doi:10.1038/nature11053

SUPPLEMENTARY MATERIALS

SUPPLEMENTARY RESULTS

Assessing the impact of depth of coverage of bacterial 16S rRNA genes

The deep coverage of the V4-16S rRNA dataset allowed us to assess the contribution of rare (arbitrarily defined as having relative abundance <0.1%) as well as dominant taxa (>0.1%) to the composition of the gut microbiota in the three human populations. When 'rare' OTUs were used to compare adult and infant microbiota by UniFrac-based PCoA, the general clustering pattern of microbiota by age and culture/geography was identical to the pattern observed when the complete dataset or the dataset of species-level OTUs with abundances >0.1% were analyzed (**Fig. S5**).

In addition, we compared the generalization error of the Random Forest classifier obtained when using Illumina HiSeq data from the V4 region for 8 different classification tasks at 6 different rarefaction depths ranging from 100 to 1,000,000 sequences/sample (**Fig. S6a**). As expected, at high rarefaction depth (1,000,000 sequences), the Illumina-based OTUs obtained consistently better predictive accuracy (lower generalization error) than at lower rarefaction depth, suggesting that increased sequencing effort improved the accuracy of the classifiers and the predictive strength of the observed OTUs. However, in cases where clear patterns of age and geography were evident (for example, differences between USA vs non-USA adults and adults vs infants), 100 sequences/sample was enough to obtain classifications with the errors much less than the baseline error (**Fig S6a**). For example, when only 100 sequences per sample were used in the analysis, the geographic separation of USA vs non-USA adults was still evident in the PCoA analysis (**Fig. S6b**).

Finally, rarefaction analysis did not reveal saturation when we included samples with at least 2,000,000 V4-16S rRNA sequences (**Fig. S6c**) in both babies (less than 6 months-old) and adults. This suggests that despite the lower complexity of the fecal microbiota in babies, complete coverage of diversity remains to be described with more advanced technologies and analytic tools.

Non-bacterial members of the fecal microbiota

Shotgun sequences were used to query the NCBI non-redundant nucleotide database (Blastn threshold E-value<10^{-5}) to identify the representation of organisms that belong to domains other than Bacteria in the 110 sampled fecal microbiomes. Across all samples, 7±8% of reads mapped to non-bacterial sequences. The majority of these sequences belonged to Archaea and Fungi. **Fig. S7** shows that the proportional representation of archaeal sequences were significantly higher in adults compared to children ≤3 years of age in Malawi and the USA (Mann-Whitney test; $p<0.05$; note that the differences between age groups were not statistically significant among Amerindians). The representation of fungi was significantly higher in adults compared to children in all populations; among adults, fungal sequences were significantly higher in Malawian and Amerindian versus USA microbiomes (**Fig. S7**). As databases of human gut-associated microbial genomes expand, it is likely that additional sequences may map to other archaea and eukaryotes.

Comparison of sequenced microbiomes to those of 124 adult Europeans from the MetaHIT study

We examined the extent to which the recent deep sampling of fecal microbiomes from 124 adults living in Spain and Denmark by the MetaHIT consortium (2-7.3 Gbp of shotgun sequence/fecal sample[2]) represents the gene content present in the microbiomes of all modern humans of all ages. Accordingly, we tested the extent to which this gene catalog recruited reads from each of our subjects, using the 90% nucleotide identity

criterion that Meta-HIT employed to identify reads as belonging to the same gene in the same microbial species[2]. On average, 91% of reads from the fecal microbiomes of adults living in the USA, 81% from Amerindian adult microbiomes, and 76% from Malawian adult microbiomes mapped to MetaHIT. The corresponding numbers for children below 3 years of age were 79%, 72% and 78% respectively (**Fig. S11**). Additionally, 70 healthy individuals from the MetaHIT European cohort cluster with the USA population we studied (**Fig. S12**).

Effects of breast milk versus formula feeding in USA twins

Epidemiologic studies have shown that formula feeding is more common in the USA than in a number of developing countries[40,41]. Random Forest analysis revealed 48 OTUs discriminating the fecal microbiota of four USA twin pairs where both co-twins were breast-fed and five age-matched (2-5 month old) USA twin pairs who were formula-fed (**Table S6e**). Six of the 48 OTUs were overrepresented in the breast-fed babies and were assigned to the genera *Bifidobacterium, Actinomyces, Erwinia and Haemophilus*. Shotgun sequences generated from the fecal microbiomes of formula-fed babies contained significantly fewer sequences that mapped to Bifidobacteria genomes, and more taxa belonging to the Firmicutes and Bacteroidetes compared to their breast-fed counterparts ($p<0.0001$; Mann-Whitney test; **Fig S20a**). Overall, the fecal microbiota of formula-fed babies was significantly more diverse than age-matched breast-fed babies ($p<0.001$, ANOVA with bonferroni post-hoc test, **Fig. S20b**).

We identified 93 KEGG ECs whose proportional representation differentiated formula- and breast-fed microbiomes (Random Forest classifier**, Table S11**). The majority of the 93 ECs that were overrepresented in formula-fed fecal microbiomes are involved in various aspects of carbohydrate metabolism (e.g., fructose, mannose) as well as nitrogen and amino acid metabolism (e.g., lysine biosynthesis). The proportional representation of genes involved in biosynthesis of cobalamin in formula-fed babies phenocopies what is observed in adults, i.e., higher than in breast-fed infants (**Fig. S14,**

S15). These findings highlight the need to use the types of biomarkers we have identified to conduct longitudinal metagenomic studies comparing the development of the microbiomes of formula- versus breast-fed individuals. The goal would be to determine whether differences between formula- and breast-fed children persist through adulthood, and the extent to which early exposure to formula heralds microbiome-encoded metabolic programs that confer human physiologic phenotypes distinct from those of breast-fed children (e.g., ref.42).

SUPPLEMENTARY REFERENCES

40 WHO. WHO Global Data Bank on Infant and Young Child Feeding (IYCF). (2006)

41 Li, R., Darling, N., Maurice, E., Barker, L. & Grummer-Strawn, L. M. Breastfeeding rates in the United States by characteristics of the child, mother, or family: the 2002 National Immunization Survey. *Pediatrics* **115,** e31-37 (2005)

42 Owen, C. G., Martin, R. M., Whincup, P. H., Smith, G. D. & Cook, D. G. Effect of infant feeding on the risk of obesity across the life course: a quantitative review of published evidence. *Pediatrics* **115**, 1367-1377 (2005)

SUPPLEMENTARY TABLES

Table S1 – Diet survey conducted in two Amerindian and four Malawian villages. (a) Platanillal. **(b)** Coromoto. (**c)** Malawi.

Table S2 – Summary of study participants and of fecal (a) bacterial 16S rRNA and (b) shotgun sequence datasets. Our analyses also included shotgun pyrosequencing data from 9 individuals representing 3 USA families, each comprised of lean adult female twins and their mother, who had been characterized in one of our earlier publications[32], plus (ii) shotgun data from 1 fecal sample obtained from a single USA child who had been the subject of report describing the assembly of that child's gut microbiota/microbiome[5].

Table S3 - P-values (Student t-test with 1000 Monte Carlo permutations) of unweighted UniFrac and Hellinger distances between the bacterial fecal communities of children and adults shown in Fig. 1b,c. UniFrac distances were calculated from the Illumina bacterial V4-16S rRNA dataset in part **(a**).

Table S4 - List of the 126 reference human gut microbial genomes.

Table S5 - Spearman correlations of relative abundances of microbial taxa in fecal microbiomes with age for each country. (a) 97%ID OTUs obtained from Illumina sequencing of the V4 region of bacterial 16S rRNA genes. Reads that map to microbial genomes in fecal microbiomes sampled at various ages from each country using (**b**) a custom reference database of 126 sequenced human gut genomes and **(c)** 1,280 microbial genomes from the KEGG database.

Table S6 - Results of Random Forests classifier of 97% ID OTUs (species-level phylotypes) that discriminate the fecal microbiota according to age and

geography/cultural tradition. The rarefaction depth was 305,631 sequences/sample. Ten even rarefactions were performed for the comparison. (**a**) Babies vs adults, (**b**) USA vs non-USA adults, (**c**) Amerindian vs Malawian adults, (**d**) Babies, (**e**) Breast-fed vs formula-fed USA babies.

Table S7 – ECs and KOs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant age-associated differences. (a) Shown are ECs with ShotgunFunctionalizeR p values < 0.0001 (adjusted for multiple comparison using Benjamini-Hochberg False Discovery Rate). ECs are sorted by p value. Mean importance of ECs identified by Random Forests is shown in the second column**. (b)** KOs with ShotgunFunctionalizeR p values < 1e-31 (adjusted for multiple comparison using Benjamini-Hochberg False Discovery Rate).

Table S8 - ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant population-specific differences in the fecal microbiomes of babies (individuals <6-months old). Shown are ECs with ShotgunFunctionalizeR p values < 0.0001 (adjusted for multiple comparisons using Benjamini-Hochberg False Discovery Rate). ECs are sorted by p value. Mean importance of ECs identified by Random Forests is shown in the second column.

Table S9 – ECs identified by Spearman correlation analysis that exhibit significant age-associated changes in their proportional representation in fecal microbiomes

Table S10 – ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant population-specific differences in their representation in the fecal microbiomes of adults. Mean importance of ECs identified by Random Forests is shown in the second column. ECs with with ShotgunFunctionalizeR p values < 0.0001 (adjusted

for multiple comparisons using Benjamini-Hochberg False Discovery Rate) are sorted by their p value.

Table S11 – ECs identified by Random Forests and ShotgunFunctionalizeR that exhibit significant differences in their representation in the fecal microbiomes of 4 breast-fed USA twin pairs versus 5 formula-fed USA twin pairs (2-5 months old). Shown are ECs with p values < 0.0001 (adjusted for multiple comparisons using Benjamini-Hochberg False Discovery Rate). ECs are sorted by their p value.

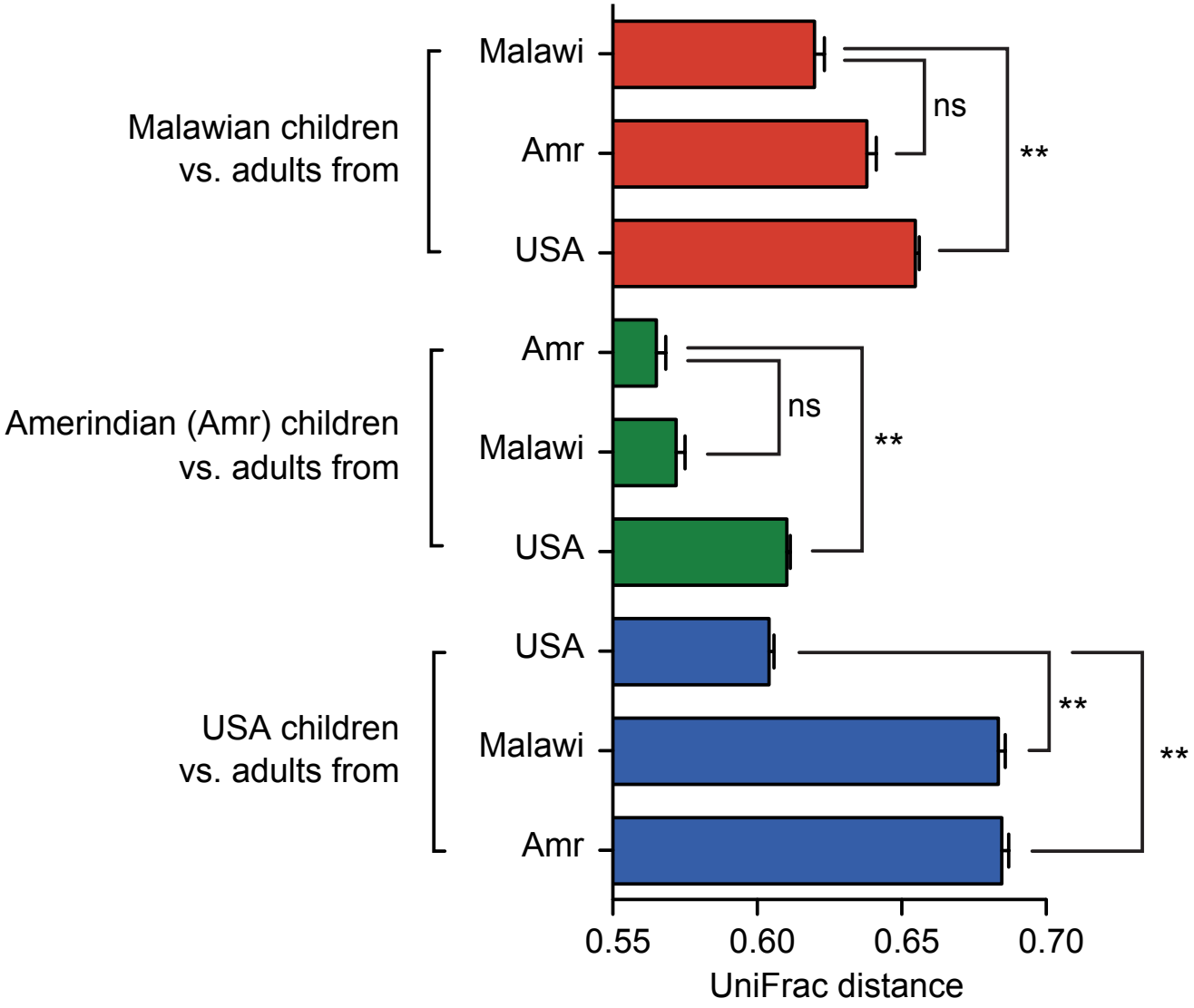


Fig. S1 – Similarity between the fecal microbiota of children and adults from the same compared to different countries. UniFrac distances calculated from the Illumina V4-16S rRNA dataset between children ≤3 years old and adults from the same population compared to adults from the other two populations. Mean values ± SEM are plotted. **p<0.001, NS, not significant based on Student's t-test with 1000 Monte Carlo simulations.

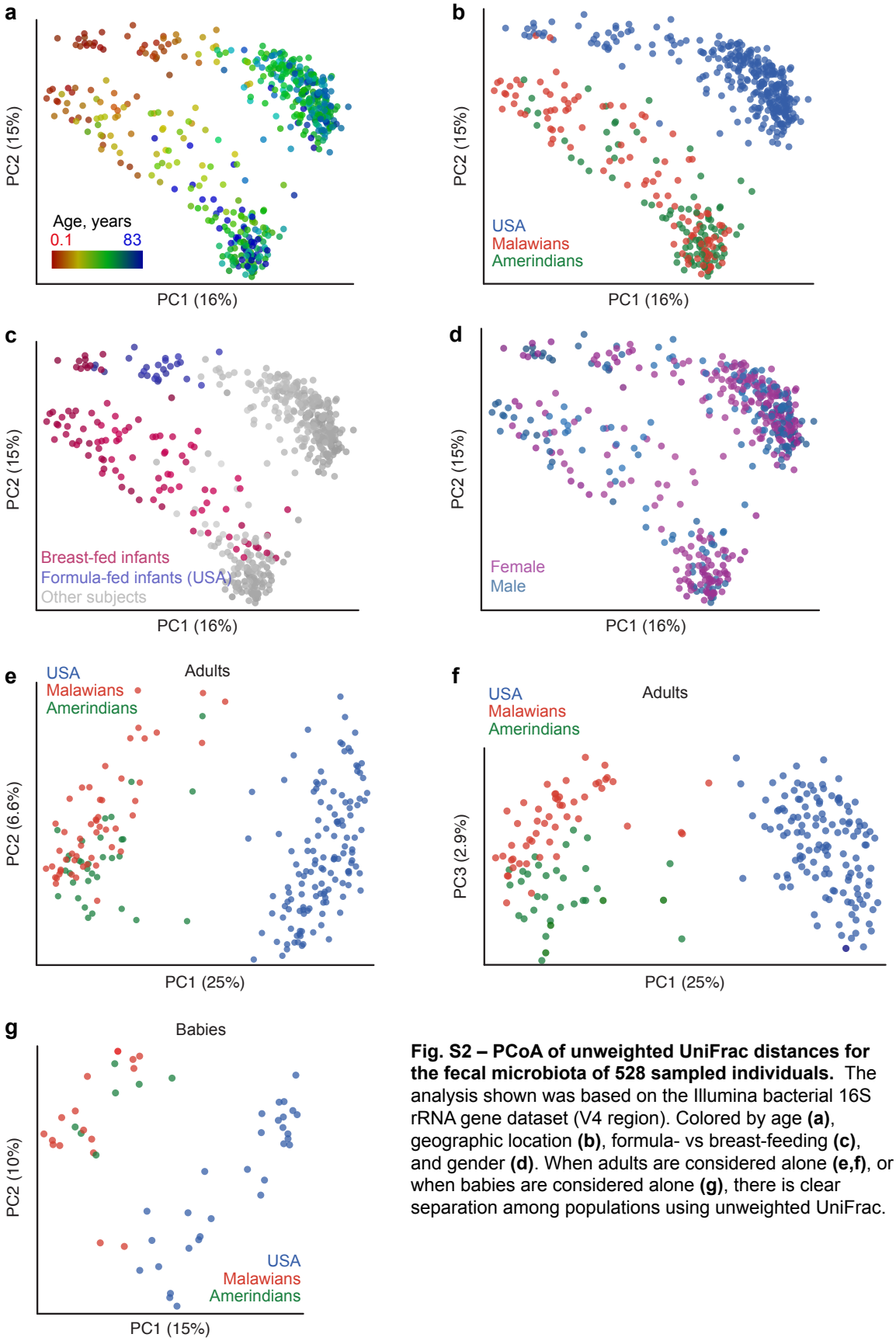


Fig. S2 – PCoA of unweighted UniFrac distances for the fecal microbiota of 528 sampled individuals. The analysis shown was based on the Illumina bacterial 16S rRNA gene dataset (V4 region). Colored by age **(a)**, geographic location **(b)**, formula- vs breast-feeding **(c)**, and gender **(d)**. When adults are considered alone **(e,f)**, or when babies are considered alone **(g)**, there is clear separation among populations using unweighted UniFrac.

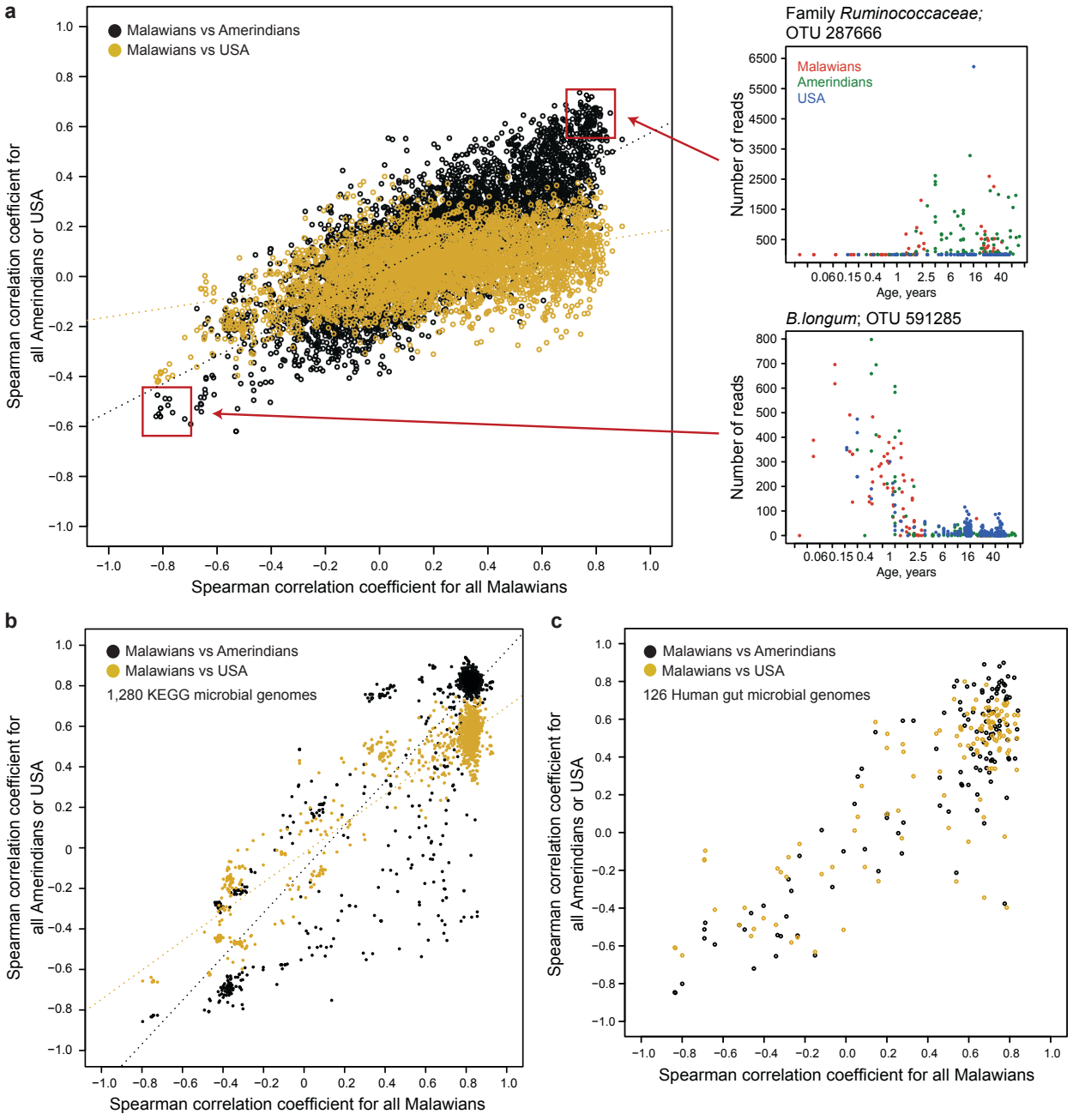


Fig. S3 – Changes in the representation of bacterial taxa in the fecal microbiota as a function of age and geographic region. (a) Spearman correlations (Rho values) were calculated for the representation of each OTU, obtained from Illumina sequencing of the V4 regions of bacterial 16S rRNA genes, against age for each population (a Rho value of ±1 indicates maximum correlation with age, a zero indicates minimum correlation). Rho values for Malawians are plotted against Rho values for Amerindians (black points) or residents of the USA (yellow-ochre points). Each point represents a 97% ID OTU; coordinates are correlations for the relative abundance of that OTU with age in Malawians (x-axis) and Amerindians or USA residents (y-axis). Spearman correlations relating populations: Malawi vs USA, Rho=0.65, $p<10^{-15}$; Amerindians vs USA Rho=0.78, $p<10^{-15}$; Malawi vs Amerindians Rho = 0.66, $p<10^{-15}$. The lower panel presents examples of the largest changes with age in all three populations (*Bifidobacterium longum*), and changes that are most pronounced in Malawi and Amerindians (OTU from the family Ruminococcaceae). Note that the OTU matrix was rarified to 290,609 sequences/sample. **(b)** An analysis similar to that shown in panel a, but using shotgun reads mapped to 1280 microbial genomes present in KEGG. **(c)** Analysis similar to **(b)** using shotgun reads mapped to 126 reference sequenced human gut microbial genomes.

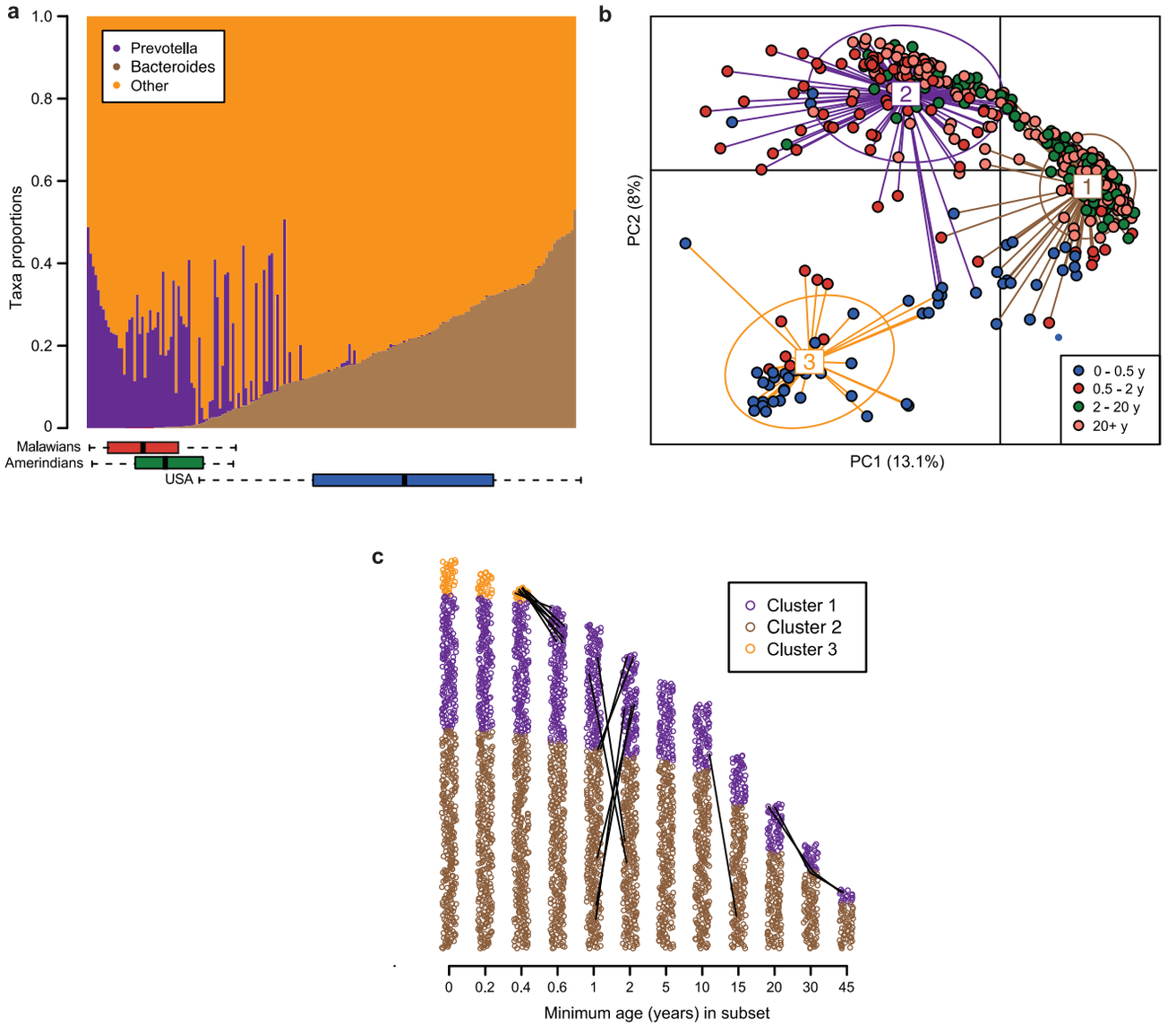


Fig. S4 – Clustering analysis of OTUs obtained from Illumina sequencing of the V4 region of bacterial 16S rRNA genes. (a) Stacked bar plot of *Bacteroides/Prevotella* gradient. Each column shows relative abundances of *Bacteroides* (brown), *Prevotella* (purple), and other genera (orange) for a single gut community. Communities are ordered according to increasing *Bacteroides* relative abundance. Box plots below show the distribution of samples from each country. **(b)** Enterotype clustering algorithm applied to samples from all countries and all ages: classical multidimensional scaling of Jensen-Shannon distances between all sampled microbial communities. Samples are colored by age group, and lines connect samples to their putative enterotype cluster centroids (silhouette index = 0.5189325). **(c)** Cluster membership for partitioned subpopulations of increasing minimum age. Samples are sorted vertically first by putative cluster number, then by age within each cluster. Lines indicate samples that switched cluster membership after a partitioning step.

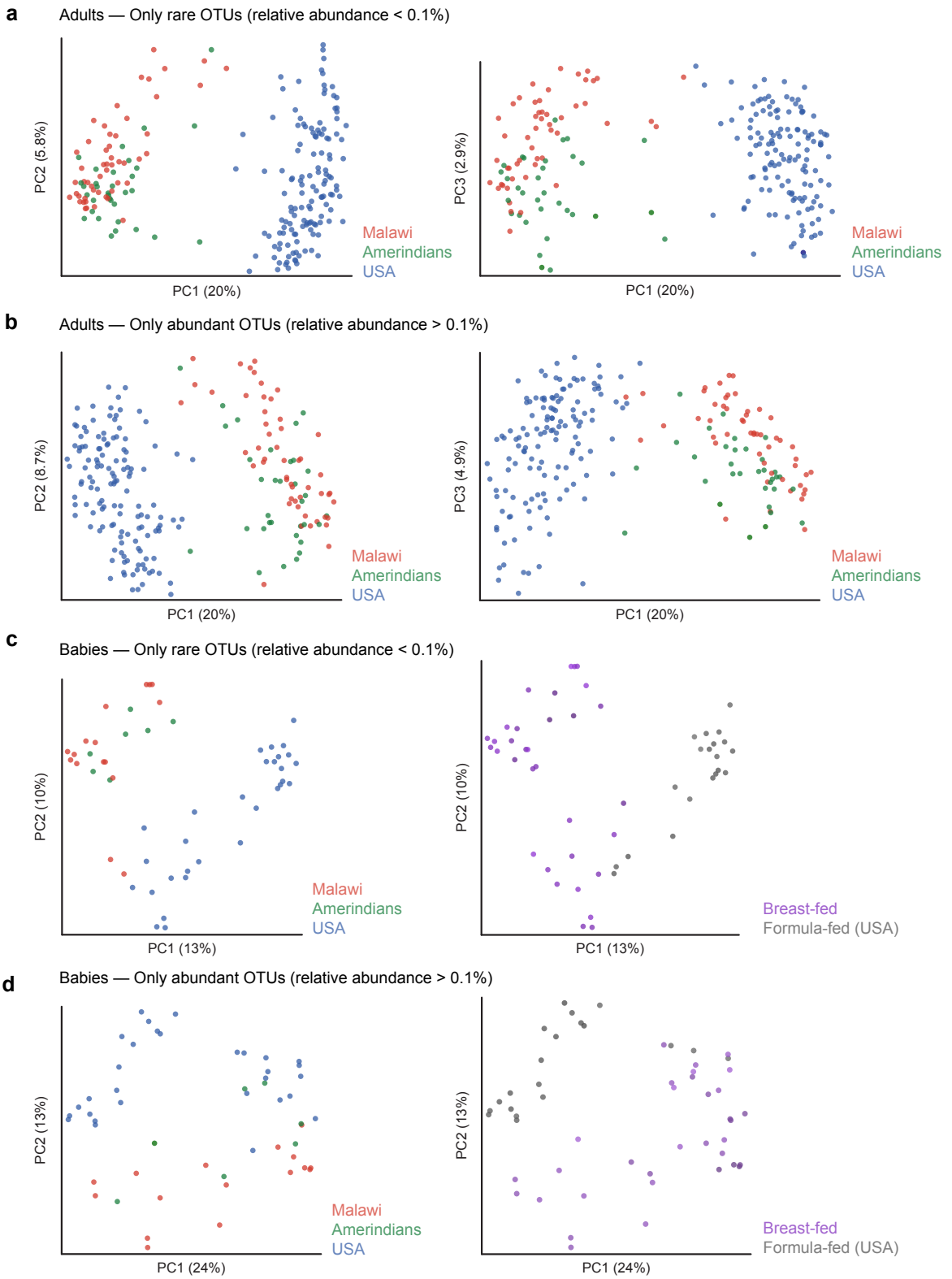


Fig. S5 – PCoA of unweighted UniFrac distances for the fecal microbiota of 59 babies and 202 adults when only rare or abundant OTUs are considered. (a) Adult microbiota when 97%ID species-level OTUs with less than 0.1% relative abundance were considered. **(b)** Same as (a) but with taxa whose relative abundance was greater than 0.1%. **(c)** Infant microbiota when taxa with less than 0.1% relative abundance were considered. **(d)** Same as **(c)** but with taxa whose relative abundance is greater than 0.1%.

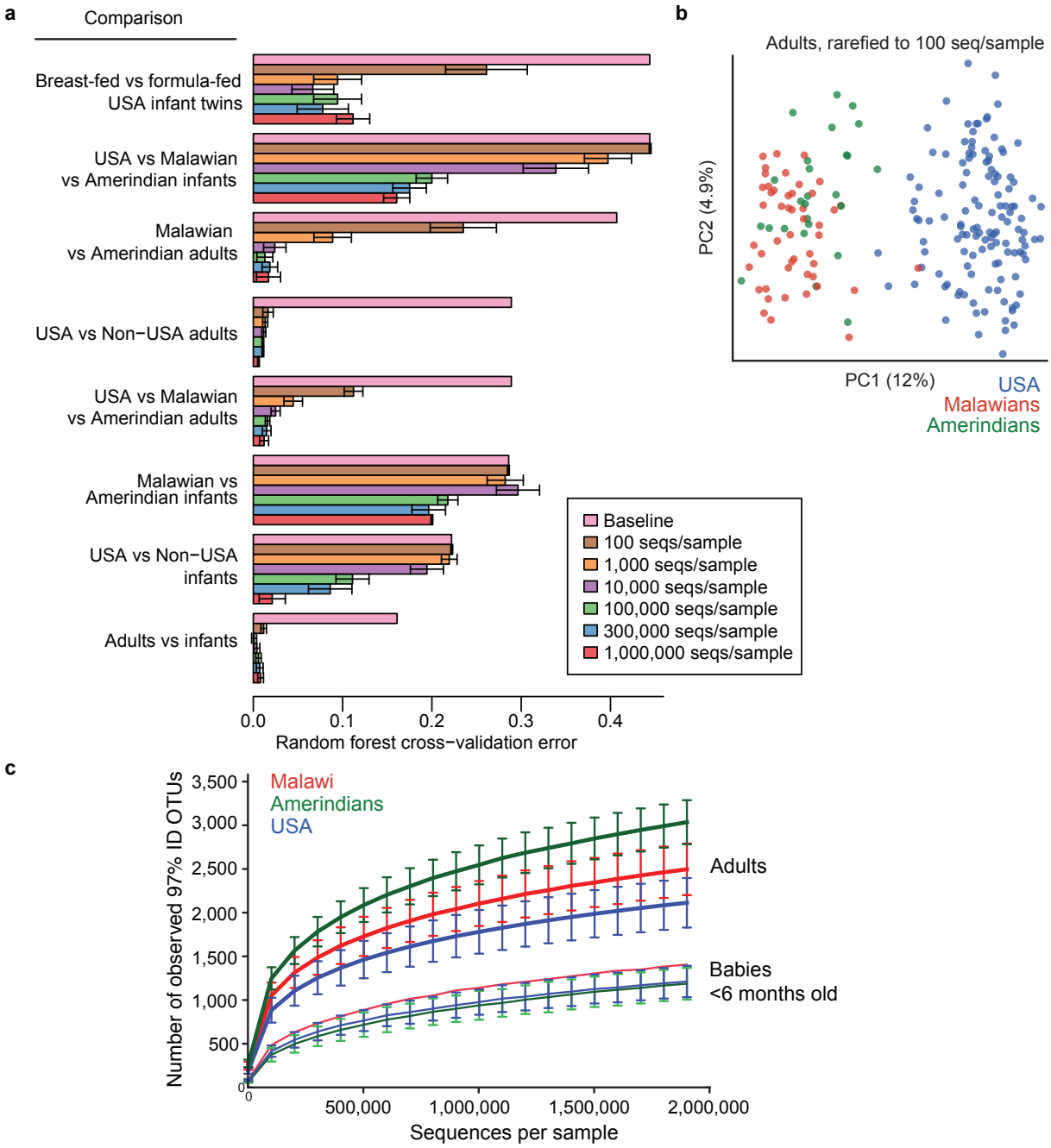


Fig. S6 – Influence of depth and type of 16S rRNA sequencing on the results of Random Forests and rarefaction analyses. (a) Generalization errors using 97%ID OTUs identified from Illumina datasets. The measure of the method's success is its ability to correctly classify unseen samples, estimated by training it on a subset of samples, and using it to classify the remaining samples (cross-validation). The cross-validation error is compared to the baseline error that would be achieved by always guessing the most common category. For each of eight comparisons, we estimated the generalization error of the Random Forests classifier using OTUs based on Illumina V4-16S rRNA sequences. Error bars show standard deviation in cross-validation error under repeated rarefactions of the data. We show the expected "Baseline" error obtained by a classifier that simply predicts the most common class label, as well as the Random Forests error obtained when different sequencing depths were used (from 100 to 1,000,000 sequences/sample). **(b)** PCoA plot of UniFrac distances between adult fecal microbiota when only 100 sequences were used from each sample. **(c)** Rarefaction curves for fecal samples, each with at least 2,000,000 V4-16S rRNA sequences. Each line connects an average number (± SD) of observed 97% ID OTUs at each rarefaction depth for adults (n=15 Malawians, 9 Amerindians, 80 USA) and babies (n=1 Malawian, 4 Amerindian, 5 USA).

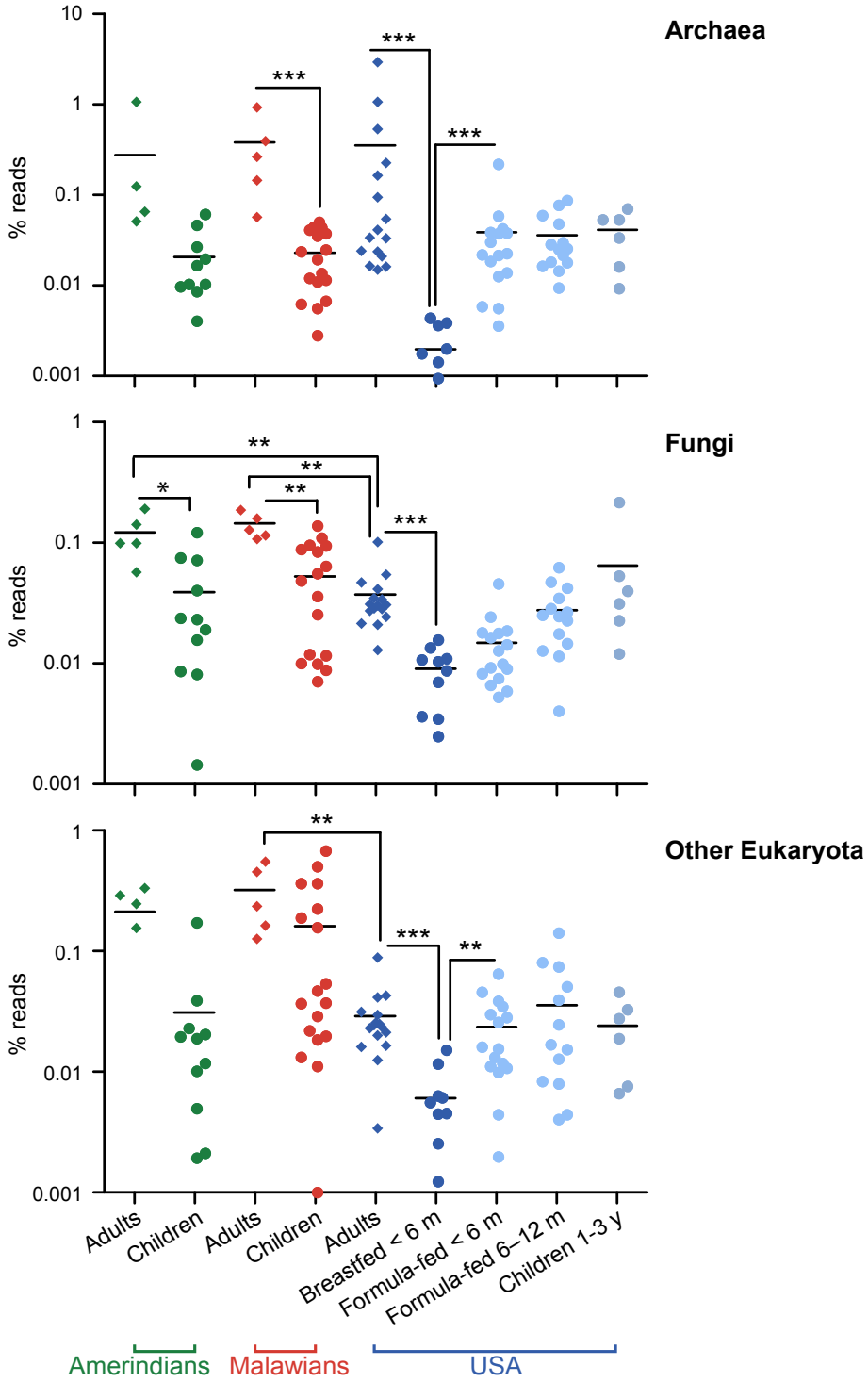


Fig. S7 – Non-bacterial members identified in fecal microbiota. Shotgun sequences were used to query the NCBI nr database (Blastn e-value threshold cutoff, 10^{-5}). The proportion of sequences that mapped to non-bacterial sequences was calculated for each age- and geographic group. The most abundant fungal sequences belong to the NCBI nr family level taxa *Ascomycota* and *Microsporidia* and were found in all three populations. In NCBI nr, 'other eukaryota' refers to sequences that do not map to fungi, plants, arthropoda, mammals, and 'other metazoa'. In USA microbiomes 'other eukaryota' was most prominently represented by *Hexamitidae*, *Trichomonadidae* families and genus *Entamoeba*, while in Malawian and Amerindian microbiomes the most abundant group was "uncultured compost protozoan". *** p<0.0005, **p<0.005, *p<0.05 (Mann-Whitney test).

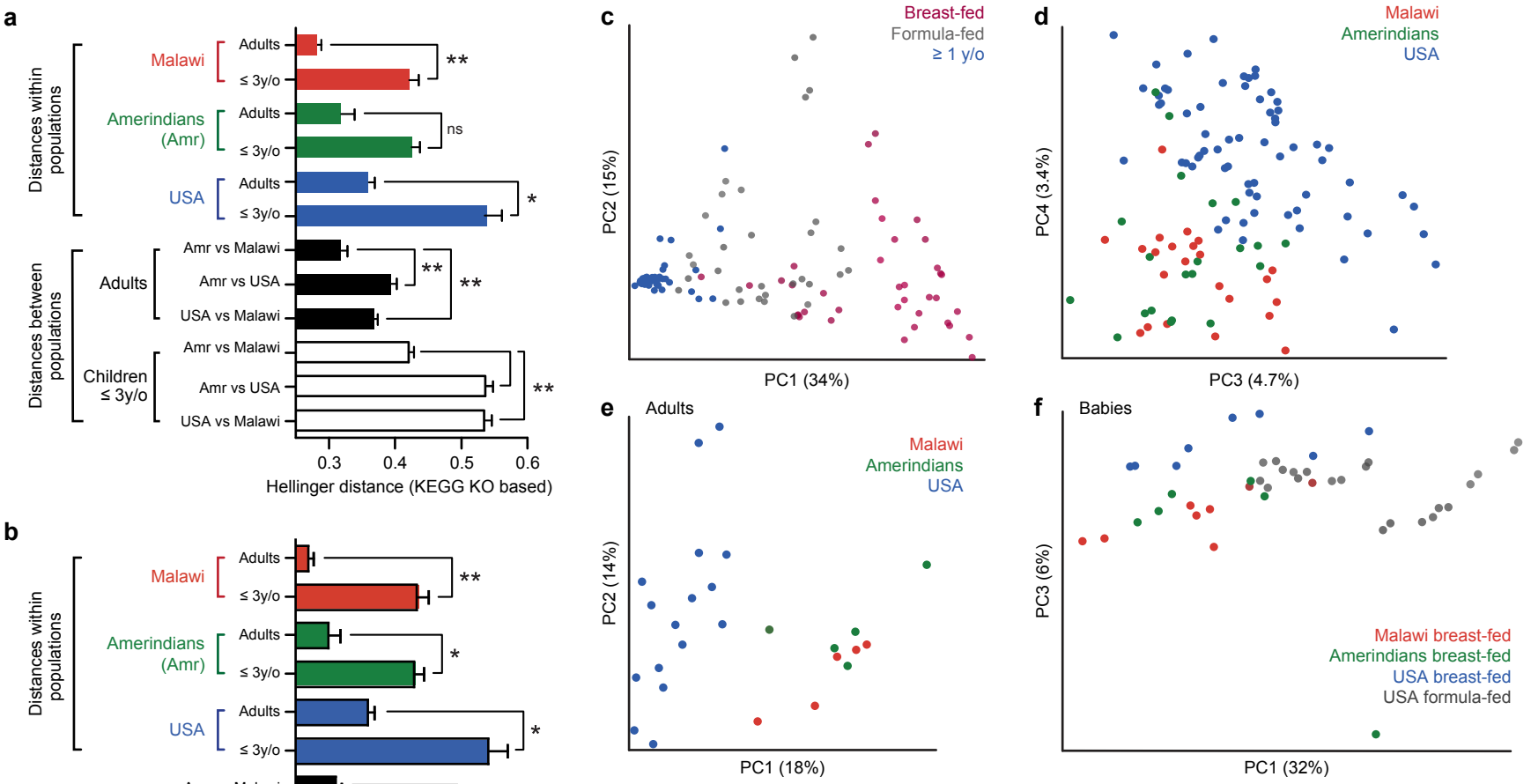


Fig. S8 – Analysis of Hellinger distances between fecal microbiome KEGG KO and EC profiles. (a) Hellinger distances derived from KO profiles are shown for unrelated children ≤3 years of age and unrelated adults (n=9 children and 5 adults from Malawi; 11 children and 5 adults from Venezuela; 10 children and 8 adults from USA). Mean values ± SEM are plotted. Abbreviations: * p<0.05; **p<0.005 (Student's t-test with 1000 Monte Carlo simulations). **(b)** Hellinger distances derived from EC profiles are shown for unrelated children ≤3 years of age and unrelated adults (n=9 children and 5 adults from Malawi; 11 children and 5 adults from Venezuela; 10 children and 8 adults from USA). Mean values ± SEM are plotted. Abbreviations: * p<0.05; **p<0.005 (Student's t-test with 1000 Monte Carlo simulations). **(c-f)** PCoA plots derived from Hellinger distances. Counts were normalized to the total number of reads for each fecal microbiome sample thus accounting for sequences unassigned to KEGG. **(c)** All fecal microbiomes colored by diet plotted along PC1 and PC2. **(d)** Geographic separation is clear along PC3 and PC4. The fecal microbiomes of adults **(e)** and babies **(f)** are colored by geographic location.

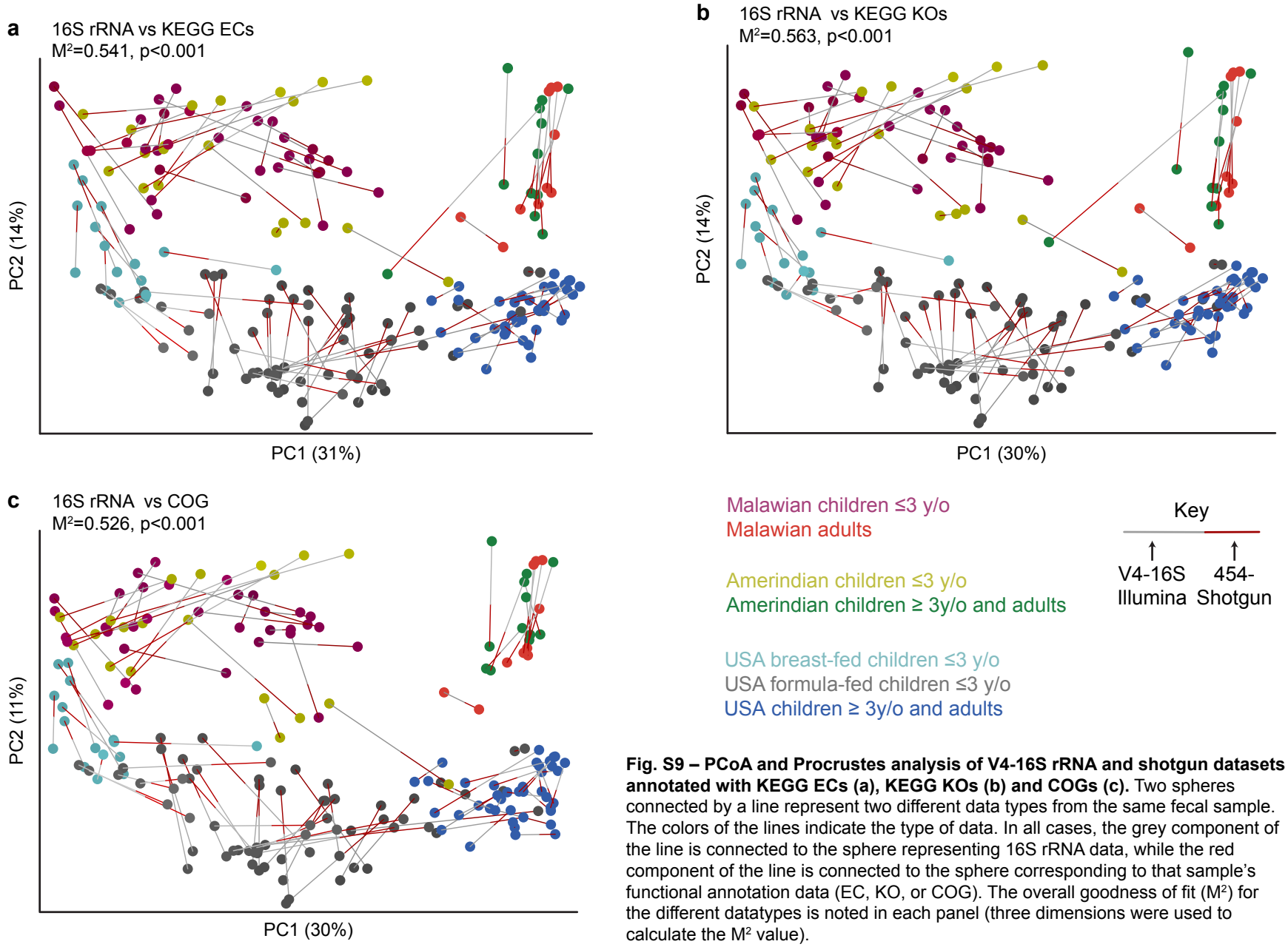


Fig. S9 – PCoA and Procrustes analysis of V4-16S rRNA and shotgun datasets annotated with KEGG ECs (a), KEGG KOs (b) and COGs (c). Two spheres connected by a line represent two different data types from the same fecal sample. The colors of the lines indicate the type of data. In all cases, the grey component of the line is connected to the sphere representing 16S rRNA data, while the red component of the line is connected to the sphere corresponding to that sample's functional annotation data (EC, KO, or COG). The overall goodness of fit (M^2) for the different datatypes is noted in each panel (three dimensions were used to calculate the M^2 value).

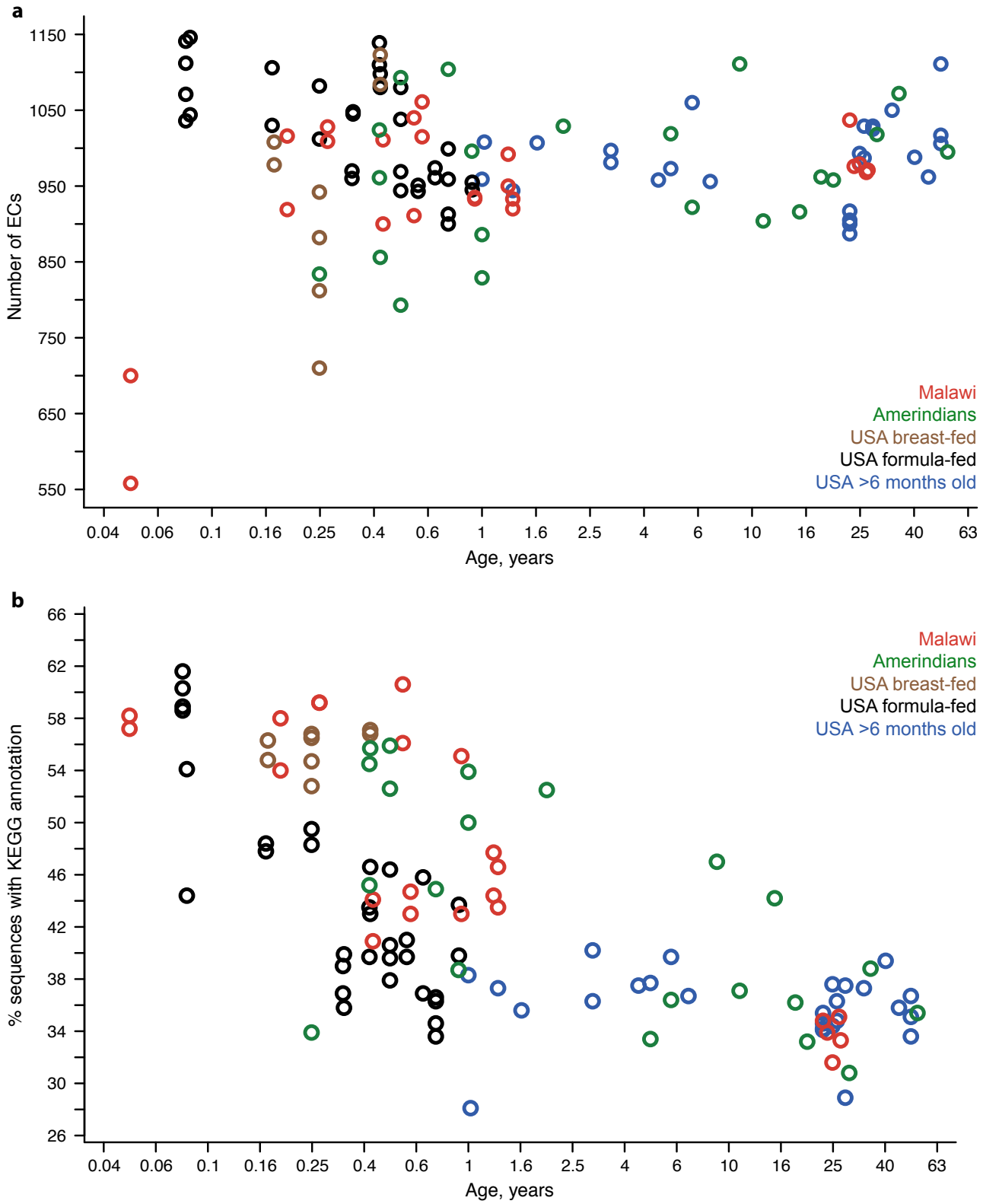


Fig. S10 – The number of KEGG ECs identified is similar in adult and infant fecal microbiomes while the fraction of microbiome shotgun reads with assignable EC annotations declines with age in all three populations. (a) The EC matrix was rarefied to 3,650 sequences per sample, and number of ECs was plotted against age for each sample. **(b)** Percent of sequences with KEGG annotation plotted against age.

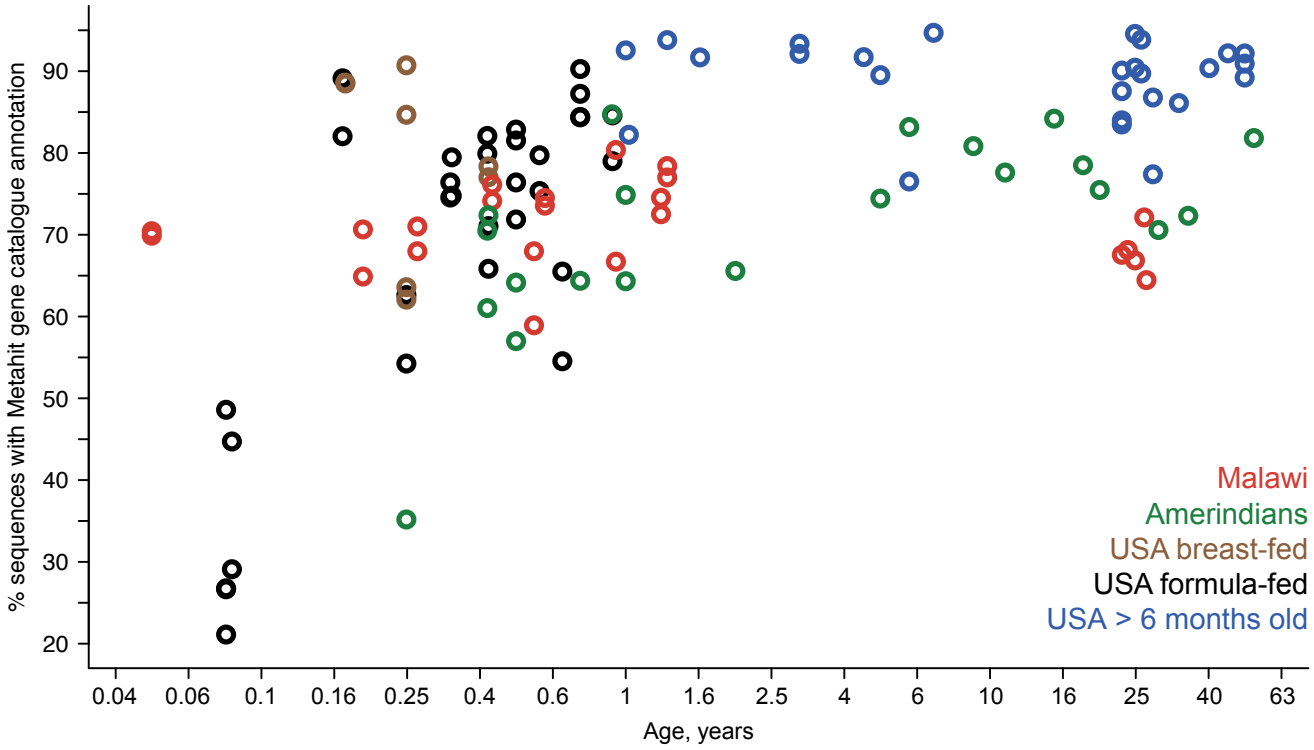


Fig. S11 – Percentage of fecal microbiome gene content in sampled members of the three populations that is also represented in the MetaHIT gene catalog generated from 124 adult Europeans. Percentage of shotgun pyrosequencing reads in each population that could be assigned to the MetaHIT gene catalog using the following Blastn parameters: ≥90% nucleotide sequence identity between the read and a member of the gene catalog, E-value $<10^{-5}$, bitscore ≥50.

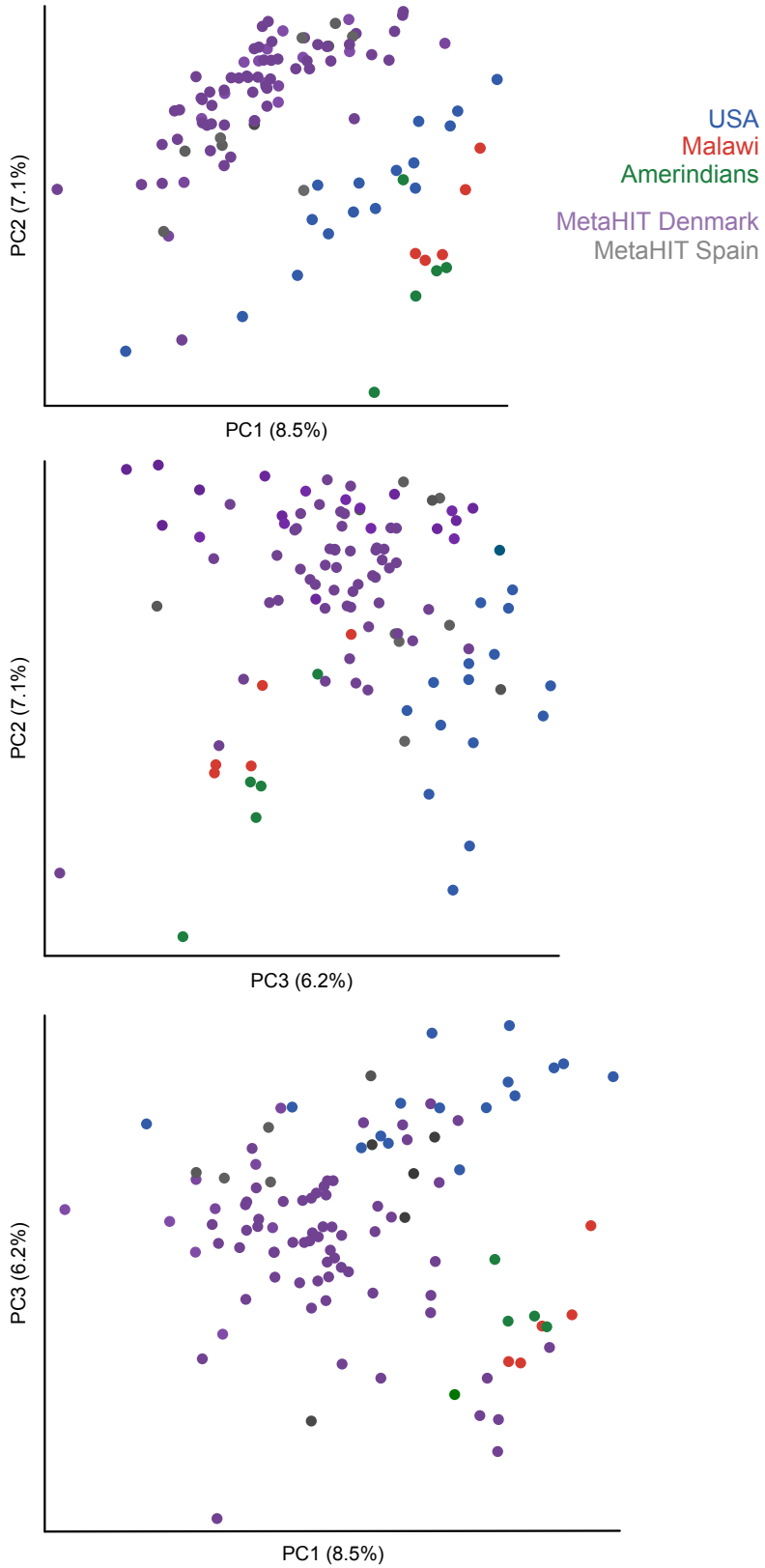


Fig. S12 – PCoA plot of Hellinger distances between the KEGG KO profiles of adult USA, Amerindian and Malawian fecal microbiomes from the present study and from 70 healthy European microbiomes in the MetaHIT dataset[2].

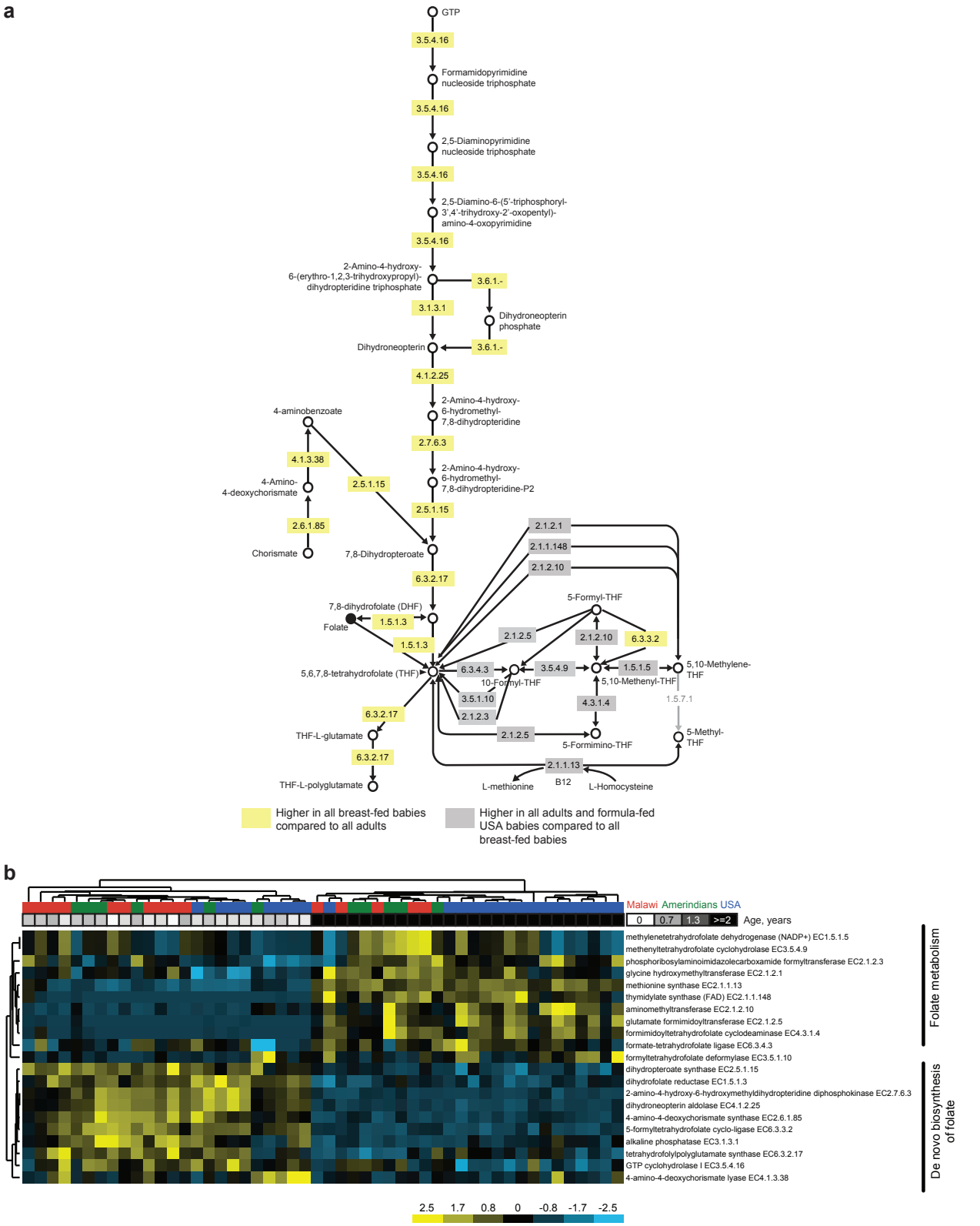


Fig. S13 – Age-related changes in the proportional representation of genes encoding ECs involved in folate metabolism. (a) KEGG pathway for folate metabolism. **(b)** UPGMA clustering (average linkage method) of fecal microbiomes of 24 babies and 26 adults based on the relative abundances of genes encoding ECs shown in **(a)**, normalized by Z-score across all datasets.

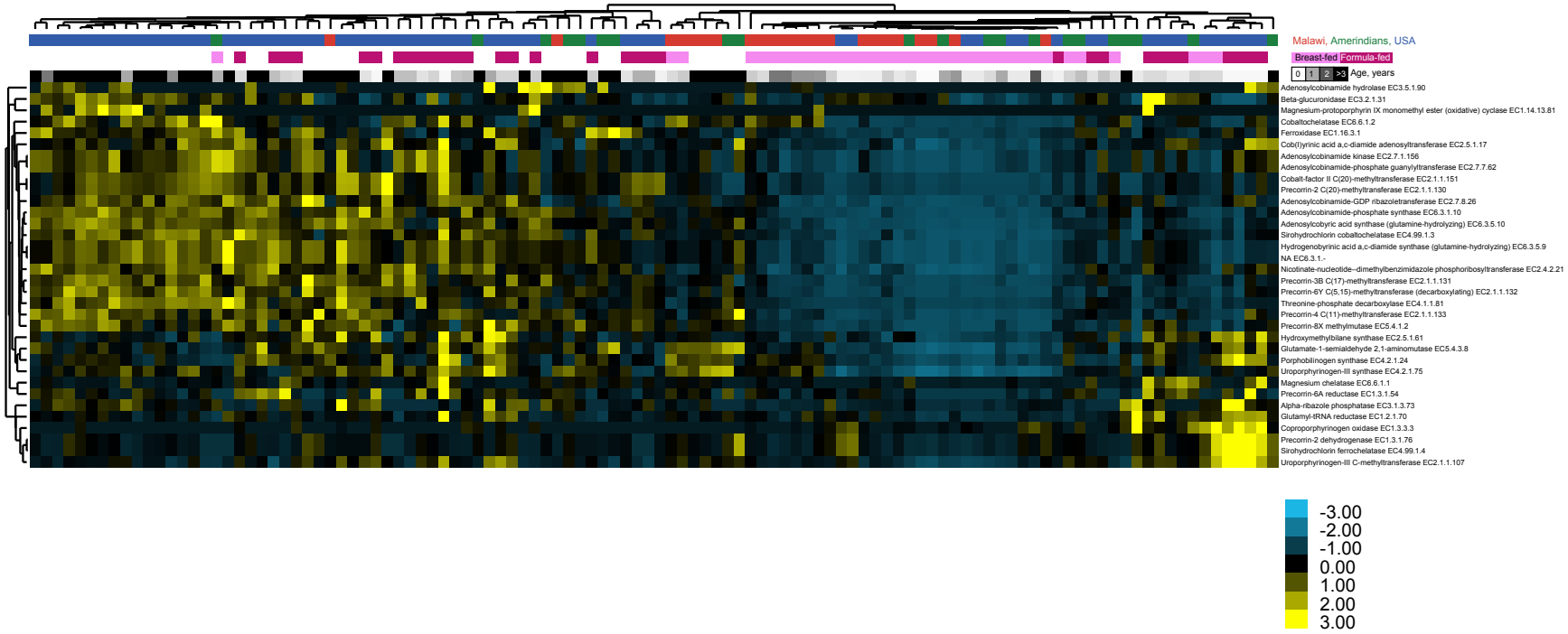


Fig. S14 – Age-related changes in the proportional representation of genes encoding ECs involved in cobalamin biosynthesis. UPGMA clustering (average linkage method) of all 110 characterized fecal microbiomes, based on the relative abundances of ECs involved in cobalamin biosynthesis (normalized by Z-score across all datasets). The bars on the top indicate the age, breastfeeding status and geographic location of each human who was sampled.

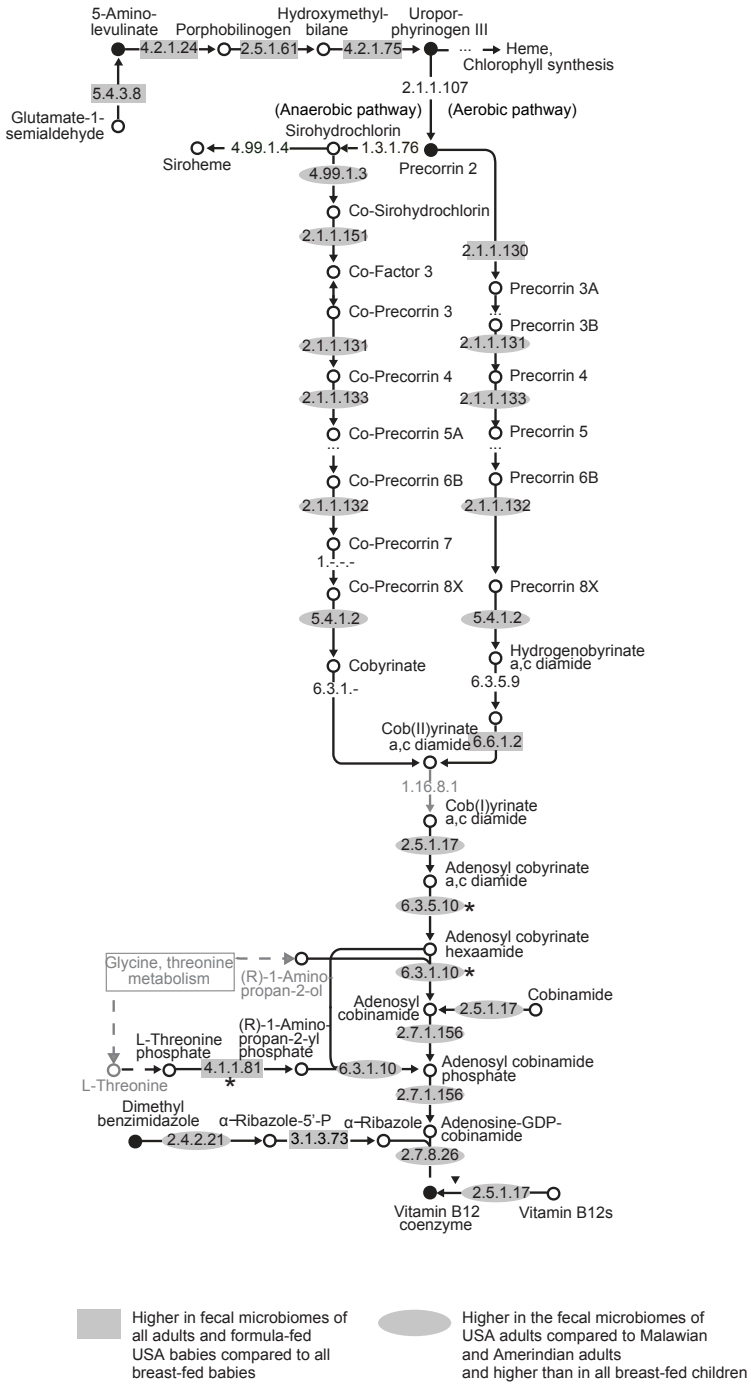


Fig. S15 – Diagram of KEGG pathway for cobalamin biosynthesis, indicating ECs whose proportional representation was higher in the fecal microbiomes of all adults and formula-fed USA infants (gray) compared to the fecal microbiomes of breast-fed babies in all populations. However, among adults, USA fecal microbiomes have higher relative representation of ECs in this pathway compared to adult Malawian/Amerindian microbiomes. p-values for the highlighted ECs can be found in **Table S7**. ECs that were discriminatory by both analyses are indicated with an asterisk.

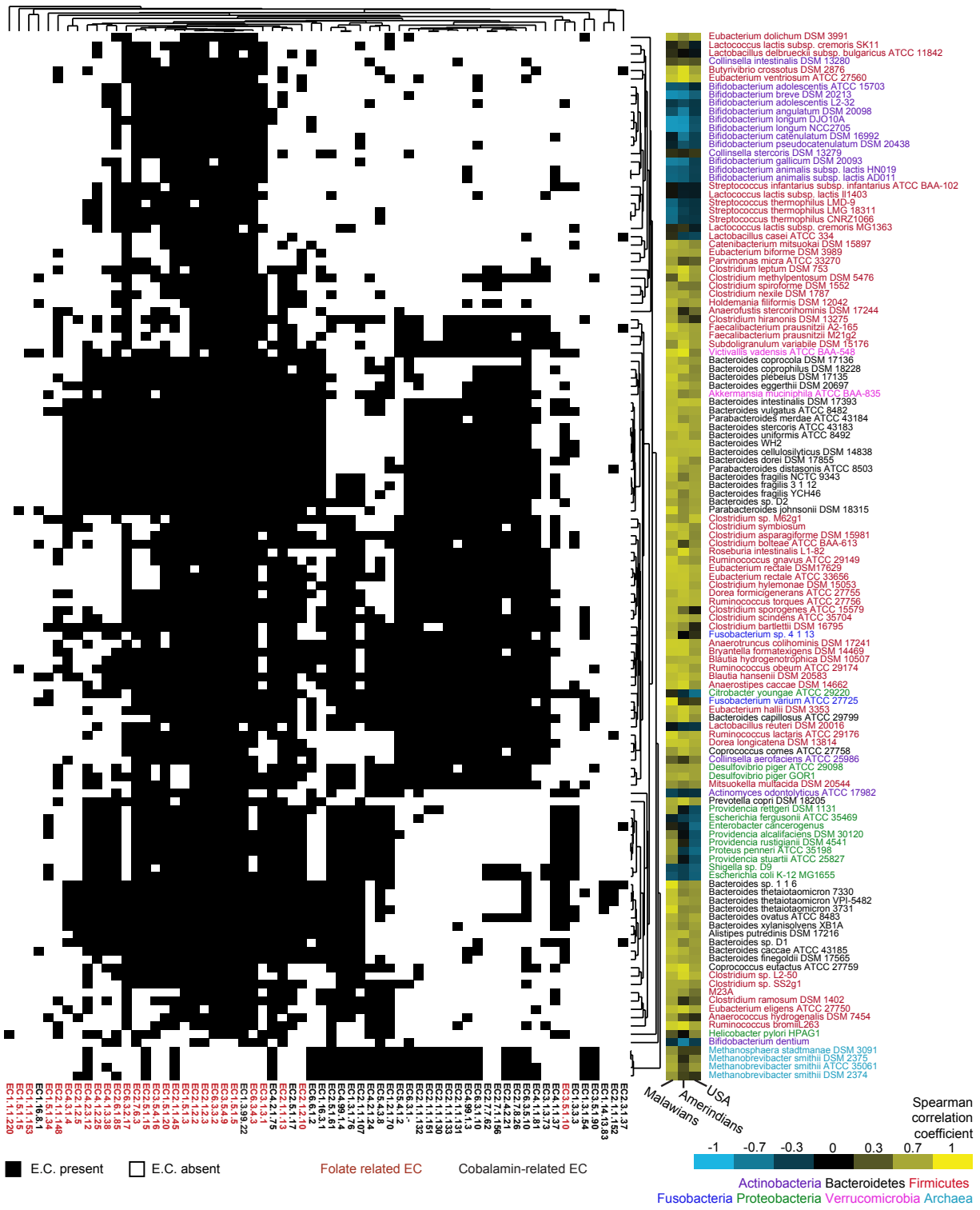


Fig. S16 – Spearman correlation between gut microbial species predicted to synthesize cobalamin and folate and their representation in fecal microbiomes at different ages and in different populations. UPGMA clustering of 126 sequenced gut genomes (average linkage method) based on the presence of the ECs involved in folate and cobalamin biosynthesis and metabolism (black squares). Spearman correlation coefficients of the proportional representation of these genomes with increasing age are shown on the right for each geographic location. A negative value indicates a decrease in the proportion of a taxon with increasing age.

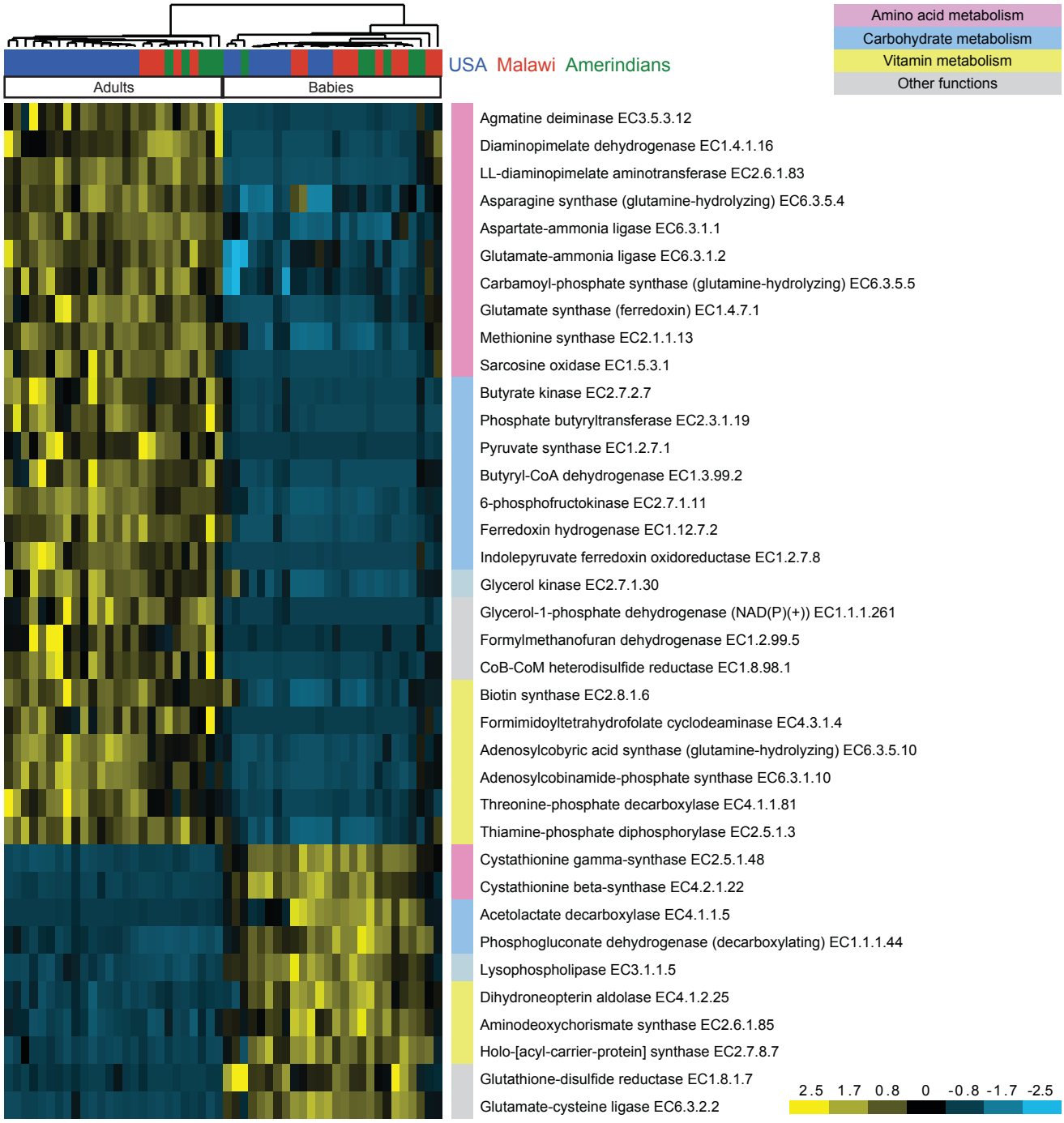


Fig. S17 – Age-related changes in the proportional representation of genes encoding ECs best discriminating baby and adult microbiomes according to both ShotgunFunctionalizeR and Random Forests analyses. UPGMA clustering (average linkage method) of fecal microbiomes, based on the relative abundances of ECs (normalized by Z-score across all datasets). The bars on the top indicate geographic location of each sampled human.

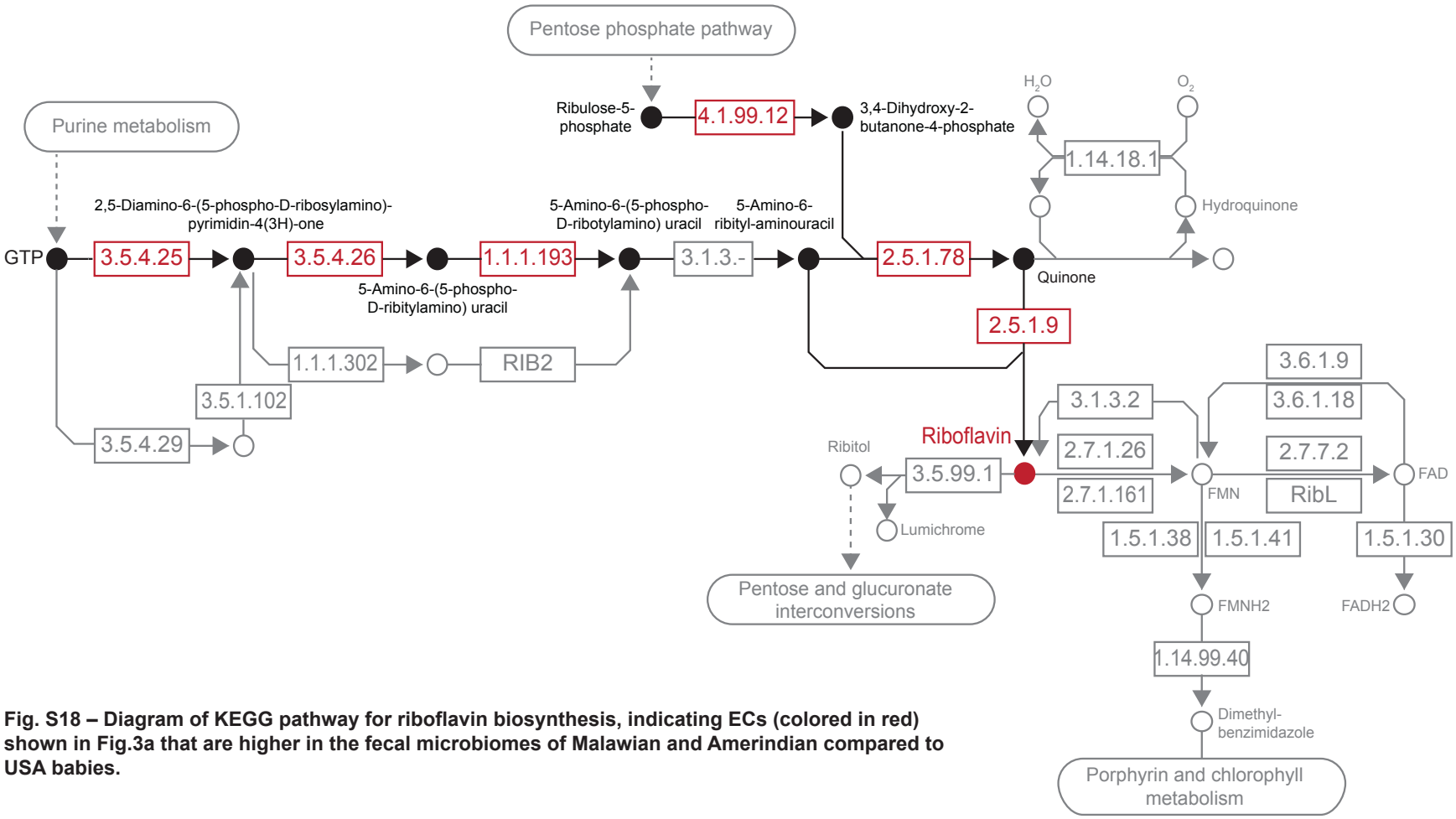
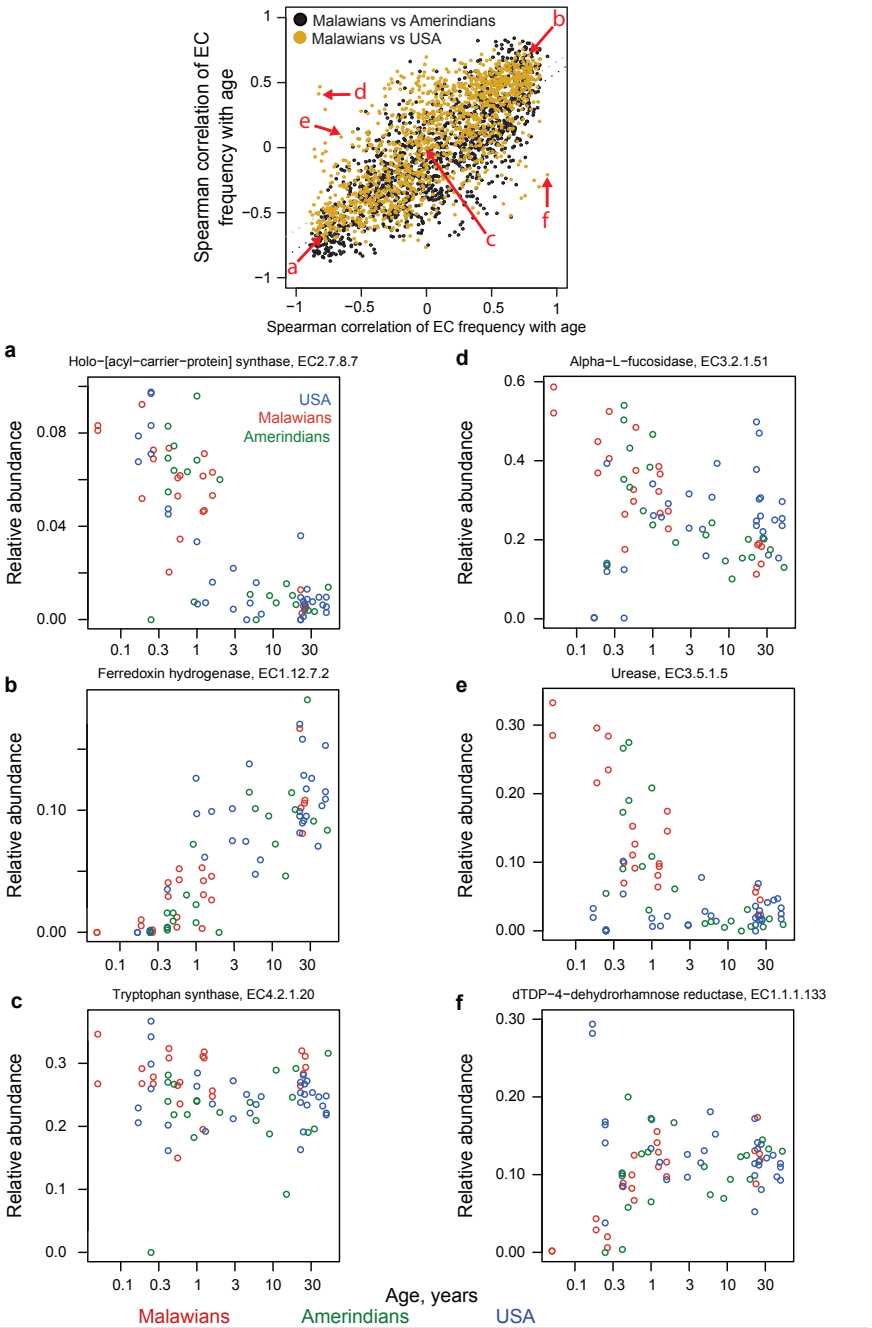


Fig. S18 – Diagram of KEGG pathway for riboflavin biosynthesis, indicating ECs (colored in red) shown in Fig.3a that are higher in the fecal microbiomes of Malawian and Amerindian compared to USA babies.



EC	Malawians Rho	Malawians p	Amerindians Rho	Amerindians p	USA Rho	USA p	Annotation
EC2.7.8.7	-0.73	1e-5	-0.49	0.004	-0.60	3e-6	Holo−[acyl−carrier−protein] synthase
EC1.12.7.2	0.83	1e-5	0.76	-0.83	0.68	2e-6	Ferredoxinhydrogenase
EC4.2.1.20	0.01	0.90	0.04	0.78	0.07	0.6	Tryptophan synthase
EC3.2.1.51	-0.75	4e-5	-0.62	0.002	0.29	0.09	Alpha−L−fucosidase
EC3.5.1.5	-0.82	2e-6	-0.74	0.0001	0.32	0.06	Urease
EC1.1.1.133	0.78	2e-5	0.34	0.13	-0.32	0.06	dTDP−4−dehydrorhamnose reductase

Fig. S19 – Changes in EC representation in fecal microbiomes as a function of age and population. Spearman correlation coefficients (Rho values) were calculated for the proportional representation of each EC against age for each human population. Plotted are Rho values for Malawians (x-axis) against Rho values for Amerindians (black points) or USA residents (yellow points). Each point represents an EC and coordinates are Rho values for that EC in Malawians (x-axis) and Amerindians or USA residents (y-axis). Spearman correlation: Malawi vs USA, Rho=0.76, $p<10^{-15}$; Amerindians vs USA Rho=0.66, $p<10^{-15}$; Malawi vs Amerindians Rho = 0.78, $p<10^{-15}$. Panels **a-f** show examples of ECs with similar or distinct Rho values for the three populations. The calculated Spearman correlation coefficient and the corresponding p-value for these examples are provided at the bottom of the figure.

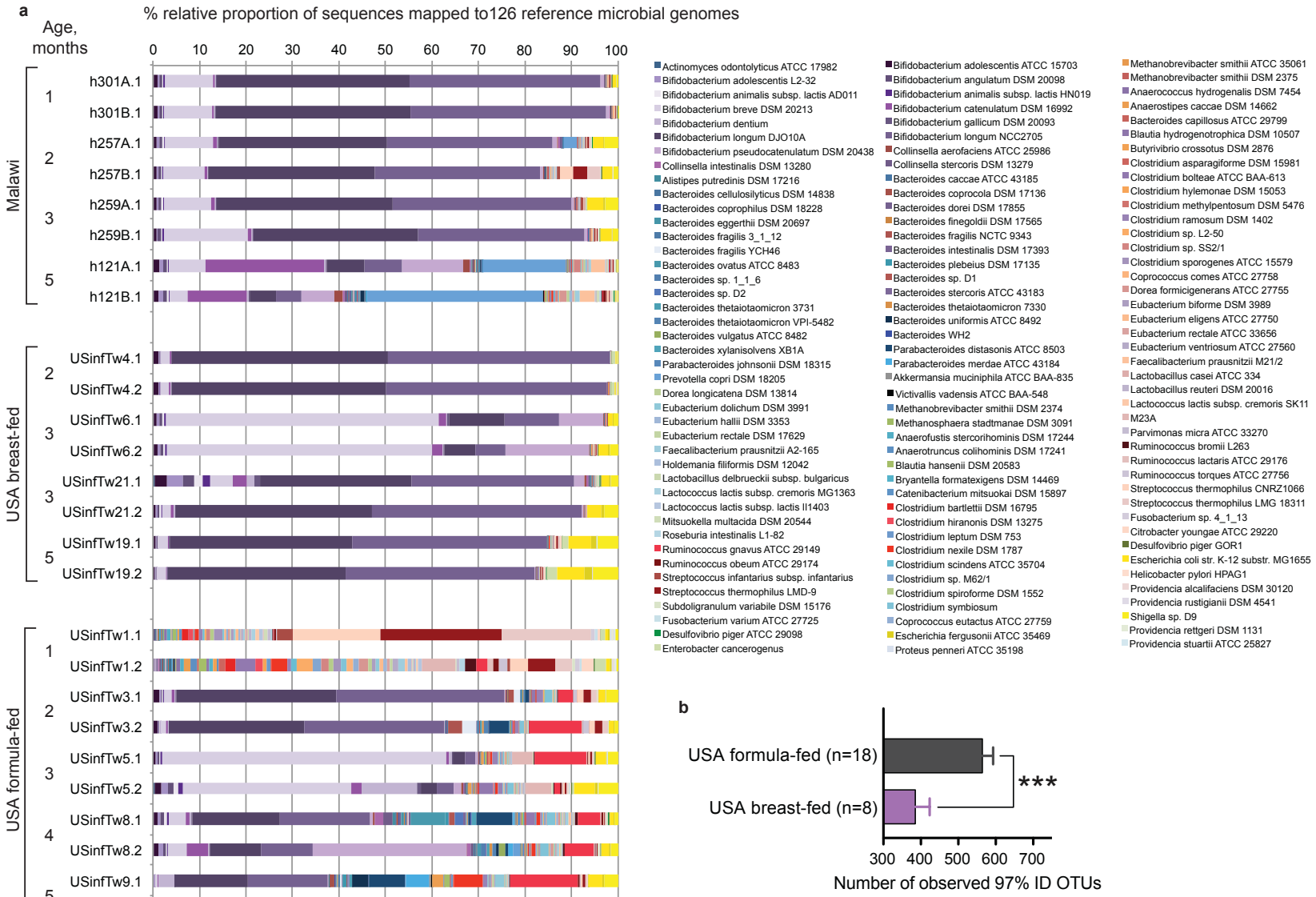


Fig. S20 – Comparison of bacterial diversity in the fecal microbiomes of breast-fed Malawian twins and breast-fed and formula-fed USA twins (1-5 months old). (a) An analysis of shotgun sequence datasets of fecal DNA using the reference database of 126 sequenced human gut genomes. **(b)** Differences in bacterial diversity between USA formula-fed and breast-fed twins, based on V4-16S rRNA data (each microbiota sampled at 281,000 sequences). ***p<0.001 (ANOVA with bonferroni post-hoc test).