RESEARCH ARTICLE

Sewage Reflects the Microbiomes of Human Populations

Ryan J. Newton,[a] Sandra L. McLellan,[a] Deborah K. Dila,[a] Joseph H. Vineis,[b] Hilary G. Morrison,[b] A. Murat Eren,[b] Mitchell L. Sogin[b]

School of Freshwater Sciences, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin, USA[a]; Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, Massachusetts, USA[b]

ABSTRACT **Molecular characterizations of the gut microbiome from individual human stool samples have identified community patterns that correlate with age, disease, diet, and other human characteristics, but resources for marker gene studies that consider microbiome trends among human populations scale with the number of individuals sampled from each population. As an alternative strategy for sampling populations, we examined whether sewage accurately reflects the microbial community of a mixture of stool samples. We used oligotyping of high-throughput 16S rRNA gene sequence data to compare the bacterial distribution in a stool data set to a sewage influent data set from 71 U.S. cities. On average, only 15% of sewage sample sequence reads were attributed to human fecal origin, but sewage recaptured most (97%) human fecal oligotypes. The most common oligotypes in stool matched the most common and abundant in sewage. After informatically separating sequences of human fecal origin, sewage samples exhibited ~3× greater diversity than stool samples. Comparisons among municipal sewage communities revealed the ubiquitous and abundant occurrence of 27 human fecal oligotypes, representing an apparent core set of organisms in U.S. populations. The fecal community variability among U.S. populations was significantly lower than among individuals. It clustered into three primary community structures distinguished by oligotypes from either: *Bacteroidaceae*, *Prevotellaceae*, or *Lachnospiraceae/Ruminococcaceae*. These distribution patterns reflected human population variation and predicted whether samples represented lean or obese populations with 81 to 89% accuracy. Our findings demonstrate that sewage represents the fecal microbial community of human populations and captures population-level traits of the human microbiome.**

IMPORTANCE **The gut microbiota serves important functions in healthy humans. Numerous projects aim to define a healthy gut microbiome and its association with health states. However, financial considerations and privacy concerns limit the number of individuals who can be screened. By analyzing sewage from 71 cities, we demonstrate that geographically distributed U.S. populations share a small set of bacteria whose members represent various common community states within U.S. adults. Cities were differentiated by their sewage bacterial communities, and the community structures were good predictors of a city's estimated level of obesity. Our approach demonstrates the use of sewage as a means to sample the fecal microbiota from millions of people and its potential to elucidate microbiome patterns associated with human demographics.**

Received 29 December 2014 **Accepted** 5 January 2015 **Published** 24 February 2015

Citation Newton RJ, McLellan SL, Dila DK, Vineis JH, Morrison HG, Eren AM, Sogin ML. 2015. Sewage reflects the microbiomes of human populations. mBio 6(2):e02574-14. doi: 10.1128/mBio.02574-14.

Editor Gary B. Huffnagle, University of Michigan Medical School

Address correspondence to Mitchell L. Sogin, mitchellsogin@gmail.com.

This article is a direct contribution from a Fellow of the American Academy of Microbiology.

The human fecal microbial community serves as a proxy for the human gut community, which exhibits considerable diversity and variability among individuals (1–3). Human microbiome data sets show that most human gut communities share specific functional gene profiles (4–6) rather than a single core set of microbial species (3, 7). Functional similarity coupled with taxonomic variability indicates niche overlap among taxa, which in the gut microbiome of healthy individuals reflects taxonomically distinct sets of cooccurring taxa or enterotypes (8) that contain similar functional gene profiles. Despite interindividual taxonomic variability, studies that include multiple samples have identified correlations between functional gene composition and taxonomic composition (5, 9, 10) and relationships between human characteristics and the gut microbial composition. For example, the gut community from an individual is more similar to itself through time than to samples collected from other individuals (7, 11, 12). Marked shifts in the gut microbial communities are also reported for the very young and very old (9, 12–14), for healthy versus disease states (15–17), across different diet regimes (6, 10, 18, 19), and in culturally isolated human populations (9, 20). The coherence of the gut microbial community within individuals and among individuals with specific characteristics suggests that gut communities maintain relatively stable equilibrium states (7). If the gut community composition tracks human characteristics, then identifying the community members or community states that differ across human population boundaries could lead to an improved understanding of how these communities influence human health.

Sampling individuals has proven to be an effective approach for identifying gut microbial community patterns that associate with human health states. However, large variation among gut microbiomes and the expense of sequencing libraries from many

TABLE 1 Sequence data set statistics

Dataset[a]	% of total sequences in sample type: Human stool	Sewage
15 most abundant bacterial families	98.0	26.1
6 most abundant bacterial families	90.8	21.6
6 most abundant bacterial families, following oligotyping sequence exclusion	78.3	18.7
6 most abundant families, oligotype data set, following exclusion of nonfecal oligotypes	78.3	11.7
Total subsampled sequence count	821,476 ($n = 137$)	11,302,794 ($n = 207$)

[a] The most abundant bacterial families are ordered according to the mean number of sequences among human stool samples ($n = 137$).

individuals limit the efficacy of microbial community comparisons from human populations over different demographic scales, e.g., city, country, or continent. Previously, we demonstrated that highly prevalent *Lachnospiraceae* organisms in a human fecal data set were the most abundant in a sewage influent data set (21) and that a single sewage sample harbored most of the *Blautia* sequence diversity identified in 10 human fecal samples (22). We hypothesize that comparison of untreated sewage samples might provide a means to assess the human fecal microbiome and by proxy the gut microbiome within and among human populations. Here, we systematically compare bacterial 16S rRNA gene profiles from healthy adult stool samples generated by the Human Microbiome Project (5) to the community profiles of >200 sewage influent samples collected from 71 U.S. cities. We used oligotyping (23), a computational method that uses positional Shannon entropy scores to decompose sequencing data into highly refined sequence-based units that make possible sensitive assessments of beta diversity. From these data, we asked (i) whether sewage influent accurately reflects a composite fecal microbiome from human populations, (ii) if "core" fecal organisms or other community trends exist across U.S. cities, and (iii) whether sewage influent microbial communities correlate with human demographic patterns.

RESULTS

Sewage influent accurately reflects a composite human stool bacterial community. The 15 most abundant bacterial families, which on average accounted for 98% of the reads in the human stool data set, represented 26% of the sequence reads in a sewage sample (Table 1). The low representation of human fecal bacteria in sewage concurs with previous reports that 80 to 90% of bacterial sequences in sewage originate from non-human-fecal sources (24, 25). Since sewage contains many organisms of nonfecal origin, our analysis focused on sequences from bacterial families that each represented >3% of total reads in the human stool data set: *Bacteroidaceae*, *Ruminococcaceae*, *Lachnospiraceae*, *Porphyromonadaceae*, *Rikenellaceae*, and *Prevotellaceae*. Five of these represented the most abundant of the human fecal matter-associated families in sewage, with the sixth, *Rikenellaceae*, contributing less than *Veillonellaceae* (Fig. 1A). Normalization of the sewage data set to the 15 most abundant families in the human stool data set (Fig. 1A) showed comparable community compositions at the family level (linear $r^2 = 0.72$, $P < 0.001$) but with an underrepresentation of *Bacteroidaceae* (Wilcoxon rank sum test, $P < 0.001$) and an overrepresentation of *Lachnospiraceae* and *Prevotellaceae* (Wilcoxon rank sum test, $P < 0.001$).

Analysis of reads that map to the human stool sample's six most abundant families identified 351 oligotypes across all human stool and sewage samples. BLAST analyses assigned 105 oligotypes to the nonfecal data set (i.e., a set of oligotypes representing 16S rRNA genes not detected in human fecal samples; see Materials and Methods for classification details). The remaining 246 oligotypes comprised the final human fecal data set and represented on average 78% of all sequences in a human stool sample and 12% of sequences in a sewage sample (Table 1). Seven of the 246 human fecal oligotypes were detected only in the human stool samples, and nine were detected only in the sewage samples.

After filtering out the nonfecal data set sequences, oligotyping revealed an overrepresentation of *Lachnospiraceae* and *Prevotellaceae* while *Bacteroidaceae* remained underrepresented in sewage compared to the human stool data set (Fig. 1B). Although there existed a broad family-level composition shift between the data

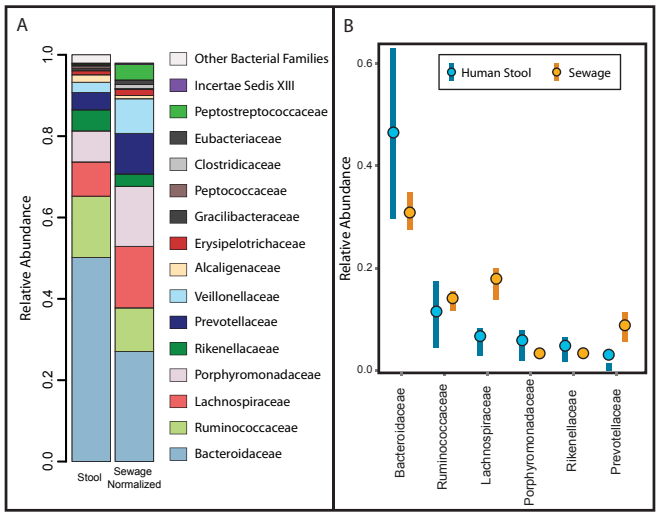

FIG 1 (A) Bacterial family taxon assignments for the 15 most abundant bacterial families in the pooled human stool data set and the pooled sewage data set. The normalized sequence counts in the sewage data set represent the total proportion of sequences (98.0%) from the 15 families in the pooled stool data set. (B) Box plot depicting the relative abundance of oligotypes classified into the six most abundant bacterial families in the human stool data set for both the pooled human stool and pooled sewage data sets after removing non-fecal-matter-associated oligotypes. The normalized sequence count in the sewage data set represents the total proportion of sequences (78.3%) assigned to those oligotypes. Circles represent sample mean values, and the line vertices represent first and third quartiles. Note that for human stool *Prevotellaceae*, the first and third quartiles do not intersect the mean.

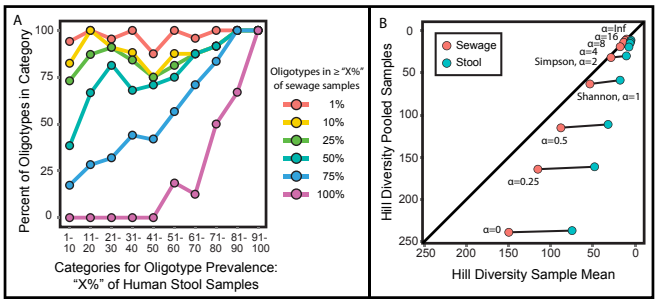

FIG 2 (A) Comparison of oligotype prevalence among the human stool samples (*x* axis) with the percentage of the oligotypes (*y* axis) that were also present at a specific prevalence level among the sewage samples (data series). Data are plotted as the percentage of oligotypes within a human stool prevalence category (e.g., 0 to 10%, 11 to 20%, etc.) that meet a specific prevalence requirement in sewage (1%, 10%, etc.). For example, 50% of human fecal oligotypes that were present in 71 to 80% of the stool samples were present in 100% of the sewage samples (see purple data series). (B) Comparison of Hill diversity values for multiple alpha parameters based on the human fecal oligotype community in the sewage data set versus the human stool data set. Higher alpha values place more weight on the most abundant organisms in the diversity calculation. Shannon and Simpson diversity indices are indicated on the plot. Sample mean diversity values are plotted on the *x* axis, and pooled diversity from all samples is plotted on the *y* axis. A one-to-one line is indicated, and lines connect equivalent alpha value results between the two data sets. For visualization, both axes are ordered from high to low diversity values.

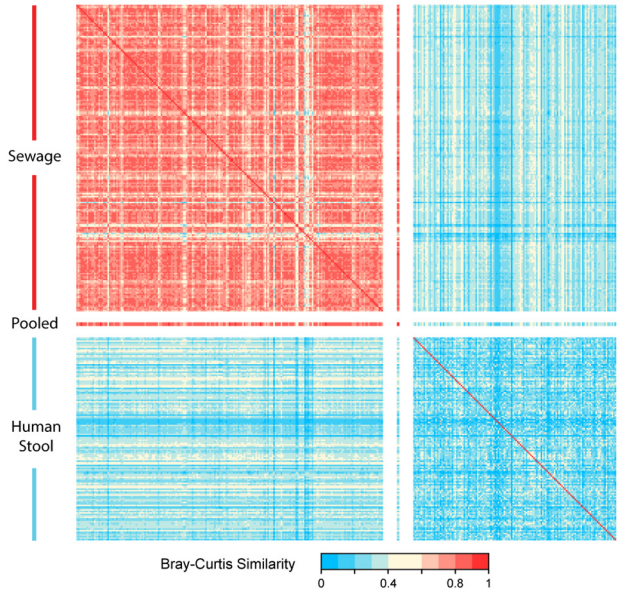

FIG 3 Heat map comparing the human fecal oligotype compositions (Bray-Curtis similarity) among samples for the sewage and human stool data sets. Comparisons for the pooled human stool data set versus all individual samples (labeled as "pooled" on the plot) are shown in the space between the sewage and human stool samples.

sets, the oligotype relative abundance profile of the pooled sewage data set resembled the pooled human stool data set (linear regression $r^2 = 0.66$, $P < 0.001$; see Fig. S1 in the supplemental material). Within families, oligotype relative abundance profiles correlated linearly ($P < 0.001$), although *Lachnospiraceae* had a much lower regression fit ($r^2 = 0.42$) than the other families (range from $r^2 = 0.76$ to 0.99; see Fig. S1).

Sewage exhibited the additive effect of combining microbial communities from many individuals into a single sample. Oligotypes that were common among individual stool samples were common among sewage samples (Fig. 2A) and typically more abundant (see Fig. S2 in the supplemental material). We also observed this additive effect in the alpha diversity measures for each data set. While the pooled human stool and pooled sewage data sets captured approximately equivalent oligotype richness ($n = 235$ for pooled stool and $n = 239$ for pooled sewage) and diversity (at several alpha values of the Hill diversity index [Fig. 2B]), the mean diversity for individual sewage samples exceeded values for individual stool samples (Wilcoxon rank sum test, $P < 0.001$ for all Hill diversity alpha values [Fig. 2B]). The high ratio for sample mean to pooled sample diversity in the sewage data set indicates that sewage had low oligotype turnover among samples relative to the low ratio (i.e., high turnover of oligotypes) among stool samples (Fig. 2B).

Homogeneity in the fecal microbial composition of human populations. Sewage samples from different U.S. cities displayed very similar human fecal oligotype compositions within and among sample periods (mean Bray-Curtis similarity = 73.2% [Fig. 3]). The oligotype composition of the pooled stool data set was strikingly similar to individual sewage samples (mean Bray-Curtis similarity = 81.0%). In contrast, the mean Bray-Curtis similarity for comparisons between individual stool samples hovered at 32.6% (Fig. 3). Hill diversity measures at alpha values of ≥1, which emphasize the contribution of more abundant taxa, indicated that each sewage sample contained the majority of the most abundant oligotypes (Fig. 2B). For example, on average, two sewage samples captured 90% of the measured fecal oligotype diversity in the sewage data set, whereas the same diversity levels required data from 71 human stool samples (see Fig. S3 in the supplemental material).

Core fecal microorganisms in the United States. No oligotype occurred in all human stool samples. The two most prevalent oligotypes were detected in 129/137 stool samples (see Table S2 in the supplemental material). In contrast, 17 oligotypes were present in all 207 sewage samples and 10 others were present in ≥205 (≥99%) of sewage samples. These 27 oligotypes also represented the most abundant amplicon sequences in the sewage samples (see Fig. S4). The 27 common and abundant oligotypes represent "core" gut microbiota among human populations in the United States. In the human stool data set, one of the core oligotypes represented the most abundant oligotype in 117 of the 137 samples. Twenty-six of the 27 core oligotypes derive from amplicon sequences that match exactly a cultivar 16S rRNA gene, and all but three oligotypes matched strains from only a single genus (see Table S2). In nearly all sewage samples (200/207), the most abundant oligotype resolved to either a *Prevotellaceae*, one of two *Bacteroidaceae*, or a *Ruminococcaceae* oligotype. These correspond to sequences that identically matched *Prevotella* sp. strain BI-42, *Bacteroides dorei* JCM 13471, *Bacteroides vulgatus* ATCC 8482, and *Faecalibacterium prausnitzii* A2-165, respectively.

Drivers of sewage community differences among cities. The human fecal oligotype composition in sewage reflected increased representation of oligotypes from one family group over the others (*Bacteroidaceae* versus *Prevotellaceae* versus *Lachnospiraceae* plus *Ruminococcaceae* [see Fig. S5 in the supplemental material]; Adonis $R^2 = 0.30$, $P < 0.01$). Of 51 treatment plant sites that had data for all three sample collections, 21 exhibited the same enrich-

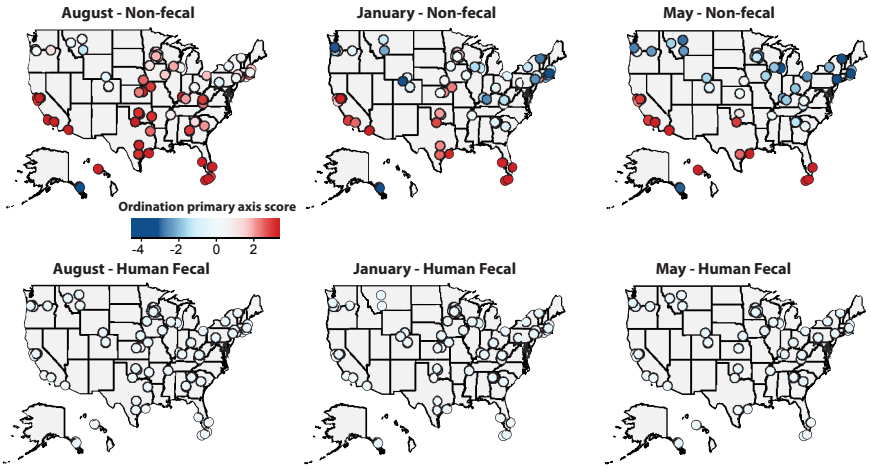

FIG 4 The primary axis scores of each sample from a constrained ordination of principal coordinates (CAP) for the temperature profile of each city are indicated via color coding on a U.S. map. Scores to the left on the ordination are plotted in shades of blue, and scores to the right on the ordination are plotted in shades of red. Samples that are ≥2 standard deviations from the primary axis origin are colored the maximum blue and red colors. Both the human fecal oligotype (bottom) and non-human-fecal oligotype (top) data sets were included in the CAP, and therefore, colors represent equivalent community variation related to the temperature profile of the cities for each data set. For comparison, sample periods are depicted in separate maps.

ment patterns, a result which exceeds random expectations (binomial $P \leq 0.01$). Cities with more than one treatment plant exhibited a higher level of consistency for paired sample comparisons between plants (14/19 paired samples were enriched for the same family; binomial $P \leq 0.01$). Principal coordinate analyses (PCoAs) did not indicate significant ($P \leq 0.01$) relationships between the geographic location of treatment plants, the plant chemical/physical measurements, or the population size served and the human fecal oligotype composition (see Table S3 for PCoA correlates). Only sample period ($r^2 = 0.09$) and city population percent obesity ($r^2 = 0.06$) explained significantly ($P < 0.01$) the variation in composition among U.S. cities (see Table S3 for PCoA correlates).

In contrast to the human fecal oligotypes, the nonfecal oligotype data set exhibited greater variation among cities and sample periods, with strong geographic and seasonal trends corresponding to air temperature and latitude differences. PCoAs revealed that the temperature profile of a city explained a significant ($P < 0.001$) proportion of the nonfecal community variation: August ($r^2 = 0.29$), January ($r^2 = 0.36$), and May ($r^2 = 0.52$). A constrained ordination for the city temperature profile parameter further illustrated that the nonfecal oligotype composition was more strongly related to a city's yearly temperature profile than the human fecal oligotype composition, and this relationship in the nonfecal community appeared as a significant divide between the northern and southern U.S. cities (Fig. 4).

Human demographics represented in sewage microbial communities. Although no measured factors explained a high percentage of the human fecal community variation among cities, the percent obesity in a city's population had explanatory power: PCoA obesity percent, $r^2 = 0.06$ in August, $r^2 = 0.08$ in January, and $r^2 = 0.11$ in May (see Table S3 in the supplemental material). A random forest classification algorithm demonstrated that human fecal oligotype composition in sewage predicted whether a sample derived from a lean or obese population with 81 to 89% accuracy (Table 2). This relationship was driven partly by an increase in the relative abundance of *Bacteroidaceae* oligotypes in samples from the most obese city populations [see Fig. S6; Wilcoxon rank sum test for relative abundance in samples categorized as lean versus obese, $W(n_1 = 51\ n_2 = 54) = 806$, $P < 0.001$], as represented by an increased ratio of two core *Bacteroidaceae* oligotypes versus two core *Ruminococcaceae* oligotypes (see Fig. S6).

DISCUSSION

Large populations with highly variable phenotypic characters (e.g., human weight and flower color) will include a greater number of variants with more even distributions than small populations. A character variant that is common among individuals in these populations will be abundant in population-level assessments and highly prevalent among populations. However, unlike weight or flower color, where a single variant represents each individual at a given moment in time, microbiomes encompass

TABLE 2 Random forest classification statistics for the sewage bacterial community composition as a predictor of obesity levels in city populations

	Classification		
	No. correct/total no.		
Data set[a]	Lean	Obese	Accuracy (% correct)
All samples,[b] quartiles[c]	44/54	45/54	82
All samples, SD[d]	26/38	44/46	83
Cities,[e] quartiles	16/21	18/21	81
Cities, SD	14/17	17/18	89

[a] City populations were classified as "lean" or "obese" based on the estimated percentage of obese people in each city.
[b] All samples for a city were included in the model and classified separately.
[c] Samples in the first (lean) and fourth (obese) quartiles for the distribution of city obesity percentages in the random forest classification model. A "lean city" versus "obese city" designation corresponds to populations with ≤22.8% obesity versus ≥30.4% obesity, respectively.
[d] Samples >1 standard deviation from the mean city obesity percentage in the random forest classification model. A city was considered to be lean at populations with ≤21.5% obesity and obese at populations with ≥31.3% obesity.
[e] Average bacterial community composition in all samples for a city.

hundreds to thousands of different kinds of microorganisms or operational taxonomic units (OTUs) that collectively define the character variant of an individual. Although this complexity confounds identification of community states (i.e., variants) that differentiate between individuals (7), it should not affect the expected distribution of community members in a population-level sample.

In support of this concept, we found that (i) the population-level (sewage) samples recaptured the majority (97%) of the oligotypes from individual stool samples, (ii) a pooled data set of human stool and sewage samples exhibited highly similar oligotype distribution patterns, (iii) sewage samples had higher richness and diversity than stool samples, and (iv) oligotypes that were more prevalent among individuals were more prevalent and more abundant in sewage. We infer that sewage influent represents the composite fecal microbiomes of many individuals and provides a metric to assess the relationship of these population-level microbial distributions with large-scale patterns in human demographics.

The complex environment of municipal sewer systems receives water and the associated microbial mélange from multiple sources, including gray water, human stools, and in some systems surface runoff. In our data set, sequences that made up on average 78% of a stool sample comprised ~12% of a sewage sample. By scaling this ratio to 100%, we estimate that only 15% of the amplicons in a typical sewage sample originate from human stool. Analyses restricted to oligotypes from nonfecal community members exhibited strong community composition relationships to geography-related differences among cities (Fig. 4). Without the human microbiome data sets to focus our analysis on human stool sequences in sewage and oligotyping to differentiate closely related sequence variants, the nonfecal community distribution patterns would have overprinted the signal from the stool samples.

Previous comparisons of individual gut microbiomes demonstrate high community composition variability among individuals and that no single core set of bacterial species dominates all human guts (7, 26). By sampling sewage, we find that U.S. populations have a much less variable fecal bacterial community composition than that of individuals. This community composition convergence among populations suggests that a finite level of composition variability is present, at least among U.S. populations, and this variability can be overcome with large sample sizes to make meaningful inferences about the gut microbiome. From the sewage sampling, we also identified a set of "core" bacteria that are both common to and abundant in U.S. populations. Although no single species dominates the fecal microbial communities among individuals, our results demonstrate consistent differential abundance in human populations for some bacterial taxa over others. Previous attempts to classify core species, using a >50% occurrence among adult individuals as the definition of core, identified *Faecalibacterium prausnitzii*, *Roseburia intestinalis*, *Bacteroides vulgatus*, *Bacteroides uniformis*, *Eubacterium rectale*, and *Ruminococcus bromii* among other undescribed species as primary members (4, 26). Except for *Roseburia intestinalis*, each of these species matched one of our core oligotypes. We also defined another 21 oligotypes as core members, most of which resolved to various *Bacteroides* spp. or *Lachnospiraceae* genera (see Table S2 in the supplemental material). The high representation of *Bacteroides* in the sewage samples is consistent with reports that adults from the United States have higher abundances of the genus *Bacteroides* than do people from non-Westernized societies (9). Since the core oligotypes were present and abundant in nearly all U.S. sewage samples, we hypothesize that these organisms represent a signature for U.S. populations that can differentiate between human gut communities from other parts of the world.

Dominance of a bacterial species in the human population may reflect its functional importance in the metabolic capacity of human guts. Despite the wide taxonomic range of the core organisms identified here and the high bacterial community variability among individuals, the functional gene composition among human gut microbiomes is fairly consistent (4, 5). Niche overlap among various members of the gut community might explain this functional consistency without requiring nearly identical microbiome compositions (7). Typically, the most abundant oligotype in a stool sample was one of the 27 core oligotypes (117 of 137 samples). The ubiquity of the core organisms in human populations and frequent dominance in individuals make these core organisms strong candidates for exploring the functional trade-offs among prominent gut bacteria and the differences that relate to human health or define stable community states.

Sewage sampling also described distinct community compositions among U.S. populations. Samples differed primarily by the increased representation of oligotypes from *Bacteroidaceae*, *Prevotellaceae*, or *Lachnospiraceae*/*Ruminococcaceae* over the other two family groups (see Fig. S5 in the supplemental material). This result resembles earlier enterotype analyses (8) and the concept that changes in dominance between taxa in these families play an important role in structuring gut communities (7). Twenty-one of 51 cities were enriched for the same bacterial family group across all three sampling periods. Although not a majority, this level of community consistency signifies that human populations at the citywide scale can have characteristic microbial community compositions.

Although we did not identify the ultimate causes of the bacterial community composition differences among U.S. cities, our single measure of lifestyle differences for individuals in these cities (obesity percent) explained a significant, albeit small, proportion of the community variation. Lifestyle differences can reproducibly alter the human gut microbiome (27), and microbial community composition is a known indicator of obesity (28–30), with up to 90% predictive accuracy for individuals (31). We observed that the obesity signal in an individual's gut microbial community composition scaled up, with nearly equivalent predictive capabilities (81 to 89% accuracy), to the level of human populations in cities. These community composition relationships to the population obesity gradient were driven in large part by increased representation of *Bacteroides* spp. and decreased representation of *Faecalibacterium* spp. in more obese populations (see Fig. S6 in the supplemental material). *Bacteroides* spp. have been found to increase in abundance in humans consuming a high-animal-fat diet (19) and are associated with low-diversity proinflammatory gut communities, while *Faecalibacterium* spp. are more prevalent in high-diversity anti-inflammatory gut communities (32). Given the relatively minor difference in population obesity percentage (as low as 9%) between city populations considered lean and obese, the observed correlations between obesity and the microbial community in sewage might reflect other, more pronounced lifestyle differences in these cities, including the influence of diet on gut microbial communities (6, 18, 19).

In summary, after filtering out overprinting sewer-associated

taxa, sewage serves as a composite proxy for population-level human fecal microbiota. Comparative sewage analysis provides a unique opportunity to explore the relationship between human fecal communities and lifestyle or demographic differences in human populations. Combined with sensitive computational approaches to analyze microbial community data, sewage sampling provided a new approach that allowed us to move beyond the large individual-based sample collections that would be needed to compare microbiomes among 71 human populations.

MATERIALS AND METHODS

Sample collection. Sewage influent samples were collected from 71 cities and 78 wastewater treatment plant (WWTP) sites from across the United States during August 2012 (7 August 2012 to 7 September 2012), January 2013 (9 January 2013 to 28 February 2013), and May 2013 (28 April 2013 to 4 June 2013). To obtain these samples, we shipped sampling supplies to each of the WWTPs, including a cooler, frozen cold packs, sterile 500-ml sample bottles, chain-of-custody forms, and sample instructions. WWTP operators at each site collected sewage influent according to their plant's standard collection procedures, which ranged from single-time-point grab samples to flow-weighted composite samples taken over a 24-h period (see Table S1 in the supplemental material for metadata details). Following sample collection, the operators transferred the influent to an autoclaved sample bottle and then placed the sample bottle in a refrigerator until shipment to our lab. Prior to overnight shipping, the operators sealed the sample with Parafilm and placed the sample in the provided cooler containing frozen cold packs. Upon receiving sewage influent from the WWTPs, we mixed the influent by shaking and collected the microbial communities by filtration of 10 25-ml subsamples onto 0.22-μm mixed cellulose ester filters (47-mm diameter; Millipore, Billerica, MA, USA). Each filter was stored in a 2-ml screw-cap freezer tube at −80°C for up to 3 months before extraction of DNA. Using a sterile spatula, we crushed the frozen filters in their 2-ml storage tube and added a bead-beating matrix plus buffers according to the standard protocol for the Fast DNA spin kit for soil (MP Biomedicals, Solon, OH, USA). Following bead beating for 1 min, we extracted DNA according to the manufacturer's instructions and purified sample DNA using the Mo Bio PowerClean DNA cleanup kit (Mo Bio Laboratories Inc., Carlsbad, CA, USA).

WWTP operators employed their routine procedures to collect separate samples for influent chemical/physical measurements (see Table S1 in the supplemental material for metadata details). In addition to the environmental parameters reported by the treatment plants, we aggregated data for the past 30-year average daily high and low temperatures on the collection date for each sample (data from NOAA National Climatic Data Center [http://www.ncdc.noaa.gov]), the latitude and longitude of each treatment plant, the 2010 median age for the county/counties in which each city lies (http://www.census.gov/2010census/data), and the 2009 age-adjusted percent obesity of the population (obesity considered body mass index of ≥30) for the county/counties in which each city lies (http://www.cdc.gov/obesity/data). For those wastewater treatment plants receiving water from more than one county, we calculated the average measurements of the counties involved (see Table S1 for metadata).

In total, we collected 219 samples at the WWTP sites. Fifty-seven of the 71 cities returned samples in all three sampling periods. Denver, CO; Honolulu, HI; Junction City, KS; Juneau, AK; Milwaukee, WI; Santa Barbara, CA; and Vancouver, WA had multiple sample sites within each city, and Palo Alto, CA, collected both grab and composite samples. In one case, the Fall River, MA, sample for the January period, a sample was collected outside the listed sample collection range (see Table S1 in the supplemental material), but for classification this sample was considered part of the January sampling period.

16S rRNA gene sequencing and processing. We amplified and sequenced the V4-V5 region of 16S rRNA genes (*Escherichia coli* positions 518 to 926, ~408 nucleotides [nt]) using primers and adaptors as described by Huse et al. (33), except for the use of a 30-cycle one-step amplification procedure with the fusion primers instead of the reported two-step amplification procedure. After size selection and quantification of the PCR products, 250 cycles on an Illumina MiSeq produced paired-end reads. To analyze the raw paired-end reads, we used the "merge-illumina-pairs" script distributed in the Illumina Utilities library (available from https://github.com/meren/illumina-utils [34]). The script removed any read-pair with more than three mismatches in the ~80-nt-long overlapping region and also required at least 66% of the nucleotides in the non-overlapping region to have greater than a Q30 score (35). The program UCHIME (36) removed chimeras according to reference 33. After eliminating samples that did not pass quality controls at the filtration, DNA extraction, or sequencing steps, we were left with 207 sewage influent samples that were represented by 16,895,573 sequences assembled from the paired-end reads. Community sequence profiles were obtained in all three sample periods for 51 of the WWTP sites. Trimmed and quality-filtered sequences under the project name SLM_NIH2_Bv4v5 can be retrieved from the website Visualization and Analysis of Microbial Population Structures (VAMPS; https://vamps.mbl.edu/) (37).

Data set construction. For comparisons between sewage influent and human gut communities, we used human stool sample data from the Human Microbiome Project (5), which sequenced amplicons from the 16S rRNA gene V3-V5 region using 454 pyrosequencing technology (see reference 38 for sample collection and processing and reference 39 for sequencing-related procedures). The algorithm Global Alignment for Sequence Taxonomy (GAST) (40) assigned taxonomy to each unique stool sample V3-V5 read and each unique sewage sample V4-V5 read. We eliminated all samples that contained fewer than 4,000 sequence reads and removed one stool sample where 98% of reads resolved to the genus *Propionibacterium*, a nontypical genus for human gut communities (4, 7, 26). Following sample control procedures, we retained data from 137 individual stool samples (see Table S2 in the supplemental material for sample list).

The use of different primers for amplification and sequencing of the 16S rRNA region V3-V5 for the human stool data set on a 454 platform versus V4-V5 for the sewage data set on an Illumina platform could contribute to observed variation in the distribution of taxa/oligotypes between but not within data sets. To compare oligotypes from these overlapping rRNA regions, we created a combined amplicon data set for each of the six most abundant bacterial families in human stool (*Bacteroidaceae*, *Lachnospiraceae*, *Ruminococcaceae*, *Porphyromonadaceae*, *Rikenellaceae*, and *Prevotellaceae*). Following amplicon alignment with the align.seqs command in mothur (41), the nonoverlapping ends were trimmed (sequence length range after trimming, 225 to 240 nt) to create alignment files with sequences of equal length (23). A high-resolution oligotype analysis was conducted with the trimmed alignments as described previously (see reference 23 for details; oligotyping.org). Oligotyping is a supervised computational method that uses position-based Shannon entropy calculations in input alignments to identify highly variable alignment positions. The nucleotide compositions of these highly variable positions are used to parse the data into groups having identical sequences at the defined positions, and these groups are known as oligotypes (23).

To mitigate the potential biases of sampling depth on oligotyping (23), we randomly subsampled each sewage and human stool data set to a maximum depth of 80,000 and 7,000 reads per sample, respectively. The mean number of reads per sample after subsampling was 54,603 for the sewage data set and 5,996 for the human stool data set. Table S2 in the supplemental material reports the number of reads for each sample. Downstream analyses used the subsampled data sets. Oligotyping analyses minimized the impact of sequencing errors by employing a minimum substantive abundance criterion (*M*) and a minimum sample criterion (*s*). Using these criteria, oligotypes had to occur in ≥5 of 344 total samples (*s*) and be present at the lesser value of ≥0.01% of a family's total sequences or 500 total occurrences (*M*) to be included in comparative analyses. These noise-filtering steps, used previously for combined data sets

(42), removed 12.5% of the human stool sequences and 2.9% of the sewage influent sequences (Table 1).

BLAST analysis against NCBI's nr database (returning the top 500 matches; July 2014) identified the highest-identity matches for each oligotype's representative sequence (unique sequence with highest count). All NCBI sequences with the highest-identity match were binned based on their NCBI database "isolation_source" record as either of human stool/gut origin or of non-stool/gut origin. We then categorized each oligotype as "human fecal" or "nonfecal" based on which bin had the highest count. For our categorization purposes and subsequent notation in the text, "nonfecal" refers to all oligotypes assigned to nonhuman gut/stool origins (i.e., these oligotypes may have highest NCBI database matches to sequences originating from fecal sources other than human). Multiple database sequence matches from the same sample were counted only once.

Oligotype categorization revealed that 105 of 351 oligotypes had best matches in NCBI's nr database to sequences of organisms or environmental samples collected from nonfecal origins (see Table S2 in the supplemental material). The remaining 246 oligotypes comprised the human fecal oligotype data set. All comparisons of sewage sequence data to the human stool samples were conducted using the human fecal oligotype data set. BLAST analyses to identify cultured isolate matches to each oligotype's representative sequence were carried out in the Ribosomal Database Project (9) sequence match program (July 2014), with "sequences ≥1200 nt" and "isolates" as exclusionary criteria.

Statistical analyses. Statistical analyses were conducted in the statistical package R (43). For all visualizations, we used the ggplot2 package (44). To calculate bacterial community similarity among samples/data sets, we used the function vegdist with the Bray-Curtis metric in the vegan package (45). Simple linear regressions were fitted with the lm function, and diversity measures were calculated with the renyiaccum function in the vegan package (45). Wilcoxon rank sum tests were used to determine whether the means between two conditions differed significantly. Heat maps were constructed using the heatmap.2 function in the gplots package (46).

We examined the relationship between the measured environmental, geographic, and demographic variables and the oligotype composition among samples with principal coordinate analysis (PCoA) using the function cmdscale in R (43). Given that temperature, latitude, and sample period are not independent factors, we ran PCoAs for each sample period in addition to analyses for all periods combined. We also created a metric that we termed the "city temperature profile" to incorporate the potential effects from climatic differences (i.e., temperature and geographic location) among cities. To calculate the city temperature profile metric, we used the previous 30-year average high and low temperatures on each of the three collection days for each city. If a city was not represented by a sample during a collection period, then the 15th day of that collection month was used in the calculation. The profile of average high and low temperatures from each of three collection days was used to construct a nonmetric multidimensional scaling (NMDS) plot (metaMDS; vegan package [45]), and each city's x axis score was used to represent the temperature profile for that city.

In order to compare community composition variation among cities between the human fecal and nonfecal oligotypes in sewage samples, we created a dummy sample matrix, where each original sewage sample was split into two samples, one containing the human fecal oligotypes and one containing the nonfecal oligotypes. The split samples were mutually exclusive for oligotypes. To account for this, each oligotype in one sample set (e.g., human fecal) was given a relative abundance of "0" in the samples in the opposing set (e.g., nonfecal). The relationship of both the human fecal and nonfecal bacterial community in each sewage sample to its city's climate was then examined in a constrained ordination of principal coordinates (CAP; capscale function, vegan package [45]) for the city temperature profile metric. The dummy relative abundance matrix allowed for the two data sets to be plotted in the same ordination and thus have comparable axis scores.

We examined the ability of the sewage bacterial community composition (human fecal oligotypes) to predict whether human populations contained a low or high percentage of obese individuals. To conduct these analyses, we created a classification model using random forests (47) implemented in randomForest v. 4.6-7 in R (48). Random forest analysis generates unbiased, "out-of-bag" error estimates for the data set without requiring the data to be split into training and test data sets. For the random forest analyses, we split the city sewage samples into three groups according to the distribution of estimated percent obese individuals for the city populations. We considered the most lean and obese human populations to be those in the first and fourth quartiles for obesity percent in our data set, which operationally resulted in a "lean" category, where ≤22.8% of individuals in a city were obese, and an "obese" category, where ≥30.4% of individuals were obese. The samples classified in the lean and obese categories were then used in the random forest model. We also compiled a second, more stringent data set using only samples that were greater than 1 standard deviation from the sewage data set's mean percent obesity. This configuration generated a lean population group consisting of samples with ≤21.5% obesity and an obese population group consisting of samples with ≥31.3% obesity.

Nucleotide sequence accession numbers. The National Center for Biotechnology Information Sequence Read Archive has archived the raw data under BioProjects PRJNA261344 and PRJNA264400.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.02574-14/-/DCSupplemental.

Figure S1, EPS file, 0.7 MB.
Figure S2, EPS file, 0.4 MB.
Figure S3, EPS file, 0.2 MB.
Figure S4, EPS file, 0.5 MB.
Figure S5, EPS file, 0.6 MB.
Figure S6, EPS file, 0.3 MB.
Table S1, XLSX file, 0.1 MB.
Table S2, XLSX file, 0.6 MB.
Table S3, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Julien Grimaud and Veolia Water North America for facilitating the involvement of a large number of the wastewater treatment plants. We thank the wastewater treatment plant operators who collected samples and ancillary metadata; Morgan Shroeder, Ian DeTuncq, Melinda Bootsma, Danielle Cloutier, Patricia Bower, and Katherine Halmo, who assisted in processing samples; and Jenny Fisher, who provided helpful comments on earlier versions of the manuscript.

Funding for this work was provided by the NIH grant R01AI091829-01A1 to S.L.M. and M.L.S.

REFERENCES

1. **Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA.** 2005. Diversity of the human intestinal microbial flora. Science **308:**1635–1638. http://dx.doi.org/10.1126/science.1110591.
2. **Claesson MJ, O'Sullivan O, Wang Q, Nikkilä J, Marchesi JR, Smidt H, de Vos WM, Ross RP, O'Toole PW.** 2009. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. PLoS One **4:**e6669. http://dx.doi.org/10.1371/journal.pone.0006669.
3. **Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, Knight R.** 2011. Moving pictures of the human microbiome. Genome Biol **12:** R50. http://dx.doi.org/10.1186/gb-2011-12-5-r50.
4. **Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, 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.** 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464:**59–65. http://dx.doi.org/10.1038/nature08821.
5. **Human Microbiome Project Consortium.** 2012. Structure, function and

diversity of the healthy human microbiome. Nature **486:**207–214. http://dx.doi.org/10.1038/nature11234.

6. **Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI.** 2009. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med **1:**6ra14. http://dx.doi.org/10.1126/scitranslmed.3000322.
7. **Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R.** 2012. Diversity, stability and resilience of the human gut microbiota. Nature **489:**220–230. http://dx.doi.org/10.1038/nature11550.
8. **Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K.** 2011. Enterotypes of the human gut microbiome. Nature **473:**174–180. http://dx.doi.org/10.1038/nature09944.
9. **Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Hildalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso J, Lozupone CA, Lauber CA, Clemente JC, Knights D, Knights R, Gordon JI.** 2012. Human gut microbiome viewed across age and geography. Nature **486:**222–227. http://dx.doi.org/10.1038/nature11053.
10. **Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, Henrissat B, Knight R, Gordon JI.** 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science **332:**970–974. http://dx.doi.org/10.1126/science.1198719.
11. **Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R.** 2009. Bacterial community variation in human body habitats across space and time. Science **326:**1694–1697. http://dx.doi.org/10.1126/science.1177486.
12. **Claesson MJ, Cusack S, O'Sullivan O, Greene-Diniz R, de Weerd H, Flannery E, Marchesi JR, Falush D, Dinan T, Fitzgerald G, Stanton C, van Sinderen D, O'Connor M, Harnedy N, O'Connor K, Henry C, O'Mahony D, Fitzgerald AP, Shanahan F, Twomey C.** 2011. Composition, variability, and temporal stability of the intestinal microbiota of the elderly. Proc Natl Acad Sci U S A **108**(Suppl 1):4586–4591. http://dx.doi.org/10.1073/pnas.1000097107.
13. **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. http://dx.doi.org/10.1073/pnas.1000081107.
14. **Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, Harris HM, Coakley M, Lakshminarayanan B, O'Sullivan O, Fitzgerald GF, Deane J, O'Connor M, Harnedy N, O'Connor K, O'Mahony D, van Sinderen D, Wallace M, Brennan L, Stanton C.** 2012. Gut microbiota composition correlates with diet and health in the elderly. Nature **488:**178–184. http://dx.doi.org/10.1038/nature11319.
15. **Frank DN, St. Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR.** 2007. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A **104:**13780–13785. http://dx.doi.org/10.1073/pnas.0706625104.
16. **Dicksved J, Halfvarson J, Rosenquist M, Järnerot G, Tysk C, Apajalahti J, Engstrand L, Jansson JK.** 2008. Molecular analysis of the gut microbiota of identical twins with Crohn's disease. ISME J **2:**716–727. http://dx.doi.org/10.1038/ismej.2008.37.
17. **Berry D, Schwab C, Milinovich G, Reichert J, Ben Mahfoudh K, Decker T, Engel M, Hai B, Hainzl E, Heider S, Kenner L, Müller M, Rauch I, Strobl B, Wagner M, Schleper C, Urich T, Loy A.** 2012. Phylotype-level 16S rRNA analysis reveals new bacterial indicators of health state in acute murine colitis. ISME J **6:**2091–2106. http://dx.doi.org/10.1038/ismej.2012.39.
18. **Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD.** 2011. Linking long-term dietary patterns with gut microbial enterotypes. Science **334:**105–108. http://dx.doi.org/10.1126/science.1208344.
19. **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. http://dx.doi.org/10.1038/nature12820.
20. **De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P.** 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci U S A **107:**14691–14696. http://dx.doi.org/10.1073/pnas.1005963107.
21. **McLellan SL, Newton RJ, Vandewalle JL, Shanks OC, Huse SM, Eren AM, Sogin ML.** 2013. Sewage reflects the distribution of human faecal *Lachnospiraceae*. Environ Microbiol **15:**2213–2227. http://dx.doi.org/10.1111/1462-2920.12092.
22. **Eren AM, Sogin ML, Morrison HG, Vineis JH, Fisher JC, Newton RJ, McLellan SL.** 2015. A single genus in the gut microbiome reflects host preference and specificity. ISME J **9:**90–100. http://dx.doi.org/10.1038/ismej.2014.97.
23. **Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML.** 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol **4:**1111–1119. http://dx.doi.org/10.1111/2041-210X.12114.
24. **Shanks OC, Newton RJ, Kelty CA, Huse SM, Sogin ML, McLellan SL.** 2013. Comparison of the microbial community structures of untreated wastewaters from different geographic locales. Appl Environ Microbiol **79:**2906–2913. http://dx.doi.org/10.1128/AEM.03448-12.
25. **Vandewalle JL, Goetz GW, Huse SM, Morrison HG, Sogin ML, Hoffmann RG, Yan K, McLellan SL.** 2012. *Acinetobacter*, *Aeromonas* and *Trichococcus* populations dominate the microbial community within urban sewer infrastructure. Environ Microbiol **14:**2538–2552. http://dx.doi.org/10.1111/j.1462-2920.2012.02757.x.
26. **Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet J.-P, Ugarte E, Muñoz-Tamayo R, Paslier DL, Nalin R, Dore J, Leclerc M.** 2009. Towards the human intestinal microbiota phylogenetic core. Environ Microbiol **11:**2574–2584. http://dx.doi.org/10.1111/j.1462-2920.2009.01982.x.
27. **David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, Erdman SE, Alm EJ.** 2014. Host lifestyle affects human microbiota on daily timescales. Genome Biol **15:**R89. http://dx.doi.org/10.1186/gb-2014-15-7-r89.
28. **Turnbaugh PJ, Bäckhed F, Fulton L, Gordon JI.** 2008. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. Cell Host Microbe **3:**213–223. http://dx.doi.org/10.1016/j.chom.2008.02.015.
29. **Ley RE, Turnbaugh PJ, Klein S, Gordon JI.** 2006. Microbial ecology: human gut microbes associated with obesity. Nature **444:**1022–1023. http://dx.doi.org/10.1038/4441022a.
30. **Moreno-Indias I, Cardona F, Tinahones FJ, Queipo-Ortuño MI.** 2014. Impact of the gut microbiota on the development of obesity and type 2 diabetes mellitus. Front Microbiol **5:**190. http://dx.doi.org/10.3389/fmicb.2014.00190.
31. **Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R.** 2011. Human-associated microbial signatures: examining their predictive value. Cell Host Microbe **10:**292–296. http://dx.doi.org/10.1016/j.chom.2011.09.003.
32. **Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto JM, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jørgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F.** 2013. Richness of human gut microbiome correlates with metabolic markers. Nature **500:**541–546. http://dx.doi.org/10.1038/nature12506.
33. **Huse SM, Young VB, Morrison HG, Antonopoulos DA, Kwon J, Dalal S, Arrieta R, Hubert NA, Shen L, Vineis JH, Koval JC, Sogin ML, Chang EB, Raffals LE.** 2014. Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects. Microbiome **2:**5. http://dx.doi.org/10.1186/2049-2618-2-5.
34. **Eren AM, Vineis JH, Morrison HG, Sogin ML.** 2013. A filtering method to generate high quality short reads using Illumina paired-end technology. PLoS One **8:**e66643. http://dx.doi.org/10.1371/journal.pone.0066643.
35. **Minoche AE, Dohm JC, Himmelbauer H.** 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol **12:**R112. http://dx.doi.org/10.1186/gb-2011-12-11-r112.
36. **Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R.** 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics **27:**2194–2220. http://dx.doi.org/10.1093/bioinformatics/btr381.
37. **Huse SM, Mark Welch DB, Voorhis A, Shipunova A, Morrison HG, Eren AM, Sogin ML.** 2014. VAMPS: a website for visualization and analysis of microbial population structures. BMC Bioinformatics **15:**41. http://dx.doi.org/10.1186/1471-2105-15-41.
38. **Aagaard K, Petrosino J, Keitel W, Watson M, Katancik J, Garcia N,**

Patel S, Cutting M, Madden T, Hamilton H, Harris E, Gevers D, Simone G, McInnes P, Versalovic J. 2013. The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. FASEB J **27:**1012–1022. http://dx.doi.org/10.1096/fj.12-220806.

39. **Human Microbiome Project Consortium.** 2012. A framework for human microbiome research. Nature **486:**215–221. http://dx.doi.org/10.1038/nature11209.
40. **Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Relman DA, Sogin ML.** 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLOS Genet **4:**e1000255. http://dx.doi.org/10.1371/journal.pgen.1000255.
41. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol **75:**7537–7541. http://dx.doi.org/10.1128/AEM.01541-09.
42. **Eren AM, Borisy GG, Huse SM, Mark Welch JL.** 2014. Oligotyping analysis of the human oral microbiome. Proc Natl Acad Sci U S A **111:**E2875–E2884. http://dx.doi.org/10.1073/pnas.1409644111.
43. **The R Core Team.** 2013. R: a language and environment for statistical computing. Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/.
44. **Wickham H.** 2009. ggplot2: elegant graphics for data analysis. Springer, New York, NY.
45. **Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Henry M, Stevens H, Wagner H.** 2013. vegan: community ecology package. R package version 2.0-8. http://CRAN.R-project.org/package=vegan.
46. **Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B.** 2013. Gplots: various R programming tools for plotting data. R package version 2.12.1.
47. **Breiman L.** 2001. Random Forests. Mach Learn **45:**5–32. http://dx.doi.org/10.1023/A:1010933404324.
48. **Liaw A, Wiener M.** 2002. Classification and regression by randomForest. R News **2:**18–22.