



FORENSIC: an Online Platform for Fecal Source Identification

Adélaïde Roguet,^a Özcan C. Esen,^b A. Murat Eren,^b Ryan J. Newton,^a Sandra L. McLellan^a

^aSchool of Freshwater Sciences, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA

^bDepartment of Medicine, University of Chicago, Chicago, Illinois, USA

ABSTRACT Sewage overflows, agricultural runoff, and stormwater discharges introduce fecal pollution into surface waters. Distinguishing these sources is critical for evaluating water quality and formulating remediation strategies. With the falling costs of sequencing, microbial community-based water quality assessment tools are under development. However, their application is limited by the need to build reference libraries, which requires extensive sampling of sources and bioinformatic expertise. Here, we introduce FORest Enteric Source IdentifiCation (FORENSIC; <https://forensic.sfs.uwm.edu/>), an online, library-independent source tracking platform based on random forest classification and 16S rRNA gene amplicon sequences to identify in environmental samples common fecal contamination sources, including humans, domestic pets, and agricultural animals. FORENSIC relies on a broad reference signature database of *Bacteroidales* and *Clostridiales*, two predominant bacterial groups that have coevolved with their hosts. As a result, these groups demonstrate cohesive and reliable assemblage patterns within mammalian species or among species sharing the same diet/physiology. We created a scalable and extensible platform that we tested for global applicability using samples collected in distant geographic locations. This Web application offers a fast and intuitive approach for fecal source identification, particularly in sewage-contaminated waters.

IMPORTANCE FORENSIC is an online platform to identify sources of fecal pollution without the need to create reference libraries. FORENSIC is based on the ability of random forest classification to extract cohesive source microbial signatures to create classifiers despite individual variability and to detect the signatures in environmental samples. We primarily focused on defining sewage signals, which are associated with a high human health risk in polluted waters. To test for fecal contamination sources, the platform only requires paired-end reads targeting the V4 or V6 regions of the 16S rRNA gene. We demonstrated that we could use V4V5 reads trimmed to the V4 positions to generate the reference signature. The systematic workflow we describe to create and validate the signatures could be applied to many disciplines. With the increasing gap between advancing technology and practical applications, this platform makes sequence-based water quality assessments accessible to the public health and water resource communities.

KEYWORDS microbial source tracking, 16S rRNA gene, high-throughput sequencing, *Bacteroidales*, *Clostridiales*, random forest classification, toolkit


Environmental fecal pollution is recognized worldwide as a major threat to human health that causes millions of deaths in developing countries (1). While the waterborne disease burden is far less in developed countries, chronic fecal pollution is regularly reported (2) and leads to waterborne illness, diminished ecosystem services, and economic loss (3). In urbanized watersheds, fecal contamination can originate from multiple sources, including failing sewage infrastructure and contaminated runoff containing waste from upstream livestock, pets, and wildlife. However, gastrointestinal

Citation Roguet A, Esen ÖC, Eren AM, Newton RJ, McLellan SL. 2020. FORENSIC: an online platform for fecal source identification. mSystems 5:e00869-19. <https://doi.org/10.1128/mSystems.00869-19>.

Editor Sarah M. Allard, University of California San Diego

Copyright © 2020 Roguet et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Sandra L. McLellan, mclellan@uwm.edu.

 In this report, we introduce FORENSIC, an online, microbial source tracking platform based on random forest classification, which relies on a broad reference signature database to identify fecal pollution in water samples (<https://forensic.sfs.uwm.edu/>).

Received 12 December 2019

Accepted 21 February 2020

Published 17 March 2020

illness risks may differ according to the pollution source. For example, human exposure to waters contaminated by sewage or cattle waste is estimated to present a greater health risk than similar exposure to avian waste (4, 5). Water quality measures based on traditional fecal indicator bacteria do not identify pollution sources (6). As a result, more advanced methods, including quantitative PCR, have been developed to distinguish fecal contamination sources by targeting marker sequences that are indicative of a specific pollution source (7).

The ever-decreasing cost of sequencing and the development of advanced computational tools has made high-throughput sequencing one of the most efficient ways to study bacterial communities, and it shows promise for microbial source tracking applications (8, 9). Recent studies have used 16S rRNA gene amplicons to successfully characterize human and animal fecal sources in surface waters (10–16). These studies rely on the detection of fecal community signatures shaped by coevolutionary dynamics between hosts and their gut microbiota (17–20). These approaches, including use of the state-of-the-art Bayesian classifier SourceTracker (21), require the creation of a reference sequence library for fecal samples, which entails sampling and sequencing both of source samples and of the water samples of interest. Such additional effort is often beyond the capabilities or resources available to a single investigator. This limitation is compounded when fecal references are geographically distant from samples tested, which can alter the accuracy of source identification (14). To date, a publicly available database that encompasses most common animal sources for fecal contamination does not exist.

Our previous work reported a source tracking method using random forest classification, which offered a rapid, sensitive, and accurate solution for identifying host-microbial signatures in environmental samples (15). In this study, we demonstrated that the two most common gut-associated bacterial orders in mammals, i.e., *Bacteroidales* and *Clostridiales* (18, 22), provide enough host-specific signal to perform accurate source identification. Similar observations were reported at the genus level (23–25), highlighting the fractal nature of specialization between gut microbiota and their hosts. Targeting taxonomically defined groups, rather than studying the whole bacterial community, may reduce the influence of large cross-phylum shifts due to diet or sequencing primer bias in the bacterial community (25) while providing relevant host-associated profiles.

Here, we introduce the Web application FORest Enteric Source Identification (FORENSIC), an online platform with a user-friendly interface that employs random forest to identify common sources responsible for fecal contamination. FORENSIC gives access to an extensible list of human and animal reference signatures, eliminating the need for users to create their own fecal reference databases. This platform responds to an emerging demand for sequence data in the advancement of water quality testing.

RESULTS

Fecal gut microbiota were explored in 132 human (sewage) and 395 animal fecal source samples collected around the world, which included but not were not limited to those from cats, cattle, dogs, deer, pigs, horses, goats, sheep, rabbits, raccoons, chicken, and geese. Amplicon sequences generated from the V4, V4V5, and V6 regions of the 16S rRNA gene were used to create and validate global FORENSIC source signatures.

Host community patterns and performance of global classifiers. We focused on *Bacteroidales* and *Clostridiales* to perform fecal source identification, since they are the most dominant fecal microbiota and have the highest genetic diversity among mammal fecal samples that we examined (Fig. 1). Despite microdiversity and geographical patterns within some sources, both taxonomic groups provided resolution to observe source-specific structures in their assemblages (Fig. 2). This structure allowed us to create and store fecal source signatures as classifiers for both bacterial orders and multiple 16S rRNA gene variable regions. A sample was considered contaminated by a source if random forest classification could identify a bacterial structure similar to the fecal signature stored in the classifier.

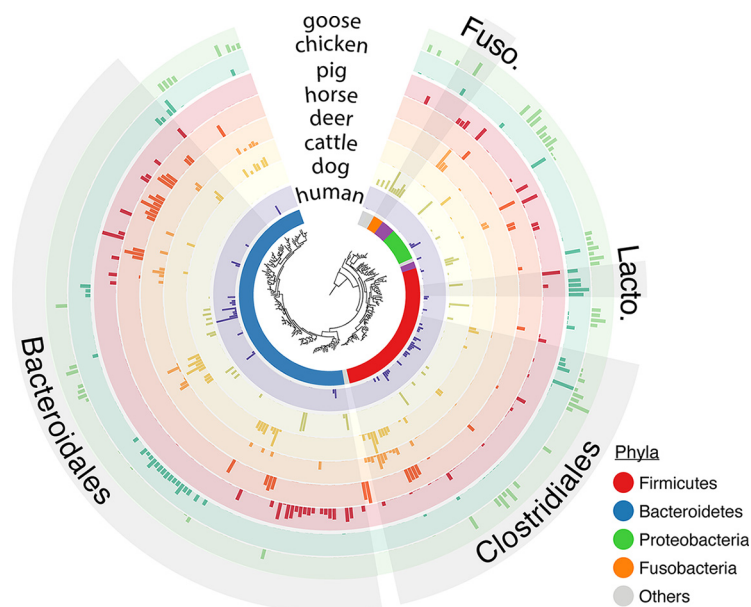


FIG 1 Phylogeny and distribution of the dominant V4 sequences across the fecal microbiota of eight host species. Ten fecal samples were averaged per host, except for dog, chicken, and goose, represented by 7, 6, and 8 samples, respectively. Only sequences ($n = 287$) with an average relative abundance per host of 0.5% are displayed. Colors in the inner circle depict phyla. Gray clades symbolize the dominant bacterial orders for at least one host. Log-transformed relative abundances were normalized by the maximum abundance for each host. Fuso., *Fusobacteriales*; Lacto., *Lactobacillales*. The tree was rooted using *Halobellus ramosii* strain S2FP14 (GenBank accession no. [NR_145608.1](https://www.ncbi.nlm.nih.gov/nuccore/NR_145608.1)). Sequences were aligned using MUSCLE implemented in MEGA (61). The tree was generated using the interactive Tree Of Life (iTOL) (62).

To create the initial classifiers, we used the 15 to 20 samples with the most similar *Bacteroidales* or *Clostridiales* sequence structure (based on Bray-Curtis intrasource comparisons) for the V4V5 and V6 variable regions. We then tested the specificity (i.e., proportion of correctly classified negative samples) and sensitivity (i.e., proportion of correctly classified positive samples) of these classifiers against samples with the most dissimilar *Bacteroidales* or *Clostridiales* sequence structure. Exceptions were applied when a strong community bifurcation was observed within an animal source assemblage (i.e., for V6 cattle and dog; see Materials and Methods). We maximized the performance of each classifier by tailoring the voting tree probability cutoff (called the decision cutoff) at which a sample was considered contaminated or not by a source (see Materials and Methods); the voting tree probability is a proxy reflecting the bacterial profile similarity between a tested sample and a fecal signature. The performances of classifiers and decision cutoffs are summarized in Table 1. The classifier amplicon sequence variants (ASVs), which are the highly resolved operational taxonomic units composing the individual source signatures, are listed in Table S1 in the supplemental material. All V4V5 and V6 sewage classifiers had a specificity of 100%. All sewage samples were correctly classified. All animal classifiers had a specificity ranging from 82 to 100%, except V4V5 dog *Bacteroidales*, for which no cutoff could be defined; this classifier was therefore discarded. Misclassifications typically involved animals with similar diets, e.g., misclassifications of horse as cattle (see Data Set S1 in the supplemental material for individual sample predictions). The fecal sources with the lowest sensitivity, e.g., dog or cattle, had a large intrasource dispersion, likely due to a larger number of atypical samples used to test the classifier performance (see nonparametric multidimensional scaling plot in Fig. S1 in the supplemental material).

We only considered classifiers to be validated as global classifiers if the specificity and the sensitivity were $\geq 70\%$ when using a voting tree decision cutoff of $> 10\%$. Reducing the voting cutoff further increases sensitivity in some cases but creates specificity tradeoffs and may generate too many misclassifications when testing envi-

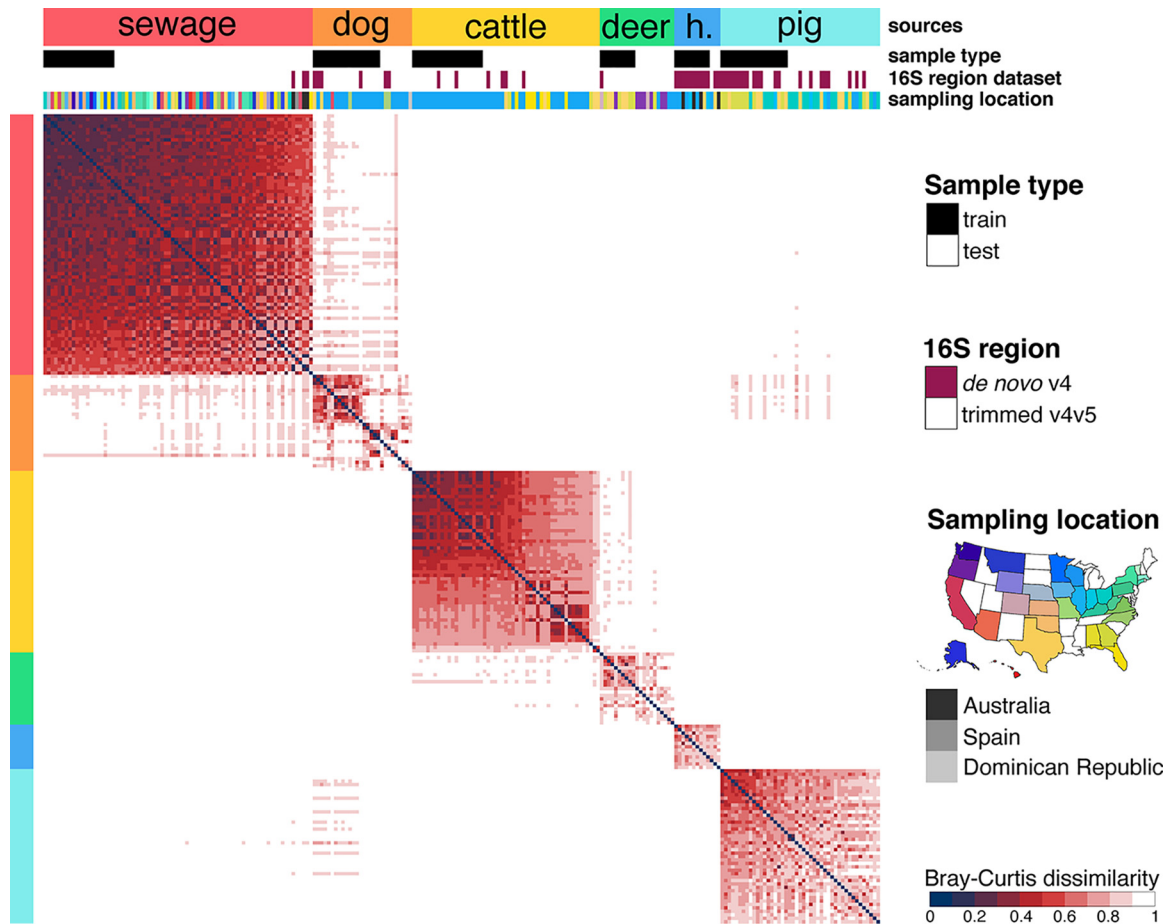


FIG 2 Bray-Curtis analysis of the *Bacteroidales* assemblage for the V4 region of the 16S rRNA gene. Color code shows the host, the type of sample (train or test the classifier), the initial 16S rRNA gene region amplified, and the geographical origin of the samples. Samples were ordered by source from the most similar samples (left) to the most dissimilar (right) based on the averaged intrasource Bray-Curtis dissimilarity comparisons. h., horse. Bray-Curtis analysis of the V4 *Clostridiales* and V6 *Bacteroidales* and *Clostridiales* assemblages are presented in Fig. S2 and S3 in the supplemental material, respectively.

ronmental samples. Classifiers with a cutoff of $\leq 10\%$ were considered “draft” in order to offer the users the ability to explore their data set while considering uncertainty. Classifiers with a decision cutoff of $\leq 5\%$ were discarded.

Scalability of the classifiers. In addition to the host-specific classifiers, we developed diet/physiology-centered classifiers, including those for herbivore and ruminant sources. These broad classifiers allowed us to extend the scope of the host-specific classifiers to capture fecal signatures shared among hosts having similar diets or physiology. This approach also allowed us to include sources with limited sample sets that prohibited creating individual host classifiers (e.g., horse or sheep). The diet/physiology-centered classifiers had a minimum specificity and sensitivity of 78% (Table 2). Interestingly, in some cases, the broader classifiers were able to identify the fecal signature in samples (e.g., *Bacteroidales* V6 deer samples) where the host-specific classifiers failed (see Data Set S1).

Importance of sequencing depth. We evaluated the performance of the source classifiers (*Bacteroidales* and *Clostridiales* for both V4V5 and V6) in seven freshwater river samples for which sewage and cattle fecal contamination had been previously identified (26) using validated quantitative PCR (qPCR) assays (Table 3). Unlike V4V5, V6 classifiers identified the sewage signature in all the samples. The sequencing of the V6 region using the Illumina HiSeq or NextSeq platform provided 4-fold higher depth than the Illumina MiSeq platform used to sequence the V4V5 region. The two V4V5 samples not classified as contaminated with sewage (MKE_18755 and MKE_18765) had the

TABLE 1 Sensitivity and specificity of the V4V5 and V6 source classifiers

Region	Source	<i>Bacteroidales</i>						<i>Clostridiales</i> ^a					
		Specificity (%) ^b	<i>n</i>	Sensitivity (%)	<i>n</i>	Decision cutoff (%)	Classifier type ^c	Specificity (%) ^b	<i>n</i>	Sensitivity (%)	<i>n</i>	Decision cutoff (%)	Classifier type ^c
V4V5	Sewage	100	82	100	52	14	Global	100	89	100	53	16	Global
	Dog	82 (cat, raccoon)	125	100	9	14	Global	NA	134	NA	8	ND	Discarded
	Pig	100	123	100	11	14	Global	100	130	100	12	39	Global
V6	Sewage	100	261	100	9	21	Global	100	265	100	9	16	Global
	Dog	91	239	84	31	1	Discarded	87 (cat)	243	100	31	34	Global
	Cattle	82 (horse, kangaroo, raccoon)	222	85	48	26	Global	86 (horse)	226	81	48	26	Global
	Deer	84 (sheep)	257	100	13	4	Discarded	95 (sheep)	261	100	13	26	Global
	Pig	99	235	100	35	14	Global	100	238	100	36	31	Global

^aNA, not applicable; ND, not defined.

^bSource(s) for which the majority of the samples were misclassified are given in parentheses.

^cType of classifier based on decision cutoff. Classifiers with a decision cutoff of >10% were considered global, and those with a cutoff of ≤5% discarded.

smallest number of sequences matching the classifiers. This increased sequencing depth appeared to provide the information needed to capture the fecal signature in the environmental samples where bacteria are already abundant (in freshwater samples, bacterial abundance is typically >10⁵ cells per ml). For this reason, we decided to focus our classifier approach on short hypervariable regions of the 16S rRNA gene to maximize sequencing depth. In addition to the V6 region, we created and tested the performance of classifiers built on the V4 region, which is targeted extensively in environmental and fecal microbiota data sets (27). We incrementally subsampled all fecal samples used to test the V4 and V6 classifiers in order to define the classifier detection limits. This analysis revealed that a minimum of 1,000 sequences was necessary to properly classify a sample. Additionally, artificial bacterial assemblages composed of several fecal signatures were also correctly classified when the fecal signature matching the classifiers was composed of at least 1,000 sequences represented by ~100 unique representative sequences (see Table S2 in the supplemental material for individual sample predictions). The relationship between the number of sequences in a sample (proxy of the percentage of contamination) and the voting tree probability was logarithmic.

Flexibility of sequence data. We developed and compared V4 classifiers from both *de novo*-produced V4 sequence data and trimmed V4V5 sequence reads. The ability to reliably use trimmed data was explored using 42 untreated sewage and 191 animal fecal samples sequenced in both the V4 and V4V5 regions. The V4 primer sites were recognized in more than 99% of *Bacteroidales* and *Clostridiales* V4V5 reads, suggesting that the V4 primer sites are highly conserved in these two bacterial orders. Since the trimmed V4V5 data set was sequenced to a lower depth, we rarefied the *de novo* V4 data set to the median count of trimmed V4V5 reads. We created three source classifiers, namely sewage, cattle, and pig, using both rarefied *de novo* V4 and trimmed V4V5 data sets. The majority of the sequences composing the classifiers were shared between the two data sets (see Table S3 in the supplemental material). Moreover, they generated comparable sensitivity and specificity metrics when tested using their

TABLE 2 Sensitivity and specificity of the V4V5 and V6 diet and physiology classifiers

Region	Classifier ^a	<i>Bacteroidales</i>						<i>Clostridiales</i>					
		Specificity (%) ^b	<i>n</i>	Sensitivity (%)	<i>n</i>	Decision cutoff (%)	Classifier type ^c	Specificity (%) ^b	<i>n</i>	Sensitivity (%)	<i>n</i>	Decision cutoff (%)	Classifier type ^c
V4V5	Ruminant	100	121	92	13	13	Global	99	123	95	19	4	Discarded
V6	Herbivore	78 (raccoon, chicken)	162	87	108	40	Global	88 (raccoon)	167	87	107	25	Global

^aClassifiers trained using cattle and deer fecal samples.

^bSource(s) for which the majority of the samples were misclassified are given in parentheses.

^cType of classifier based on decision cutoff. Classifiers with a decision cutoff of >10% were considered global, and those with a cutoff of ≤5% discarded.

TABLE 3 Random forest classifications using the V4V5 and V6 global classifiers of seven freshwater river samples contaminated by sewage and cattle fecal pollution^a

Sample identifier	Human marker ^b	Ruminant marker ^b	Total no. of sequences (no. of sequences matching the V4V5 classifiers)		Classifier for:		Total no. of sequences (no. of sequences matching the V6 classifiers)		Classifier for:	
					<i>Bacteroidales</i>	<i>Clostridiales</i>	<i>Bacteroidales</i>	<i>Clostridiales</i>		
MKE_18755	4.26	4.54	54,495 (25)					212,620 (818)		
MKE_18756	5.71	4.61	51,879 (363)		Sewage		Sewage	296,677 (3,171)		Sewage
MKE_18757	5.63	4.66	75,024 (588)		Sewage		Sewage	276,911 (2,585)		Sewage
MKE_18759	5.65	5.30	76,611 (732)		Sewage		Sewage	258,374 (2,863)		Sewage
MKE_18761	5.63	5.33	72,991 (526)		Sewage		Sewage	345,635 (3,849)		Sewage
MKE_18763	5.15	5.78	65,949 (288)		Sewage		Sewage	249,932 (1,442)		Sewage/herbivore
MKE_18765	5.02	6.22	68,772 (220)		Ruminant		Sewage	332,866 (1,779)		Sewage/cattle/deer/herbivore

^aV4V5 global classifiers include sewage, dog (*Bacteroidales*), pig, and ruminant (*Bacteroidales*). V6 global classifiers include sewage, dog (*Clostridiales*), cattle, pig, deer (*Clostridiales*), and herbivore.

^bCopy number per 100 ml (log₁₀-transformed). Human and ruminant quantitative PCR marker data from Olds et al. (26).

TABLE 4 Sensitivity and specificity of the combined V4 classifiers generated using *de novo* V4 and trimmed V4V5 sequences to the V4 primer positions

Classifier	<i>Bacteroidales</i>					<i>Clostridiales</i>						
	Specificity (%) ^a	<i>n</i>	Sensitivity (%)	<i>n</i>	Decision cutoff (%)	Classifier type (%) ^b	Specificity (%) ^a	<i>n</i>	Sensitivity (%)	<i>n</i>	Decision cutoff (%)	Classifier type ^b
Sewage	100	138	100	56	19	Global	100	142	100	57	17	Global
Dog	87 (cat)	185	100	9	22	Global	75 (cat, raccoon)	190	78	9	10	Draft
Cattle	100	161	100	33	9	Draft	99	162	100	37	16	Global
Pig	99	168	100	26	13	Global	99	174	100	25	29	Global
Ruminant ^c	93 (kangaroo)	150	100	44	5	Discarded	97 (kangaroo)	151	100	48	10	Draft
Herbivore ^d	88 (raccoon, goose)	135	95	59	33	Global	99	136	89	63	22	Global

^aSource(s) for which the majority of the samples were misclassified are given in parentheses.

^bType of classifier based on their decision cutoff. Classifiers with a decision cutoff of >10% were considered global, those with a decision cutoff of >5% were considered draft, and those with a cutoff of ≤5% were discarded.

^cRuminant classifiers trained using cattle and deer fecal samples.

^dHerbivore classifiers trained using cattle, deer, and horse fecal samples.

respective data set and the reciprocal data set (i.e., rarified *de novo* V4 classifiers tested using trimmed samples and *vice versa*). Interestingly, a nonrarefied *de novo* V4 classifier had lower sensitivity when using trimmed V4V5 data, likely because of the lower sequence depth of the test data set. The performance of each configuration is described in Table S4 in the supplemental material. Thereafter, V4 classifiers exploiting both rarefied *de novo* V4 and trimmed V4V5 data sets were implemented in FORENSIC. The performance is described in Table 4.

FORENSIC implementation and user interface. FORENSIC integrates the validated classifiers to perform fecal source identification and can be accessed online at <https://forensic.sfs.uwm.edu/>. It requires a FASTA file (see Materials and Methods) and yields source contamination predictions in raw downloadable text and interactive visualizations. Classification relies upon an exact match with fecal signature sequences, and therefore trimming to exact V4 or V6 primer positions is crucial. Most FORENSIC analysis takes a few seconds to a few minutes, depending on the number of input sequences. FORENSIC currently encompasses a total of five global source classifiers and two physiology/diet classifiers for V4 and/or V6 regions (Tables 1, 2, and 4).

FORENSIC visualizes its predictions in multiple ways, including a color-coded table of source category predictions for each submitted sample (Fig. 3a) and an interactive association networks of ASVs and sources that allows the user to explore and disentangle the different fecal profiles recovered from the samples (Fig. 3b). This interactive visualization strategy can also help to detect potentially false-positive predictions by discerning the prevailing fecal signature(s) from other trace source signatures and/or from signatures sharing sequences with the dominant one. Users can access raw data outputs such as counts of sequences matching to each classifier, voting tree probabilities, and simplified identifications. Warning flags are displayed if samples do not meet certain criteria, such as the number of sequences in the submitted FASTA or the number of sequences (or unique representative sequences) matching to the classifiers being too low indicating insufficient sequencing depth or, in the case of no matching sequences, technical issues in data processing.

DISCUSSION

FORENSIC is an online platform for performing fecal source identification in water from 16S rRNA gene amplicons. It leverages *Bacteroidales* and *Clostridiales*, two bacterial orders with many members that show remarkable host specificity in mammalian guts (15, 23, 28, 29). Microbial community differences that distinguish animals are substantially diet driven due to the specialization of bacterial members for specific metabolic functions (30–32). Here, we used these host-microbiota association patterns to discriminate fecal contamination with both host-based (e.g., human, cattle, pig) and/or diet/physiology-based (e.g., ruminant, herbivore) classifiers. Specific markers for both *Bacteroides* (33, 34) and *Lachnospiraceae* (35) have shown a broad geographic distribution.

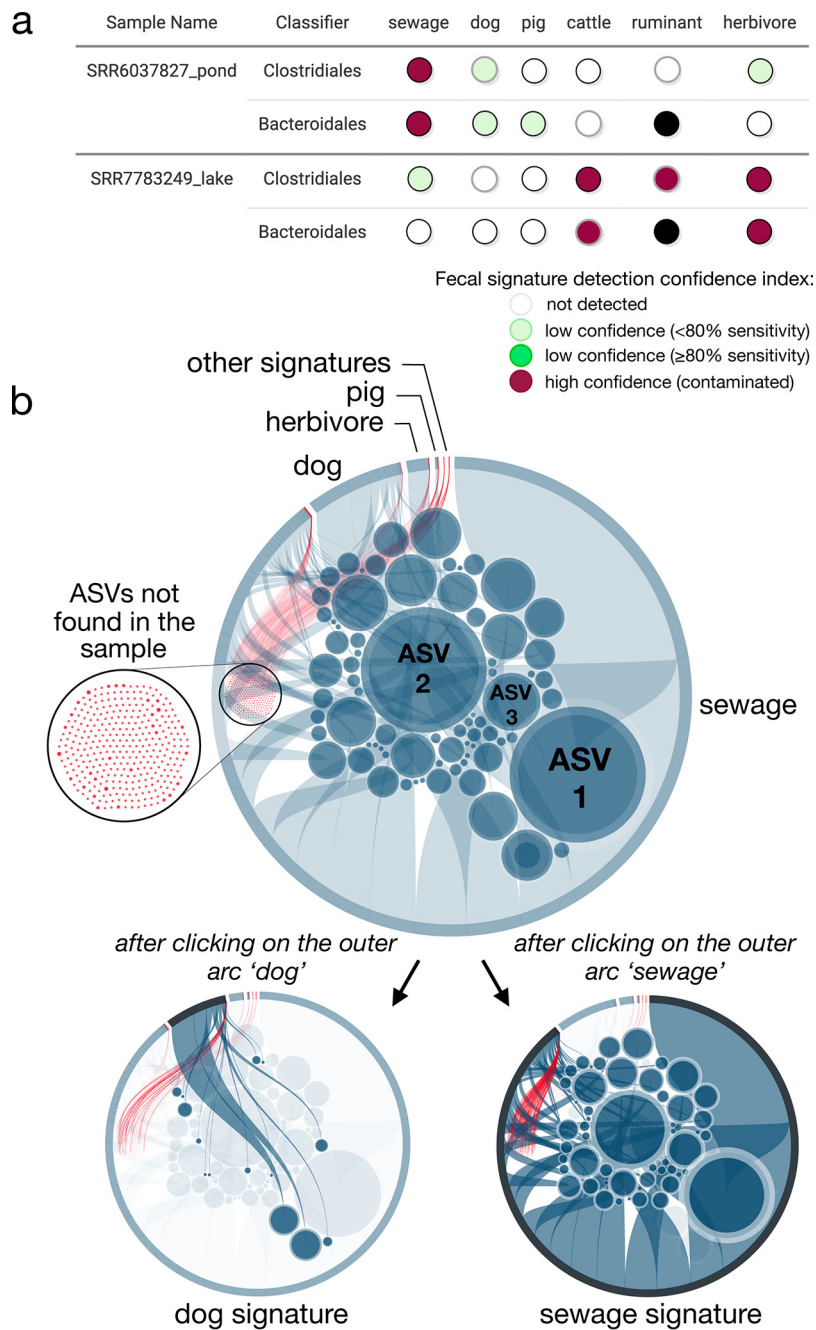


FIG 3 Forensic interactive report, including (a) a source identification predictions table and (b) bubble plot to characterize the fecal signature in the submitted samples. (a) Colors indicate if the fecal source signature was characterized (red) or not (white or green) in the tested sample. Dark green sources indicate that at the voting tree probability observed, the specificity of the classifier was at least 80%; light green sources indicate a specificity of <80%. Global classifiers are represented by black rings. Discarded classifiers are symbolized using black circles. (b) Bubbles represent all the amplicon sequence variants (ASVs) that compose the classifiers for a given bacterial group. Blue bubbles show the ASVs recovered from the sample tested, unlike red bubbles. The sizes of the bubbles are proportional to the relative abundances of the ASV in the tested sample. Sources are represented by the outer arcs. The longer the arc is, the more that source contributed to the fecal pollution (among the sources investigated). The example shows the V4 *Clostridiales* profile recovered from a Chinese surface-water pond (SRA accession number [SRR6037827](https://www.ncbi.nlm.nih.gov/sra/SRR6037827)) and classified as sewage by random forest. A total of 74 out of the 385 ASVs that compose all the classifiers were found in the sample; 55 and 14 ASVs were associated with the sewage and dog classifiers, respectively. The largest arc (representing 91% of the relative abundance of all the fecally associated ASVs) was associated with the sewage signature (bottom right). The second largest was associated with the dog and the third with the pig signature. The interactive version of this figure is available at <https://forensic.sfs.uwm.edu/result/example>. Data from Hägglund et al. and Hu et al. (16, 63). See Table S5 in the supplemental material for the full list of the individual predictions.

Consistent with these findings, we observed minimal geographic sample partitioning for the whole *Bacteroidales* and *Clostridiales* assemblages, which demonstrates that these classifiers could have global relevance.

The FORENSIC sewage classifiers had a sensitivity of 100%, meaning that all tested sewage samples were correctly classified. These samples were collected from three continents, which confirms that sewage carries a robust human fecal signature (36, 37) and that the classifiers are potentially applicable to cities worldwide with modern sewer infrastructure. The sewage signature also is unique and differentiated from the other nonintegrative (individual animal) sources, as no animal fecal samples were identified as sewage and no sewage samples were classified as an animal source.

Given that the microbial community composition of a single fecal sample can be quite distinct from that of an integrative sample (36), we wanted to test whether this approach could correctly identify sources from samples encompassing the variability expected among individual sources. By creating classifiers using only the 15 to 20 most similar samples and then testing their performance using the most dissimilar samples, we were able to examine this scenario. The random forest approach generated high sensitivity and specificity for most fecal pollution sources we tested, demonstrating the consistency and cohesion of the host fecal signature as characterized by a minimum number of samples. Misclassifications generally involved sources with similar diets or samples with atypical fecal microbiota. Given that fecal pollution in waterways is generally derived from multiple animals, we anticipate that the application of FORENSIC for actual water sample testing would produce few false-negative results and that sporadic individual atypical animals would not influence the results.

For two animal sources (cattle and dog), the community composition data split into two very distinct community types, possibly driven by the diet of the animal or other factors such as cohabitation, age, or individual variation. For instance, a large dissimilarity in the cattle microbial community composition has been observed between cattle that are grass fed versus those that are grain fed (38, 39). Therefore, specific animal groups, such as cattle, that are fed distinct diets or have other distinct lifestyle characteristics may not be suitable for a single-marker-gene or classifier approach for tracking their impact. Instead, multiple classifiers used in combination may be needed to cover the diversity of a single animal group.

Global classifiers implemented in FORENSIC were characterized from a minimum of 15 source samples and had a specificity and sensitivity of 70% and a decision cutoff of $>10\%$. Classifiers that did not meet these requirements were either discarded or designated draft classifiers. Cats and dogs had microbial communities similar to that of sewage, indicating shared sequences (Fig. S1). This may be due to shared diet or to actual pet or wildlife inputs into the sewer conveyance system. We did not attempt to make a cat classifier, and two of the four dog classifiers were either discarded or designated draft.

FORENSIC is implemented with random forest classification, which is particularly suitable for an online application, especially since the classifiers do not require each new investigation to reanalyze *de novo* bacterial assemblages in all trained and tested samples. The global classifiers of FORENSIC are based on the shared/common source signatures, increasing access to sequence-based analysis for users who do not have source samples. SourceTracker (21), a well-established tool, relies upon direct comparison of a source community to the sink community and requires both communities to be used as inputs. Users would need to provide representative samples, or chose samples from publicly available data. Analysis is performed *de novo*, i.e., without a reference database of what to expect, and therefore does not require that a source contain a universal signature. SourceTracker is highly sensitive, particularly for atypical sources or uncharacterized sources, when the source community is available. Furthermore, since the entire community is used for comparison, SourceTracker may be more suitable to identify animals where additional taxonomic groups provide a more distinctive signature than *Bacteroidales* or *Clostridiales*. Closely related animals (cattle and horses), unique geographic signatures, or the presence of uncharacterized sources, particularly wildlife, should be considered when choosing a tool.

The scalability of FORENSIC makes it possible to add more samples to the reference database as they become available. This characteristic will allow for systematic verification of the classifier performance, with a goal of continuous improvement. Overall, future directions include (i) extending the source classifiers to include farm animals or waterfowl and targeting predominant bacterial groups appropriate to these sources, (ii) strengthening broad diet/physiology classifiers, or, inversely, (iii) characterizing finer subsource clusters to improve the sensitivity of the classifiers.

Our study also highlights that sequencing depth is crucial to capturing fecal signatures within environmental samples. Even at relatively significant contamination levels, the fecal microorganisms can be outnumbered by the native microorganisms by >2 orders of magnitude. We recommend implementing our approach on deeply sequenced data (i.e., >100,000 reads) and using sequencing platforms that generate shorter reads (e.g., V4 or V6 regions of the 16S rRNA gene) and therefore greater depth. However, FORENSIC correctly identified the source of fecal pollution in environmental samples with lower sequencing depth where the fecal signature was represented with a minimum of 1,000 sequences and 100 unique representative sequences. This rate of correct identification equates to correctly identifying the fecal signature in 100,000 reads if the signature comprises 1% of the community.

When generating the classifier, the short V4 reads performed as well as the longer V4V5 reads, which supports previous reports that minimal information was needed for resolution of microbial members (40, 41). Although primers for the region targeted are often cited for introducing bias in the assessment of microbial communities (42, 43), we demonstrated the equivalence of using V4V5 sequences trimmed to the Earth Microbiome Project V4 primer positions and of using *de novo* V4 sequences for the specific FORENSIC application. This result is congruent with previous studies that reported a limited bias for several phyla, including *Bacteroidetes* and *Firmicutes* (44, 45), likely because primer sites are conserved within closely related groups.

In summary, FORENSIC offers a comprehensive open-source platform that enables rapid and straightforward screening of large sequence data sets to identify fecal contamination. In the era of affordable and accessible high-throughput sequencing, FORENSIC is a valuable tool for community-based water quality assessments to prospect the main sources of fecal pollution and guide appropriate management actions.

MATERIALS AND METHODS

Sample collection and processing. Animal stool samples were collected in the United States, France, the Dominican Republic, and Australia. Samples collected in the United States were preserved in lysis buffer from the stool extraction kit and shipped to the laboratory on ice and stored at -80°C upon arrival. Dominican Republic samples were stored at -20°C until extracted. Australian fecal sample processing and storage were described previously, and extracted DNA was provided (46). Animal stool samples from France were freeze-dried prior to DNA extraction. A human fecal signature was characterized from sewage influent samples from 71 cities across the United States (36), Reus in Spain (47), and Sydney in Australia. Finally, seven river samples collected after heavy rainfall in the Milwaukee River Basin, Milwaukee, WI, USA, were used to test the sensitivity of the classifiers (26). Sewage and animal fecal sample details are reported in Data Set S1.

Sample processing and DNA extraction. For the U.S. and Dominican Republic fecal samples, bacterial DNA was extracted from approximately 1 g of material using a QIAmp DNA stool minikit according to the manufacturer's instructions (Qiagen, Valencia, CA) as described (25). About 200 mg of Australian stool samples was extracted using the QIAmp stool DNA kit, while about 500 mg of dry fecal sample for French samples was extracted using the FastDNA spin kit for soil (MP Biomedicals, Carlsbad, CA). Australian and French extracted DNA was freeze-dried before being shipped to the United States.

In total, 25 ml for sewage influent and between 200 and 400 ml for freshwater samples was filtered onto 0.22- μm mixed cellulose ester filters with a 47-mm diameter (Millipore, USA). DNA from filters was then extracted using the FastDNA spin kit for soil (MP Biomedicals, Solon, OH) as previously described (26, 36). DNA concentration was assessed on a Qubit 2.0 fluorometer (Invitrogen, Carlsbad, CA).

16S rRNA gene sequencing and library construction. Amplicon libraries were constructed at the Josephine Bay Paul Center in the Marine Biological Laboratory (Woods Hole, MA) and/or the Great Lake Genomics Center (Milwaukee, WI) using the MiSeq Illumina platform for the V4 and V4V5 hypervariable regions and the HiSeq and/or NextSeq Illumina platforms for V6 hypervariable regions. Details for amplicon library construction and sequencing procedures for the V4V5 and V6 regions were followed as previously described (48, 49). The construction of the V4 amplicon library using the Earth Microbiome primers is described in Text S1 in the supplemental material.

Reads were trimmed using cutadapt v1.14 (50), allowing for two to four mismatches in the primer sequence depending on the region. Forward and reverse reads were merged using PEAR v0.9.10 (51) with the default parameters. Merged reads with a length shorter or longer than 5% of the V4V5 median (372 bp) or 10% of the V4 and V6 median (i.e., 253 and 60 bp, respectively) were removed during this step. Quality filtering was assessed using mothur 1.39.5 (52). Assembled reads containing ambiguous bases or with more than eight successive homopolymers were discarded. Sequences were taxonomically assigned based on the best match in a Global Alignment for Sequence Taxonomy (GAST) (53) process using the Silva 132 database (54). Only sequences assigned to *Bacteroidales* and *Clostridiales* were selected for further analysis.

Minimum entropy decomposition analysis. Minimum entropy decomposition (MED) analysis was performed using the oligotyping pipeline version 2.1 (55). MED uses nucleotide entropy (nucleotide variant variability) along DNA sequences to distinguish differences in nucleotides originating as a result of true genetic variation among organisms from noise due to sequencing errors (56). MED partitions DNA sequences into amplicon sequence variants (ASVs) according to the position in the DNA sequence with the highest entropy. This step iteratively lasts until each final ASV satisfies the maximum entropy criterion. ASVs that do not meet the minimum substantive abundance (M) criterion were discarded. M was set to $N/30,000$, $N/40,000$, and $N/60,000$, for the V4V5, V4, and V6 data sets, respectively, where N was the total number of sequences in the data set.

Selection of the samples to train and test the classifiers. Sequencing quality was systematically evaluated for each sewage and animal fecal sample before creating or testing the classifiers. For that, a prospective MED analysis was performed for both *Bacteroidales* and *Clostridiales* using all the samples of a given source. Samples in which more than 50% of the reads were discarded during the MED analysis for at least one of the bacterial groups were deleted from further analysis. The total numbers of amplicon sequence variants (ASVs) were \log_{10} transformed and plotted in a box plot. Lower outlier samples, i.e., those lower than 1.5 times the interquartile, were considered to have low sequencing depth and were discarded. We also verified for each source that there was no sample with an extremely atypical bacterial assemblage using Bray-Curtis dissimilarities computed separately for each source and bacterial group via the *vegan* package (57) implemented in R. Intrasource dissimilarities were averaged per sample and plotted in a box plot; upper outliers were not used to evaluate the performance of the classifier but are reported separately in Data Set S1. These extreme outliers comprised less than 8% of the total samples and, in some cases, had Bray-Curtis dissimilarity scores of >0.95 compared to other samples from a given host, which could have resulted from sample handling, external contamination, or poor sequencing quality.

Each classifier was built using 15 to 20 samples having the most similar intrasource bacterial assemblages (based on Bray-Curtis dissimilarity index). The validation of the classifier predictions was assessed with a minimum of six samples (30% minimum of the samples per source) having the most dissimilar intrasource bacterial assemblages to ensure a stringent assessment of performance. We systematically explored the intrasource variability of the bacterial group assemblages using a hierarchical clustering analysis (based on the Bray-Curtis dissimilarity and Ward linkage) in order to take into account environmental factors that may result in bifurcation of gut microbiota structure within sources. We defined subgroups within sources when the branch length between two clusters for a given source was greater than five. A relevant shift was observed in the V6 cattle (*Bacteroidales* and *Clostridiales*) and V6 dog (*Bacteroidales*) bacterial assemblages. A balanced number of samples from each cluster was used to build the V6 cattle and dog (*Bacteroidales*) classifiers. Diet and physiology classifiers were trained and tested using the collection of samples for each respective category (e.g., ruminant trained and tested using cattle and deer samples).

Random forest classifications. To characterize the fecal source signature for a classifier, bacterial assemblages from a given source (e.g., sewage) were compared to the assemblages from all other source samples (e.g., nonsewage samples from dog, cattle, etc.). For that, 100 random forests consisting of 10,000 trees were computed per source and per bacterial group using the default settings of the 'randomForest' function implemented in the *randomForest* R package (58). Mean decreases in Gini were averaged for each ASV among the 100 random forest replicates. Mean decreases in Gini were plotted in a scree plot. ASVs with mean decreases in Gini above the breakpoint curve were chosen to be part of the classifier. Breakpoints were estimated using the "breakpoints" function included in the *strucchange* R package (59). Classifiers were trained with 100 forests constituted of 1,000 trees to recognize the fecal source signatures from the ASVs selected upstream (per source and bacterial group, the total relative abundance of the ASVs was of 100%). Replicates were pooled using the "combine" function. One classifier was created for each source, bacterial group, and region of the 16S rRNA gene. To identify the fecal source signature in a test sample, sequences matching the ones in the classifiers were first extracted. Bacterial profiles were then compared to the ones stored in the classifiers using the "predict" function, which generated a voting tree probability, translating the degree of similarity between the test and the classifier profiles. We defined the voting tree probability cutoff (called the decision cutoff) to maximize the sensitivity and specificity with a minimum of 70% for each classifier. Samples associated with a voting tree higher than the decision cutoff for a given source were considered to be contaminated by that source. In addition to the source-specific classifiers (e.g., sewage, cattle), we created diet/physiology classifiers from multiple fecal source samples: V4/V4V5 ruminant classifiers were created using cattle and deer fecal samples, and a V4 herbivore classifier was created using cattle, deer, and horse fecal samples. Although V6 herbivore classifiers were created using ruminant fecal samples (i.e., cattle and deer), their specificity to only detect the ruminant fecal signature was low compared to that for

detecting the herbivore signature (data not shown). Thus, these classifiers were considered herbivore classifiers. All analyses were conducted using the statistical environment R (version 3.6.0) (60).

Sensitivity, specificity, and library size detection limit of random forest classifications. Sensitivity was calculated as the number of true positives (TPs) (i.e., when a source was correctly detected) divided by the sum of TPs and false negatives (FNs) (i.e., when a source was not detected as expected). In contrast, specificity was defined as the number of true negatives (TNs) (i.e., when a source did not have a classifier and thus should have been classified as “unknown”) divided by the sum of TNs and false positives (FPs) (i.e., when an unexpected source was detected). Classifier performance for specificity was evaluated using the test samples from host groups that were the focus of the classifier, but with also seven additional animal fecal sources; see Data Set S1 for details. The minimum number of sequences at which a classifier can correctly classify a fecal sample was assessed using an *in silico* analysis. All V4 and V6 test samples were subsampled at 90, 80, 70, 60, 50, 40, 30, 20, 10, 5, 1, 0.5, 0.1, 0.05, 0.01, 0.005, and 0.001%. The subsampling was performed using the function “sub.sample” in mothur 1.39.5. Additionally, the minimum number of sequences belonging to the classifier signatures at which a classifier can correctly identify multiple fecal sources was assessed using artificial bacterial assemblages generated *in silico*. Defined sequence proportions from pristine surface freshwater and fecal samples (i.e., sewage, cattle, pig, and sheep) not used to train the classifiers were combined to create V4 and V6 artificial bacterial community mixes (see Table S2 in the supplemental material). Sequences from each sample were selected by randomly subsampling (99 repeats) using the “rrarefy” function included in the *vegan* R package.

Implementation of the online interface. The FORENSIC Web interface (<https://forensic.sfs.uwm.edu/>) was designed using free and open software development practices in mind. The source code of FORENSIC was implemented in Python programming language using the Flask Web Framework (<https://palletsprojects.com/p/flask/>) for easy maintenance. FORENSIC can run as a Web application behind a reverse proxy (such as NGinx) and on Web server (such as Apache; <http://apache.org/>), and can be deployed on any Unix-based operating system. The reverse proxy handles Hypertext Transfer Protocol Secure (HTTPS) encryption for security, sanitization of malformed inputs, and efficient data upload. FORENSIC requires the upload of a multi-FASTA file and assumes the sequencing reads to have been trimmed and merged. The Python Web application detects the 16S rRNA region targeted in sequences and generates a count table of the ASVs matching the sequences in the classifiers. After the classification, the Web platform generates a static report using the data generated. The interactive report page visualizations are created using D3.js (<https://d3js.org/>) and our *ad hoc* JavaScript code. The data used for visualizations can also be directly downloaded from the interface as tab-separated data files for user-directed analyses. The web platform also notifies the user via e-mail when analysis is complete if an e-mail address has been provided during submission. A unique identifier (ID) is also provided during the submission to review the results for up to 1 year.

Data availability. The source code of FORENSIC is available online (https://github.com/mclellanlab/web_forensic) and licensed with the General Public License to allow local and institutional deployments. The R scripts to create, train, and test the classifiers are available on figshare, as well as the script to determine the decision cutoff (<https://doi.org/10.6084/m9.figshare.11231627>). Raw read files were uploaded to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and are accessible under the BioProject accession numbers [PRJNA261344](#), [PRJNA264400](#), [PRJNA433408](#), [PRJNA591968](#), and [PRJNA591970](#) for V4V5; [PRJNA591976](#) and [PRJNA591975](#) for V4; and [PRJNA235337](#), [PRJNA433407](#), [PRJNA591949](#), and [PRJNA505345](#) for V6. Data Set S1 contains the list of samples and their corresponding SRA study accession numbers.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TEXT S1, DOCX file, 0.02 MB.

FIG S1, TIF file, 2.2 MB.

FIG S2, TIF file, 2.2 MB.

FIG S3, TIF file, 2.9 MB.

TABLE S1, XLSX file, 0.1 MB.

TABLE S2, XLSX file, 0.01 MB.

TABLE S3, XLSX file, 0.05 MB.

TABLE S4, XLSX file, 0.01 MB.

TABLE S5, XLSX file, 0.03 MB.

DATA SET S1, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Warish Ahmed and Lucas Françoise for providing us with animal fecal sample DNA. We also thank Mahdi Belcaid and Jill S. McClary for their insightful discussions.

This project was funded under National Institutes of Health (NIH) grant R01 AI091829 to S.L.M.

REFERENCES

- World Health Organization. 2011. Guidelines for drinking-water quality, 4th ed. World Health Organization, Geneva, Switzerland.
- Santo Domingo JW, Ashbolt NJ. 2008. Fecal pollution of water, 1–12. In Cleveland CJ (ed), Encyclopedia of Earth. National Council for Science and the Environment, Washington, DC.
- DeFlorio-Barker S, Wing C, Jones RM, Dorevitch S. 2018. Estimate of incidence and cost of recreational waterborne illness on United States surface waters. *Environ Heal A Glob Access Sci Source* 17:3.
- Soller JA, Schoen ME, Bartrand T, Ravenscroft JE, Ashbolt NJ. 2010. Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. *Water Res* 44:4674–4691. <https://doi.org/10.1016/j.watres.2010.06.049>.
- Schoen ME, Ashbolt NJ. 2010. Assessing pathogen risk to swimmers at non-sewage impacted recreational beaches. *Environ Sci Technol* 44:2286–2291. <https://doi.org/10.1021/es903523q>.
- Willey JM, Sherwood LM, Woolverton CJ. 2017. Proteobacteria, p 504–534. In Prescott's microbiology, 10th ed. McGraw-Hill, New York, NY.
- Harwood VJ, Staley C, Badgley BD, Borges K, Korajkic A. 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. *FEMS Microbiol Rev* 38:1–40. <https://doi.org/10.1111/1574-6976.12031>.
- Unno T, Staley C, Brown CM, Han D, Sadowsky MJ, Hur H. 2018. Fecal pollution: new trends and challenges in microbial source tracking using next-generation sequencing. *Environ Microbiol* 20:3132–3140. <https://doi.org/10.1111/1462-2920.14281>.
- Vierheilig J, Savio D, Farnleitner AH, Reischer GH, Ley RE, Mach RL, Farnleitner AH, Reischer GH. 2015. Potential applications of next generation DNA sequencing of 16S rRNA gene amplicons in microbial water quality monitoring. *Water Sci Technol* 72:1962–1972. <https://doi.org/10.2166/wst.2015.407>.
- Henry R, Schang C, Coutts S, Kolotelo P, Prosser T, Crosbie N, Grant T, Cottam D, O'Brien P, Deletic A, McCarthy D. 2016. Into the deep: evaluation of SourceTracker for assessment of faecal contamination of coastal waters. *Water Res* 93:242–253. <https://doi.org/10.1016/j.watres.2016.02.029>.
- Brown CM, Staley C, Wang P, Dalzell B, Chun CL, Sadowsky MJ. 2017. A high-throughput DNA sequencing approach to determine sources of fecal bacteria in a Lake Superior estuary. *Environ Sci Technol* 51:8263–8271. <https://doi.org/10.1021/acs.est.7b01353>.
- Comte J, Berga M, Severin I, Logue JB, Lindström ES. 2017. Contribution of different bacterial dispersal sources to lakes: population and community effects in different seasons. *Environ Microbiol* 19:2391–2404. <https://doi.org/10.1111/1462-2920.13749>.
- McCarthy D, Jovanovic D, Lintern A, Teakle I, Barnes M, Deletic A, Coleman R, Rooney G, Prosser T, Coutts S, Hipsey MR, Bruce LC, Henry R. 2017. Source tracking using microbial community fingerprints: method comparison with hydrodynamic modelling. *Water Res* 109:253–265. <https://doi.org/10.1016/j.watres.2016.11.043>.
- Staley C, Kaiser T, Lobos A, Ahmed W, Harwood VJ, Brown CM, Sadowsky MJ. 2018. Application of SourceTracker for accurate identification of fecal pollution in recreational freshwater: a double-blinded study. *Environ Sci Technol* 52:4207–4217. <https://doi.org/10.1021/acs.est.7b05401>.
- Roguet A, Eren AM, Newton RJ, McLellan SL. 2018. Fecal source identification using random forest. *Microbiome* 6:185. <https://doi.org/10.1186/s40168-018-0568-3>.
- Hägglund M, Bäckman S, Macellaro A, Lindgren P, Borgmästars E, Jacobsson K, Dryselius R, Stenberg P, Sjödin A, Forsman M, Ahlinder J. 2018. Accounting for bacterial overlap between raw water communities and contaminating sources improves the accuracy of signature-based microbial source tracking. *Front Microbiol* 9:2364. <https://doi.org/10.3389/fmicb.2018.02364>.
- Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, Gordon JI. 2008. Evolution of mammals and their gut microbes. *Science* 320:1647–1652. <https://doi.org/10.1126/science.1155725>.
- Kostic AD, Howitt MR, Garrett WS. 2013. Exploring host-microbiota interactions in animal models and humans. *Genes Dev* 27:701–718. <https://doi.org/10.1101/gad.212522.112>.
- Yoon SS, Kim EK, Lee WJ. 2015. Functional genomic and metagenomic approaches to understanding gut microbiota-animal mutualism. *Curr Opin Microbiol* 24:38–46. <https://doi.org/10.1016/j.mib.2015.01.007>.
- Nishida AH, Ochman H. 2018. Rates of gut microbiome divergence in mammals. *Mol Ecol* 27:1884–1897. <https://doi.org/10.1111/mec.14473>.
- Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST. 2011. Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* 8:761–763. <https://doi.org/10.1038/nmeth.1650>.
- Dethlefsen L, McFall-Ngai MJ, Relman DA. 2007. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* 449:811–818. <https://doi.org/10.1038/nature06245>.
- Fisher JC, Eren AM, Green HC, Shanks OC, Morrison HG, Vineis JH, Sogin ML, McLellan SL. 2015. Comparison of sewage and animal fecal microbiomes by using oligotyping reveals potential human fecal indicators in multiple taxonomic groups. *Appl Environ Microbiol* 81:7023–7033. <https://doi.org/10.1128/AEM.01524-15>.
- 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. <https://doi.org/10.1038/ismej.2014.97>.
- Feng S, Ahmed W, McLellan SL. 2020. Ecological and technical mechanisms for cross reaction of human fecal indicators with animal hosts. *Appl Environ Microbiol* 86:e02319-19. <https://doi.org/10.1128/AEM.02319-19>.
- Olds HT, Corsi SR, Dila DK, Halmo KM, Bootsma MJ, McLellan SL. 2018. High levels of sewage contamination released from urban areas after storm events: a quantitative survey with sewage specific bacterial indicators. *PLoS Med* 15:e1002614. <https://doi.org/10.1371/journal.pmed.1002614>.
- 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, Kosciolk 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, Consortium TEMP. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551:457–463. <https://doi.org/10.1038/nature24621>.
- McLellan SL, Eren AM. 2014. Discovering new indicators of fecal pollution. *Trends Microbiol* 22:697–706. <https://doi.org/10.1016/j.tim.2014.08.002>.
- 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. <https://doi.org/10.1038/nrmicro1978>.
- Zou Y, Xue W, Luo G, Deng Z, Qin P, Guo R, Sun H, Xia Y, Liang S, Dai Y, Wan D, Jiang R, Su L, Feng Q, Jie Z, Guo T, Xia Z, Liu C, Yu J, Lin Y, Tang S, Huo G, Xu X, Hou Y, Liu X, Wang J, Yang H, Kristiansen K, Li J, Jia H, Xiao L. 2019. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat Biotechnol* 37:179–185. <https://doi.org/10.1038/s41587-018-0008-8>.
- La Reau AJ, Meier-Kolthoff JP, Suen G. 2016. Sequence-based analysis of the genus *Ruminococcus* resolves its phylogeny and reveals strong host association. *Microb Genom* 2:e000099. <https://doi.org/10.1099/mgen.0.000099>.
- Krishnan S, Alden N, Lee K. 2015. Pathways and functions of gut microbiota metabolism impacting host physiology. *Curr Opin Biotechnol* 36:137–145. <https://doi.org/10.1016/j.copbio.2015.08.015>.
- Reischer GH, Kasper DC, Steinborn R, Farnleitner AH, Mach RL. 2007. A quantitative real-time PCR assay for the highly sensitive and specific detection of human faecal influence in spring water from a large alpine catchment area. *Lett Appl Microbiol* 44:351–356. <https://doi.org/10.1111/j.1472-765X.2006.02094.x>.
- Feng S, McLellan SL. 2019. Highly specific sewage-derived *Bacteroides* quantitative PCR assays target sewage-polluted waters. *Appl Environ Microbiol* 85:e02696-18. <https://doi.org/10.1128/AEM.02696-18>.
- Feng S, Bootsma M, McLellan SL. 2018. Novel human-associated *Lachnospiraceae* genetic markers improve detection of fecal pollution sources in urban waters. *Appl Environ Microbiol* 84:e00309-18. <https://doi.org/10.1128/AEM.00309-18>.
- Newton RJ, McLellan SL, Dila DK, Vineis JH, Morrison HG, Murat Eren A,

- Sogin ML. 2015. Sewage reflects the microbiomes of human populations. *mBio* 6:e02574–14. <https://doi.org/10.1128/mBio.02574-14>.
37. 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. <https://doi.org/10.1111/1462-2920.12092>.
 38. Shanks OC, Kely CA, Archibeque S, Jenkins M, Newton RJ, McLellan SL, Huse SM, Sogin ML. 2011. Community structures of fecal bacteria in cattle from different animal feeding operations. *Appl Environ Microbiol* 77:2992–3001. <https://doi.org/10.1128/AEM.02988-10>.
 39. Hagey JV, Bhatnagar S, Heguy JM, Karle BM, Price PL, Meyer D, Maga EA. 2019. Fecal microbial communities in a large representative cohort of California dairy cows. *Front Microbiol* 10:1093. <https://doi.org/10.3389/fmicb.2019.01093>.
 40. Jeraldo P, Chia N, Goldenfeld N. 2011. On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys. *Environ Microbiol* 13:3000–3009. <https://doi.org/10.1111/j.1462-2920.2011.02577.x>.
 41. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* 35:e120. <https://doi.org/10.1093/nar/gkm541>.
 42. Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG. 2015. Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol* 6:771. <https://doi.org/10.3389/fmicb.2015.00771>.
 43. Rintala A, Pietilä S, Munukka E, Eerola E, Pursiheimo JP, Laiho A, Pekkala S, Huovinen P. 2017. Gut microbiota analysis results are highly dependent on the 16S rRNA gene target region, whereas the impact of DNA extraction is minor. *J Biomol Tech* 28:19–30. <https://doi.org/10.7171/jbt.17-2801-003>.
 44. Mao DP, Zhou Q, Chen CY, Quan ZX. 2012. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC Microbiol* 12:66. <https://doi.org/10.1186/1471-2180-12-66>.
 45. Ghyselinck J, Pfeiffer S, Heylen K, Sessitsch A, De Vos P. 2013. The effect of primer choice and short read sequences on the outcome of 16S rRNA gene based diversity studies. *PLoS One* 8:e71360. <https://doi.org/10.1371/journal.pone.0071360>.
 46. Ahmed W, Goonetilleke A, Powell D, Chauhan K, Gardner T. 2009. Comparison of molecular markers to detect fresh sewage in environmental waters. *Water Res* 43:4908–4917. <https://doi.org/10.1016/j.watres.2009.09.047>.
 47. Fisher JC, Levican A, Figueras MJ, McLellan SL. 2014. Population dynamics and ecology of *Arcobacter* in sewage. *Front Microbiol* 5:525. <https://doi.org/10.3389/fmicb.2014.00525>.
 48. Morrison HG, Grim SL, Vineis JH, Sogin ML. 2013. 16S amplicon Illumina sequencing methods. *figshare* <https://doi.org/10.6084/m9.figshare.833944.v1>. Retrieved 12 Dec 2017.
 49. 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. <https://doi.org/10.1371/journal.pone.0066643>.
 50. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17:10–12. <https://doi.org/10.14806/ej.17.1.200>.
 51. Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30:614–620. <https://doi.org/10.1093/bioinformatics/btt593>.
 52. 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. <https://doi.org/10.1128/AEM.01541-09>.
 53. Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Welch DM, Relman DA, Sogin ML. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* 4:e1000255. <https://doi.org/10.1371/journal.pgen.1000255>.
 54. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:590–596. <https://doi.org/10.1093/nar/gks1219>.
 55. 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. <https://doi.org/10.1111/2041-210X.12114>.
 56. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. 2015. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* 9:968–979. <https://doi.org/10.1038/ismej.2014.195>.
 57. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin RP, O'Hara RB, Simpson GL, Solymos P, Stevens MH, Szoecs E, Wagner H. 2017. *vegan: community ecology package version 2.5-6*.
 58. Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* 2/3:18–22.
 59. Zeileis A, Leisch F, Homik K, Kleiber C. 2002. strucchange: an R package for testing for structural change in linear regression models. *J Stat Softw* 7:1–38.
 60. R Core Team. 2019. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
 61. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874. <https://doi.org/10.1093/molbev/msw054>.
 62. Letunic I, Bork P. 2016. Interactive Tree of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44:W242–W245. <https://doi.org/10.1093/nar/gkw290>.
 63. Hu A, Li S, Zhang L, Wang H, Yang J, Luo Z, Rashid A, Chen S, Huang W, Yu C-P. 2018. Prokaryotic footprints in urban water ecosystems: a case study of urban landscape ponds in a coastal city, China. *Environ Pollut* 242:1729–1739. <https://doi.org/10.1016/j.envpol.2018.07.097>.