ARTICLE

doi:10.1038/nature11053

Human gut microbiome viewed across age and geography

Tanya Yatsunenko[1], Federico E. Rey[1], Mark J. Manary[2,3], Indi Trehan[2,4], Maria Gloria Dominguez-Bello[5], Monica Contreras[6], Magda Magris[7], Glida Hidalgo[7], Robert N. Baldassano[8], Andrey P. Anokhin[9], Andrew C. Heath[9], Barbara Warner[2], Jens Reeder[10], Justin Kuczynski[10], J. Gregory Caporaso[11], Catherine A. Lozupone[10], Christian Lauber[10], Jose Carlos Clemente[10], Dan Knights[10], Rob Knight[10,12] & Jeffrey I. Gordon[1]

Gut microbial communities represent one source of human genetic and metabolic diversity. To examine how gut microbiomes differ among human populations, here we characterize bacterial species in fecal samples from 531 individuals, plus the gene content of 110 of them. The cohort encompassed healthy children and adults from the Amazonas of Venezuela, rural Malawi and US metropolitan areas and included mono- and dizygotic twins. Shared features of the functional maturation of the gut microbiome were identified during the first three years of life in all three populations, including age-associated changes in the genes involved in vitamin biosynthesis and metabolism. Pronounced differences in bacterial assemblages and functional gene repertoires were noted between US residents and those in the other two countries. These distinctive features are evident in early infancy as well as adulthood. Our findings underscore the need to consider the microbiome when evaluating human development, nutritional needs, physiological variations and the impact of westernization.

Genetic variation between human populations is typically viewed as differences in the allele frequencies of shared *Homo sapiens* genes. Another source of genetic and metabolic diversity resides in differences in the representation of the millions of genes and myriad gene functions within our gut microbial communities[1–3]. Sampling a broad population of healthy humans representing different ages and cultural traditions offers an opportunity to discover how our gut microbiomes evolve within a lifespan, vary between populations, and respond to our changing lifestyles[1,4–9]. Therefore, we conducted a demonstration project to address the question of whether there are discernible patterns of functional maturation of the gut communities of healthy infants and children living in geographically and culturally distinct settings.

Fecal samples were obtained from individuals in families of Guahibo Amerindians residing in two villages (Platanillal and Coromoto), separated by 10 miles, and located near Puerto Ayacucho in the Amazonas State of Venezuela (see Supplementary Table 1a, b for information about their diets). Fecal samples were also procured from members of families living in four rural communities of Malawi located within 10–70 miles of one another (Chamba, Makwhira, Mayaka and Mbiza). Lifestyles in these villages are very similar, and diets are relatively monotonous, dominated by maize (Supplementary Table 1c). In addition, we sampled families distributed across the United States, including the greater metropolitan areas of St Louis, Philadelphia and Boulder. The sampled populations included parents and siblings, and, in the United States and Malawi, monozygotic and dizygotic twin pairs. A total of 531 individuals (151 families) were studied: 115 individuals (34 families) from Malawi; 100 individuals (19 families) from Venezuela; and 316 individuals (98 families) from the United States (see Supplementary Table 2 for subject characteristics; note that all except 35 adults and one child from the United States were explicitly recruited for this study).

DNA was extracted from a single fecal sample donated by each person. Variable region 4 (V4) of bacterial 16S ribosomal RNA genes present in each fecal community was amplified by PCR, and the resulting amplicons were sequenced on an Illumina HiSeq 2000 instrument ($n = 1{,}803{,}250 \pm 562{,}877$ (mean $\pm$ s.d.) reads per fecal sample; 1,093,740,274 total reads; Supplementary Table 2a) to define the phylogenetic types (phylotypes) present. Species-level bacterial phylotypes were defined as organisms sharing $\geq$97% nucleotide sequence identity in the V4 regions of their 16S rRNA genes[10]. In addition, we characterized functions encoded in community DNA by performing multiplex shotgun 454 pyrosequencing of fecal DNA from a subset of 110 fecal samples, encompassing 43 families with members matched as closely as possible for age ($155{,}890 \pm 87{,}083$ reads per sample; total size of data set, 5.9 Gb; Supplementary Table 2b). The resulting shotgun reads were annotated with Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology group (KO) assignments and with Enzyme Commision (EC) numbers (KEGG version 58).

Taxonomic changes as a function of age and population

Many reports have examined the bacterial species content of the gastrointestinal tracts of infants and children within one population using culture-based methods. Far fewer studies have attempted to compare the gut communities of humans living in markedly different socio-economic, geographic and cultural settings[11,12]. Culture-independent techniques have been used to define the gut microbiota at various points in postnatal development[6,13], but have been limited

[1]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, Missouri 63108, USA. [2]Department of Pediatrics, Washington University School of Medicine, St Louis, Missouri 63110, USA. [3]Department of Community Health, University of Malawi College of Medicine, Blantyre, Malawi. [4]Department of Paediatrics and Child Health, University of Malawi College of Medicine, Blantyre, Malawi. [5]Department of Biology, University of Puerto Rico - Rio Piedras, Puerto Rico 00931-3360. [6]Venezuelan Institute of Scientific Research (IVIC), Carretera Panamericana, Km 11, Altos de Pipe, Venezuela. [7]Amazonic Center for Research and Control of Tropical Diseases (CAICET), Puerto Ayacucho 7101, Amazonas, Venezuela. [8]Division of Gastroenterology and Nutrition, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [9]Department of Psychiatry, Washington University School of Medicine, St Louis, Missouri 63110, USA. [10]Department of Chemistry and Biochemistry, University of Colorado, Boulder 80309, USA. [11]Department of Computer Science, Northern Arizona University, Flagstaff, Arizona 86001, USA. [12]Howard Hughes Medical Institute, University of Colorado, Boulder 80309, USA.

by the analytic methods used, by the low number of subjects examined, or by the scope of the populations surveyed. These studies have nonetheless provided important insights. Using 16S rRNA gene-based microarrays[14], a recent study found considerable intra- and interpersonal variation in fecal bacterial community structures during the first year of life in 12 unrelated children and 1 twin pair. Interpersonal variation was less within the twin pair, and intrapersonal variation decreased as a function of age. A quantitative PCR study of five bacterial taxa in the fecal microbiota of 1,032 Dutch infants at 1 month of age[15] documented differences based on birth mode (Caesarian versus vaginal; also see ref. 8).

We collected bacterial V4 16S rRNA data from 326 individuals aged 0–17 years (83 Malawian, 65 Amerindian and 178 US residents), plus 202 adults aged 18–70 years (31 Malawians, 35 Amerindians and 136 US residents). The 16S rRNA data sets were first analysed using UniFrac, an algorithm that measures similarity among microbial communities based on the degree to which their component taxa share branch length on a bacterial tree of life[16]. There were several notable findings. First, the phylogenetic composition of the bacterial communities evolved towards an adult-like configuration within the three-year period after birth in all three populations (Fig. 1a and Supplementary Fig. 1). Second, interpersonal variation was significantly greater among children than among adults; this finding was robust to geography (Fig. 1b; see also ref. 4). Third, there were significant differences in the phylogenetic composition of fecal microbiota between individuals living in the different countries, with especially pronounced separation occurring between the US and the Malawian and Amerindian gut communities; this was true for individuals aged 0–3 years, 3–17 years, and for adults (Fig. 1b and Supplementary Table 3). Unsupervised clustering using principal coordinates analysis (PCoA) of UniFrac distance matrices indicated that age and geography/cultural traditions primarily explain the variation in our data set, in which US microbiota clustered separately from non-US microbiota along principal coordinate 1 (Fig. 1c and Supplementary Fig. 2). However, within the non-US populations, separation between Malawians and Amerindians was also observed (along principal coordinate 3 in the case of adults; Supplementary Fig. 2f). We did not find any significant clustering by village for Malawians and Amerindians or by region within the United States. Fourth, bacterial diversity increased with age in all three populations (Fig. 2a, b). The fecal microbiota of US adults was the least diverse compared with the two other populations (Fig. 2c, $P < 0.005$, analysis of variance (ANOVA) with Bonferroni post-hoc test): these differences were evident in children older than 3 years of age ($P < 0.005$, ANOVA with Bonferroni post-hoc test), but not in younger subjects.

We next used the non-parametric Spearman rank correlation to determine which bacterial taxa change monotonically with increasing age within and between the three sampled populations. We only considered children who were breastfed and used data sets obtained from the V4 region of the 16S rRNA gene as well as data sets of shotgun pyrosequencing reads from the fecal microbiomes of the 110 sampled individuals (24 babies (0.6–5 months old), 60 children and adolescents (6 months to 17 years old) and 26 adults). Shotgun reads were mapped to 126 sequenced human gut-derived microbial species (Supplementary Table 4). The advantage of using these 126 gut microbes as a reference database is that spurious hits of shotgun microbiome reads to taxa that are not present in the gut are minimized. Nonetheless, when we repeated the entire analysis, blasting against 1,280 genomes in KEGG, the results were similar (Supplementary Fig. 3). Phylotypes belonging to *Bifidobacterium longum* exhibited a significant decline in proportional representation with increasing age in all three populations (Supplementary Fig. 3a). Most (75 ± 20%) shotgun and 16S rRNA V4 sequences in all babies mapped to members of the *Bifidobacterium* genus. Bifidobacteria continued to dominate fecal communities throughout the first year of life, although their

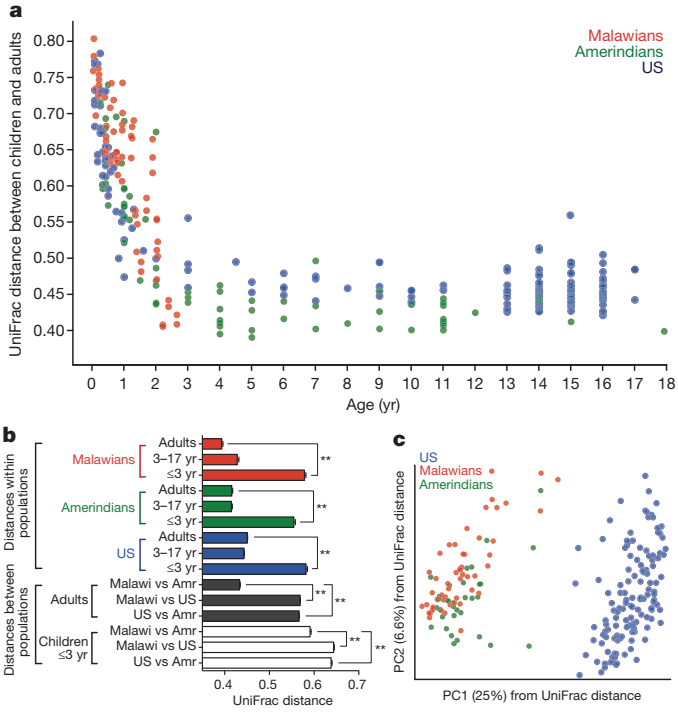

Figure 1 | Differences in the fecal microbial communities of Malawians, Amerindians and US children and adults. **a**, UniFrac distances between children and adults decrease with increasing age of children in each population. Each point shows the average distance between a child and all adults unrelated to that child but from the same country. Results are derived from bacterial V4 16S rRNA data sets. **b**, Large interpersonal variations are observed in the phylogenetic configurations of fecal microbial communities at early ages. Malawian and Amerindian (Amr) children and adults are more similar to one another than to US children and adults. UniFrac distances were defined from bacterial V4 16S rRNA data generated from the microbiota of 181 unrelated adults (≥18 years old) and 204 unrelated children ($n = 31$ Malawians 0.03–3 years old, 21 3–17 years old; 30 Amerindians 0.08–3 years old, 29 3–17 years old; 31 US residents 0.08–3 years old, 62 sampled at 3–17 years of age). $^*P < 0.05$, $^{**}P < 0.005$ (Student's *t*-test with 1,000 Monte Carlo simulations). See Supplementary Table 3 for a complete description of the statistical significance of all comparisons shown. **c**, PCoA of unweighted UniFrac distances for the fecal microbiota of adults. PC, principal coordinate.

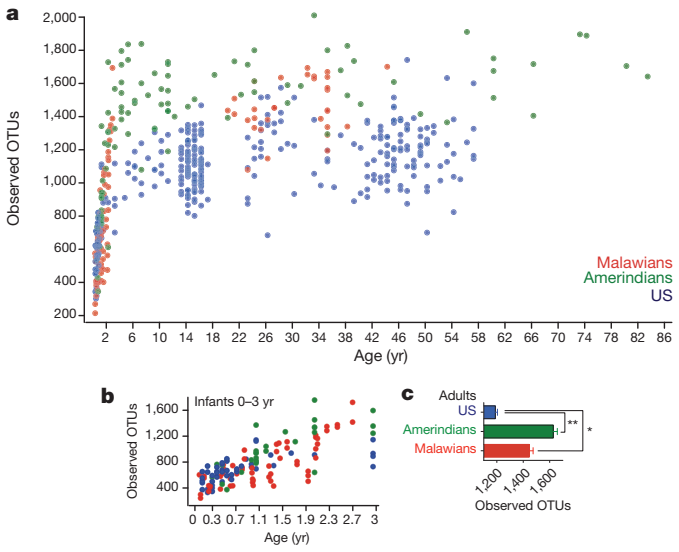

Figure 2 | Bacterial diversity increases with age in each population. **a**–**c**, The number of observed OTUs sharing ≥97% nucleotide sequence identity plotted against age for all subjects (**a**), during the first 3 years of life (**b**), and adults (**c**). Mean ± s.e.m. are shown in **c**. $^*P < 0.05$, $^{**}P < 0.005$ (ANOVA with Bonferroni post-hoc test).

proportional representation diminished during this period, in agreement with the results of several studies of small numbers of children[4,6,7] (Supplementary Fig. 3). Supplementary Table 5 lists the species-level bacterial taxa whose representation changes significantly with age in all three populations, as well as species that change in a population-specific manner as defined from analysis of the shotgun sequencing data that were available from 110 of the 531 individuals.

We also used Random Forests, a supervised machine-learning technique[17], and the V4 16S rRNA data sets obtained from 528 individuals to identify bacterial species-level operational taxonomic units (OTUs) that differentiate fecal community composition in children and adults within and between the three populations. The purpose of a classifier such as Random Forests is to learn a function that maps a set of input values or predictors (here, relative OTU abundances in a community) to a discrete output value (here, US versus non-US microbiota). Random Forests is a particularly powerful classifier that can exploit nonlinear relationships and complex dependencies between OTUs. The measure of the success of the method is its ability to classify unseen samples correctly, 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 with the baseline error that would be achieved by always guessing the most common category. As an added benefit, Random Forests assigns an importance score to each OTU by estimating the increase in error caused by removing that OTU from the set of predictors. In our analysis, we considered an OTU to be highly predictive if its importance score was at least 0.001; all error estimates and OTU importance scores were averaged over 100 rarefactions at the same sample size for each community (305,631 sequences) to control for sequencing effort.

Random Forests analysis confirmed the dominance of *Bifidobacterium* in the baby microbiota (Supplementary Table 6a). For adults, Random Forests revealed distinct community signatures for Western (US) and non-Western individuals (baseline error = 0.289, cross-validation error = 0.011 ± 0.000). Of the 92 highly predictive species-level OTUs shown in Supplementary Table 6b, 73 were over-represented in non-US adults, and 23 out of the 73 were assigned to the *Prevotella* genus. Malawians and Amerindians could also be distinguished from each other, although the difference was less extreme than the US versus non-US comparison (baseline error = 0.407, cross-validation error = 0.018 ± 0.009, 56 highly predictive OTUs; Supplementary Table 6c). Only 28 OTUs distinguished US and non-US infants (Supplementary Table 6d). Intriguingly, three OTUs assigned to the *Prevotella* genus were overrepresented in the US infant microbiota, unlike the result observed in adults (Supplementary Table 6d). Twenty-three OTUs discriminated Malawian and Amerindian baby microbiomes, 20 of which were overrepresented in the latter: most belonged to the Enterococcaceae family (Supplementary Table 6d). Thus, a Western (US) lifestyle seems to affect the bacterial component of the gut microbiota systematically, although this influence is subtle compared with the high degree of variability observed in infants and children within each population (perhaps analogous to human genetic variability, in which variation among populations is small compared to variation within populations).

Confirming the importance of *Prevotella* as a discriminatory taxon, a recent study also showed that this genus was present in higher abundance in the fecal microbiota of children living in West Africa (Burkina Faso) compared with children living in Europe (Italy)[11]. Furthermore, a member of this genus is one of three bacterial species that, in European adults, distinguishes strongly among three clusters, or enterotypes, of gut microbiota configurations that are claimed to be reproducible across Western adult populations[18]. Therefore, we asked whether the fecal microbiota of infants and adults in each of our three geographically distinct populations fell into natural discrete clusters[19]. We did not find strong evidence for discrete clustering (see Methods), but rather for variation driven in adults by a trade-off between *Prevotella* and *Bacteroides* (Supplementary Fig. 4a). Including infants introduces a new, strongly supported gradient driven by *Bifidobacteria*, generally orthogonal to the *Bacteroides*/*Prevotella* trade-off (Supplementary Fig. 4b). Clustering of sub-populations of increasing minimum age indicates that adult cluster membership is generally consistent, but that children between 0.6 and 1 year of age may be clustered with adults or with younger children, depending on whether the younger children are included in the analysis (Supplementary Fig. 4c).

This clustering analysis suggests that some features of normal variation in the bacterial composition of the gut microbiota, such as the *Prevotella*/*Bacteroides* trade-off, are highly reproducible even in human population subsets of reduced variability. However, a complete description of variation in the human gut microbiota will require a substantially broader cross-cultural and cross-age sampling. Importantly, the observed age-related and geographic patterns were also detected with lower sequencing coverage (see Supplementary Results for a discussion of the influence of sequencing depth on the performance of Random Forests and PCoA analyses (Supplementary Figs 5 and 6), and our analysis of non-bacterial taxa that vary with age and population (Supplementary Fig. 7)).

Shared functional changes over time

Few studies have described changes in the gene content of the gut microbiome as a function of age: the largest study reported so far was carried out in 13 healthy Japanese individuals (5 children, the youngest 3 months old, plus 8 adults)[4]. Our shotgun sequencing data set from 110 individuals allowed us to characterize the representation of functional gene groups (KEGG KO annotations and EC numbers) in microbiomes representing broader age groups (youngest 3 weeks), and several distinct geographic locations and cultural traditions. We used Hellinger distance measurements to show that just as children are significantly more different from one another than are adults in terms of their fecal bacterial community phylogenetic structure, they are also more different in terms of their repertoires of microbiome-encoded functions, as defined by the proportional representation of EC and KO assignments. Moreover, as with UniFrac distances, Hellinger distances were greater between the US and the other two populations at all ages sampled (Supplementary Fig. 8). Of interest is the concordance of patterns of covariance between the two data types: Procrustes analysis disclosed that the goodness of fit was significant ($P < 0.001$ with 1,000 iterations) whether UniFrac (the most appropriate metric for 16S rRNA data) or Hellinger distances (for consistency with the method used on the KEGG EC and KEGG Orthology data) were used to reduce the OTU table (Supplementary Fig. 9a, b and data not shown). Annotation of shotgun reads from the microbiomes using the Clusters of Orthologous Groups (COG) database produced similar concordance with 16S rRNA data sets (Supplementary Fig. 9c).

When examining KEGG EC profiles across 110 fecal microbiomes, we obtained the remarkable result that there were no ECs identified as being unique to adults ($n = 26$) or babies (less than 6 months old, $n = 24$). Moreover, the total number of ECs found in adults was not significantly different than the total number of ECs scored in babies (sampling normalized to coverage in Supplementary Fig. 10a). This finding was robust to geography. The fraction of sequences with assignable KEGG EC annotations declined with increasing age in all three populations (Supplementary Fig. 10b). This may be due to the increased complexity of the adult microbiome, with fewer representative species characterized by genome sequencing, genetic manipulation or biochemically (also see Supplementary Results and Supplementary Figs 11 and 12 for a comparison of our data set to a published data set of fecal microbiomes sampled from 124 adults living in Denmark and Spain[2]).

We used ShotgunFunctionalizerR[20], a software tool designed for metagenomic analysis and based on a Poisson model, to identify 1,008 ECs whose proportional representation in fecal microbiomes differed significantly between all sampled breastfed babies and all adults irrespective of their geographic location; 530 were significantly higher

in adults ($P < 0.0001$, Supplementary Table 7). A prominent example of these shared age-related changes involves the metabolism of vitamins B12 (cobalamin) and folate. In contrast to folate, which is synthesized by microbes and plants, cobalamin is primarily produced by microbes[21]. The gut microbiomes of babies are enriched in genes involved in the *de novo* biosynthesis of folate, whereas those of adults have a significantly higher representation of genes that metabolize dietary folate and its reduced form tetrahydrofolate (THF; Supplementary Fig. 13 and Supplementary Table 7). Unlike *de novo* folate biosynthetic pathway components, which decrease with age, the proportional representation of genes encoding most enzymes involved in cobalamin biosynthesis increases with age (Supplementary Figs 14, 15 and Supplementary Table 7). The folate and cobalamin pathways are linked functionally by methionine synthase (EC2.1.1.13), which catalyses the formation of THF from 5-methyl-THF and L-homocysteine, requiring cobalamin as a cofactor; the representation of this enzyme also increases with age (Supplementary Fig. 13).

The low relative abundance of ECs involved in cobalamin biosynthesis in the fecal microbiomes of babies correlates with the lower representation of members of Bacteroidetes, Firmicutes and Archaea in their microbiota (see Supplementary Fig. 16 for Spearman correlation coefficients). Although the biosynthetic pathway for cobalamin is well represented in the genomes of these organisms (Supplementary Fig. 16), *Bifidobacterium*, *Streptococcus*, *Lactococcus* and *Lactobacillus*, which dominate the baby gut microbiota (Supplementary Table 5 and Supplementary Fig. 3), are deficient in these genes (Supplementary Fig. 16). By contrast, several of these early gut colonizers contain ECs involved in folate biosynthesis and metabolism (Supplementary Fig. 16). The conventional view of the developing infant gut is that the main change is in the representation of Bifidobacteria. Although differences in the representation of Bifidobacteria contribute to this effect, differences in vitamin metabolism among the rest of the bacteria remain even when all Bifidobacteria reads were excluded (data not shown). These changes in vitamin biosynthetic pathway representation in the microbiome correlate with published reports indicating that blood levels of folate decrease and cobalamin increase with age[22].

Besides cobalamin and folate, the relative abundance of ECs involved in the biosynthesis of vitamins B7 (biotin) (biotin synthase, EC2.8.16) and B1 (thiamine) (thiamine-phosphate diphosphorylase, EC2.5.1.3) are significantly higher in adult microbiomes than the microbiomes of babies (Supplementary Fig. 17 and Supplementary Table 7). Together, these findings suggest that the microbiota should be considered when assessing the nutritional needs of humans at various stages of development.

Random Forests analysis asks a different statistical question from ShotgunFunctionalizeR: that is, which genes or species are most discriminatory among different class labels, rather than which are most over/underrepresented, and tends to identify fewer features than ShotgunFunctionalizeR when applied to the same data. Random Forests analysis yielded 107 ECs that best discriminate the adult and baby microbiomes (Supplementary Table 7). These predictive ECs were among the most significantly different ECs determined by ShotgunFunctionalizeR and included ECs involved in the metabolism of cobalamin and folate (Supplementary Table 7). In addition, Random Forests showed that ECs involved in fermentation, methanogenesis and the metabolism of arginine, glutamate, aspartate and lysine were higher in the adult microbiomes, whereas ECs involved in the metabolism of cysteine and fermentation pathways found in lactic acid bacteria (acetolactate decarboxylase (EC4.1.1.5) and 6-phosphogluconate dehydrogenase (EC1.1.1.44)) were mainly represented in baby microbiomes (Supplementary Fig. 17).

Comparison of the representation of KEGG KOs between baby and adult microbiomes yielded essentially the same results as those reported with ECs. The only new finding was the overrepresentation of KEGG KOs assigned to a wide variety of ATP-binding cassette (ABC) transporters in baby microbiomes (Supplementary Table 7b).

Population- and age-specific differences

ShotgunFunctionalizeR, Random Forests and Spearman rank correlation analyses were all used to compare EC representation in fecal microbiomes as a function of predefined categories of geographic location and age. A total of 476 ECs were identified as being significantly different in US versus Malawian and Amerindian breastfed babies ($P < 0.0001$, ShotgunFunctionalizeR; Supplementary Table 8). The most prominent differences involved pathways related to vitamin biosynthesis and carbohydrate metabolism. Malawian and Amerindian babies had higher representation of ECs that were components of the vitamin B2 (riboflavin) biosynthetic pathway (Fig. 3a and Supplementary Fig. 18). These differences were not evident in adults (Supplementary Table 7). Riboflavin is found in human milk and in meat and dairy products. We did not measure the levels of these vitamins in mothers and in their breastmilk in the sampled populations, although it is tempting to speculate that the observed differences in baby microbiomes may represent an adaptive response to vitamin availability.

Studies in gnotobiotic mouse models indicate that the ability of members of the microbiota to access host-derived glycans plays a key part in establishing a gut microbial community[23,24]. As expected[4,5], compared with adults, baby microbiomes were enriched in ECs involved in the foraging of glycans represented in breastmilk and the intestinal mucosa (mannans, sialylated glycans, galactose and fucosyloligosaccharides; Supplementary Table 7). Several genes involved in using these host glycans are significantly overrepresented in Amerindian and Malawian baby microbiomes compared with US

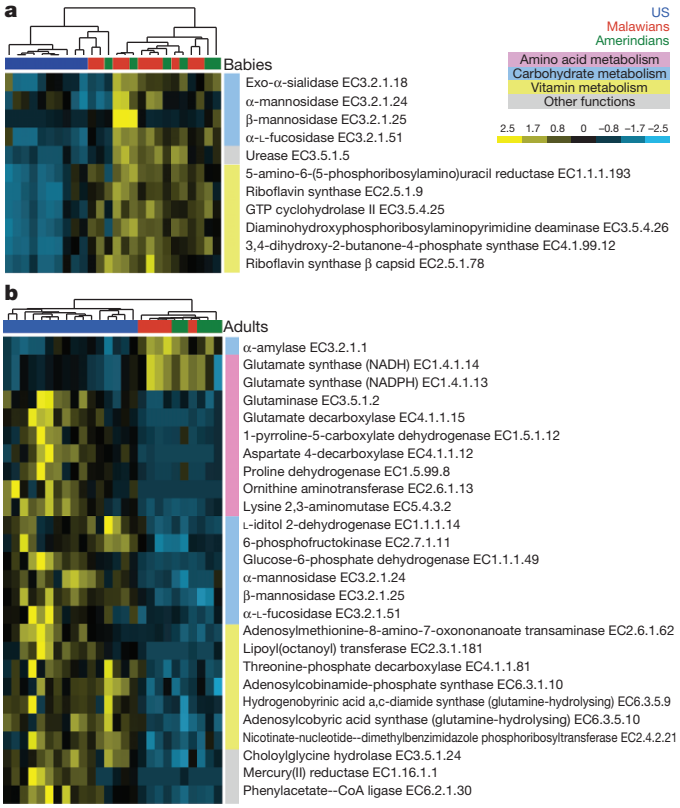

Figure 3 | Differences in the functional profiles of fecal microbiomes in the three study populations. Examples of KEGG ECs that showed the largest differences, as determined by Random Forests and ShotgunFunctionalizeR analyses, in proportional representation between US and Malawian/Amerindian populations. Shown are the relative abundances of genes encoding the indicated ECs (normalized by *Z*-score across all data sets). **a**, UPGMA (unweighted pair group method with arithmetic mean) clustering of 10 US, 10 Malawian and 6 Amerindian baby fecal microbiomes. **b**, UPGMA clustering of 16 US, 5 Malawian and 5 Amerindian adult fecal microbiomes.

baby microbiomes, most notably exo-α-sialidase and α-L-fucosidase (Fig. 3a and Supplementary Table 8). These population-specific biomarkers may reflect differences in the glycan content of breastmilk. In fact, the representation of these glycoside hydrolases decreases as Malawian and Amerindian babies mature and transition to a diet dominated by maize, cassava and other plant-derived polysaccharides. By contrast, α-fucosidase gene representation in the US infants increases with age and as they become exposed to diets rich in readily absorbed sugars (Supplementary Fig. 19d and Supplementary Table 9).

Another biomarker that distinguishes microbiomes based on age and geography is urease (EC3.5.1.5). Urease gene representation is significantly higher in Malawian and Amerindian baby microbiomes and decreases with age in these two populations, unlike in the United States, where it remains low from infancy to adulthood (Fig. 3a and Supplementary Fig. 19e). Urea comprises up to 15% of the nitrogen present in human breastmilk[25]. Urease releases ammonia that can be used for microbial biosynthesis of essential and nonessential amino acids[26,27]. Furthermore, urease has a crucial involvement in nitrogen recycling, particularly when diets are deficient in protein[28,29]. Under conditions in which dietary nitrogen is limiting, the ability of the microbiome to use urea would presumably be advantageous to both the microbial community and host. Although urease is generally attributed to *Helicobacter* and *Proteus* spp., the relative abundance of members of these two genera was low (<0.05%) and not significantly different between the three populations. Urease activity has been characterized previously in *Streptococcus thermophilus*[30]. Our analysis of shotgun reads that matched to the 126 reference gut genomes showed that the representation of five species that possess EC3.5.1.5 (*Bacteroides cellulosilyticus* WH2, *Coprococcus comes*, *Roseburia intestinalis*, *Streptococcus infantarius* and *S. thermophilus*) was significantly higher in Malawian and Amerindian baby microbiomes than in US baby microbiomes (Supplementary Table 5).

Further support of the role of diet in shaping the infant gut microbiome comes from the differences detected between breastfed and formula-fed babies who were part of the US infant twin cohort (see Supplementary Results and Supplementary Figs 2c, 8 and 20).

Differences in adult fecal microbiomes

Annotation of the shotgun sequencing data sets yielded a total of 1,349 ECs in the 26 adults surveyed: ShotgunFunctionalizeR showed that the representation of genes encoding 893 of these ECs was significantly different in US versus Malawian/Amerindian fecal microbiomes ($P < 0.005$ after multiple comparison correction; 433 overrepresented in US samples). By contrast, at this threshold only 445 ECs were identified as different between Malawian and Amerindian adults (see Supplementary Table 10 for a complete list). The Random Forests classifier revealed 52 ECs that were best at discriminating US versus non-US adult fecal microbiomes. These ECs were also identified by ShotgunFunctionalizeR as the most significantly different (Supplementary Table 10).

A typical US diet is rich in protein, whereas diets in Malawi and Amerindian populations are dominated by corn and cassava (Supplementary Table 1). The differences between US and Malawian/Amerindian microbiomes can be related to these differences in diet. The ECs that were the most significantly enriched in US fecal microbiomes parallel differences observed in carnivorous versus herbivorous mammals[31]. ECs encoding glutamate synthase have higher proportional representation in Malawian and Amerindian adult microbiomes and are also higher in herbivorous mammalian microbiomes[31] (Fig. 3b), whereas the degradation of glutamine was overrepresented in US as well as carnivorous mammalian microbiomes. Several ECs involved in the degradation of other amino acids were overrepresented in adult US fecal microbiomes: aspartate (EC4.1.1.12), proline (EC1.5.99.8), ornithine (EC2.6.1.13) and lysine (EC5.4.3.2) (Fig. 3b), as were ECs involved in the catabolism of simple sugars (glucose-6-phosphate dehydrogenase and 6-phosphofructokinase), sugar substitutes (L-iditol 2-dehydrogenase, which degrades sorbitol), and host glycans (α-mannosidase, β-mannosidase and α-fucosidase; Fig. 3b). By contrast, α-amylase (EC3.2.1.1), which participates in the degradation of starch, was overrepresented in the Malawian and Amerindian microbiomes, reflecting their corn-rich diet.

US microbiomes also had significant overrepresentation of ECs involved in vitamin biosynthesis (cobalamin (Fig. 3b and Supplementary Fig. 15), biotin and lipoic acid (Fig. 3b)), in the metabolism of xenobiotics (phenylacetate CoA ligase (EC6.2.1.30), which participates in the metabolism of aromatic compounds, and mercury reductase (EC1.16.1.1)), and in bile salt metabolism (choloylglycine hydrolase (EC3.5.1.24), perhaps reflecting a diet richer in fats (Fig. 3b)).

Effects of kinship on the microbiome across countries

Differences in social structures may influence the extent of vertical transmission of the microbiota and the flow of microbes and microbial genes among members of a household. Differences in cultural traditions also affect food, exposure to pets and livestock, and many other factors that could influence how and from where a gut microbiota/microbiome is acquired. We previously observed that adult monozygotic twins are no more similar to one another in terms of their gut bacterial community structure than are adult dizygotic twins[32]. This result suggests that the overall heritability of the microbiome is low. We confirmed that the phylogenetic architecture of the fecal microbiota of monozygotic Malawian co-twins ≤3 years of age is

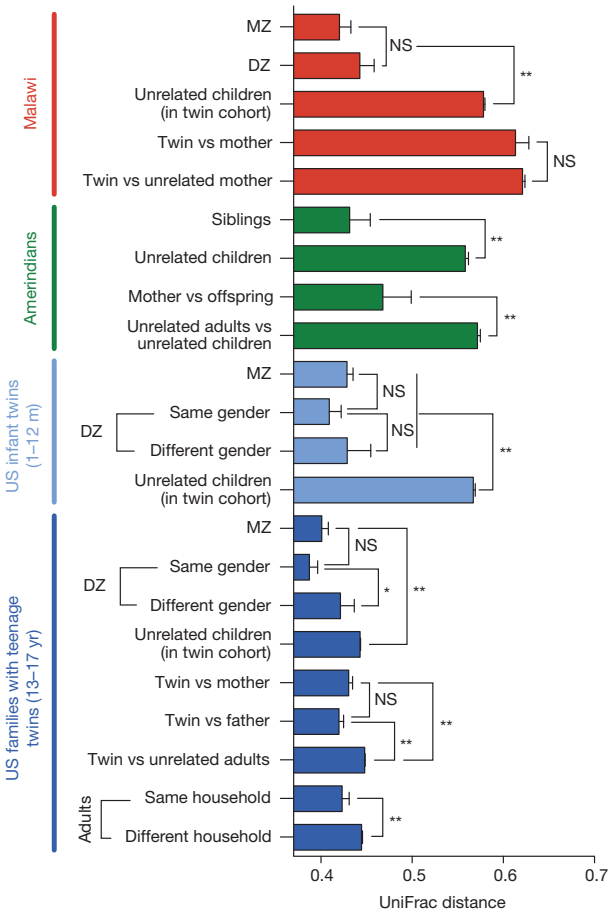

Figure 4 | Differences in the fecal microbiota between family members across the three populations studied. UniFrac distances between the fecal bacterial communities of family members were calculated ($n = 19$ Amerindian, 34 Malawian families, and 54 US families with teenage twins). DZ, dizygotic; MZ, monozygotic. Mean and s.e.m. values are plotted. The UniFrac matrix was permuted 1,000 times; P values represent the fraction of times permuted differences were greater than real differences. m, months; NS (not significant; $P > 0.05$), *$P < 0.05$, **$P < 0.005$.

no more similar than the microbiota of similarly aged dizygotic co-twins ($n = 15$ monozygotic and 6 dizygotic twin pairs). We found that this is also true for monozygotic and dizygotic twin pairs aged 1–12 months ($n = 16$ twin pairs), as well as teenaged twins (13–17 years old; $n = 50$ pairs) living together in the United States (Fig. 4).

Although biological mothers are in a unique position to transmit an initial inoculum of microbes to their infants during and after birth, our analysis of mothers of teenage US twins showed that their fecal microbiota were no more similar to their children than were those of biological fathers, and that genetically unrelated but co-habiting mothers and fathers were significantly more similar to one another microbially than were members of different families (Fig. 4; note that no fathers were sampled in Malawi and only four fathers in the Amerindian cohort). These latter observations emphasize the importance of a history of numerous common environmental exposures in shaping gut microbial ecology. Moreover, the similarity in the overall pattern of the effects of kinship on microbial community structure suggests that despite the large influence of cultural factors on which microbes are present in both children and adults in each population, the bases for the degree of similarity among members of a family are consistent across the three populations studied.

Prospectus

Our results emphasize that it is essential to sample a broad population of healthy humans over time, both in terms of their age, geography and cultural traditions, to discover features of gut microbiomes that are unique to different locations and lifestyles. In addition, we need to understand how the pressures of westernization are changing the microbial parts of our genetic landscape—changes that potentially mediate the suite of pathophysiological states correlated with westernization. Finally, given the need for global policies about sustainable agriculture and improved nutrition, it will be important to understand how we can match these policies not only to our varying cultural traditions but also to our varied gut microbiomes.

METHODS SUMMARY

Sample collection. Subjects were recruited for the present study using procedures approved by Human Studies Committees from Washington University in St Louis, Children's Hospital of Philadelphia, the University of Colorado, Boulder, the University of Malawi, the University of Puerto Rico, and the Venezuelan Institute for Scientific Research. Each fecal sample was frozen within 30 min of donation.

Multiplex DNA sequencing. Extracted genomic DNA was subjected to multiplex Illumina sequencing of the V4 region of bacterial 16S rRNA genes, as well as multiplex 454 pyrosequencing of total community DNA. See Methods for further details about the analysis.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 February 2011; accepted 20 March 2012.
Published online 9 May 2012.

1. Mueller, S. *et al.* Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. *Appl. Environ. Microbiol.* **72,** 1027–1033 (2006).
2. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
3. Li, M. *et al.* Symbiotic gut microbes modulate human metabolic phenotypes. *Proc. Natl Acad. Sci. USA* **105,** 2117–2122 (2008).
4. Kurokawa, K. *et al.* Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* **14,** 169–181 (2007).
5. Koenig, J. E. *et al.* Microbes and Health Sackler Colloquium: Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl Acad. Sci. USA* **108** (suppl. 1), 4578–4585 (2011).
6. Favier, C. F., Vaughan, E. E., De Vos, W. M. & Akkermans, A. D. Molecular monitoring of succession of bacterial communities in human neonates. *Appl. Environ. Microbiol.* **68,** 219–226 (2002).
7. Tannock, G. W. What immunologists should know about bacterial communities of the human bowel. *Semin. Immunol.* **19,** 94–105 (2007).
8. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl Acad. Sci. USA* **107,** 11971–11975 (2010).
9. Blaser, M. J. & Falkow, S. What are the consequences of the disappearing human microbiota? *Nature Rev. Microbiol.* **7,** 887–894 (2009).
10. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335–336 (2010).
11. 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. USA* **107,** 14691–14696 (2010).
12. Peach, S., Fernandez, F., Johnson, K. & Drasar, B. S. The non-sporing anaerobic bacteria in human faeces. *J. Med. Microbiol.* **7,** 213–221 (1974).
13. Mackie, R. I., Sghir, A. & Gaskins, H. R. Developmental microbial ecology of the neonatal gastrointestinal tract. *Am. J. Clin. Nutr.* **69,** 1035S–1045S (1999).
14. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5,** e177 (2007).
15. Penders, J. *et al.* Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* **118,** 511–521 (2006).
16. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).
17. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35,** 343–359 (2011).
18. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
19. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–108 (2011).
20. Kristiansson, E., Hugenholtz, P. & Dalevi, D. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* **25,** 2737–2738 (2009).
21. Kräutler, B. Vitamin B_{12}: chemistry and biochemistry. *Biochem. Soc. Trans.* **33,** 806–810 (2005).
22. Monsen, A. L., Refsum, H., Markestad, T. & Ueland, P. M. Cobalamin status and its biochemical markers methylmalonic acid and homocysteine in different age groups from 4 days to 19 years. *Clin. Chem.* **49,** 2067–2075 (2003).
23. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4,** 447–457 (2008).
24. Hooper, L. V., Xu, J., Falk, P. G., Midtvedt, T. & Gordon, J. I. A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem. *Proc. Natl Acad. Sci. USA* **96,** 9833–9838 (1999).
25. Harzer, G., Franzke, V. & Bindels, J. G. Human milk nonprotein nitrogen components: changing patterns of free amino acids and urea in the course of early lactation. *Am. J. Clin. Nutr.* **40,** 303–309 (1984).
26. Metges, C. C. *et al.* Incorporation of urea and ammonia nitrogen into ileal and fecal microbial proteins and plasma free amino acids in normal men and ileostomates. *Am. J. Clin. Nutr.* **70,** 1046–1058 (1999).
27. Millward, D. J. *et al.* The transfer of 15N from urea to lysine in the human infant. *Br. J. Nutr.* **83,** 505–512 (2000).
28. Meakins, T. S. & Jackson, A. A. Salvage of exogenous urea nitrogen enhances nitrogen balance in normal men consuming marginally inadequate protein diets. *Clin. Sci. (Lond.)* **90,** 215–225 (1996).
29. Langran, M., Moran, B. J., Murphy, J. L. & Jackson, A. A. Adaptation to a diet low in protein: effect of complex carbohydrate upon urea kinetics in normal man. *Clin. Sci. (Lond.)* **82,** 191–198 (1992).
30. Mora, D. *et al.* Characterization of urease genes cluster of *Streptococcus thermophilus*. *J. Appl. Microbiol.* **96,** 209–219 (2004).
31. Muegge, B. D. *et al.* Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332,** 970–974 (2011).
32. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Wagoner and J. Manchester for superb technical assistance, plus B. Muegge, A. Grimm, A. Hsiao, N. Griffin and P. Tarr for suggestions, and M. Ndao, T. Tinnin and R. Mkakosya for patient recruitment and/or technical assistance. This work was supported in part by grants from the National Institutes of Health (DK078669, T32-HD049338), St. Louis Children's Discovery Institute (MD112009-201), the Howard Hughes Medical Institute, the Crohn's and Colitis Foundation of America, and the Bill and Melinda Gates Foundation. Parts of this work used the Janus supercomputer, which is supported by National Science Foundation grant CNS-0821794, the University of Colorado, Boulder, the University of Colorado, Denver, and the National Center for Atmospheric Research.

Author Contributions T.Y., R.K. and J.I.G. designed the experiments, M.J.M., I.T., M.G.D.-B., M.C., M.M., G.H., A.C.H., A.P.A., R.K., R.N.B., C.A.L., C.L. and B.W. participated in patient recruitment, T.Y. generated the data, T.Y., F.E.R., J.R., J.K., J.G.C., J.C.C., D.K., R.K. and J.I.G. analysed the results, T.Y., R.K. and J.I.G. wrote the paper.

Author Information DNA sequences have been deposited in MG-RAST (http://metagenomics.anl.gov/) under accession numbers 'qiime:850' for Illumina V4 16S rRNA data sets, and 'qiime:621' for fecal microbiome shotgun sequencing data sets. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.I.G. (jgordon@wustl.edu).

METHODS

Isolation of fecal DNA and multiplex sequencing. Each participant provided a fecal specimen that was frozen within 30 min. All samples were stored at −80 °C and subjected to a common protocol for DNA extraction. Fecal samples were pulverized with a mortar and pestle at −80 °C. Genomic DNA was extracted from 400 mg aliquots of frozen pulverized samples. Methods for multiplex Illumina sequencing of V4 amplicons have been described[33].

For multiplex shotgun 454 Titanium FLX pyrosequencing, each fecal community DNA sample was randomly fragmented by nebulization (500–800 base pairs) and then labelled with a distinct Multiplex IDentifier (MID; Roche) according the manufacturer's protocol (Rapid Library preparation for FLX Titanium, Roche). Equivalent amounts of 12 MID-labelled samples were pooled before each pyrosequencer run.

Data analysis. 16S rRNA OTUs were picked from the Illumina reads using a closed-reference OTU picking protocol against the Greengenes database clustered at 97% identity (that is, uclust_ref: sequences are clustered against a reference database, and reads that do not match the database are excluded from further analyses) with uclust using the QIIME suite of software tools[10] version 1.3.0-dev (pick_otus.py parameters: –max_accepts 1 –max_rejects 8 –stepwords 8 –word_length 8). Of the 1,093,740,274 Illumina reads from the V4 region of bacterial 16S rRNA genes that passed the QIIME quality filters, 87.1% (952,115,802) matched a reference sequence at ≥97% nucleotide sequence identity. Taxonomy assignments were associated with OTUs based on the taxonomy associated with the Greengenes reference sequence defining each OTU. UniFrac distances between samples were calculated using the Greengenes reference tree. Greengenes reference sequences, trees and taxonomy data used in this analysis can be found at: http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/Reference_OTUs_for_Pipelines/Caporaso_Reference_OTUs/gg_otus_4feb2011.tgz.

A table of OTU counts per sample was generated and used in combination with the tree to calculate α and β diversity. To generate unweighted UniFrac distance matrices, all communities were rarefied to 290,609 V4 16S rRNA reads per sample. Unweighted UniFrac rather than weighted UniFrac was used for analyses owing to the large differences in taxonomic representation among the samples. Nonetheless, the patterns were similar with weighted UniFrac (data not shown). Rarefaction analysis was conducted using the QIIME scripts multiple_rarefaction.py, alpha_diversity.py and collate_alpha.py. The QIIME metric 'observed species' was used to estimate α diversity in the data set.

Clustering analysis. Testing for discrete clusters was performed on the rarefied versions of the 16S rRNA OTU relative abundance tables. OTU counts were binned into genus-level taxonomic groups according to the taxonomic assignments described earlier. Several distance measures were considered, including Jensen–Shannon divergence, Bray–Curtis and weighted/unweighted UniFrac distances. Clustering was performed via partitioning around medoids in the R package cluster[34]. The choice of number of clusters and quality of the resulting clusters were assessed by maximizing the silhouette index[35]. Traditionally, silhouette indices of 0.5 or above have been considered evidence of reasonable clustering structure. Although some silhouette scores above 0.5 were found in this data set (for example, for two clusters when clustering all adult populations with Jensen–Shannon divergence), reclustering within different subpopulations (for example, individual countries) introduced new cluster boundaries with silhouette scores still near or above 0.5, indicating that silhouette index scores may need to be substantially above 0.5 to claim clustering structure for microbial enterotype testing. We also tested for discrete clusters using the prediction strength measure[36], which showed negative results for all distances measures but unweighted UniFrac (prediction strength = 0.963 ± 0.012 (mean ± s.d.)). We estimated the s.d. by tenfold jackknifing.

Shotgun sequences from fecal microbiomes. Shotgun reads were filtered using custom Perl scripts and publicly available software to remove (1) all reads <60 nucleotides; (2) Titanium pyrosequencing reads with two continuous and/or three total degenerate bases (N); (3) all duplicates (a known artefact of pyrosequencing), defined as sequences in which the initial 20 nucleotides are identical and that share an overall identity of >97% throughout the length of the shortest read[37]; and (4) all sequences with significant similarity to human reference genomes (Blastn *E*-value threshold ≤ 10^{-5}, bitscore ≥ 50, percentage identity ≥ 75%) to ensure the continued de-identification of samples.

Searches against the database of 126 human gut bacterial genomes were conducted with Blastn. A sequence read was annotated as the best hit in the database if the *E*-value was ≤10^{-5}, the bit score was ≥50, and the alignment was at least 95% identical between query and subject. Relative abundances of reads mapped to each of the 126 genomes were adjusted to genome sizes. Searches against the protein-coding component of the KEGG database (v58) and against COG (v8.3) were conducted with BLASTX. (Note that when we performed searches against a separate KEGG database of intergenic regions alone, very few hits were observed.) Counts were normalized to the mapped reads. In total, 40 ± 8% reads were mapped to KEGG KOs and 56 ± 11% to COG; 44 ± 16% of the reads mapped to the 126 gut genomes using 95% sequence similarity cut-off. Unmapped reads were excluded from the analyses shown in the main text, although repeating the analyses including these reads had little effect on the results. To quantify the differences in KEGG EC profiles among fecal microbiomes, evenly rarefied matrices of EC counts were created with all samples, and Hellinger distances were calculated using QIIME.

Spearman rank correlations were carried out using the R statistical software[38]. To identify bacterial taxa that change with increasing age in each population, the proportion of reads that map to each of the 126 reference sequenced human gut genomes in each fecal microbiome was identified. The relative abundance of reads from each genome was then correlated with age (years) for each geographic region. To identify genes encoding ECs that change with age, the proportion of reads annotated with each EC in each fecal microbiome was identified. The relative abundance of each EC was subsequently correlated with age (years) for each geographic region.

Random Forests analysis. Random Forests analysis was applied as described in ref. 8, using the randomForest package in R[39] with 500 trees and all default settings. The generalization error was estimated using fivefold cross-validation for all comparisons involving adults from the 16S rRNA data; leave-one-out cross-validation was used for all other comparisons. For each comparison, the relevant subset of samples was extracted from the table of OTU or EC counts, and all singleton OTUs/ECs (or all OTUs/ECs present in fewer than 5 samples for the 16S rRNA comparisons involving adults) were subsequently removed. Random Forests analysis was performed for each comparison on 100 rarefied versions of the data, and the average cross-validation error estimates and OTU/EC importance estimates were reported. Rarefaction depths were chosen manually to include all samples without exceptionally low total sequences. The chosen depth for each comparison and the resulting number of samples are shown in Supplementary Tables 6–8 and Supplementary Fig. 6. For the analysis shown in Supplementary Fig. 6a, we compared the generalization errors obtained when using 16S rRNA-based OTUs from the Illumina V4 data sets at various sequencing depths. For direct evaluation of the predictive strength of the Illumina-based OTUs, we rarefied at the lowest observed depth of 305,631 sequences for each classification task, as well as at sequencing depths of 100, 1,000, 10,000, 100,000 and 1,000,000 reads per sample. The mean and s.d. of the cross-validation error were estimated for each classification tasking using ten independent rarefactions of the data. We also included the expected 'baseline' error obtained by a classifier that simply predicts the most common class label.

Data deposition. DNA sequences have been deposited in MG-RAST (http://metagenomics.anl.gov/) under accession numbers 'qiime:850' for Illumina V4 16S rRNA data sets, and 'qiime:621' for fecal microbiome shotgun sequencing data sets.

33. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12,** R50 (2011).
34. Kaufman, L. & Rousseeuw, P. J. *Finding Groups in Data: an Introduction to Cluster Analysis* Ch. 2 68–125 (Wiley, 1990).
35. Rousseeuw, P. J. Silhouettes — a graphical aid to the interpretation and validation of cluster-analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).
36. Tibshirani, R. & Walther, G. Cluster validation by prediction strength. *J. Comput. Graph. Statist.* **14,** 511–528 (2005).
37. Teal, T. K. & Schmidt, T. M. Identifying and removing artificial replicates from 454 pyrosequencing data. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5409 (2010).
38. R Development Core Team.. R: A Language and Envirnoment for Statistical Compuiting (R Foundation for Statistical Computing, 2010).
39. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2,** 18–22 (2002).