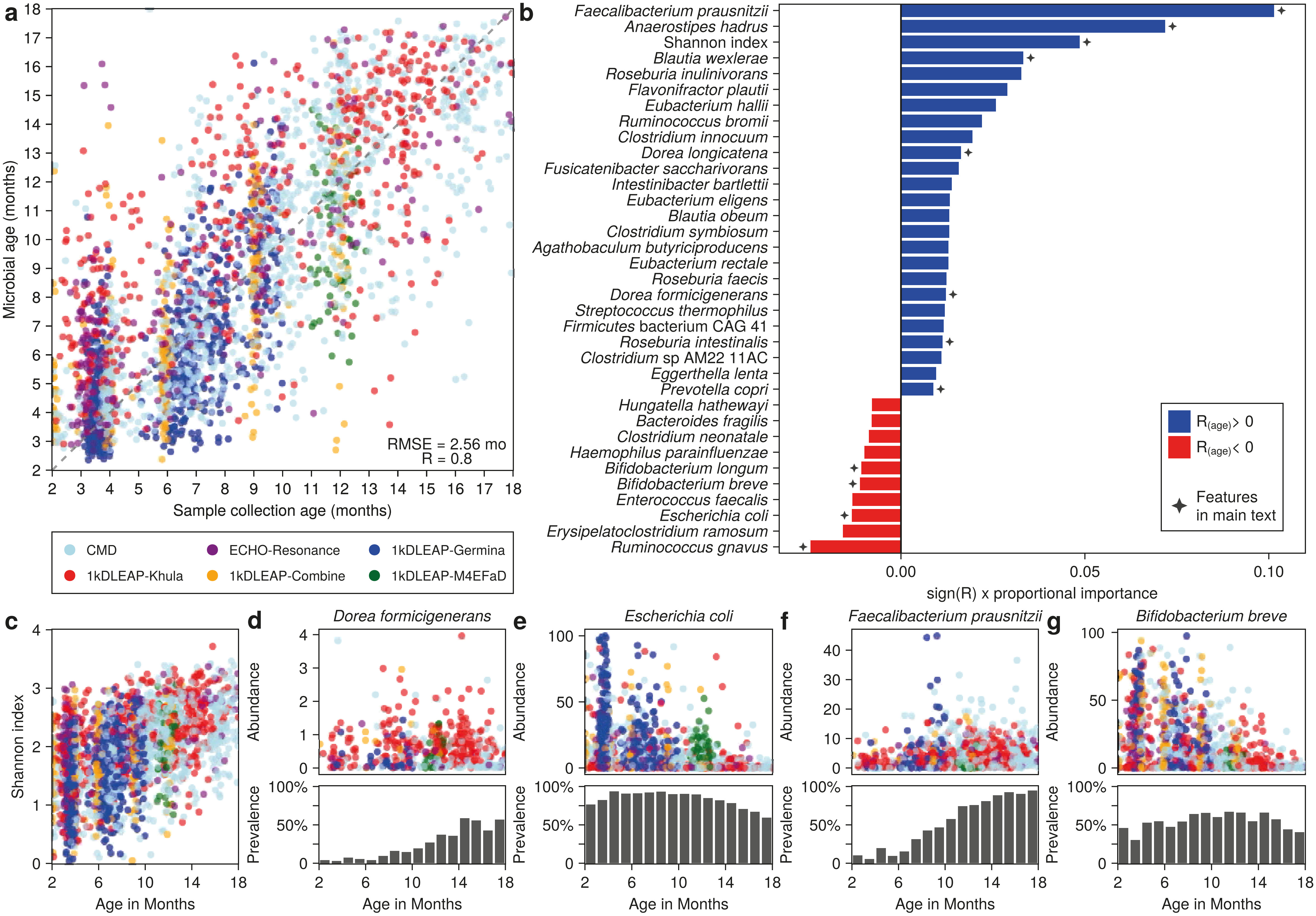
[b] The South African Educational System was formerly divided into years called standards, similarly to the way the United States Educational System is divided into grades. The equivalent in terms of standards is provided in parentheses next to each mentioned grade. “University/College/Post-Matric Education” refers to tertiary or post-secondary education as defined by the World Bank.

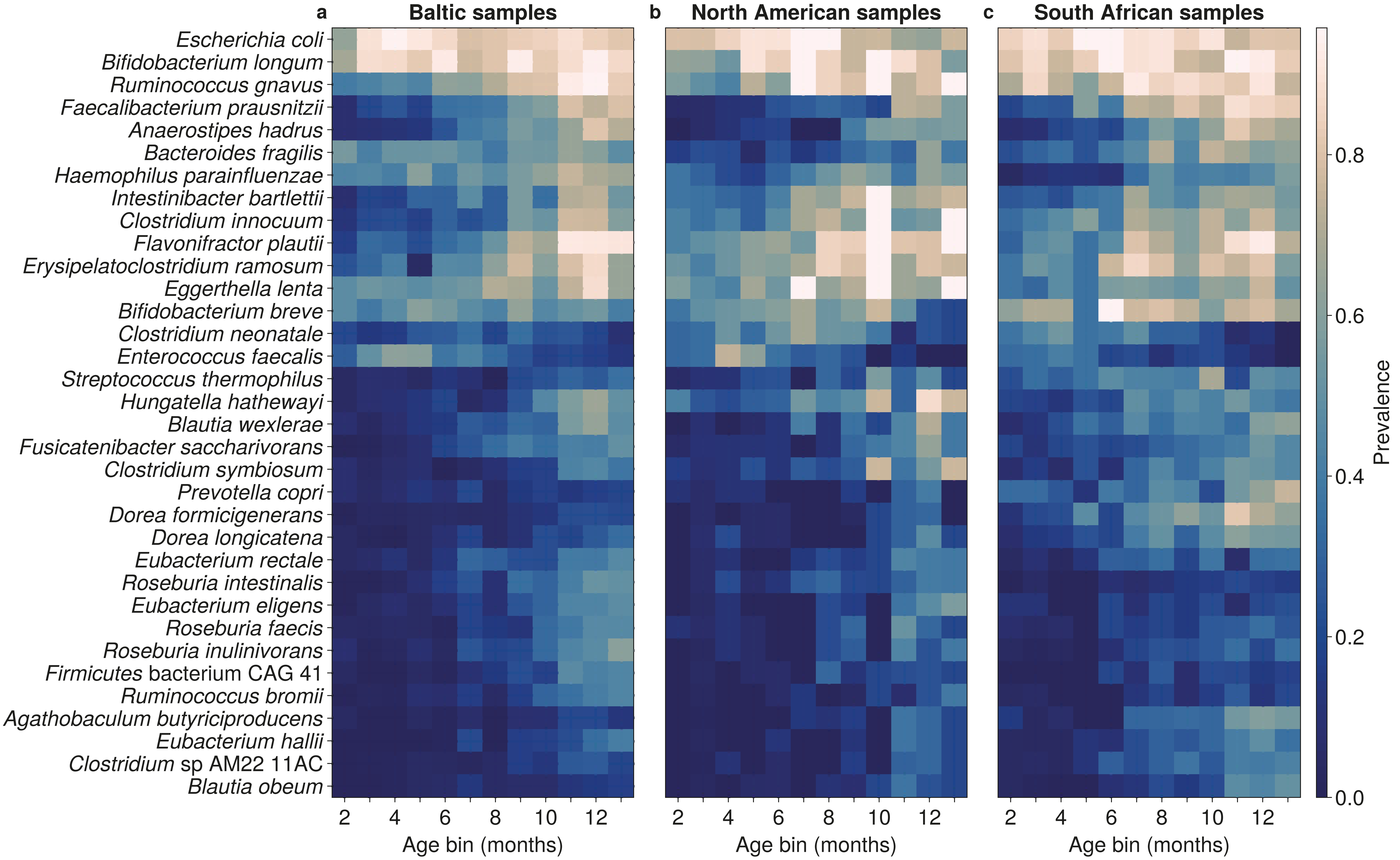
[c] At the time of writing (JUN 15, 2024), 1 US Dollar = 18.35 South African Rand (ZAR).

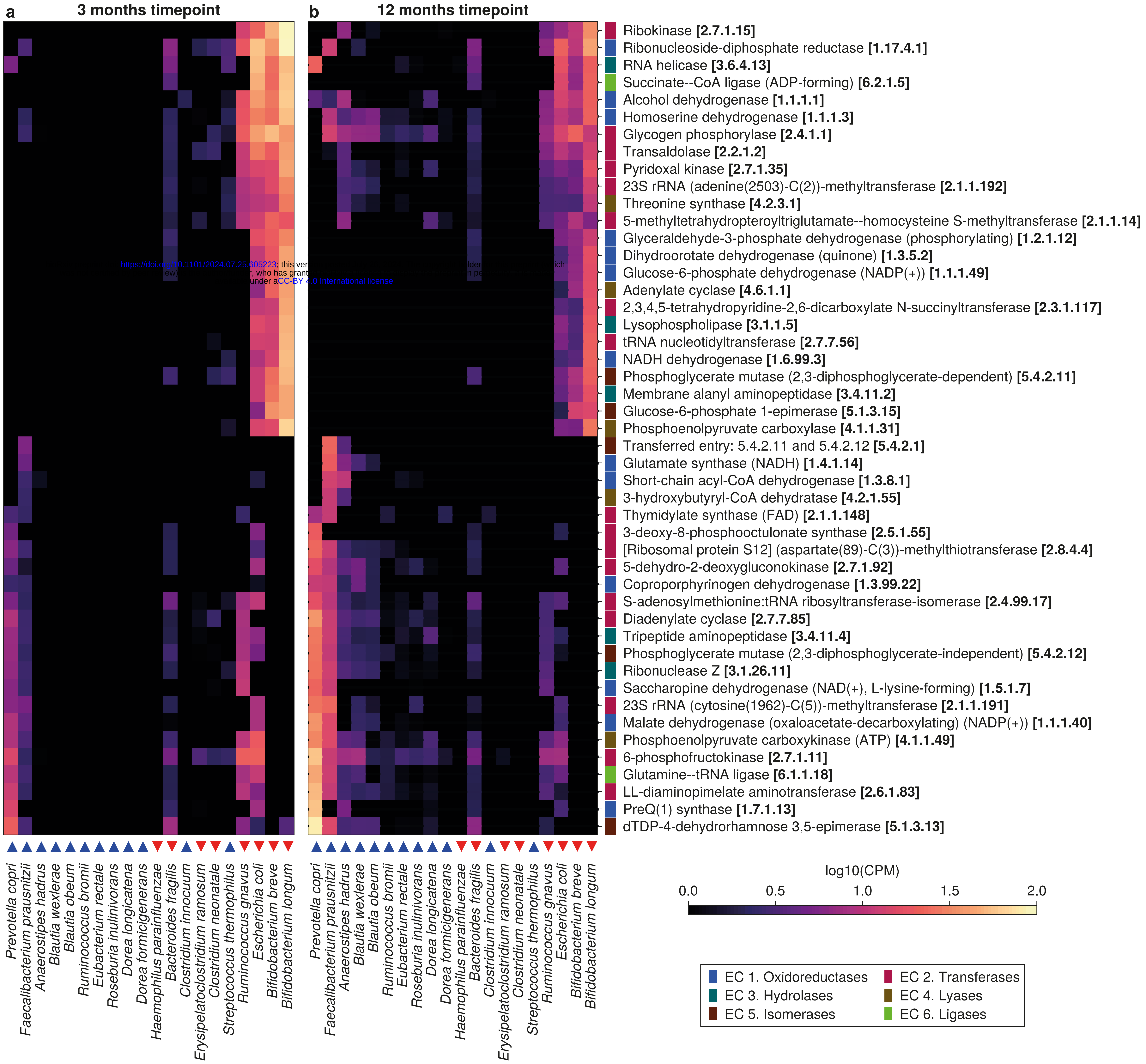
[d] Depression was measured using the Edinburgh Postnatal Depression Scale (EPDS) at enrollment.

692


a
CMD
ECHO-Resonance
1kDLEAP-Germina
1kDLEAP-Khula
1kDLEAP-Combine
1kDLEAP-M4EFaD
b
Number of samples
800
600
400
200
0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Age in months
Participants
253
173
154
74
629
544
Samples
427
224
353
74
1113
963
c
Public Data
Download
Samples Collected
Sequencing
Contributed Data
Uniformized processing pipeline
(Biobakery v3)
KneadData
MetaPhlAn
HUMAnN
Taxonomic profiles + age metadata
Random Forest Regression
(3-fold repeated CV)
Benchmark + Downstream analysis
d
MDS2 (11.23%)
0.04
0.02
0.00
−0.02
−0.025
0.000
0.025
MDS1 (15.95%)
R² = 0.05
p = 0.001
e
MDS2 (11.23%)
0.04
0.02
0.00
−0.02
−0.025
0.000
0.025
MDS1 (15.95%)
R² = 0.034
p = 0.001
Age in Months
15
10
5


a
Microbial age (months)
Sample collection age (months)
RMSE = 2.56 mo
R = 0.8
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
● CMD
● ECHO-Resonance
● 1kDLEAP-Germina
● 1kDLEAP-Khula
● 1kDLEAP-Combine
● 1kDLEAP-M4EFaD
b
Faecalibacterium prausnitzii
Anaerostipes hadrus
Shannon index
Blautia wexlerae
Roseburia inulinivorans
Flavonifractor plautii
Eubacterium hallii
Ruminococcus bromii
Clostridium innocuum
Dorea longicatena
Fusicatenibacter saccharivorans
Intestinibacter bartlettii
Eubacterium eligens
Blautia obeum
Clostridium symbiosum
Agathobaculum butyriciproducens
Eubacterium rectale
Roseburia faecis
Dorea formicigenerans
Streptococcus thermophilus
Firmicutes bacterium CAG 41
Roseburia intestinalis
Clostridium sp AM22 11AC
Eggerthella lenta
Prevotella copri
Hungatella hathewayi
Bacteroides fragilis
Clostridium neonatale
Haemophilus parainfluenzae
Bifidobacterium longum
Bifidobacterium breve
Enterococcus faecalis
Escherichia coli
Erysipelatoclostridium ramosum
Ruminococcus gnavus
🟦 $R_{(age)} > 0$
🟥 $R_{(age)} < 0$
✦ Features in main text
0.00
0.05
0.10
sign(R) x proportional importance
c
Shannon index
0
1
2
3
4
Age in Months
d
Dorea formicigenerans
Abundance
0
1
2
3
4
Prevalence
0
50%
100%
Age in Months
e
Escherichia coli
Abundance
0
50
100
Prevalence
0
50%
100%
Age in Months
f
Faecalibacterium prausnitzii
Abundance
0
10
20
30
40
Prevalence
0
50%
100%
Age in Months
g
Bifidobacterium breve
Abundance
0
50
100
Prevalence
0
50%
100%
Age in Months
2
6
10
14
18


a
Baltic samples
b
North American samples
c
South African samples
Escherichia coli
Bifidobacterium longum
Ruminococcus gnavus
Faecalibacterium prausnitzii
Anaerostipes hadrus
Bacteroides fragilis
Haemophilus parainfluenzae
Intestinibacter bartlettii
Clostridium innocuum
Flavonifractor plautii
Erysipelatoclostridium ramosum
Eggerthella lenta
Bifidobacterium breve
Clostridium neonatale
Enterococcus faecalis
Streptococcus thermophilus
Hungatella hathewayi
Blautia wexlerae
Fusicatenibacter saccharivorans
Clostridium symbiosum
Prevotella copri
Dorea formicigenerans
Dorea longicatena
Eubacterium rectale
Roseburia intestinalis
Eubacterium eligens
Roseburia faecis
Roseburia inulinivorans
Firmicutes bacterium CAG 41
Ruminococcus bromii
Agathobaculum butyriciproducens
Eubacterium hallii
Clostridium sp AM22 11AC
Blautia obeum
2
4
6
8
10
12
Age bin (months)
Prevalence
0.0
0.2
0.4
0.6
0.8


a
3 months timepoint
b
12 months timepoint
Ribokinase [2.7.1.15]
Ribonucleoside-diphosphate reductase [1.17.4.1]
RNA helicase [3.6.4.13]
Succinate--CoA ligase (ADP-forming) [6.2.1.5]
Alcohol dehydrogenase [1.1.1.1]
Homoserine dehydrogenase [1.1.1.3]
Glycogen phosphorylase [2.4.1.1]
Transaldolase [2.2.1.2]
Pyridoxal kinase [2.7.1.35]
23S rRNA (adenine(2503)-C(2))-methyltransferase [2.1.1.192]
Threonine synthase [4.2.3.1]
5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase [2.1.1.14]
Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) [1.2.1.12]
Dihydroorotate dehydrogenase (quinone) [1.3.5.2]
Glucose-6-phosphate dehydrogenase (NADP(+)) [1.1.1.49]
Adenylate cyclase [4.6.1.1]
2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase [2.3.1.117]
Lysophospholipase [3.1.1.5]
tRNA nucleotidyltransferase [2.7.7.56]
NADH dehydrogenase [1.6.99.3]
Phosphoglycerate mutase (2,3-diphosphoglycerate-dependent) [5.4.2.11]
Membrane alanyl aminopeptidase [3.4.11.2]
Glucose-6-phosphate 1-epimerase [5.1.3.15]
Phosphoenolpyruvate carboxylase [4.1.1.31]
Transferred entry: 5.4.2.11 and 5.4.2.12 [5.4.2.1]
Glutamate synthase (NADH) [1.4.1.14]
Short-chain acyl-CoA dehydrogenase [1.3.8.1]
3-hydroxybutyryl-CoA dehydratase [4.2.1.55]
Thymidylate synthase (FAD) [2.1.1.148]
3-deoxy-8-phosphooctulonate synthase [2.5.1.55]
[Ribosomal protein S12] (aspartate(89)-C(3))-methylthiotransferase [2.8.4.4]
5-dehydro-2-deoxygluconokinase [2.7.1.92]
Coproporphyrinogen dehydrogenase [1.3.99.22]
S-adenosylmethionine:tRNA ribosyltransferase-isomerase [2.4.99.17]
Diadenylate cyclase [2.7.7.85]
Tripeptide aminopeptidase [3.4.11.4]
Phosphoglycerate mutase (2,3-diphosphoglycerate-independent) [5.4.2.12]
Ribonuclease Z [3.1.26.11]
Saccharopine dehydrogenase (NAD(+), L-lysine-forming) [1.5.1.7]
23S rRNA (cytosine(1962)-C(5))-methyltransferase [2.1.1.191]
Malate dehydrogenase (oxaloacetate-decarboxylating) (NADP(+)) [1.1.1.40]
Phosphoenolpyruvate carboxykinase (ATP) [4.1.1.49]
6-phosphofructokinase [2.7.1.11]
Glutamine--tRNA ligase [6.1.1.18]
LL-diaminopimelate aminotransferase [2.6.1.83]
PreQ(1) synthase [1.7.1.13]
dTDP-4-dehydrorhamnose 3,5-epimerase [5.1.3.13]
Prevotella copri
Faecalibacterium prausnitzii
Anaerostipes hadrus
Blautia wexlerae
Blautia obeum
Ruminococcus bromii
Eubacterium rectale
Roseburia inulinivorans
Dorea longicatena
Dorea formicigenerans
Haemophilus parainfluenzae
Bacteroides fragilis
Clostridium innocuum
Erysipelatoclostridium ramosum
Clostridium neonatale
Streptococcus thermophilus
Ruminococcus gnavus
Escherichia coli
Bifidobacterium breve
Bifidobacterium longum
log10(CPM)
0.0
0.5
1.0
1.5
2.0
EC 1. Oxidoreductases
EC 2. Transferases
EC 3. Hydrolases
EC 4. Lyases
EC 5. Isomerases
EC 6. Ligases