American Gut: an Open Platform for Citizen-Science Microbiome Research

Daniel McDonald[a,%], Embriette Hyde[a,%], Justine W. Debelius[a], James T. Morton[a], Antonio Gonzalez[a], Gail Ackermann[a], Alexander A. Aksenov[b,c], Bahar Behsaz[d], Caitriona Brennan[a], Yingfeng Chen[e], Lindsay DeRight Goldasich[a], Pieter C. Dorrestein[b,c], Robert R. Dunn[f], Ashkaan K. Fahimipour[g], James Gaffney[a], Jack A Gilbert[h,i,j,k], Grant Gogul[a], Jessica L. Green[g], Philip Hugenholtz[l], Greg Humphrey[a], Curtis Huttenhower[m,n], Matthew A. Jackson[o], Stefan Janssen[a], Dilip V. Jeste[p,q], Lingjing Jiang[a], Scott T. Kelley[14], Dan Knights[r,s], Tomasz Kosciolek[a], Joshua Ladau[t], Jeff Leach[u], Clarisse Marotz[a], Dmitry Meleshko[v], Alexey V. Melnik[b,c], Jessica L. Metcalf[w], Hosein Mohimani[x], Emmanuel Montassier[r,y], Jose Navas-Molina[a], Tanya T. Nguyen[p,q], Shyamal Peddada[z], Pavel Pevzner[b,d,aa], Katherine S. Pollard[t], Gholamali Rahnavard[m,n], Adam Robbins-Pianka[bb], Naseer Sangwan[j], Joshua Shorenstein[a], Larry Smarr[d,aa,cc], Se Jin Song[a], Timothy Spector[o], Austin D. Swafford[aa], Varykina G. Thackray[dd], Luke R. Thompson[ee], Anupriya Tripathi[a], Yoshiki Vazquez-Baeza[a], Alison Vrbanac[a], Paul Wischmeyer[ff,gg], Elaine Wolfe[a], Qiyun Zhu[a], The American Gut Consortium, Rob Knight[a,d,aa,#]

[a] Department of Pediatrics, University of California San Diego, La Jolla, CA, USA

[b] Collaborative Mass Spectrometry Innovation Center, University of California, San Diego, La Jolla, CA, USA

[c] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA

[d] Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA

[e] Department of Biology, San Diego State University, San Diego, CA, USA

[f] Department of Applied Ecology, North Carolina State University, Raleigh, NC, USA

[g] Biology and the Built Environment Center; University of Oregon, Eugene, OR, USA

[h] Department of Surgery, University of Chicago, Chicago, IL, USA

[i] Institute for Genomic and Systems Biology, University of Chicago, Chicago, IL, USA

[j] Department of Biosciences, Argonne National Laboratory, Chicago, IL, USA

[k] Marine Biology Laboratory, University of Chicago, Chicago, IL, USA

[l] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD, Australia

[m] Harvard T. H. Chan School of Public Health, Boston, MA, USA

[n] The Broad Institute of MIT and Harvard, Boston, MA, USA

[o] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK

[p] Departments of Psychiatry and Neurosciences, University of California San Diego, La Jolla, CA, USA

[q] Sam and Rose Stein Institute for Research on Aging, and Center for Healthy Aging, University of California San Diego, La Jolla, CA, USA

[r] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA

[s] Biotechnology Institute, University of Minnesota, Minneapolis, MN, USA

[t] The Gladstone Institutes, University of California, San Francisco, CA, USA

[u] Human Food Project, Terlingua, TX, USA

[v] St. Petersburg State University, Center for Algorithmic Biotechnology, Saint Petersburg, Russia

[w] Department of Animal Science, Colorado State University, Fort Collins, CO, USA

[x] Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA

[y] Université de Nantes, Microbiotas Hosts Antibiotics and Bacterial Resistances (MiHAR), Nantes, France

[z] Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA.

[aa] Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA

[bb] Department of Computer Science, University of Colorado Boulder, Boulder, Colorado, USA

[cc] California Institute for Telecommunications and Information Technology (Calit2), University of California San Diego, La Jolla, CA, USA

[dd] Department of Reproductive Medicine, University of California San Diego, La Jolla, CA, USA

[ee] Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, La Jolla, CA, USA

[ff] Department of Anesthesiology and Surgery, Duke University School of Medicine, Durham, NC, USA

[gg] Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC, USA

Running Title (max 54c): American Gut: an Open Platform for Microbiome Research

# Address correspondence to Rob Knight, robknight@ucsd.edu

% D.M. and E.H. contributed equally to this work.

Abstract word count: 203

Main text word count: 3280

Abstract: Although much work has linked the human microbiome to specific phenotypes and lifestyle variables, data from different projects have been challenging to integrate and the extent of microbial and molecular diversity in human stool remains unknown. Using standardized protocols from the Earth Microbiome Project and sample contributions from over 10,000 citizen-scientists, together with an open research network, we compare human microbiome specimens primarily from the USA, UK, and Australia to one another and to environmental samples. Our results show an unexpected range of beta-diversity in human stool microbiomes as compared to environmental samples, demonstrate the utility of procedures for removing the effects of overgrowth during room-temperature shipping for revealing phenotype correlations, uncover new molecules and kinds of molecular communities in the human stool metabolome, and examine emergent associations among the microbiome, metabolome, and the diversity of plants that are consumed (rather than relying on reductive categorical variables such as veganism, which have little or no explanatory power). We also demonstrate the utility of the living data resource and cross-cohort comparison to confirm existing associations between the microbiome and psychiatric illness, and to reveal the extent of microbiome change within one individual during surgery, providing a paradigm for open microbiome research and education.

Importance: We show that a citizen-science, self-selected cohort shipping samples through the mail at room temperature recaptures many known microbiome results from clinically collected cohorts and reveals new ones. Of particular interest is integrating n=1 study data with the population data, showing that the extent of microbiome change after events such as surgery can

exceed differences between distinct environmental biomes, and the effect of diverse plants in the diet which we confirm with untargeted metabolomics on hundreds of samples.

Introduction

The human microbiome plays a fundamental role in human health and disease. While many studies link microbiome composition to phenotypes, we lack understanding of the boundaries of bacterial diversity within the human population, and the relative importance of lifestyle, health conditions, and diet, to underpin precision medicine or to educate the broader community about this key aspect of human health.

We launched the American Gut Project (AGP; http://americangut.org) in November of 2012 as a collaboration between the Earth Microbiome Project (EMP) (1) and the Human Food Project (HFP; http://humanfoodproject.com/) to discover the kinds of microbes and microbiomes "in the wild” via a self-selected citizen-scientist cohort. The EMP is tasked with characterizing the global microbial taxonomic and functional diversity, and the HFP is focused on understanding microbial diversity across human populations. As of May 2017, the AGP included microbial sequence data from 15,096 samples from 11,336 human participants, totaling over 467 million (48,599 unique) 16S rRNA V4 gene fragments (“16S”). Our project informs citizen-scientist participants about their own microbiomes by providing a standard report (fig 1A) and resources to support human microbiome research, including an online course (Gut Check: Exploring Your Microbiome; *https://www.coursera.org/learn/microbiome*). AGP deposits all de-identified data into the public domain on an ongoing basis without access restrictions (table S1). This reference database characterizes the diversity of the industrialized human gut microbiome on an unprecedented scale, reveals novel relationships with health, lifestyle, and dietary factors, and establishes the AGP resource and infrastructure as a living platform for discovery.

Results

Cohort characteristics. AGP participants primarily reside in the United States (n=7,860). However, interest in the AGP rapidly expanded beyond the US to United Kingdom (n=2,518), and Australia (*n*=321), with 42 other countries or territories also represented (fig 1A; table S1). Participants in the US inhabit urban (*n*=7,317), rural (*n*=29), and mixed (*n*=98) communities (2010 US Census data based on participant zip codes), and span greater ranges of age, race, and ethnicity than other large-scale microbiome projects (2–6). Because the AGP is crowdsourced and self-selected, and subjects generally support the cost of sample processing, the population is unrepresentative in several important respects, including having lower prevalence of smoking and obesity, higher education and income (fig S1A), and underrepresentation of Hispanic and African American communities (table S1); generalization of the results is cautioned. Targeted and population-based studies will be crucial for filling these cohort gaps (Supplemental text).

Using a survey modified from (7, 8), participants reported general health status, disease history, and lifestyle data (table S2, supplemental text). In accordance with our IRB, all survey questions were optional (median per-question response 70.9%; table S2). Additionally, 14.8% of participants completed a validated picture-based food frequency questionnaire (FFQ) (VioScreen; *http://www.viocare.com/vioscreen.html*), and responses correlated well with primary survey diet responses (table S2).

We sought to minimize errors and misclassifications well-known to occur in self-reported data (9). Survey responses relied on controlled vocabularies. For analyses, we trimmed numeric entries at extremes (e.g., weight over 200kg or below 2.5kg) and excluded obviously incorrect answers (e.g., infants drinking alcohol) and samples for which necessary data were not supplied (e.g., missing zip code data for spatial analyses); see supplement for details. We focused our primary investigative efforts on a "healthy adult" subset (*n*=3,942) of individuals aged 20-69 with BMIs ranging between 18.5–30 kg/m^2, no self-reported history of inflammatory bowel disease, diabetes, or antibiotic use in the past year, and at least 1,250 16S sequences/sample (fig 1B, S1B).

The two largest populations in the dataset (US and UK) differed significantly in alpha-diversity, with Faith's phylogenetic diversity (PD) higher in UK samples (*13*) (Mann Whitney $p<1\text{x}10^{-15}$; fig 1C). One balance (10) (a log-ratio compositional transform) explained most of the taxonomic separation between US and UK samples (AUC=77.7% ANOVA p=1.01x10^{-78}, F=386.85) (fig S1C, table S3). To understand how these two populations differed from others, we compared adult AGP samples (predominantly from industrialized regions) to samples from adults living traditional lifestyles (6, 11, 12). As previously observed (6), samples from industrial and traditional populations separated in Principal Coordinates Analysis (PCoA) space of unweighted UniFrac distances (13) (fig S1D). They show greater variation within industrial populations than within traditional populations (2) and facile separation based on microbial taxonomy (industrial vs. non-industrial agrarian: AUC=98.9%, ANOVA p=1.52x10^{-260}, F=1265.8; industrial vs. hunter-gatherer: AUC=99.5%, ANOVA p=4.48x10^{-227}, F=1092.35) (fig 1D, table S3).

Removal of bacterial blooms. An important practical question is whether self-collected microbiome samples can match those from better-controlled studies. Most AGP samples are stools collected on dry swabs and shipped without preservative to minimize costs and avoid exposure to toxic preservatives. *E. coli* and a few other taxa grow in transit, so based on data from controlled storage studies as previously described (14) we removed sOTUs (sub-OTUs (15); median of 7.9% of sequences removed per sample) shown to bloom.

We further characterized the impact of these organisms through culturing, HPLC-MS analysis of cultured isolates, and shotgun metagenomics of the primary samples and storage controls (16). Culturing primary specimens stored at -80°C (US: *n*=116; UK: *n*=73; other: *n*=25) showed a strong correlation between the fraction of sequences reported as blooms in 16S sequencing and positive microbial growth following overnight incubation in aerobic conditions (fig 2A). Culture supernatants were characterized using HPLC-MS; most metabolites in these supernatants were absent from the primary specimens (fig 2B, C, method details in SI). We sequenced draft genomes of 169 isolates; of these, 65 contained the exact *E. coli* 16S sequence in the published bloom filter (14). To characterize the impact of the 16S bloom filter, we computed effect sizes over the participant covariates and technical parameters for 9,511 individual participant samples, including and excluding blooms (complete list table S2; comparisons to (17, 18) in supplementary text), and observed tight correlations for both unweighted (fig 2D, Pearson r=0.91, p=3.76x10^{-57}; Spearman r=0.90, p=9.45x10^{-55}) and weighted UniFrac (fig 2E, Pearson r=0.42, p=1.71x10^{-6}; Spearman r=0.58, p=1.03x10^{-9}). An outlier on the quantitative metric (weighted UniFrac) is present and corresponds to a variable representing the fraction of bloom reads in a sample.

Novel taxa and microbiome configurations. To understand human microbiome diversity, we placed AGP samples in the context of the EMP (1). Building on earlier work revealing a striking difference between host-associated and environmental microbiomes (19), we found that the diversity of microbiomes associated with the human gut (just one vertebrate) occupies a vast extent of the microbiome diversity of the planet (fig 3A).

Inserting the sOTU fragments of AGP and EMP samples into a Greengenes (20) reference phylogenetic tree using SEPP (21) (fig 3B) showed that the AGP population harbored much broader microbial diversity than the Human Microbiome Project (5). Both datasets are dwarfed by the breadth of bacterial and archaeal phylogenetic diversity in environmental samples. Examining sOTUs over increasing numbers of samples, we observed a reduction in the discovery rate of novel sOTUs starting around 3,000 samples, emphasizing the need for focused sampling efforts outside the present AGP population (fig 3C). The importance of sample size for detecting novel microbes and microbiomes is apparent when contrasted against Yatsunenko et al. (6), which contained hundreds of samples from three distinct human populations at ~1 million sequences/sample (fig 3D). This effect is magnified in beta-diversity analysis, where the AGP has saturated the configuration space, and new samples are not “distant” from existing samples (fig 3E). To encourage community engagement with sOTUs found in the AGP, we adapted the EMP “trading cards” for sOTUs (figs 3F, S2).

Temporal and spatial analyses. Longitudinal samples are required for understanding human microbiome dynamics (22). We examined 565 individuals who contributed multiple samples and observed an increasing trend of intrapersonal divergence with time. Still, over time individuals resemble themselves more than others, even after one year (fig 4A).

We tested whether patterns in individual longitudinal sample sets could be better explained when placed in the context of the AGP by integrating samples collected from: a) a time series of 58 time points from one subject (described as “LS”), prior to and following a large bowel resection, b) 2 time points from 121 patients in an intensive care unit (ICU) (23), c) samples from the "extreme" diet study from David et al. (24), and d) samples from the Hadza hunter-gatherers for additional context (25). Through the longitudinal sampling of LS, dramatic pre- and post-microbial configuration changes that exceeded the span of microbial diversity associated with the AGP population were observed (fig 4E, animated in (26)). After surgery, subject's samples more closely resembled those of ICU patients (Kruskal Wallis H=79.774, $p=4.197x^{-19}$, fig S2A-C), and showed a persistent state change upon return to the AGP fecal space. Remarkably, the UniFrac distance between the samples immediately prior to and following the surgery was almost identical to the distance between a marine sediment sample and a plant rhizosphere sample (unweighted UniFrac distance of 0.78). Furthermore, the observed state change in LS is not systematically observed in the extreme diet study (fig S2D; PERMANOVA n.s. when controlling for individual). Despite extensive dietary shifts, these subjects do not deviate from the background AGP context.

Recent reports suggest that the microbes of bodies (8), like those of homes (27), are influenced mostly by local phenomena rather than regional biogeography (28), and accordingly we observed only weak geographic associations with sOTUs (fig 4B), no significant distance-decay relationships (fig 4C), and, with Bray-Curtis distance, only a weak effect at neighborhood sizes of ca. 100km (Mantel r=0.036, Benjamini-Hochberg adjusted p=0.03) to 1,000km (Mantel r=0.016, Benjamini-Hochberg adjusted p=0.03).

Dietary plant diversity. The self-reported dietary data suggested, unexpectedly, that the number of unique plant species a subject consumes is associated microbial diversity, rather than self-reported categories such as "vegan" or "omnivore" (fig 2D, E). Principal Components Analysis of FFQ responses (fig 5A) revealed clusters associated with diet types such as "vegan."

However, these dietary clusters did not significantly relate to microbiome configurations (fig 5B; Procrustes fig 5A, M^2=0.988). We therefore characterized the impact of dietary plant diversity on the microbial community.

Using balances (10), we identified several putative short-chain fatty acid (SCFA) fermenters associated with eating more than 30 types of plants, including sOTUs putatively of the species *Faecalibacterium prausnitzii* and of the genus *Oscillospira* (29) (AUC=68.5%, ANOVA p=8.9x10^{-39}, F=177.2) (fig 5E, table S3). These data suggest community-level changes associated with microbial fermentation of undigested plant components. Because bacteria differ in their carbohydrate-binding modules and enzymes that hydrolyze diverse substrates in the gut (30), a diet containing various types of dietary fibers and resistant starches likely supports a more diverse microbial community (31, 32).

To test these effects in the stool metabolome, we performed HPLC-MS annotation and annotation propagation (33, 34) on a subset of fecal samples (n=219) preferentially selecting individuals at the extremes of plant type consumption, i.e. eating <10 or >30 different types of plants per week. Several fecal metabolites differed between the two groups, with one key discriminating feature annotated as octadecadienoic acid (annotation level 2 according to the 2007 metabolomics initiative, (35)). Further investigation using authentic standards revealed that the detected feature was comprised of multiple isomers, including linoleic acid (LA) and conjugated linoleic acid (CLA). CLA abundance was significantly higher in individuals consuming > 30 types of plants, and those consuming more fruits and vegetables generally, (fig 5D, 1-sided t-test; p < 10^{-5}), but did not correlate with dietary CLA consumption as determined by the FFQ (dietary fig 5C; Spearman r < 0.16; p > 0.15). CLA is a known end-product of LA conversion by lactic acid bacteria in the gut, such as *Lactobacillus plantarum* (36) and

Bifidobacterium spp. (37). FFQ-based dietary levels of LA and MS-detected LA did not differ significantly between groups (fig S3), suggesting that their different microbiomes may differentially convert LA to CLA. Several other putative octadecadienoic acid isomers were also detected (fig 5F), some strongly correlated with plant consumption. Determining these compounds' identities as well as their origin and function may uncover new links between the diet, microbiome, and health.

Molecular novelty in the human gut metabolome. Our untargeted HPLC-MS approach allowed us to search for novel molecules in the human stool metabolome, parallel to our search for novelty in microbes and microbiome configurations described above. Bacterial N-acyl amides were recently shown to regulate host metabolism by interacting with G-protein-coupled receptors (GPCRs) in the murine gastrointestinal tract, mimicking host-derived signaling molecules (38). These agonistic molecules regulate metabolic hormones and glucose homeostasis as efficiently as host ligands. Manipulating microbial genes that encode metabolites eliciting host cellular responses could enable new drugs or treatment strategies for many major diseases, including diabetes, obesity, and Alzheimer's disease: roughly 34% of all marketed drugs target GPCRs (39). We observed N-acyl amide molecules previously hypothesized but unproven to be present in the gut (38) (fig 6, S4), as well as new N-acyl amides (fig 6).

Levels of two N-acyl amides, annotated as commendamide (*m/z* 330.2635, fig S4B) and N-3-OH-palmitoyl ornithine (*m/z* 387.3220, fig S4C), positively correlated with a self-reported medical diagnosis of thyroid disease (Kruskal–Wallis, FDR p=0.032, p=2.48x10^{-3}, χ2=11.99; N-3-OH-palmitoyl ornithine; Kruskal–Wallis, FDR p=0.048, p=5.63x10^{-3}, χ2=10.35). Conversely, glycodeoxycholic acid (m/z 450.3187) was significantly higher in individuals not reporting thyroid disease diagnosis (Kruskal–Wallis; FDR p=1.28x10^{-4}, p= 4.41x10^{-7}, χ2=29.27). This cholic acid is produced through microbial dehydroxylation, again linking gut microbiota to endocrine function (40, 41).

Finally, we compared metabolome diversity to 16S diversity in the samples selected for dietary plant diversity and a second set of samples selected to explore antibiotic effects (*n*=256 individuals who self-reported not having taken antibiotics in the past year (*n*=117), or having taken antibiotics in the past month (*n*=139); participants were matched for age, BMI, and country). By computing a collector's curve of observed molecular features in both cohorts (fig 6K, 6M), we observe that, paradoxically, individuals who had taken antibiotics in the past month (*n*=139) had significantly greater molecular diversity (Kruskal Wallis, *H*=255.240, *p*=1.87x10^{-57}) than those who had not taken antibiotics in the past year (*n*=117), and differed in molecular beta-diversity (fig 6K inset), suggesting that antibiotics promote unique metabolomes that result from differing chemical and microbial environments in the gut. Notably, the diversity relationships of this set are not reflected in 16S diversity (fig 6L, 6N), where antibiotic use shows decreased diversity (Kruskal Wallis *H*=3983.839, *p*=0.0). Within the dietary plant diversity cohort, we observed a significant increase (Kruskal Wallis, *H*=897.106, *p*=4.17x10^{-197}) in molecular alpha diversity associated with a high diversity of plant consumption (*n*=42) compared to low plant diversity (*n*=43), a relationship also observed in 16S diversity, where high dietary plant diversity increased 16S alpha diversity (Kruskal Wallis, *H*=65.817, *p*=4.947x^{-16}).

A living dataset. The AGP is dynamic, with samples arriving from around the world daily. This allows a living analysis, similar to continuous molecular identification and annotation revision in the Global Natural Products Molecular Networking (GNPS) database (34). Although the analysis presented here represents a single snapshot, samples continued to arrive during manuscript preparation. For example, after we defined the core "healthy" sample set, an exploratory analysis using matched controls was performed by collaborators to test for correlations between mental illness and microbiome composition (as reported in (42, 43)). By analyzing mental illness status (depression, schizophrenia, post-traumatic stress disorder (PTSD) and bipolar disorder – four of the most disabling illnesses per World Health Organization (44)) reported by AGP participants (*n*=125) against matched 1:1 healthy controls (*n*=125), we observed a significant partitioning using PERMANOVA in weighted UniFrac (*p*=0.05, *pseudo-F*=2.36). These findings were reproducible within US residents (*n*=122, p=0.05, *pseudo-F*=2.58), UK residents (*n*=112, *p*=0.05, *pseudo-F*=2.16), women (*n*=152, *p*=0.04, *pseudo-F*=2.35), and people 45 years of age or younger (*n*=122, *p*=0.05, *pseudo-F*=2.45). We also reproduce some previously reported differentially abundant taxa in Chinese populations using our UK subset (42, 45)(table S3). This shows that multi-cohort replication is possible within the AGP (additional detail supplemental text).

Discussion

The AGP provides an example of a successful crowdfunded citizen science project that facilitates human microbiome hypothesis generation and testing on an unprecedented scale, provides a free data resource derived from over 10,000 human-associated microbial samples, and both recaptures known microbiome results and yields new ones. Ongoing living data efforts, such as the AGP, will allow researchers to document and potentially mitigate the effects of a slow but steady global homogenization driven by increased travel, lifespans, and access to similar diets and therapies, including antibiotics. Because the AGP is a subproject of the EMP (1), all samples were processed using the publicly available and widely used EMP protocols to facilitate meta-analyses, as highlighted above. Further example applications include assessing the stability of AGP runs over time, comparing the AGP population to fecal samples collected from a fecal transplant study (46) and an infant microbiome time series (47), the latter using different DNA sequencing technology, to highlight how this context can provide insight (48).

A unique aspect of the AGP is the open community process of assembling the Research Network and analyzing these data, which are released immediately on data generation. Analysis details are shared through a public forum (GitHub, https://github.com/knightlab-analyses/american-gut-analyses). Scientific contributions to the project were made through a geographically diverse Research Network represented herein as the American Gut Consortium, established prior to project launch and which has grown over time. This model allows a “living analysis” approach, embracing new researchers and analytical tools on an ongoing basis (e.g., Qiita (*Web:http://qiita.microbio.me*) and GNPS (34)). Examples of users of the AGP as a research platform include educators at several universities, UC San Diego Athletics, and the American Gastroenterological Association (AGA). Details on projects using the AGP infrastructure can be found in the supplement.

To promote public data engagement, we aimed to broaden the citizen science experience obtained by participating in AGP by “gamifying” the data and separately by developing an online forum for microbiome data discussion and discovery. The gamification introduces concepts of beta-diversity and challenges users to identify clusters of data in principal coordinates space (http://csb.cs.mcgill.ca/colonyb/). The forum, called Gut Instinct (http://gutinstinct.ucsd.edu), enables participants to share lifestyle-based insights with one another. Participants also have the option to share their AGP sample barcodes, which will help us uncover novel contextual knowledge. Gut Instinct now has over 1,050 participants who have collectively created over 250 questions. Participants will soon design and run their own investigations using controlled experiments to further understand their own lifestyle and the AGP data.

The AGP therefore represents a unique citizen-science dataset and resource, providing a rich characterization of microbiome and metabolome diversity at the population level. We believe the community process for involving participants from sample collection through data analysis and deposition will be adopted by many projects harnessing the power of citizen science to understand the world around and within our own bodies.

Materials and methods

Participant Recruitment and Sample Processing. Participants signed up for the project through Indiegogo (https://www.indiegogo.com/) and later, FundRazr (http://fundrazr.com/). A contribution to the project was made to help offset the cost of sample processing and sequencing (typically $99 per sample; no requirement to contribute if another party was covering the contribution). All participants were consented under an approved Institutional Review Board human research subjects protocol, either from the University of Colorado Boulder (protocol #12-0582; December 2012 - March 2015) or the University of California, San Diego (protocol #141853; February 2015 - present). The IRB-approved protocol specifically allows for public deposition of all data that is not personally identifying and for return of results to participants (fig. 1A).

Self-reported metadata were collected through a web portal (http://www.microbio.me/americangut). Samples were collected using BBL Culture Swabs (Becton, Dickinson and Company; Sparks, MD) and returned by mail. Samples were processed using the EMP protocols. Briefly, the V4 region of the 16S rRNA gene was amplified with barcoded primers and sequenced as previously described (49). Sequencing prior to August 2014 was done using the 515f/806r primer pair with the barcode on the reverse primer (50); subsequent rounds were sequenced with the updated 515f/806rB primer pair with the barcode on the forward read (51). Sequencing batches 1-19 and 23-49 were sequenced using an Illumina MiSeq; sequencing for 20 and 21 were performed with an Illumina HiSeq Rapid Run and round 22 was sequenced with an Illumina HiSeq High Output.

16S Data Processing. The 16S sequence data were processed using a sequence variant method, Deblur v1.0.2 (52) trimming to 125nt (otherwise default parameters), to maximize the specificity of 16S data; a trim of 125nt was used because one sequencing round in the American Gut used 125 cycles while the rest used 150. Following processing by Deblur, previously recognized bloom sequences were removed (14). The Deblur sub Operational Taxonomic Units (sOTUs) were inserted into the Greengenes 13_8 (53) 99% reference tree using SEPP (54). Taxonomy was assigned using an implementation of the RDP classifier (55) as implemented in QIIME2 (56). Multiple rarefactions were computed, with the minimum being 1250 sequences per sample with the analyses using the 1250 set except where noted explicitly. Diversity calculations were computed using scikit-bio 0.5.1 with the exception of UniFrac (57) which was computed using an unpublished algorithmic variant, Striped UniFrac (*https://github.com/biocore/unifrac)*, which scales to larger datasets and produces identical results to previously published UniFrac algorithms.

Metadata Curation. To address the self-reported nature of the AGP data and ongoing nature of the project, basic filtering was performed on the age, height, weight, and body mass index (BMI). Height and weight were gated to only consider heights between 48 cm and 210 cm, and weight between 2.5 kg and 200 kg. BMI calculations using values outside this range were not considered. We assumed age was misreported by any individual who reported a birth date after their sample was collected. We also assumed age was misreported for participants who reported an age of less than 4 years, but height over 105 cm, weight over 20 kg, or any alcohol consumption. Values assumed to be incorrect were dropped from analyses (fig S1B).

Sample Selection. Analyses in the manuscript were performed on a subset of the total AGP samples. A single fecal sample was selected for each participant with at least one fecal sample that amplified to 1250 sequences per sample unless otherwise noted. Priority was given to samples that were associated with VioScreen (*http://www.viocare.com/vioscreen.html)* metadata.

The samples used for analysis and subsets used in various analyses are described in table S2. Briefly, we defined the healthy subset (*n*=3,942) as adults aged 20-69 years with a BMI between 18.5 and 30 kg/m^2 who reported no history of inflammatory bowel disease or diabetes and no antibiotic use in the last year. There were 1,762 participants who provided results for the VioScreen Food Frequency Questionnaire (FFQ; *http://www.viocare.com/vioscreen.html*). The meta-analysis with non-Western samples (*n*=4,643) included children over the age of 3, adults with a BMI of between 18.5 and 30 kg/m^2, and no reported history of inflammatory bowel disease, diabetes, or antibiotic use in the last year.

Population Level Comparisons. Population level comparisons were calculated for all American Gut participants living in the United States. BMI categorization was only considered for adults over the age of twenty, since the description of BMI in children is based on their age and sex. Education level was considered for adults over the age of 25. This threshold was used to match the available data from the US Census Bureau (*https://www.census.gov/content/dam/Census/library/publications/2016/demo/p20- 578.pdf*). The percentage of the American Gut participants was calculated as the fraction of individuals who reported results for that variable. US population data is from the 2010 census (*https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf*), US Census bureau reports (*https://www.census.gov/content/dam/Census/library/publications/2016/demo/p20- 578.pdf*), Centers for Disease Control reports on obesity (*https://www.cdc.gov/nchs/data/hus/2015/058.pdf*), diabetes (57, 58), IBD (*http://www.cdc.gov/ibd/ibd-epidemiology.htm*), smoking (*https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm*), and a report from the Williams Institute (*http://williamsinstitute.law.ucla.edu/wp-content/uploads/How-Many-Adults-Identify-as-Transgender-in-the-United-States.pdf*) (table S2).

Within American Gut Alpha- and Beta-Diversity Analyses. OTU tables generated in the primary processing step were rarefied to 1,250 sequences per sample. Shannon, Observed OTU, and PD whole tree diversity metrics were calculated as the mean of ten rarefactions using QIIME (56, 59). Alpha-diversity for single metadata categories was compared with a Kruskal-Wallis test. Unweighted UniFrac distance between samples was tested with PERMANOVA (60) and permuted *t*-tests in QIIME.

Balances. The goal of this analysis was to design two-way classifiers to classify samples and sOTUs. This will allow us to identify sOTUs that are strongly associated with a given environment. To do this while accounting for issues due to compositionality, we used balances (61) constructed from Partial Least Squares (62).

First the sOTU table was centered log-ratio (CLR) transformed with a pseudocount of 1. Partial least squares discriminant analysis (PLS-DA) was then performed on this sOTU table using a single PLS component, using a binary categorical variable as the response and the CLR transformed sOTU table as the predictor. This PLS component represented an axis, which assigns scores to each OTU according to how strongly associated they are to each class. An sOTU with a strong negative score indicates an association for the one category, which we will denote as the negative category. An sOTU with a strong positive score indicates that sOTU is strongly associated with the other category, which we will denote as the positive category.

We assumed that PLS scores associated with each OTU were normally distributed. Specifically

$$score(x_{pos}^{(i)}) \sim N(\mu_{pos}, \sigma_{pos}^2)$$
$$score(x_{neg}^{(i)}) \sim N(\mu_{neg}, \sigma_{neg}^2)$$
$$score(x_{null}^{(i)}) \sim N(\mu_{null}, \sigma_{null}^2)$$

Where $\mu_{null} \approx 0$, $\mu_{neg} < 0$ and $\mu_{pos} > 0$. To obtain estimates of these normal distributions, Gaussian Mixture Models with three Gaussians were fitted from the PLS scores. Thresholds were determined from the intersection of Gaussians. The OTUs with PLS scores less than the intersection $N(\mu_{null}, \sigma^2_{null})$ and $N(\mu_{neg}, \sigma^2_{neg})$ are classified to be associated with the negative category. The OTUs with PLS scores greater than the intersection $N(\mu_{null}, \sigma^2_{null})$ and $N(\mu_{pos}, \sigma^2_{pos})$ are classified to be associated with the positive category.

The balance was constructed as follows

$$b = \sqrt{\frac{|x_{pos}||x_{neg}|}{|x_{pos}| + |x_{neg}|}} log\,(\frac{g(x_{pos})}{g(x_{neg})})$$

From this balance, we calculated receiver operator characteristic (ROC) curves and AUC to assess the classification accuracy, and ran ANOVA to assess the statistical significance. The dimensionality was shrunk through some initial filtering (an sOTU must have at least 50 reads, must exist in at least 20 samples except where noted, and have a variance over 10 to remove sOTUs that do not appear to change), so that the number of samples is greater than the number of sOTUs to reduce the likelihood of over-fitting. This technique was used to investigate differences due to plant consumption, country of residence and western vs non-western and was consistently applied with the exception that a filter of 5 samples was used for the western vs. non-western analysis due to group sample sizes.

Balances on plant consumption were constructed using Partial Least Squares. Only samples from people who consumed less than 10 types of plants a week or more than 30 types of plants a week were considered.

Meta-analysis of samples from the American Gut and from individuals living agrarian and hunter-gatherer lifestyles. A meta-analysis compared fecal samples collected from healthy individuals that were 3 years of age or older and included in the AGP data set to a previously published 16S rRNA V4 region data set that included healthy people living an industrialized, remote agrarian or hunter-gatherer lifestyle (63–65). The AGP subset of healthy individuals was determined by filtering by the metadata columns "subset_antibiotic", "subset_ibd", "subset_diabetes", and for individuals over the age of 16 years "subset_bmi". All datasets were processed using the Deblur pipeline as noted above, with the exception that all reads in the meta-analysis, including AGP data, were trimmed to 100nt to accommodate the read length in Yatsunenko et al (63). Bloom reads as described above were removed from all samples. We used Striped UniFrac as noted above to estimate beta-diversity (unweighted UniFrac) and EMPeror software (66) version 0.9 to visualize principal coordinates. We used a non-parametric PERMANOVA with 999 permutations to test for significant differences in fecal microbiomes associated with industrialized, remote agrarian, and hunter-gatherer lifestyles. All AGP samples were considered to be from people living an industrialized lifestyle. Balances were constructed from Partial Least Squares to assess the differences between the hunter-gather vs. industrialized populations and the remote farmers vs industrialized populations.

Spatial Autocorrelation. We sought to investigate distance-decay patterns – the relationship between microbial community similarity and spatial proximity – among American Gut participants, to determine the extent to which geographical distances could explain variation in microbial community taxonomic compositions between participant pairs. The correlation between community-level Bray-Curtis (67) distances and participants' spatial proximities (i.e., great-circle distances, km) was assessed using a Mantel test (68) with 1000 matrix permutations. Analyses were conducted using the subset of participants located in the continental United States that had not received antibiotics in the last year. Different neighborhood sizes were investigated in order to detect the relevant spatial scale on which significant distance-decay patterns in microbial community compositions emerged. To accomplish this, we computed distance-decay relationships for a series of model adjacencies corresponding to neighborhood radiuses of 50, 100, 500, 1000, 2500, and 4500 km among participants, and adjusted *p*-values for multiple comparisons using the Benjamini-Hochberg procedure (69). We also studied spatial correlations in phylogenetic community dissimilarities, calculated as weighted normalized UniFrac distances, using the procedure described above. Analyses were conducted in R statistical programming environment.

The spatial autocorrelation of each individual taxon was assessed using Moran's *I* statistic (70). Taxa present in less than 10 samples were filtered, since these would not be sufficiently powered. Analyses were conducted using binary spatial weight matrices, with neighborhoods of 0 – 50 km, 50 – 100 km, and 100 – 250 km. The different neighborhoods were useful for detecting spatial autocorrelation at different scales. All spatial weights matrices were row-standardized. We checked for spatial autocorrelation at three taxonomic ranks: class, genus, and OTU. We also considered whether there was autocorrelation within subsets of individuals who were under 20 years old and between 20 and 70 years old; those having IBD, no IBD, diabetes, and no diabetes; and those who had taken antibiotics within the past week, year, or not within the past year. The results presented above did not qualitatively depend on the subset of individuals considered. Statistical significance was assessed using permutation tests, which were implemented using a Markov Chain Monte Carlo algorithm. To assess each *p*-value, 100 chains were run each starting from a different random permutation. Each chain had 1000 iterations. We used Bonferroni corrections to correct for multiple comparisons, with an overall significance level set to 0.05. Analyses were run using custom Java code, optimized for running many spatial autocorrelation analyses on large data sets (71).

Metadata cross-correlation. To account for covariance among metadata for effect size and variation analyses, we examined the correlation between individual metadata variables including technical parameters. Groups in ordinal variables were combined if there were insufficient sample size (e.g. people who reported sleeping less than 5 hours were combined with those who reported sleeping 5 to 6 hours into a variable described as “Less than 6”). The same transformations were used for effect size analysis. Any group with less than 25 total observations was ignored during analysis; if this resulted in a metadata column having no groups, the column was removed from analysis. The relationship between continuous and ordinal covariates was calculated using Pearson’s correlation. Ordinal and categorical covariates were compared using a modified Cramer’s V statistic (72). Continuous and categorical covariates were compared with a Welch’s T test (73). We treated used *1-R* as a distance between the covariates. Traversing the resulting binary, weighted cluster tree starting at tip level into the direction of the root, i.e. bottom-up, we grouped tips together that are members of the same subtree after covering a distance of approximately 0.5 (branch length 0.29). A representative variable from each cluster was selected for analysis (table S2).

Effect Size Calculations. Effect size was calculated on 179 covariates (including technical parameters), selected from the cross-correlation (table S2). Ordinal groups with small sample sizes at the extreme were collapsed as noted above. Individuals who reported self-diagnosis or diagnosis from alternative practitioner for medical conditions were excluded from the analysis. Any metadata variable with less than 50 observations per group or that made up less than 3% of the total number of respondents was also excluded from the effect size analysis. Continuous covariates were categorized into quartiles. For each one of the 179 variables, we applied the mdFDR (74) methodology to test for the significance of each pairwise comparison among the groups. For each significant pairwise comparison, we computed the effect size using Cohen's *d* (75), or the absolute difference between the mean of each group divided by the pooled standard deviation. For analysis of diversity, we used Faith's Phylogenetic Diversity (alpha-diversity) and weighted and unweighted UniFrac distances (beta-diversity).

Variation analysis. Using the methodology reported in the supplemental material for (76), we computed Adonis (77) using 1000 permutations, over the sample sets used in the effect size calculations as noted above, and applied Benjamini-Hochberg correction (FDR<0.1) to assess drivers of variation in beta-diversity.

Meta-analysis movie. American Gut samples from all body sites were combined with data from an infant time series (78), a fecal transplant study (79), and recent work characterizing the microbiome of patients in the intensive care unit (80). The combination of the datasets in movie S2 required that all sequences were trimmed to an even length of 125 nucleotides. All projects except for the infant time series were sequenced using an Illumina instrument. In order to combine the data, we expressed the Illumina and non-Illumina data through a common reference database. Specifically, the Deblur sOTUs from the Illumina data were mapped against the Greengenes (53) database (13_8 release) using 99% similarity; the associations between the input sOTUs, and their cluster memberships, were used to construct an OTU table based on the original sOTU per sample sequences counts (i.e., summing the counts for all sOTUs in a common OTU). The infant time series data were picked using a closed reference OTU picking approach against the same reference at the same similarity. The infant time series dataset followed a closed reference OTU picking approach using 99% similarity. The resulting two tables (from Illumina-generated data and the ITS dataset) were merged and analyzed using the Greengenes 99% tree. The table was rarefied to 1,250 sequences per sample. Principal coordinates projections were calculated based on unweighted UniFrac distance (57). The principal coordinates analysis was visualized and animated in EMPeror 1.0.0-beta8-dev (66, 81). The movie was captured in QuickTime (Apple, Cupertino, CA), and edited with Premiere Pro (Adobe, San Jose, CA).

Integration with the Earth Microbiome Project. A precomputed 100nt Deblur BIOM table representing the data in (82) was obtained from (ftp://ftp.microbio.me/emp/release1/otu_tables/deblur/). 100nt Deblur tables were also obtained from Qiita for Hadza fecal samples (Qiita study ID 11358, (83)), ICU microbiome samples (Qiita study ID 2136, (80)), and a longitudinal series which includes samples immediately prior to and following a large bowel resection (Qiita study ID 10283, EBI accession ERP105968, unpublished); all samples were processed using the EMP Illumina 16S V4 protocol. The EMP dataset used a minimum sOTU count of 25; the same threshold was applied to the other datasets included prior to merge. Blooms as identified by (84) were removed from all samples. This collection of BIOM tables was then merged yielding an OTU table representing 40,600 samples. sOTUs were restricted to those already present in the EMP 100nt fragment insertion tree, which represents 329,712 sOTUs. The table was then rarefied to 1000 sequences per sample, and unweighted UniFrac computed using 768 processors with the aforementioned Striped algorithm. Visualizations and animations were performed using EMPeror v1.0.0b12.dev0.

Extreme diet study state assessment. The sequence data from (85) were processed by Deblur to assess 16S sOTUs in common with the AGP processing above. In order to assess a state difference with PERMANOVA, we needed to control for sample independence within the longitudinal sampling. To do so, we randomly selected one sample from each individual per diet, computed PERMANOVA, and repeated the process 100 times. None of the trials produced a *p*-value below 0.05.

Vioscreen PCA and diet type Procrustes analysis. Before performing Principal Component Analysis (PCA) on the informal diet questions, Vioscreen variables that are categorical or receive less than 90% response among the 1762 participants were excluded leaving 1596 participants. PCA was then performed using the Vioscreen information from these participants' responses over 207 Vioscreen questions, and then colored by their types of diet as answered in the AGP informal food survey. The coordinates from the PCA were extracted. For the same samples, PCoA of unweighted UniFrac distances was computed on the 16S data subset from the primary processing set. The coordinates from the PCA and the PCoA were assayed for a measure of fitness using Procrustes as implemented in QIIME v1.9.1.

Beta-diversity added. To assess added beta-diversity, we applied the technique used in (86) figure 3. Specifically, we randomly sampled *N* samples from the distance matrix 10 times, over an increasing value of *N*. For each set of sampled distances, we computed the minimum observed distance.

sOTU novelty. To assess sOTU novelty, we randomly sampled *N* samples from an sOTU table 10 times, over an increasing value of *N*. At each sampling, we computed the number of sOTUs observed with read counts within minimum thresholds. In other words, a minimum threshold of 1 is the number of singletons observed in the sampled set, a minimum threshold of 2 is the number of singletons and doubletons, etc.

Within-individual beta-diversity. Many of the individuals in the American Gut Project contributed multiple samples, but at uneven time intervals. In order to explore intrapersonal variation, we replicated the analysis in Lloyd-Price et al. figure 3 (87). Specifically, we determined all time deltas between a subjects samples, and gathered the distributions of beta-diversity between any two samples binned by month. An individual is only represented a single time in a given month, but may be represented in multiple months if they had, for instance, contributed samples over the course of a year.

High Performance Liquid Chromatography Mass Spectrometry (HPLC-MS) Analysis. A total of 498 samples were selected for analysis via mass spectrometry. Specifically, two groups were chosen. First, given the large body of primary literature describing the negative impact of antibiotics on the gut microbiome, and the general interest in this topic from many American Gut participants, we chose 279 samples from individuals (age, BMI, and country matched) who self-reported not having taken antibiotics in the past year, or having taken antibiotics in the past month or week. We chose a second group of 219 samples collected from individuals who answered the question "In an average week, how many different plants do you eat? (e.g., if you consume a can of soup that contains carrots, potatoes and onion, you can count this as 3 different plants; If you consume multi-grain bread, each different grain counts as a plant. Include all fruits in the total)" on the main American Gut Project main survey and who had also completed the VioScreen Food Frequency Questionnaire. When American Gut participants collect samples, they do so on a double headed swab; therefore, all samples chosen for this analysis had one remaining swab head (the first had been used for DNA extraction and microbiome sequencing).

Cell cultures sample preparation for metabolomics analysis. The supernatant collected from cell cultures (see "expanded bloom assessment" below) were processed to make them compatible with HPLC-MS analysis. The solid phase extraction with wash was carried out to reduce impact of cell culture media, which is highly detrimental for the ESI. The 30 mg sorbent Oasis HLB (Waters, Waltham, MA) SPE cartridges were used to achieve broad metabolite coverage. The cell samples were stored at -80°C and thawed at room temperature immediately prior to extraction. The thawed samples were then centrifuged for 10 minutes at 1200 rpm and extracted. For the SPE extraction, the Oasis HLB SPE cartridge was conditioned with 700µL of 100% HPLC-grade methanol and equilibrated with 700µL of HPLC-grade DI water. The cell supernatant (~350-400µL) was loaded into cartridge and allowed to slowly elute. The loaded SPE wells were then washed with 800µL of 5% methanol in water and the absorbed material was slowly eluted with 200 µL of 100% methanol. Vacuum up to ~ 20 psi was applied for the wells that did not elute within an hour. The collected eluent was stored at -20°C until the HPLC-MS analysis.

Fecal sample preparation for metabolomics analysis. The swab tubes scheduled for analysis were removed from the -80°C freezer and placed on dry ice for the duration of sample processing. Each tube with swab was logged by reading the barcode with barcode scanned and the swab was removed from tube and placed onto a ThermoFisher Scientific (ThermoFisher Scientific, Waltham, MA) 2 ml deep well 96-well plate set on top of dry ice coolant. The top part of each swab's stick was snapped off and discarded. Immediately after filling all of the wells with swabs, 200 μL of HPLC-grade 90% v:v ethanol:water solvent was added to each well using multichannel pipette. Four blanks of unused swabs and extraction solvent were included onto each plate. Each plate was then sealed with 96-well plate lid, sonicated for 10 minutes and placed into the refrigerator at 2 °C to extract samples overnight. After extraction, the swabs were removed from wells and discarded, the plates were placed into a lyophilizer, and the entire sample was dried down and then re-suspended in 200 μL 90% v:v ethanol:water. The plates were resealed and centrifuged at 2000 rpm for 10 minutes. The 100 μL aliquots of sample were then transferred onto a Falcon 96-well MS plate using a multichannel pipette, and each plate was immediately sealed with sealing film. The MS plates were centrifuged at 2000 rpm for 10 minutes and stored at 2 °C until analysis.

HPLC-MS analysis. The metabolomics analysis of samples was conducted using reverse phase (RP) high performance liquid chromatography mass spectrometry (HPLC-MS). The HPLC-MS analysis was performed on a Dionex UltiMate 3000 ThermoFisher Scientific high-performance liquid chromatography system (ThermoFisher Scientific, Waltham, MA) coupled to a Bruker impact HD qTOF mass spectrometer. The chromatographic separation was carried out on a

Kinetex C18 1.7 µm, 100Å UHPLC column (50 mm x 2.1 mm) (Phenomenex, Torrance, CA), held at 40 ºC during analysis. A total of 5 µL of each sample was injected. Mobile phase A was water, mobile phase B was acetonitrile, both with added 0.1% v:v formic acid. The solvent gradient table was set as follows: initial mobile phase composition was 5% B for 1 min, increased to 40% B over 1 min, then to 100% B over 6 min, held at 100% B for 1 min, decreased back to 5% B in 0.1 min, followed by a washout cycle and equilibration for a total analysis time of 13 min. The scanned m/z range was 80-2000 Th, the capillary voltage was 4500 V, the nebulizer gas pressure was 2 bar, the drying gas flow rate was 9 L/min, and the temperature was 200 °C. Each full MS scan was followed by MS/MS using collision-induced dissociation (CID) fragmentation of the seven most abundant ions in the spectrum. For MS/MS, the collision cell collision energy was set at 3 eV and the collision energy was stepped 50%, 75%, 150% and 200% to obtain optimal fragmentation for differentially sized ions of different sizes. The scan rate was 3 Hz. A HP-921 lock mass compound was infused during the analysis to carry out post-processing mass correction. All of the raw data are publicly available at the UCSD Center for Computational Mass Spectrometry (*111*) (dataset ID: MassIVE MSV000080179).

MS data analysis. The collected HPLC-MS raw data files were first converted from Bruker's *d* to mzXML format and then processed with the open source OpenMS 2.0 software (88) in order to deconvolve and align each peak across different chromatograms (feature detection). The alignment window was set at 0.5 minutes, the noise threshold at 1000 counts, the chromatographic peak FWHM value at 20, and the mass error at 30 ppm. All of the peaks that were present in any of the blanks with S/N below 10:1 were removed from the final feature table. The number of features with corresponding MS/MS was as follows: Vioscreen study sample cohort: 5144 total MS2 features; antibiotics study samples cohort: 8288 total MS2 features. The number of MS1 features is difficult to estimate exactly as it depends on feature detection settings and the number of samples, but it is typically about 4-5 fold greater depending on the sample. For all of the MS1 features detected across all samples, only ~1-5% are present in an individual sample.

Chemical annotations were carried out by automatic matching fragmentation spectra to multiple databases using Global Natural Product Social Molecular Networking (GNPS) (89) and then examining the data at the MS/MS level by molecular networking (90). The goal is to retrieve spectra with identical and similar fragmentation patterns and combine them into consensus nodes and clusters, respectively. The consensus node spectra are then compared against public MS/MS libraries to provide molecular annotations (91). Further annotations could be suggested by examining the molecular network (90) (so called propagated annotations). Annotations obtained with precursor and MS/MS matching are considered level two annotations according to the 2007 metabolomics standards initiative (92). All molecular networking analysis and annotations are available here: antibiotic use subset (93); types of plants subset (94), cell cultures of isolates (95) and fecal samples co-networked with the cell cultures (96). The raw data contain a significant number of abundant features originating from swab polymers. Therefore, selective background peak removal was carried out specifically for the polymer compounds originating from swabs that were used for the sample collection. The m/z shifts that correspond to the polymer repeating units (44.0262, 88.0524, 132.0786, 176.1049) were identified with GNPS m/z differences frequency plot. The network clusters that contained nodes with the corresponding mass differences were deemed to belong to polymers and all member nodes of the network clusters were removed from the feature table (a total of 1632 features/nodes). Principal Coordinates Analysis (PCoA) using a Hellinger distance (97) matrix was used to confirm that the batch effect corresponding to the batches of swabs was mitigated prior to further analysis. To confirm putative annotations, authentic standards were purchased for the linoleic acid (LA; Spectrum Laboratory Products, Inc., USA), conjugated linoleic acid (CLA; mixture of 4 isomers: 9,11 and 10,12 isomers, E and Z) (Sigma-Aldrich, USA), and selected antibiotics: tetracycline, oxytetracycline, and doxycyclin (Abcam Inc., USA). For level one identifications, each authentic compound was analyzed under identical experimental conditions and retention time and MS/MS spectra were compared with putatively annotated compounds.

Selective feature detection. Selective feature extraction was performed with open source MZmine2 software (98). To separate closely eluting LA and CLA isomers as well as separate various N-acyl amides, crop filtering with RT range of 5.4-6.0 minutes and m/z range of 281.246 - 281.248 was applied to all chromatograms. Mass detection was performed with a signal threshold of 1.0E2 and a 0.6 s minimum peak width. The mass tolerance was set to 20 ppm and the maximum allowed retention time deviation was set to 5 s. For chromatographic deconvolution, the baseline cutoff algorithm with a 5.0E1 signal threshold was used. The maximum peak width was set to 0.5 min. Similarly, the MS feature for reference compound stercobilin was extracted with a crop filter RT range of 2.0-4.0 minutes and m/z range of 595.345-595.355. The stercobilin reference compound was used to assess variability of chromatographic retention times to ensure that the compounds of interest (LA and CLA in particular) retention times were correctly identified. After isotope peak removal, the peak lists of all samples were aligned within the corresponding retention time and mass tolerances. Gap filling was performed on the aligned peak list using the peak finder module with 1% intensity, 10 ppm m/z tolerance, and 0.05 min RT tolerance, respectively. After the creation and export of a feature matrix containing the feature retention times, exact mass, and peak areas of the corresponding extracted ion chromatograms, we added sample metadata to the feature matrix metadata of the samples.

The selective feature extraction with the same settings has been performed for all of the detected compounds listed on the Figure 6A-I (the m/z range crop filter window was set for corresponding m/z for each compound).

Molecular Networking. Raw data files were converted to the .mzXML format using Bruker Data Analysis software and uploaded to the GNPS (https://gnps.ucsd.edu/) MassIVE mass spectrometry database (https://massive.ucsd.edu/). Molecular networking was performed to identify spectra shared between different sample types and to identify known molecules in the data set. All annotations are at level 2 according to the proposed minimum standards in metabolomics (92). The data were filtered by removing all MS/MS peaks within +/- 17 Da of the precursor m/z. MS/MS spectra were window-filtered by choosing only the top 6 peaks in the +/- 50 Da window throughout the spectrum. The MS spectra were then clustered with MS-Cluster algorithm with a parent mass tolerance of 0.02 Da and a MS/MS fragment ion tolerance of 0.02 Da to create consensus spectra (89). Further, consensus spectra that contained less than 4 spectra were discarded. A network was then created where edges were filtered to have a cosine score above 0.65 and more than 5 matched peaks. The edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. The spectra in the network were then searched against GNPS spectral libraries. The library spectra were filtered in the same manner as the input data. All library matches were required to have a score above 0.7 and at least 6 matched peaks. Molecular networks were visualized and mined using the Cytoscape software (www.cytoscape.org/).

Molecular networking-based propagation of annotations. The annotation of GPCR agonist compounds was not possible via direct library matching, as their spectra are not present in any MS libraries, but direct comparison with fragmentation patterns presented in (99) allowed us to establish these compounds' identity with level 3 identification (92). Consequently, manual annotation of compounds was carried out in two steps. The exact mass of compounds and their MS/MS fragmentation spectra were matched to the reference spectra found in supplementary info of (99) (fig S4A). Compound m/z 611.5357 was identified in this fashion. In addition, commendamide (330.2640) and its analogue (m/z 344.2799) were identified by matching exact mass of the corresponding ion and by in silico prediction of the MS/MS fragmentation spectra with the CSI:FingerID (100) (fig S4B). For novel molecules that were found within clusters of compounds of interest, but were not described in the literature previously, the structure was postulated using annotation propagation from adjacent annotated nodes in the cluster as described in (89) by assessing differences in parent mass and fragmentation patterns. The key structure, m/z 387.322 has been annotated as N-3-OH-palmitoyl ornithine based on the exact mass and previous annotation (99) as well as analysis of fragmentation pattern to confirm structural moieties of fragments (fig S4C). The rest of the structural assignments have been propagated from that structure. The ornithine moiety has been determined to be present in each structure (due to presence of the signature ion with m/z 115.09), and acylation of the hydroxyl is not possible due to insufficient mass of the structures; thus, the changing mass was postulated to correspond to different length of the alkyl substituent (fig 6, in the main text).

Correlations of Metabolites with Metadata. We have investigated correlations between metabolites (especially those of interest, such as N-Acyl amides) and all of the categories in the metadata. The data were subsetted into the Vioscreen and Antibiotics cohorts and normalized using probabilistic quotient normalization (101). In order to test the association of the metabolites to the categorical metadata fields we performed the Kruskal–Wallis test followed by Benjamini & Hochberg FDR correction to all metabolites. The significant metabolite-metadata associations (*p*-value adjusted < 0.05) were further connected to GNPS spectral library matches associating the MS1 feature to the MS2 precursor ion in a 10 ppm mass window and 20 seconds retention time window. The results are summarized in table S5.

Data pretreatment for statistical analysis. A PCoA plot using Hellinger distance (distance matrix: Hellinger; grouping: HCA) was built with all samples in the subset; one sample was found to be an outlier and removed. The data were then filtered to remove features with near-constant, very small values and values with low repeatability using the inter-quartile range estimate. Detailed description of methodology is given in (102). The samples were normalized by sum total of peak intensities, an important step due to large variability of the fecal material load on different swabs. To reduce the effect of background signal and make the sum normalization appropriate, the subtraction of blank and polymer peak features was conducted prior to analysis, as described above. The data were further scaled by mean centering and dividing by standard deviation for each feature.

The data were split into two groups for downstream analysis. Group one contained samples from individuals answering “More than 30” (*n*=41) and “Less that 10” (*n*=44) to the main American Gut Project survey question “In an average week, how many different plants do you eat?” Group two contained samples from individuals answering “antibiotic use within last week” (*n*=56) and “I have not taken antibiotics in the past year” (*n*=115) to the main American Gut Project survey question “I have taken antibiotics in the last ____.” for the Antibiotic history study, correspondingly.

The resultant features tables were used as input for the Metabonalist software (103). Partial least squares Discriminant Analysis (PLS-DA) (62) was used to explore and visualize variance within data and differences among experimental categories. Random forests (104) (RF) supervised analysis was used to further verify validity of determined discriminating features.

Expanded bloom assessment. The American Gut Project dataset now spans multiple-omics types, and include data that were unavailable during the analysis described in Amir et al. (14). To better understand how the blooming organisms impacted the samples in the American Gut, we 1) performed an additional set of 16S-based experiments; 2) cultured historical samples covering a range of bloom fractions, characterized their metabolites and sequenced the isolates; 3) performed shotgun metagenomics sequencing on the “high bloom” samples; 4) ran the set of samples previously run for HPLC-MS (e.g., the plants and antibiotics cohorts) for shotgun metagenomics, and 5) ran the storage samples from (105) for shotgun metagenomics. The additional sequencing effort was to provide a basis to assess whether functional potential driven by the blooms was impacting any of the biological results discussed in the manuscript. The additional HPLC-MS work was to characterize the metabolites specific to the blooms to remove them from analysis. The additional sequence data generated from the American Gut samples were deposited in EBI under the American Gut accession (ERP012803), and the storage sample data under its accession (ERP015155).

16S-based bloom experiments. Effect size calculations were computed prior to and following the removal of bloom reads using the procedure described by Amir et al., 2017 (84). The fraction of reads recruiting to blooms was included as a covariate. Effect sizes were assessed over Faith's Phylogenetic Diversity (59), unweighted UniFrac (57) and weighted UniFrac (106). We then computed Pearson and Spearman correlations of the effect sizes, per metric, between the bloom and bloom-removed result (fig 2D, E). In addition to the effect size calculations, we also tested whether the bloom fraction was correlated to any metadata category and did not observe significant correlations.

We then tested the removal of blooms from other studies in which room temperature shipping was not performed by retrieving a wide variety of human fecal studies from Qiita. UniFrac distance matrices were computed prior to and following bloom removal, followed by Mantel tests. The results of this procedure are outlined in table S4.

Finally, we correlated the relative intensities of the HPLC-MS data associated with the antibiotics and plants cohorts against the fraction of blooming reads. Critically, we observed a set of spectra that are significantly correlated (table S5) to this fraction. On annotation using molecular networking (discussed in detail the HPLC-MS section), we observed these metabolites to putatively be LysoPE, lysophospholipid (LPL), which has previously been associated with the release of colicin (107). These metabolites were removed from subsequent analyses.

Culturing. Primary specimens (*n*=214) were selected from three plates based off of the median fraction of reads recruiting to the blooms across the plate, whether the primary specimen still existed, and as to gather samples from at least the US (*n*=116) and UK (*n*=73); additional countries were included in smaller sample sizes and include Australia (*n*=7), Germany (*n*=7), Canada (*n*=3), Croatia (*n*=2), Belgium (*n*=2), France (*n*=1), Austria (*n*=1), Sweden (*n*=1), and the Czech Republic (*n*=1). The bloom typically observed in these samples (and in the full AGP dataset) is an *E. coli* (ID: 04195686f2b70585790ec75320de0d6f from (84)), although a few of the other bloom sequences were represented at high read fraction as well. Samples were retrieved from -80°C and thawed on ice. The swab head was broken off into 500 µl sterile 1x Dulbecco's Phosphate-Buffered Saline and vortexed vigorously for 30 seconds. Serial dilutions from this initial stock were made including 1:10,000 and 1:1,000,000. 10µl of the 1:10,000 dilution were inoculated into 1.5 ml sterile Tryptic Soy broth (TSB, BD cat#2253534) in sterile 96-deep-well plates (community cultures, CC) and incubated overnight at 37°C on an orbital shaker at 500 rpm. OD600 values above 0.1 (TSB controls measured ~0.08) were counted as positive growth. Samples with high bloom fraction tended to grow overnight in ambient conditions, samples with a low bloom fraction tended to not grow in these conditions (fig 2A). Additionally, 100 µl of each dilution were plated onto Tryptic Soy agar using sterile glass beads and incubated overnight at 37°C. The following morning, a picture of the best dilution was captured and the most representative colony was selected from each plate and inoculated into 1.5 ml sterile TSB for overnight incubation as above (isolates, IS). The following morning, OD600 measurements were taken and the cultures were pelleted at 3,000 g for 5 min. The supernatant and cell pellets were stored at -20°C for metabolomic analysis and DNA extraction, respectively.

Shotgun sequencing was performed on all isolates and community cultures using a 1:10 miniaturized Nextera library prep with 1 ng gDNA input or up to 1 µl and a 15 cycle PCR amplification. Libraries were quantified with PicoGreen™ dsDNA Assay Kit and 50 ng of each library (or 4 µl maximum) was pooled. The library was size-selected for 200-700 bp using the Sage Bioscience Pippin Prep and sequenced as a paired end 150 cycle run on an Illumina HiSeq 2500 v2 in Rapid Run mode at the UCSD IGM Genomics Center. Sequence processing including assembly performed as in the metagenomic processing section below with the exception that "--meta" was not used with SPAdes (108), and read binning against the resulting contigs was not performed. For each isolate, contigs with abnormally high or low coverage as defined by the 1.5 × IQR rule were dropped. The characterization of the metabolites from the supernatant using HPLC-MS is discussed in the HPLC-MS section above.

Following assembly of the draft genomes, taxonomic assessment by Kraken (109) revealed that of the 119 successfully sequenced colony isolate cultures, 95 matched the bloom organisms identified by Amir et al., 2017. Compellingly, 70 of these isolate genomes contained exact 16S sequence matches to a bloom organism identified by (84), including 65 of which matched the dominant *E. coli* bloom in the American Gut (table S4).

The read data for the isolates were then assessed for predicted biosynthetic gene clusters (BGCs). We used biosyntheticSPAdes (110) to analyze BGCs in the assembly graph of individual genomes. Below we focus on the longest BGCs that are particularly difficult to reconstruct based on ad hoc analysis of contigs and reveal their variations (that likely translate into variations of their natural products). Some of the reconstructed long BGC are ubiquitous (shared by many isolates, albeit with some variations), while others are unique, e.g., present in a single or small number of isolates. We identified BGCs, representing in the alphabet of their domains (table S4), and uncovered variations in their sequence across multiple isolates. Specifically, a ubiquitous BGC similar to the elusive peptide-polyketide genotoxin colibactin and a unique surfactin-like BGC. Colibactin triggers DNA double-strand breaks in eukaryotic cells (111, 112) and induces cellular senescence and metabolic reprogramming in affected mammalian cells (113). Of the 11 samples containing the longest colibactin-like BGC, 10 of them contained the exact *E. coli* bloom 16S sequence described above; the 11th isolate was actually a canine fecal sample plated alongside human (as the AGP allows participants to submit pet samples).

Although colibactin is frequently harbored by various *E. coli* strains, the variations of colibactin BGCs across various isolates have not been studied before. Genomic analysis revealed wide variations in colibactin-like BGCs suggesting that various strains produce related but not identical variants of natural products (114). These variations may give rise to the suite of LysoPE-associated spectra identified between the 16S and HPLC-MS datasets.

Shotgun sequencing of the high bloom and storage samples. Previously extracted DNA from the "high bloom" samples used for culturing was obtained, as was previously extracted DNA from Song et al. (105). Shotgun sequencing libraries from a total of 5 ng (or 3.5 μl maximum) gDNA was used in a 1:10 miniaturized KAPA HyperPlus protocol with a 15 cycle PCR amplification. Libraries were quantified with PicoGreen™ dsDNA Assay Kit and 50 ng (or 1 μl maximum) of each library was pooled. The pool was size-selected for 300-700 bp and sequenced as a paired end 150 cycle run on an Illumina HiSeq 2500 v2 in Rapid Run mode at the UCSD IGM Genomics Center. Sequence processing including assembly was performed as in the metagenomic processing section below.

Functional assessment of conjugated and non-conjugated linoleic acid. To investigate the metabolic potential of gut microbiome for producing conjugated linoleic acid from linoleic acid, we estimated the abundance of linoleic acid isomerase (LAI) in the fecal metagenome. We focused this investigation on the "plants" cohort, which were samples selected to maximize the difference between the number of types of plants metadata category as discussed in the main text. First, we translated the assembled metagenomes to metaproteomes using Prodigal gene prediction software. To map LAI to these metaproteomes, we used a representative LAI protein sequence (UniProt: D2BQ64), which was matched against UniProtKB (via https://www.ebi.ac.uk/Tools/hmmer/) for multiple sequence alignment (MSA). The resulting MSA file in clustal format was then used to generate a hidden Markov model (HMM) profile for LAI using hmmbuild in HMMER software (115). Subsequently, we mapped the resulting HMM profile to sample metaproteomes using hmmsearch with an E-value threshold of 10E-5. We calculated abundances of LAI per sample based on abundance (coverage x length) of LAI containing contigs in each sample, normalized to total sample biomass and performed linear regression between LAI abundances and bloom fraction. We did not note any correlation between metabolic potential of gut metagenome to produce LAI and the fraction of blooming bacteria (samples with no LAI hits were removed from this analysis). Similarly, there was no correlation between CLA abundances and bloom fraction in the samples. These results suggest that our report on the differential abundance of CLA in subjects with different dietary practices (with respect to the number of different types of plants consumed) is unlikely to be confounded by the presence of blooming bacteria.

Storage sample assessment. Metagenomic reads from the storage samples were mapped to the 169 isolate assemblies. We then ran model comparison tests on each to determine which mappings were significantly different between frozen samples and samples left out at ambient temperatures for various periods of time. Using the 'lme' package (116) in R (v3.3.3. R Core Team 2017), linear mixed effects models were applied to the abundances, with individual treated as the random effect. Mappings were considered to be significantly associated with temperature if the model was significantly improved (ANOVA $p<=0.05$) by incorporating a fixed effect of temperature. Seven mappings to isolates were found to be significantly increased in samples stored in ambient temperatures compared to frozen samples in both storage studies, of which 3 contained the 16S of the dominant *E. coli* bloom in the AGP samples, and 2 contained the 16S from other blooms recognized by (84).

Shotgun sequence processing. Raw FastQ files were processed using Atropos v1.1.5 (117) to remove adapters and low-quality regions. Putative human genome contaminations were identified and removed by using Bowtie2 v2.3.0 (118) with the “--very-sensitive” option against the human reference genome GRCh37/hg19.

Sequences were assigned taxonomy using Kraken v1.0.0 (109) against the “standard” database built following the Kraken manual, which contains all complete bacterial, archeal, and viral genomes available from NCBI RefSeq as of Aug. 3, 2017. Results were processed using Bracken v1.0.0 (119) to estimate the relative abundance of species-level taxa.

Metagenome sequencing data were assembled using SPAdes v3.11.1 (108) with the “--meta” flag enabled. Contigs ≥ 1 kb in length were retained and fed to the prokaryotic genome annotation pipeline Prokka v1.12 (120). putatively individual genomes were inferred using MaxBin2 v2.2.4 (121).

In parallel, contigs were sheared into 200-bp fragments and taxonomy was assigned using Kraken (see above). For each contig, the most assigned taxon at each taxonomic rank and the proportion of sequences assigned to it was inferred.

A total of 3725 genome bins were identified from 677 out of 780 AGP metagenomes, with 5.50 ± 4.05 bins per sample, and a maximum bin number of 30. Bins with completeness < 50% were dropped, leaving 1029 bins from 464 samples (2.22 ± 1.97 bins per sample, maximum bins = 19).

Filtering Bacterial Blooms for Metabolomics Analysis. To assess and account for the impact of the metabolites contributed by these organisms, we have performed HPLC-MS analysis of cultures of blooming organisms to establish possible contributions, as described above. The raw data are publicly available at the UCSD Center for Computational Mass Spectrometry (*http://massive.ucsd.edu/*, dataset ID: MassIVE MSV000081777). It was found that there is a negligible overlap of the bloom-associated metabolites with the compounds detected in AGP samples (fig 2B). Furthermore, we have verified that none of the compounds discussed in this work (LA, CLA, compounds on Fig 6A-I) are present in these bloom cultures. The main organism implicated in bloom was determined to be *E. coli*, as described earlier and MS data corroborate these findings (fig 2C).

Considering that the metabolites resulting from microbial activity in cultures can differ significantly from those in vivo (e.g. many of the metabolites could originate not from de novo synthesis, but rather from microbial modifications of external compounds that are not present in media, e.g. from the host), we also explored associations of metabolites in AGP metabolomics samples and blooms. Spearman rank correlation analysis of the fraction of 16S reads in a sample reporting as bloom to metabolites observed in the same samples revealed several features that correlate significantly (table S5). There exists a significant overlap between the Antibiotics and Vioscreen studies subsets, indicating potential common origin of these features. The strongest correlation was found for the feature m/z 480.3106 with multiple bloom organisms ($\boldsymbol{\rho}$^2 > 0.25 for *E. coli* at $p < 1\text{e-}40$). This feature was found to also significantly correlate with the principle coordinates of the PCoA, with and without blooms in the UniFrac matrices for both subsets. The tentative annotation of this feature is lysoPE, a lysophospholipid (LPL). The LPLs production in vivo is a result of phospholipase A enzymatic activity associated with Gram-negative bacteria. It is known that lysoPE is essential for release of colicin (107). Colicin (by itself not detectable with the MS methodology in this study due to very high molecular mass) is a bacteriocin related to microbial warfare and is known to be produced by *E. coli*, the major bloomer in AGP. It can be suggested that the blooming of an organism is related to attempting to kill competitors to maximize nutrient availability. Importantly, removal of all of the features associated with bloom does not alter the metabolomics results at all, which indicates that all of the observed biological trends reported here are not related to blooms.

Mental health in the American Gut Project. From AGP cohort, we selected subjects who endorsed a mental health disorder (depression, schizophrenia, PTSD, and/or bipolar disorder). This resulted in 1,140 subjects. 636 subjects endorsed at least one of the exclusion criteria (antibiotic use in the last year, IBD, *C. difficile* infection, pregnancy, Alzheimer's, anorexia or bulimia, history of substance use disorder, epilepsy or seizure disorder, kidney disease, phenylketonuria). Out of the remaining 504 subjects, 319 did not provide information regarding country of residence, hence forming a case cohort of 185 subjects. The remaining samples were further filtered down to 125 samples to include only high quality fecal microbiome data (at least 1,250 sequences/sample) at a single time point per subject. For those cases, we created a 1:1 matched sample of patients and non-psychiatric comparison (NC) participants based on age (±5 years), BMI, history of diabetes, smoking frequency, country of residence, census region (if in US), and sequencing plate. For each of the cohorts we calculated beta-diversity distance matrices using Bray-Curtis dissimilarity and weighted UniFrac. On resulting matrices we ran pairwise PERMANOVA with 999 permutations between "cases" (people who reported mental illness) and NCs (out matched control dataset). Differential abundance testing was performed using permutive mean difference test at 10,000 permutations, with discrete FDR (122) correction at alpha=0.1.

Acknowledgments: The authors would like to thank first all the citizen science volunteers around the world. We also thank Katherine Amato, Margaret K. Butler, Gabriela Surdulescu, Luke Ursell, Jeff DeReus, Evguenia Kopylova, John Graham, Victoria Vazquez, Jai Ram Rideout, Merete Eggesbo, Steve Green, Kathy Holt, Jim Huntley, Zehra Esra Ilhan, Daniel Mayer, Catherine Nicholas, Laura Parfrey, Juanma Peralta, Dorota Porazinska, Jennifer Smilowitz, Matt Gebert, Se Jin Song, Kumar Thurimella, Sam Way, Tony Walters, Sophie Weiss, Ulla Westermann, Shawnelle White, Doug Woodhams, Jerry Kennedy, Scott Handley, Chandni Desai, Carolina Carpenter, Gabriella Fanelli, Deborah L. Bright and Xochitl C. Morgan for their contributions to the American Gut Project. Some sequencing was conducted by Mahdieh Khosroheidari and Clifford L. Green at the IGM Genomics Center, University of California, San Diego, La Jolla, CA.The de-identified participant and sequence data for the American Gut are available from the European Bioinformatics Institute under accession number ERP012803; the HPLC-MS data are available in MassIVE under accession MSV000080179. Per sample accession numbers are included in the supplemental information (table S1). The EBI accession includes more samples than presented here as some samples were consented after the analysis was completed. Crowdfunding support was provided through Indiegogo and FundRazr. Some consortium members were supported by NSF IGERT award 1144807, the Sloan Foundation Microbiology of the Built Environment Program, the Wyss Institute, the Howard Hughes Medical Institute, the U.S. Dept. of Energy under Contract DE-AC02-06CH11357.NS, the ERC Starting Grant 2013 (European Research Council, Starting grant 336452-ENIGMO), WELBIO WELBIO-CR-2012S-02R and Robert Tournut award (Société Nationale Française de Gastro-Entérologie). The British Gut collection was carried out by the Department of Twin

Research at KCL funded by the Wellcome Trust, MRC, JPI Dinamic grant and the National Institute for Health Research (NIHR) Clinical Research Facility at Guy's & St Thomas' NHS Foundation Trust and NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. Tim Spector is an NIHR senior Investigator. This research was performed in accordance with the University of Colorado Boulder's Institutional Review Board protocol #12-0582 and the University of California San Diego's Human Research Protection Program protocol #141853. Parts of the computations were performed using the San Diego Supercomputer Center (SDSC) through XSEDE allocations, which is supported by the NSF grant ACI-1053575. We thank the BioFrontiers Institute for their administrative support for the project while it was based at the University of Colorado Boulder.

References

1. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463.

2. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, Tito RY, Chaffron S, Rymenans L, Verspecht C, De Sutter L, Lima-Mendez G, D'hoe K, Jonckheere K, Homola D, Garcia R, Tigchelaar EF, Eeckhaudt L, Fu J, Henckaerts L, Zhernakova A, Wijmenga C, Raes J. 2016. Population-

level analysis of gut microbiome variation. Science 352:560–564.

3. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, MetaHIT Consortium, Bork P, Ehrlich SD, Wang J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59–65.

4. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, Mujagic Z, Vila AV, Falony G, Vieira-Silva S, Wang J, Imhann F, Brandsma E, Jankipersadsing SA, Joossens M, Cenit MC, Deelen P, Swertz MA, LifeLines cohort study, Weersma RK, Feskens EJM, Netea MG, Gevers D, Jonkers D, Franke L, Aulchenko YS, Huttenhower C, Raes J, Hofker MH, Xavier RJ, Wijmenga C, Fu J. 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352:565–569.

5. Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214.

6. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI. 2012. Human gut microbiome viewed across age and geography. Nature 486:222–227.

7. Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, Leff JW,

Vázquez-Baeza Y, Gonzalez A, Knight R, Dunn RR, Fierer N. 2014. Temporal variability is a personalized feature of the human microbiome. Genome Biol 15:531.

8. Song SJ, Lauber C, Costello EK, Lozupone CA, Humphrey G, Berg-Lyons D, Caporaso JG, Knights D, Clemente JC, Nakielny S, Gordon JI, Fierer N, Knight R. 2013. Cohabiting family members share microbiota with one another and with their dogs. Elife 2:e00458.

9. Smith B, Chu LK, Smith TC, Amoroso PJ, Boyko EJ, Hooper TI, Gackstetter GD, Ryan MAK, Millennium Cohort Study Team. 2008. Challenges of self-reported medical conditions and electronic medical records among members of a large military cohort. BMC Med Res Methodol 8:37.

10. Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, Navas-Molina JA, Song SJ, Metcalf JL, Hyde ER, Lladser M, Dorrestein PC, Knight R. 2017. Balance Trees Reveal Microbial Niche Differentiation. mSystems 2.

11. Clemente JC, Pehrsson EC, Blaser MJ, Sandhu K, Gao Z, Wang B, Magris M, Hidalgo G, Contreras M, Noya-Alarcón Ó, Lander O, McDonald J, Cox M, Walter J, Oh PL, Ruiz JF, Rodriguez S, Shen N, Song SJ, Metcalf J, Knight R, Dantas G, Dominguez-Bello MG. 2015. The microbiome of uncontacted Amerindians. Sci Adv 1.

12. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, Spicer P, Lawson P, Marin-Reyes L, Trujillo-Villarroel O, Foster M, Guija-Poma E, Troncoso-Corzo L, Warinner C, Ozga AT, Lewis CM. 2015. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun 6:6505.

13. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71:8228–8235.

14. Amir A, McDonald D, Navas-Molina JA, Debelius J, Morton JT, Hyde E, Robbins-Pianka A, Knight R. 2017. Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping. mSystems 2.

15. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2.

16. Song SJ, Amir A, Metcalf JL, Amato KR, Xu ZZ, Humphrey G, Knight R. 2016. Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. mSystems 1.

17. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, Tito RY, Chaffron S, Rymenans L, Verspecht C, De Sutter L, Lima-Mendez G, D’hoe K, Jonckheere K, Homola D, Garcia R, Tigchelaar EF, Eeckhaudt L, Fu J, Henckaerts L, Zhernakova A, Wijmenga C, Raes J. 2016. Population-level analysis of gut microbiome variation. Science 352:560–564.

18. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, Mujagic Z, Vila AV, Falony G, Vieira-Silva S, Wang J, Imhann F, Brandsma E, Jankipersadsing SA, Joossens M, Cenit MC, Deelen P, Swertz MA, LifeLines cohort study, Weersma RK, Feskens EJM, Netea MG, Gevers D, Jonkers D, Franke L, Aulchenko YS, Huttenhower C, Raes J, Hofker MH, Xavier RJ, Wijmenga C, Fu J. 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352:565–569.

19. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. 2008. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol 6:776–788.

20. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618.

21. Mirarab S, Nguyen N, Warnow T. 2012. SEPP: SATé-enabled phylogenetic placement. Pac Symp Biocomput 247–258.

22. Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, D'Amato M, Bonfiglio F, McDonald D, Gonzalez A, McClure EE, Dunklebarger MF, Knight R, Jansson JK. 2017. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol 2:17004.

23. McDonald D, Ackermann G, Khailova L, Baird C, Heyland D, Kozar R, Lemieux M, Derenski K, King J, Vis-Kampen C, Knight R, Wischmeyer PE. 2016. Extreme Dysbiosis of the Microbiome in Critical Illness. mSphere 1.

24. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

25. Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, Knight R, Manjurano A, Changalucha J, Elias JE, Dominguez-Bello MG, Sonnenburg JL. 2017. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357:802–806.

26. McDonald D, Robbins-Pianka A, Mann AE, Vrbanac A, Amir A, Frazier A, Gonzalez A, Tripathi A, Fahimipour AK, Brennen C, Martino C, Lebrilla C, Lozupone C, M. Lewis C Jr, Raison C, Zhang C, L. Lauber C, Warinner C, A. Lowry C, Callewaert C, Bloss C, Huttenhower C, Knights D, Willner D, Galzerani DD, Gonzalez DJ, Mills DA, Chopra D,

Gevers D, Berg-Lyons D, D. Sears D, Wendel D, Wolfe E, Lovelace E, R. Hyde E, Pierce E, TerAvest E, Montassier E, Bolyen E, Bushman FD, Ackermann G, D. Wu G, Church GM, Rahnavard G, Saxe G, Gogul G, Humphrey G, D. Holscher H, Ugrina I, German JB, Gregory Caporaso J, Gilbert J, Wozniak JM, Kerr J, Ravel J, Gaffney J, D. Lewis J, Morton JT, Suchodolski JS, Jansson JK, Hampton-Marcell JT, Bobe J, Leach J, Raes J, L. Green J, Metcalf JL, H. Chase J, A. Eisen J, Monk J, Navas-Molina JA, C. Clemente J, Petrosino J, Ladau J, Shorenstein J, Goodrich J, Gauglitz J, Jacobs J, W. Debelius J, Zengler K, S. Pollard K, Swanson KS, Lewis K, Mayer K, Bittinger K, Goldasich LD, Dillon L, S. Zaramela L, R. Thompson L, M. Schriml L, Dominguez-Bello MG, Jankowska MM, Blaser M, Jackson MA, Pirrung M, Minson M, Kurisu M, Ajami N, Sangwan N, Gottel NR, Chia N, Fierer N, White O, D. Cani P, Wischmeyer P, Gajer P, Strandwitz P, Hugenholtz P, Dorrestein PC, Kashyap P, Dutton R, Park RS, Xavier RJ, Knight R, R. Dunn R, Mills RH, Krajmalnik-Brown R, Ley R, M. Owens S, T. Kelley S, Klemmer S, Matamoros S, Peddada S, Mirarab S, Janssen S, Moorman S, Holmes S, Schwartz T, Eshoo-Anton TW, Vigers T, Spector T, Kosciolek T, G. Thackray V, Pandey V, Van Treuren W, Fang X, Chen Y, Vázquez-Baeza Y, Zech Xu Z, Aksenov AA, Melnik AV, Jarmusch A, Geier J, Pevzner P, Meleshko D, Behsaz B, Reeve N, Silva R, Zhu Q, Kopylova E, Marotz C, Smarr L, Nguyen T, Mohimani H, Jiang L, Swafford A, Nguyen D, Song SJ, Sanders K, Benitez RAS, Heale AC, V. Jeste D, Nguyen TT, Brennan C, Abramson M. 2018. movie_s1.mp4. figshare.

27. Barberán A, Dunn RR, Reich BJ, Pacifici K, Laber EB, Menninger HL, Morton JM, Henley JB, Leff JW, Miller SL, Fierer N. 2015. The ecology of microscopic life in household dust. Proc Biol Sci 282.

28. Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O’Dwyer J, Koeppel AF, Green JL, Pollard KS. 2013. Global marine bacterial diversity peaks at high latitudes in winter. ISME J 7:1669–1677.

29. Gophna U, Konikoff T, Nielsen HB. 2017. Oscillospira and related bacteria - From metagenomic species to metabolic features. Environ Microbiol 19:835–841.

30. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42:D490–5.

31. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B. 2013. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nat Rev Microbiol 11:497–504.

32. Cummings JH, Macfarlane GT. 1991. The control and consequences of bacterial fermentation in the human colon. J Appl Bacteriol 70:443–459.

33. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, Moore BS, Laskin J, Bandeira N, Dorrestein PC. 2012. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A 109:E1743–52.

34. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail

KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837.

35. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TW-M, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR. 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 3:211–221.

36. Kishino S, Takeuchi M, Park S-B, Hirata A, Kitamura N, Kunisawa J, Kiyono H, Iwamoto R, Isobe Y, Arita M, Arai H, Ueda K, Shima J, Takahashi S, Yokozeki K, Shimizu S, Ogawa J. 2013. Polyunsaturated fatty acid saturation by gut lactic acid bacteria affecting host lipid composition. Proc Natl Acad Sci U S A 110:17808–17813.

37. Coakley M, Ross RP, Nordgren M, Fitzgerald G, Devery R, Stanton C. 2003. Conjugated linoleic acid biosynthesis by human-derived Bifidobacterium species. J Appl Microbiol 94:138–145.

38. Cohen LJ, Esterhazy D, Kim S-H, Lemetre C, Aguilar RR, Gordon EA, Pickard AJ, Cross JR, Emiliano AB, Han SM, Chu J, Vila-Farres X, Kaplitt J, Rogoz A, Calle PY, Hunter C, Bitok JK, Brady SF. 2017. Commensal bacteria make GPCR ligands that mimic human

signalling molecules. Nature 549:48–53.

39. Hauser AS, Attwood MM, Rask-Andersen M, Schiöth HB, Gloriam DE. 2017. Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov 16:829–842.

40. Cani PD, Knauf C. 2016. How gut microbes talk to organs: The role of endocrine and nervous routes. Mol Metab 5:743–752.

41. Clarke G, Stilling RM, Kennedy PJ, Stanton C, Cryan JF, Dinan TG. 2014. Minireview: Gut microbiota: the neglected endocrine organ. Mol Endocrinol 28:1221–1238.

42. Jiang H, Ling Z, Zhang Y, Mao H, Ma Z, Yin Y, Wang W, Tang W, Tan Z, Shi J, Li L, Ruan B. 2015. Altered fecal microbiota composition in patients with major depressive disorder. Brain Behav Immun 48:186–194.

43. Lin P, Ding B, Feng C, Yin S, Zhang T, Qi X, Lv H, Guo X, Dong K, Zhu Y, Li Q. 2017. Prevotella and Klebsiella proportions in fecal microbial communities are potential characteristic parameters for patients with major depressive disorder. J Affect Disord 207:300–304.

44. Saraceno B. 2002. The WHO World Health Report 2001 on mental health. Epidemiol Psichiatr Soc 11:83–87.

45. Zheng P, Zeng B, Zhou C, Liu M, Fang Z, Xu X, Zeng L, Chen J, Fan S, Du X, Zhang X, Yang D, Yang Y, Meng H, Li W, Melgiri ND, Licinio J, Wei H, Xie P. 2016. Gut microbiome remodeling induces depressive-like behaviors through a pathway mediated by the host's metabolism. Mol Psychiatry 21:786–796.

46. Weingarden A, González A, Vázquez-Baeza Y, Weiss S, Humphry G, Berg-Lyons D,

Knights D, Unno T, Bobr A, Kang J, Khoruts A, Knight R, Sadowsky MJ. 2015. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection. Microbiome 3:10.

47. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE. 2011. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A 108 Suppl 1:4578–4585.

48. McDonald D, Robbins-Pianka A, Mann AE, Vrbanac A, Amir A, Frazier A, Gonzalez A, Tripathi A, Fahimipour AK, Brennen C, Martino C, Lebrilla C, Lozupone C, M. Lewis C Jr, Raison C, Zhang C, L. Lauber C, Warinner C, A. Lowry C, Callewaert C, Bloss C, Huttenhower C, Knights D, Willner D, Galzerani DD, Gonzalez DJ, Mills DA, Chopra D, Gevers D, Berg-Lyons D, D. Sears D, Wendel D, Wolfe E, Lovelace E, R. Hyde E, Pierce E, TerAvest E, Montassier E, Bolyen E, Bushman FD, Ackermann G, D. Wu G, Church GM, Rahnavard G, Saxe G, Gogul G, Humphrey G, D. Holscher H, Ugrina I, German JB, Gregory Caporaso J, Gilbert J, Wozniak JM, Kerr J, Ravel J, Gaffney J, D. Lewis J, Morton JT, Suchodolski JS, Jansson JK, Hampton-Marcell JT, Bobe J, Leach J, Raes J, L. Green J, Metcalf JL, H. Chase J, A. Eisen J, Monk J, Navas-Molina JA, C. Clemente J, Petrosino J, Ladau J, Shorenstein J, Goodrich J, Gauglitz J, Jacobs J, W. Debelius J, Zengler K, S. Pollard K, Swanson KS, Lewis K, Mayer K, Bittinger K, Goldasich LD, Dillon L, S. Zaramela L, R. Thompson L, M. Schriml L, Dominguez-Bello MG, Jankowska MM, Blaser M, Jackson MA, Pirrung M, Minson M, Kurisu M, Ajami N, Sangwan N, Gottel NR, Chia N, Fierer N, White O, D. Cani P, Wischmeyer P, Gajer P, Strandwitz P, Hugenholtz P, Dorrestein PC, Kashyap P, Dutton R, Park RS, Xavier RJ, Knight R, R. Dunn R, Mills RH, Krajmalnik-Brown R, Ley R, M. Owens S, T. Kelley S, Klemmer S, Matamoros S, Peddada S, Mirarab S, Janssen S, Moorman S, Holmes S, Schwartz T, Eshoo-Anton TW, Vigers T, Spector T, Kosciolek T, G. Thackray V, Pandey V, Van Treuren W, Fang X, Chen Y, Vázquez-Baeza

Y, Zech Xu Z, Aksenov AA, Melnik AV, Jarmusch A, Geier J, Pevzner P, Meleshko D, Behsaz B, Reeve N, Silva R, Zhu Q, Kopylova E, Marotz C, Smarr L, Nguyen T, Mohimani H, Jiang L, Swafford A, Nguyen D, Song SJ, Sanders K, Benitez RAS, Heale AC, V. Jeste D, Nguyen TT, Brennan C, Abramson M. 2018. movie_s2.mp4. figshare.

49. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621–1624.

50. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 108 Suppl 1:4516–4522.

51. Apprill A, McNally S, Parsons R, Weber L. 2015. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquat Microb Ecol 75:129–137.

52. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2.

53. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J 6:610–618.

54. Mirarab S, Nguyen N, Warnow T. 2012. SEPP: SATé-enabled phylogenetic placement. Pac Symp Biocomput 247–258.

55. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid

assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267.

56. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336.

57. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71:8228–8235.

58. Gao HX, Regier EE, Close KL. 2016. Prevalence of and trends in diabetes among adults in the United States, 1988-2012. J Diabetes 8:8–9.

59. Faith DP. 1992. Conservation evaluation and phylogenetic diversity. Biol Conserv 61:1–10.

60. Anderson MJ. 2008. A new method for non-parametric multivariate analysis of variance. Austral Ecol 26:32–46.

61. Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, Navas-Molina JA, Song SJ, Metcalf JL, Hyde ER, Lladser M, Dorrestein PC, Knight R. 2017. Balance Trees Reveal Microbial Niche Differentiation. mSystems 2.

62. Wold S, Sjöström M, Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. Chemometrics Intellig Lab Syst 58:109–130.

63. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI.

2012. Human gut microbiome viewed across age and geography. Nature 486:222–227.

64. Clemente JC, Pehrsson EC, Blaser MJ, Sandhu K, Gao Z, Wang B, Magris M, Hidalgo G, Contreras M, Noya-Alarcón Ó, Lander O, McDonald J, Cox M, Walter J, Oh PL, Ruiz JF, Rodriguez S, Shen N, Song SJ, Metcalf J, Knight R, Dantas G, Dominguez-Bello MG. 2015. The microbiome of uncontacted Amerindians. Sci Adv 1.

65. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, Zech Xu Z, Van Treuren W, Knight R, Gaffney PM, Spicer P, Lawson P, Marin-Reyes L, Trujillo-Villarroel O, Foster M, Guija-Poma E, Troncoso-Corzo L, Warinner C, Ozga AT, Lewis CM. 2015. Subsistence strategies in traditional societies distinguish gut microbiomes. Nat Commun 6:6505.

66. Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience 2:16.

67. Bray JR, Curtis JT. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol Monogr 27:325–349.

68. Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220.

69. Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc 57:289–300.

70. Moran PAP. 1950. Notes on continuous stochastic phenomena. Biometrika 37:17–23.

71. Maurice CF, Knowles SCL, Ladau J, Pollard KS, Fenton A, Pedersen AB, Turnbaugh PJ. 2015. Marked seasonal variation in the wild mouse gut microbiota. ISME J 9:2423–2434.

72. Cramér H. 1946. Mathematical Methods of Statistics (PMS-9).

73. Welch BL. 1947. The generalisation of student's problems when several different population variances are involved. Biometrika 34:28–35.

74. Guo W, Sarkar SK, Peddada SD. 2010. Controlling false discoveries in multidimensional directional decisions, with applications to gene expression data on ordered categories. Biometrics 66:485–492.

75. Cohen J. 1992. A power primer. Psychol Bull 112:155–159.

76. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, Mujagic Z, Vila AV, Falony G, Vieira-Silva S, Wang J, Imhann F, Brandsma E, Jankipersadsing SA, Joossens M, Cenit MC, Deelen P, Swertz MA, LifeLines cohort study, Weersma RK, Feskens EJM, Netea MG, Gevers D, Jonkers D, Franke L, Aulchenko YS, Huttenhower C, Raes J, Hofker MH, Xavier RJ, Wijmenga C, Fu J. 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352:565–569.

77. Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance: NON-PARAMETRIC MANOVA FOR ECOLOGY. Austral Ecol 26:32–46.

78. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE. 2011. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A 108 Suppl 1:4578–4585.

79. Weingarden A, González A, Vázquez-Baeza Y, Weiss S, Humphry G, Berg-Lyons D, Knights D, Unno T, Bobr A, Kang J, Khoruts A, Knight R, Sadowsky MJ. 2015. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection. Microbiome 3:10.

80. McDonald D, Ackermann G, Khailova L, Baird C, Heyland D, Kozar R, Lemieux M,

Derenski K, King J, Vis-Kampen C, Knight R, Wischmeyer PE. 2016. Extreme Dysbiosis of the Microbiome in Critical Illness. mSphere 1.

81. Vázquez-Baeza Y, Gonzalez A, Smarr L, McDonald D, Morton JT, Navas-Molina JA, Knight R. 2017. Bringing the Dynamic Microbiome to Life with Animations. Cell Host Microbe 21:7–10.

82. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth’s multiscale microbial diversity. Nature 551:457–463.

83. Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, Knight R, Manjurano A, Changalucha J, Elias JE, Dominguez-Bello MG, Sonnenburg JL. 2017. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357:802–806.

84. Amir A, McDonald D, Navas-Molina JA, Debelius J, Morton JT, Hyde E, Robbins-Pianka A, Knight R. 2017. Correcting for Microbial Blooms in Fecal Samples during Room-Temperature Shipping. mSystems 2.

85. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

86. Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214.

87. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550:61–66.

88. Röst HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich H-C, Gutenbrunner P, Kenar E, Liang X, Nahnsen S, Nilse L, Pfeuffer J, Rosenberger G, Rurik M, Schmitt U, Veit J, Walzer M, Wojnar D, Wolski WE, Schilling O, Choudhary JS, Malmström L, Aebersold R, Reinert K, Kohlbacher O. 2016. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods 13:741–748.

89. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan

BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837.

90. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, Moore BS, Laskin J, Bandeira N, Dorrestein PC. 2012. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A 109:E1743–52.

91. Nguyen DD, Melnik AV, Koyama N, Lu X, Schorn M, Fang J, Aguinaldo K, Lincecum TL Jr, Ghequire MGK, Carrion VJ, Cheng TL, Duggan BM, Malone JG, Mauchline TH, Sanchez LM, Kilpatrick AM, Raaijmakers JM, Mot RD, Moore BS, Medema MH, Dorrestein PC. 2016. Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. Nat Microbiol 2:16197.

92. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TW-M, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR. 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 3:211–221.

93. GNPS antibiotic use subset. https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9bd16822c8d448f59a03e6cc8f017f43

94. GNPS plants subset.

https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d26ae082b1154f73ac050796fcaa6bda

95. GNPS isolate supernatant. https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=23f0f5e5c70f4163b445de71d086d186

96. GNPS fecal samples co-networked. https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=adcfbba9b4ca448f8b2133559b16d954

97. 1909. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. J Reine Angew Math 1909.

98. Pluskal T, Castillo S, Villar-Briones A, Oresic M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11:395.

99. Cohen LJ, Esterhazy D, Kim S-H, Lemetre C, Aguilar RR, Gordon EA, Pickard AJ, Cross JR, Emiliano AB, Han SM, Chu J, Vila-Farres X, Kaplitt J, Rogoz A, Calle PY, Hunter C, Bitok JK, Brady SF. 2017. Commensal bacteria make GPCR ligands that mimic human signalling molecules. Nature 549:48–53.

100. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci U S A 112:12580–12585.

101. Ejigu BA, Valkenborg D, Baggerman G, Vanaerschot M, Witters E, Dujardin J-C, Burzykowski T, Berg M. 2013. Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. OMICS 17:473–485.

102. Hackstadt AJ, Hess AM. 2009. Filtering for increased power for microarray data

analysis. BMC Bioinformatics 10:11.

103. Xia J, Wishart DS. 2016. Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis. Curr Protoc Bioinformatics 55:14.10.1–14.10.91.

104. Breiman L. 2001. 10.1023/A:1010933404324. Machine Learning.

105. Song SJ, Amir A, Metcalf JL, Amato KR, Xu ZZ, Humphrey G, Knight R. 2016. Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. mSystems 1.

106. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 73:1576–1585.

107. Pugsley AP, Goldzahl N, Barker RM. 1985. Colicin E2 production and release by Escherichia coli K12 and other Enterobacteriaceae. J Gen Microbiol 131:2673–2686.

108. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

109. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.

110. Meleshko D, Mohimani H, Hajirasouliha I, Medema MH, Korobeynikov A, Pevzner P. BiosyntheticSPAdes: Reconstructing Biosynthetic Gene Clusters From Assembly Graphs. submitted.

111. Nougayrède J-P, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G,

Buchrieser C, Hacker J, Dobrindt U, Oswald E. 2006. Escherichia coli induces DNA double-strand breaks in eukaryotic cells. Science 313:848–851.

112. Putze J, Hennequin C, Nougayrède J-P, Zhang W, Homburg S, Karch H, Bringer M-A, Fayolle C, Carniel E, Rabsch W, Oelschlaeger TA, Oswald E, Forestier C, Hacker J, Dobrindt U. 2009. Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae. Infect Immun 77:4696–4703.

113. Secher T, Samba-Louaka A, Oswald E, Nougayrède J-P. 2013. Escherichia coli producing colibactin triggers premature and transmissible senescence in mammalian cells. PLoS One 8:e77157.

114. Gurevich A, Mikheenko A, Shlemov A, Korobeynikov A, Mohimani H, Pevzner P. 2018. Modification-tolerant database search reveals surprising diversity of peptidic natural products. Nature Microbiology.

115. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–37.

116. Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw 67.

117. Didion JP, Martin M, Collins FS. 2017. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ 5:e3720.

118. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

119. Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in metagenomics data.

120. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

121. Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607.

122. Jiang L, Amir A, Morton JT, Heller R, Arias-Castro E, Knight R. 2017. Discrete False-Discovery Rate Improves Identification of Differentially Abundant Microbes. mSystems 2.

123. Cohen LJ, Kang H-S, Chu J, Huang Y-H, Gordon EA, Reddy BVB, Ternei MA, Craig JW, Brady SF. 2015. Functional metagenomic discovery of bacterial effectors in the human microbiome and isolation of commendamide, a GPCR G2A/132 agonist. Proc Natl Acad Sci U S A 112:E4825–34.

Figure 1. Population characteristics. **(A)** Participants across the world have sent in samples to the American Gut, although the primary geographic regions of participation are in North America and the United Kingdom; the report a participant receives is depicted. **(B)** The primary sample breakdown for subsequent analyses. Red denotes reasons samples were removed. **(C)** Between the two largest populations, the US (*n*=6,634) and the UK (*n*=2,071), we observe a significant difference in alpha diversity. **(D)** In a meta-analysis, the largely industrialized population that makes up the American Gut exhibits significant differential abundances to non-industrialized populations.

Figure 2. Blooms and effect sizes. **(A)** The fraction of 16S reads that recruit to bloom reads defined by Amir et al. 2017 is strongly associated with the likelihood for microbial growth under aerobic culture conditions on rich media. **(B)** Molecular network of the metabolites observed in

the supernatant from cultures (*n*=217) derived from fecal samples. The nodes in red (n=239) are metabolites associated with *E. coli*. **(C)** Overlap of metabolites between AGP samples and blooms. **(D)** Unweighted UniFrac effect sizes. The inset shows the correlation of effect sizes when including or excluding the bloom 16S reads (Pearson *r*=0.91, *p*=3.76x10^{-57}). **(E)** Weighted UniFrac effect sizes. The inset shows the correlation of the effect sizes when including or excluding bloom 16S reads (Pearson *r*=0.42, *p*=1.71x10^{-6}); the outlier is the 16S bloom fraction of the sample.

Figure 3. OTU and beta-diversity novelty. **(A)** The AGP data placed into the context of extant microbial diversity at a global scale. **(B)** A phylogenetic tree showing the diversity spanned by the AGP, and the HMP in the context of Greengenes and the EMP. **(C)** sOTU novelty over increasing numbers of samples in the AGP; the AGP appears to have begun to reach saturation and is contrasted with **(D)** Yatsunenko et al. 2012 which unlike the AGP had extremely deep sequencing per sample. **(E)** The minimum observed UniFrac distance between samples over increasing numbers of samples for the AGP and the HMP; inset is from 0-500 samples. **(F)** An AGP "trading card" of an sOTU of interest (shown in full in fig S2).

Figure 4. Temporal and spatial patterns. **(A)** 565 individuals had multiple samples. Distances between samples within an individual shown at 1 month, 2 months, etc out to over 1 year; between subject distances shown in "BSD." Even at one year, the median distance between a participant's samples is less than the median between participant distance. **(B)** Within the US, spatial processes of sOTUs appear driven by stochastic processes as few sOTUs exhibit spatial autocorrelation (Moran's *I*) on the full dataset or partitions (e.g., participants older than 20). **(C)**

Distance-decay relationship for Bray-Curtis dissimilarities between subject pairs that are within 100km (great-circle distance) radius of one another (Mantel test; r=0.036, adjusted p=0.03). Inset shows the largest radius (i.e., the contiguous US). Darker colors indicate higher-frequency bins. Dashed lines represent fits from linear models to raw data. **(D)** Mantel correlogram of estimated r coefficients, significance of distance-decay relationships, and radius (x-axis). Red points represent neighborhood sizes that were significant (adjusted p-values < 0.05). **(E)** Characterizing a large bowel resection using the AGP, the EMP, a hunter-gatherer population, and ICU patients in an unweighted UniFrac principal coordinates plot. A state change was observed in the resulting microbial community. The change in the microbial community immediately following surgery is the same as the distance between a marine sediment sample and a plant rhizosphere sample.

Figure 5. Diversity of plants in a diet. **(A)** Procrustes analysis of fecal samples from (n=1,596) individuals using Principal Components of the Vioscreen FFQ responses and Principal Coordinates of the unweighted UniFrac distances (M^2=0.988) colored by diet; Procrustes tests the fit of one ordination space to another. PCA shows grouping by diets such as Vegan suggesting self-reported diet type is consistent with differences in micro and macro nutrients as recorded by the FFQ, however these dietary differences do not explain relationships between the samples in 16S space. **(B)** The full AGP dataset including skin and oral samples through unweighted UniFrac and Principal Coordinates Analysis highlighting a lack of apparent clustering by diet type. **(C)** Dietary conjugated linoleic acid levels as reported by the FFQ between the extremes of plant diversity consumption, and **(D)** the observed levels of CLA by HPLC-MS. **(E)** Differential abundances of sOTUs (showing the most specific taxon name per

sOTU) between those who eat fewer than 10 plants per week vs. those who eat over 30 per week. **(F)** The molecules, linoleic acid (LA) and conjugated linoleic acid (CLA) (only trans-, trans-isomers are shown) were found to comprise the octadecadienoic acid found to be the key feature in this difference in number of plants consumption.

Figure 6. Molecular novelty in the gut microbiome. **(A-I)** Molecular sub-network of N-acyl amides. Cluster/nodes of microbially-derived G protein-coupled receptor agonistic molecules detected in human fecal samples are shown. Molecules B, G and H have been described (compounds 1, 2 & 4b (38) and commendamide (123)); molecules A, C, D, E and I are previously not reported (proposed structures are shown). **(J)** Compound occurrence frequency plot. Examples of compounds originating from food (piperine, black pepper alkaloid), host (stercobilin, heme catabolism product), bacterial activity (lithocholic acid, microbially-modified bile acid) or exogenous compounds such as antibiotics (rifaximin) or other drugs (lisinopril, high blood pressure medication) are shown. **(K-N)** Alpha and beta-diversity assessments of antibiotic and plants cohorts; insets depict minimum observed beta-diversity over increasing samples.

Supplementary Text:

Effect size comparisons

Multi-cohort replication detail

Projects using the American Gut infrastructure

American Gut Survey

Supplemental references

Supplementary Figures:

Figure S1. Workflow and population scale analyses. **(A)** Heatmap of income levels from the US Census and American Gut participant locations. **(B)** Sample flowchart for what sample sets correspond to each analysis. **(C)** Using PLS-DA we observed separation between US (*n*=6,634) and UK (*n*=2,071) fecal samples. **(D)** We performed a Principal Coordinates analysis comparing children over the age of 3 and adults from industrialized (*n*=4,643 AGP samples, *n*=4,927 samples total), remote farming (*n*=131), and hunter-gatherer (*n*=30) lifestyles.

Figure S2. Trading cards and LS's samples compared to ICU patients and AGP participants and diet state change analysis. **(A)** Unweighted UniFrac distance distributions for the sample immediate prior to surgery vs. all ICU fecal samples, and distances of the sample immediately following surgery vs. all ICU fecal samples (Kruskal Wallis H=79.774, p=4.198x^{-19}). **(B)** Same as panel **(A)** except comparing against all AGP fecal samples (Kruskal Wallis H=8117.734, p=0.0). **(C)** The median distances of each sample in Larry's longitudinal dataset compared to both ICU and AGP. The last pre-surgery sample is on day 25 and the first post-surgery sample is day 27. **(D)** A principal coordinates analysis of UniFrac distances of the American Gut Project, samples from the "extreme" diet study by David et al. (85), and the Earth Microbiome Project. No obvious state change by the diet of the participants in David et al. is observed.

Figure S3. Dietary levels of linoleic acid based on validated food frequency questionnaire responses, and the detected linoleic acid by mass spectrometry did not differ significantly between groups consuming few or many types of plants per week.

Figure S4. Metabolomic identification and annotation. **(A)** Manual annotation via comparison of experimental MS fragmentation patterns to those given in (99). Top panel: reference spectrum for the "Compound 2" in (99); bottom panel: experimental MS/MS spectrum for the parent ion m/z 611.5357. The compound is annotated as 3-(myristoyloxy)palmitoyl lysine. **(B)** I*n silico* annotation using CSI:FingerID (100) for the ion with m/z 330.2640. Top panel: experimental fragmentation pattern explained by the putative fragmentation tree; bottom panel: the possible candidate structures ranked by match %. The top structure with 71.02% match corresponds to commendamide. **(C)** Manual annotation via comparison of experimental exact mass to that of identified compound in (100), N-3-OH-palmitoyl ornithine. The peaks in experimental MS/MS spectrum are examined and compared to theoretical fragments that would result from breaking bonds in the proposed structure. The structure is deemed to be consistent with the N-3-OH-palmitoyl ornithine annotation.

Supplemental Tables:

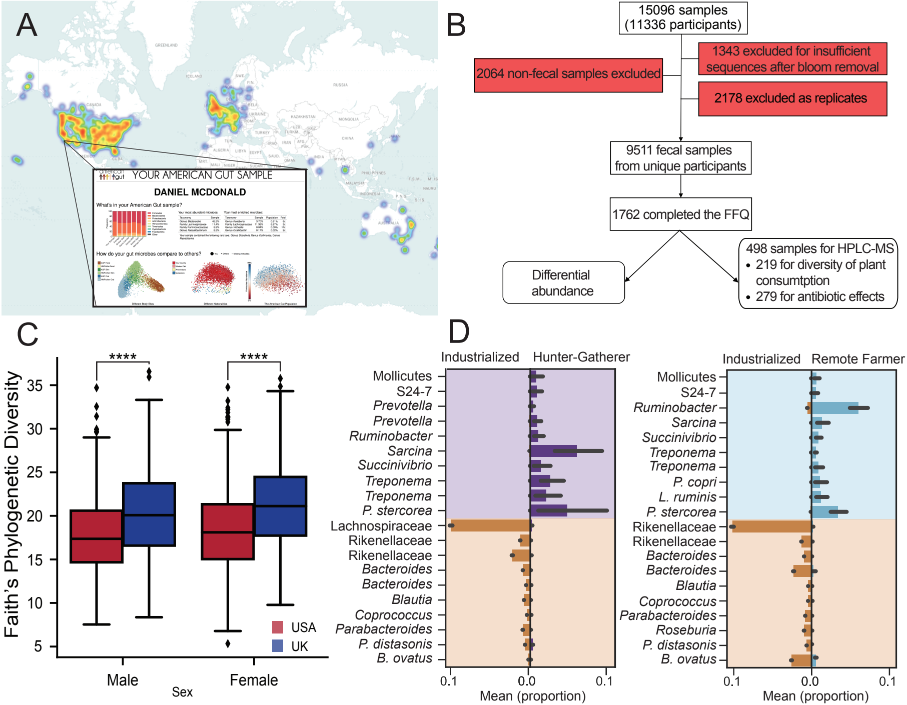
Table S1. Summary of sample numbers and type in the American Gut other studies, sample distributions by country and territory, sample distributions by US state, US participant demographics and per sequencing round sample accessions in EBI.

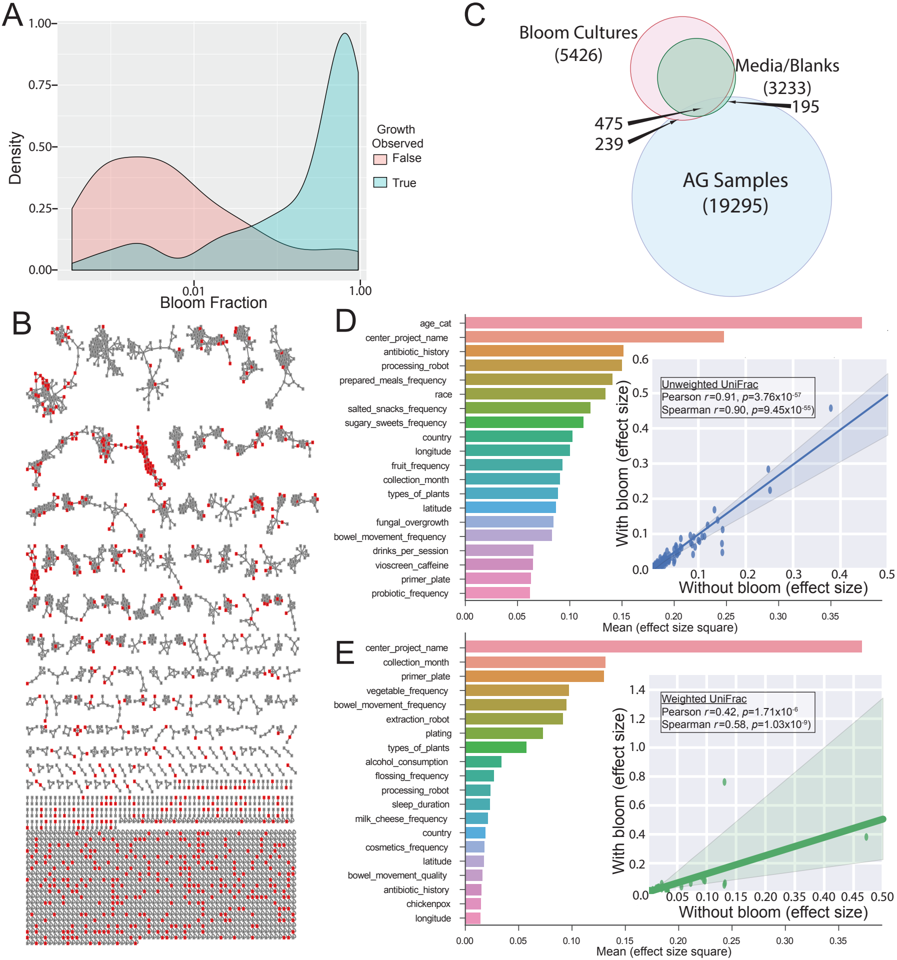
Table S2. American Gut data dictionary, proportion of responses per AG survey question that are represented as a single question; multiselect responses were omitted as these are stored in the metadata as per response type, informal dietary questions and correlations to the food frequency questionnaire, effect size results without bloom sOTUs, variable mapping with Falony et al. 2016 Science.

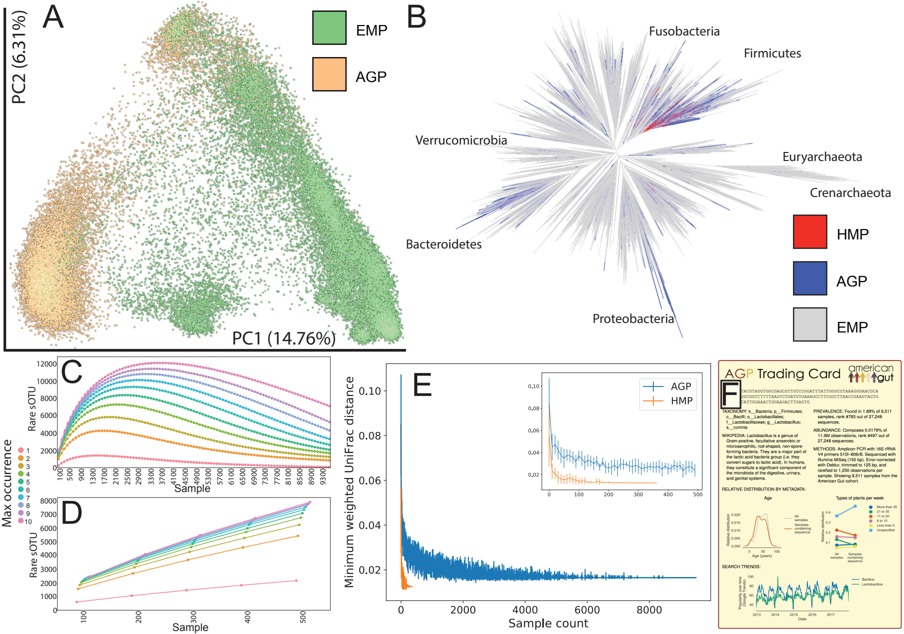
Table S3. sOTUs relevant to the balance analyses, and summary of differentially abundant taxa in UK cohort (negative effect size indicated the taxon is more prevalent in control (NC) subjects).

Table S4. Application of the filter for blooms to other human fecal studies which were not subjected to room temperature shipping, taxonomy of the draft isolate genomes, the specific bloom 16S sOTUs observed, and ubiquitous colibactin-like biosynthetic gene clusters (top) and a unique surfactin-like biosynthetic gene cluster observed in the bloom isolates.

Table S5. A set of molecular features which appeared to significantly correlate to the bloom fraction, and Kruskal–Wallis tests for metabolites in the Antibiotics and Vioscreen cohorts of samples.


A
YOUR AMERICAN GUT SAMPLE
DANIEL MCDONALD
What's in your American Gut sample?
How do your gut microbes compare to others?
B
15096 samples (11336 participants)
2064 non-fecal samples excluded
1343 excluded for insufficient sequences after bloom removal
2178 excluded as replicates
9511 fecal samples from unique participants
1762 completed the FFQ
Differential abundance
498 samples for HPLC-MS
• 219 for diversity of plant consumtption
• 279 for antibiotic effects
C
Faith's Phylogenetic Diversity
35
30
25
20
15
10
5
****
****
Male
Female
Sex
USA
UK
D
Industrialized
Hunter-Gatherer
Mollicutes
S24-7
Prevotella
Prevotella
Ruminobacter
Sarcina
Succinivibrio
Treponema
Treponema
P. stercorea
Lachnospiraceae
Rikenellaceae
Rikenellaceae
Bacteroides
Bacteroides
Blautia
Coprococcus
Parabacteroides
P. distasonis
B. ovatus
0.1
0.0
0.1
Mean (proportion)
Industrialized
Remote Farmer
Mollicutes
S24-7
Ruminobacter
Sarcina
Succinivibrio
Treponema
Treponema
P. copri
L. ruminis
P. stercorea
Rikenellaceae
Rikenellaceae
Bacteroides
Bacteroides
Blautia
Coprococcus
Parabacteroides
Roseburia
P. distasonis
B. ovatus
0.1
0.0
0.1
Mean (proportion)


A
Density
1.00
0.75
0.50
0.25
0.00
0.01
1.00
Bloom Fraction
Growth Observed
False
True
C
Bloom Cultures (5426)
Media/Blanks (3233)
195
475
239
AG Samples (19295)
B
D
age_cat
center_project_name
antibiotic_history
processing_robot
prepared_meals_frequency
race
salted_snacks_frequency
sugary_sweets_frequency
country
longitude
fruit_frequency
collection_month
types_of_plants
latitude
fungal_overgrowth
bowel_movement_frequency
drinks_per_session
vioscreen_caffeine
primer_plate
probiotic_frequency
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Mean (effect size square)
Unweighted UniFrac
Pearson r=0.91, p=3.76x10^{-57}
Spearman r=0.90, p=9.45x10^{-55})
With bloom (effect size)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
Without bloom (effect size)
E
center_project_name
collection_month
primer_plate
vegetable_frequency
bowel_movement_frequency
extraction_robot
plating
types_of_plants
alcohol_consumption
flossing_frequency
processing_robot
sleep_duration
milk_cheese_frequency
country
cosmetics_frequency
latitude
bowel_movement_quality
antibiotic_history
chickenpox
longitude
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Mean (effect size square)
Weighted UniFrac
Pearson r=0.42, p=1.71x10^{-6}
Spearman r=0.58, p=1.03x10^{-9})
With bloom (effect size)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
Without bloom (effect size)



AGP Trading Card american gut

TACGTAGGTGGCGAGCGTTGTCCGGATTTATTGGGCGTAAAGGGAACGCA
GGCGGTCTTTTAAGTCTGATGTGAAAGCCTTCGGCTTAACCGAAGTAGTG
CATTGGAAACTGGAAGACTTGAGTG

TAXONOMY: k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g__Lactobacillus; s__ruminis

WIKIPEDIA: Lactobacillus is a genus of Gram-positive, facultative anaerobic or microaerophilic, rod-shaped, non-spore-forming bacteria. They are a major part of the lactic acid bacteria group (i.e. they convert sugars to lactic acid). In humans, they constitute a significant component of the microbiota of the digestive, urinary, and genital systems.

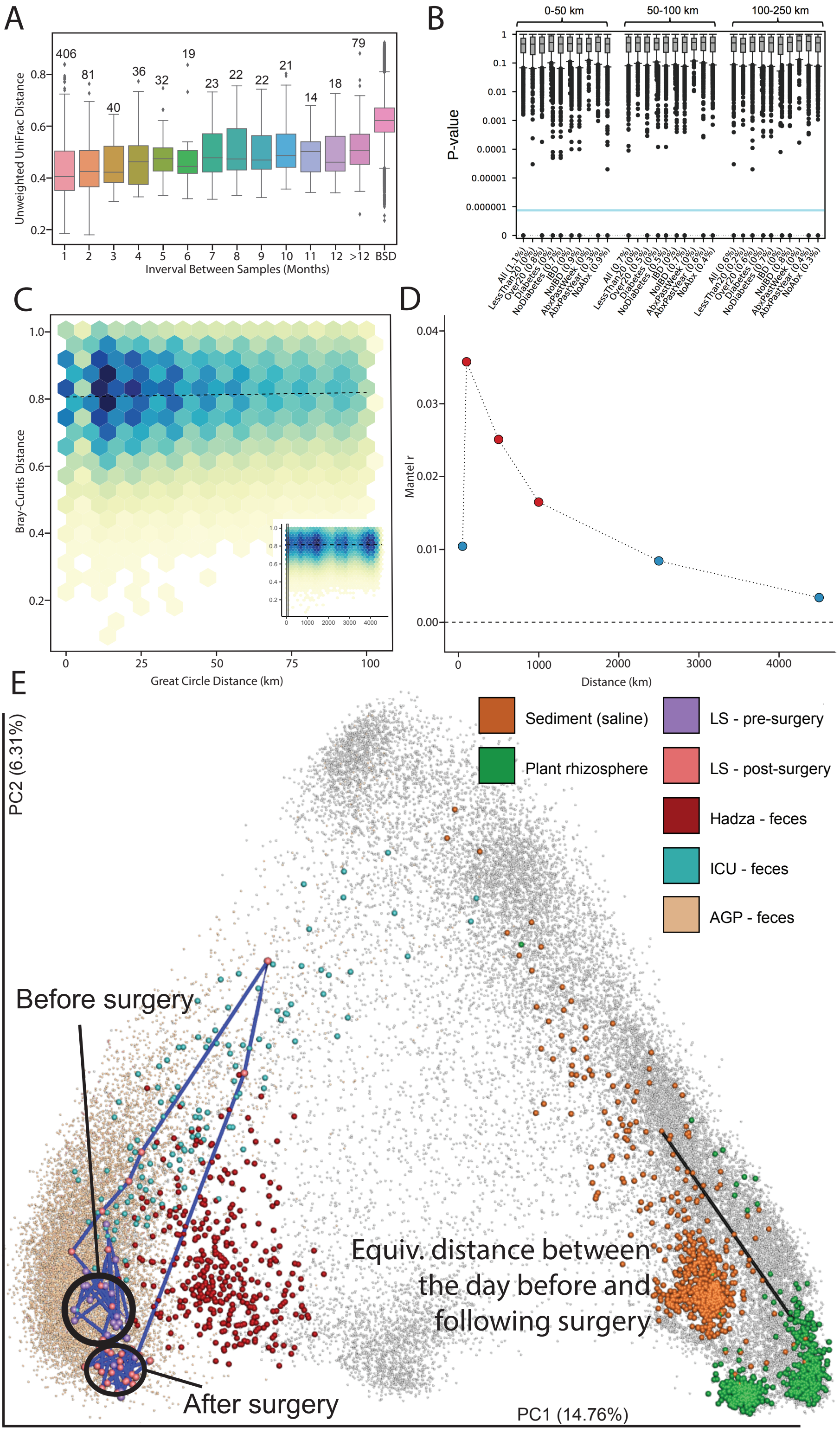
PREVALENCE: Found in 1.89% of 9,511 samples, rank #783 out of 27,248 sequences.

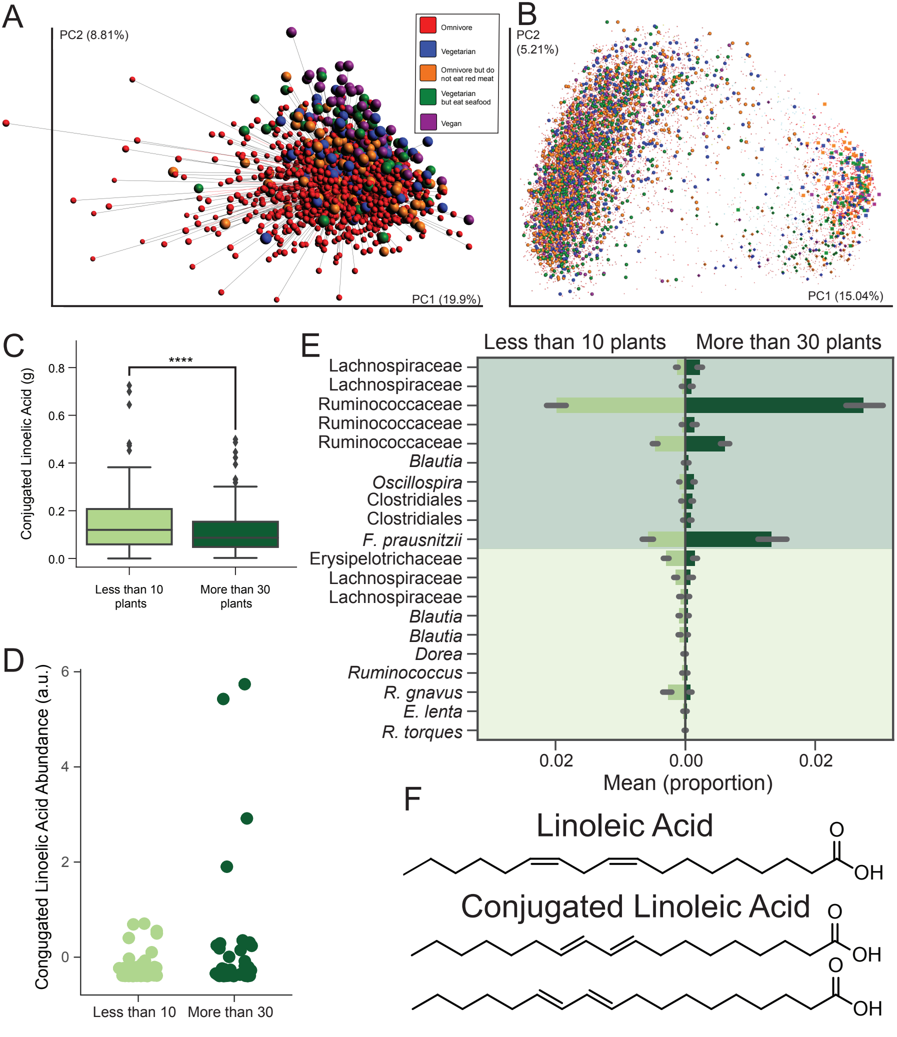
ABUNDANCE: Composes 0.0179% of 11.9M observations, rank #497 out of 27,248 sequences.

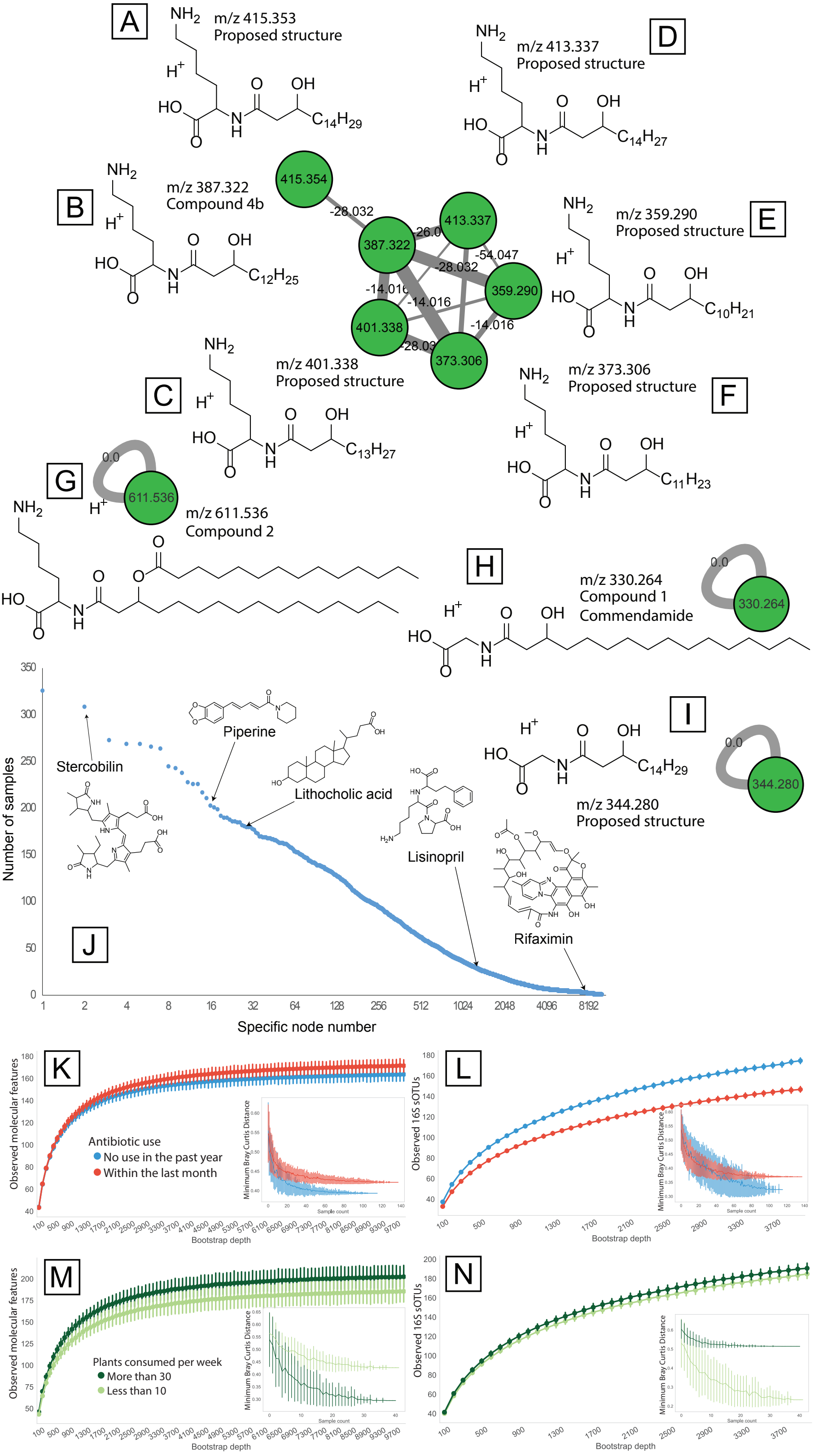
METHODS: Amplicon PCR with 16S rRNA V4 primers 515f–806rB. Sequenced with Illumina MiSeq (150 bp). Error-corrected with Deblur, trimmed to 125 bp, and rarefied to 1,250 observations per sample. Showing 9,511 samples from the American Gut cohort.

RELATIVE DISTRIBUTION BY METADATA:

SEARCH TRENDS:


A
Unweighted UniFrac Distance
0.2
0.4
0.6
0.8
406
81
40
36
32
19
23
22
22
21
14
18
79
1
2
3
4
5
6
7
8
9
10
11
12
>12
BSD
Inverval Between Samples (Months)
B
0-50 km
50-100 km
100-250 km
P-value
1
0.1
0.01
0.001
0.0001
0.00001
0.000001
0
All (1.1%)
LessThan20 (0%)
Over20 (0.8%)
Diabetes (0%)
NoDiabetes (0.7%)
IBD (0%)
NoIBD (0.9%)
AbxPastWeek (0%)
AbxPastYear (0.3%)
NoAbx (0.9%)
All (0.7%)
LessThan20 (0%)
Over20 (0.5%)
Diabetes (0%)
NoDiabetes (0.5%)
IBD (0%)
NoIBD (0.7%)
AbxPastWeek (0%)
AbxPastYear (0.6%)
NoAbx (0.4%)
All (0.6%)
LessThan20 (0.2%)
Over20 (0.6%)
Diabetes (0%)
NoDiabetes (0.7%)
IBD (0%)
NoIBD (0.8%)
AbxPastWeek (0%)
AbxPastYear (0.4%)
NoAbx (0.3%)
C
Bray-Curtis Distance
1.0
0.8
0.6
0.4
0.2
0
25
50
75
100
Great Circle Distance (km)
1.0
0.8
0.6
0.4
0.2
0
1000
2000
3000
4000
D
Mantel r
0.04
0.03
0.02
0.01
0.00
0
1000
2000
3000
4000
Distance (km)
E
PC2 (6.31%)
PC1 (14.76%)
Sediment (saline)
Plant rhizosphere
LS - pre-surgery
LS - post-surgery
Hadza - feces
ICU - feces
AGP - feces
Before surgery
After surgery
Equiv. distance between the day before and following surgery


A
PC2 (8.81%)
PC1 (19.9%)
🟥 Omnivore
🟦 Vegetarian
🟧 Omnivore but do not eat red meat
🟩 Vegetarian but eat seafood
🟪 Vegan
B
PC2 (5.21%)
PC1 (15.04%)
C
****
Conjugated Linoelic Acid (g)
0.8
0.6
0.4
0.2
0.0
Less than 10 plants
More than 30 plants
D
Congugated Linoelic Acid Abundance (a.u.)
6
4
2
0
Less than 10
More than 30
E
Less than 10 plants
More than 30 plants
Lachnospiraceae
Lachnospiraceae
Ruminococcaceae
Ruminococcaceae
Ruminococcaceae
Blautia
Oscillospira
Clostridiales
Clostridiales
F. prausnitzii
Erysipelotrichaceae
Lachnospiraceae
Lachnospiraceae
Blautia
Blautia
Dorea
Ruminococcus
R. gnavus
E. lenta
R. torques
0.02
0.00
0.02
Mean (proportion)
F
Linoleic Acid
O
OH
Conjugated Linoleic Acid
O
OH
O
OH



A
m/z 415.353
Proposed structure
NH_2
H^+
HO
O
N
H
O
OH
$C_{14}H_{29}$
D
m/z 413.337
Proposed structure
$C_{14}H_{27}$
B
m/z 387.322
Compound 4b
$C_{12}H_{25}$
E
m/z 359.290
Proposed structure
$C_{10}H_{21}$
C
m/z 401.338
Proposed structure
$C_{13}H_{27}$
F
m/z 373.306
Proposed structure
$C_{11}H_{23}$
415.354
413.337
387.322
359.290
401.338
373.306
-28.032
-26.0
-54.047
-28.032
-14.016
-14.016
-14.016
-28.03
G
0.0
611.536
m/z 611.536
Compound 2
H
m/z 330.264
Compound 1
Commendamide
0.0
330.264
I
0.0
344.280
m/z 344.280
Proposed structure
$C_{14}H_{29}$
J
Number of samples
Specific node number
0
50
100
150
200
250
300
350
1
2
4
8
16
32
64
128
256
512
1024
2048
4096
8192
Stercobilin
Piperine
Lithocholic acid
Lisinopril
Rifaximin
K
Observed molecular features
Bootstrap depth
Antibiotic use
● No use in the past year
● Within the last month
Minimum Bray Curtis Distance
Sample count
L
Observed 16S sOTUs
Bootstrap depth
Minimum Bray Curtis Distance
Sample count
M
Observed molecular features
Bootstrap depth
Plants consumed per week
● More than 30
● Less than 10
Minimum Bray Curtis Distance
Sample count
N
Observed 16S sOTUs
Bootstrap depth
Minimum Bray Curtis Distance
Sample count