AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology

Metagenomic Approach for Identification of the Pathogens Associated with Diarrhea in Stool Specimens

Yanjiao Zhou,[a]* Kristine M. Wylie,[a] Rana E. El Feghaly,[b] Kathie A. Mihindukulasuriya,[c] Alexis Elward,[a] David B. Haslam,[d] Gregory A. Storch,[a] George M. Weinstock[c]*

Department of Pediatrics, Washington University School of Medicine, St. Louis, Missouri, USA[a]; Department of Pediatrics, University of Mississippi Medical Center, Jackson, Mississippi, USA[b]; McDonnell Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA[c]; Division of Infectious Disease, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA[d]

The potential to rapidly capture the entire microbial community structure and/or gene content makes metagenomic sequencing an attractive tool for pathogen identification and the detection of resistance/virulence genes in clinical settings. Here, we assessed the consistency between PCR from a diagnostic laboratory, quantitative PCR (qPCR) from a research laboratory, 16S rRNA gene sequencing, and metagenomic shotgun sequencing (MSS) for *Clostridium difficile* identification in diarrhea stool samples. Twenty-two *C. difficile*-positive diarrhea samples identified by PCR and qPCR and five *C. difficile*-negative diarrhea controls were studied. *C. difficile* was detected in 90.9% of *C. difficile*-positive samples using 16S rRNA gene sequencing, and *C. difficile* was detected in 86.3% of *C. difficile*-positive samples using MSS. CFU inferred from qPCR analysis were positively correlated with the relative abundance of *C. difficile* from 16S rRNA gene sequencing ($r^2 = -0.60$) and MSS ($r^2 = -0.55$). *C. difficile* was codetected with *Clostridium perfringens*, norovirus, sapovirus, parechovirus, and anellovirus in 3.7% to 27.3% of the samples. A high load of *Candida* spp. was found in a symptomatic control sample in which no causative agents for diarrhea were identified in routine clinical testing. Beta-lactamase and tetracycline resistance genes were the most prevalent (25.9%) antibiotic resistance genes in these samples. In summary, the proof-of-concept study demonstrated that next-generation sequencing (NGS) in pathogen detection is moderately correlated with laboratory testing and is advantageous in detecting pathogens without *a priori* knowledge.

Sequencing technology has revolutionized infectious diseases research over the past decade. Whole-genome sequencing (WGS) of pure cultures has been widely used for pathogen characterization, evolutionary studies, transmission investigations, and outbreak detection (1, 2, 3). WGS of cultured isolates is now moving from the proof-of-concept phase to implementation. The two major applications of WGS of cultured strains in clinical diagnostic microbiology are molecular epidemiology and antibiotic resistance gene prediction (4). In contrast to the WGS sequencing of cultured isolates, metagenomics assesses a community of organisms but eliminates the isolation step. This can be done by focusing on a specific conserved gene, such as the 16S rRNA gene, or by the metagenomic shotgun sequencing (MSS) of total microbial nucleic acids within samples. For the purpose of this study, metagenomics sequencing refers to either 16S rRNA gene sequencing or MSS; 16S rRNA gene sequencing and MSS using next-generation sequencing (NGS) platforms produce large quantities of data in a relatively short time. Although 16S rRNA gene sequencing is less expensive than MSS, it suffers from potential PCR-related bias. Taxonomical classification based on partial 16S rRNA gene sequencing is generally limited to phylum to genus level specificity. Nevertheless, highly heterogeneous species within certain genera can be distinguished (5). In addition to the relatively high cost, the large amount of sequence data generated by MSS requires significant computing resources for data processing and storage. However, MSS, without the bias inherent to PCR, is capable of classifying bacteria to the species or strain level. It can also detect viruses, fungi, and other microbial components, some of which cannot yet be cultured (6). MSS has been used to identify pathogens, including known and novel viruses that cause diarrhea or fever (7, 8). Recently, MSS of cerebrospinal fluid from a comatose patient with a congenital immunodeficiency revealed the uncommon pathogen *Leptospira santarosai* after extensive standard testing had not yielded an etiologic agent (9).

Using metagenomic sequencing, the Human Microbiome Project (HMP) demonstrated that millions of microbes coexist with their healthy hosts (10). In such individuals, many of these microbes maintain symbiotic relationships with their hosts, including assisting in food digestion and immune modulation (11). Some microbes may be residing latently or subclinically in a healthy host but may cause disease at a later time. For example, opportunistic organisms, such as *Clostridium difficile*, *Staphylococcus aureus*, *Acinetobacter baumannii*, and *Candida albicans*, can affect people with compromised immune systems but often colonize without causing disease. Similarly, many viruses were detected in the healthy subjects from the HMP cohort, including herpesviruses and papillomaviruses (12). While not causing overt

Received 22 July 2015 **Returned for modification** 19 August 2015
Accepted 23 November 2015

Accepted manuscript posted online 4 December 2015

Citation Zhou Y, Wylie KM, El Feghaly RE, Mihindukulasuriya KA, Elward A, Haslam DB, Storch GA, Weinstock GM. 2016. Metagenomic approach for identification of the pathogens associated with diarrhea in stool specimens. J Clin Microbiol 54:368–375. doi:10.1128/JCM.01965-15.

Editor: P. H. Gilligan

Address correspondence to George M. Weinstock, george.weinstock@jax.org.

* Present address: Yanjiao Zhou, The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA; George M. Weinstock, The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA.

disease at the time of sampling, these viruses can become problematic if the subject becomes immunocompromised or if cofactors predispose to cancer. Additionally, MSS has uncovered alterations in microbial communities that are associated with a wide range of disease (13, 14, 15). For example, compared to healthy controls, intestinal dysbiosis is evident in patients with diarrhea caused by *C. difficile* or other intestinal pathogens (16).

Another attribute of whole microbial profiling is that it can be used to identify coinfecting agents within a clinical specimen. Current clinical approaches for investigating the copresence of pathogens use multiplex PCR. However, MSS is not limited to targeted organisms and can potentially identify the cooccurrence of a wide panel of organisms. To date, few clinical metagenomic sequencing studies have investigated pathogen cooccurrence in clinical specimens.

Here, we conducted a proof-of-concept study with the goal of evaluating the concordance of metagenomic sequencing and diagnostic and research laboratory testing in pathogen identification. Twenty-two *C. difficile*-positive and 5 *C. difficile*-negative diarrhea stool samples (by laboratory testing) were sequenced using 16S rRNA gene sequencing and MSS. *C. difficile* and viruses identified from different approaches were compared.

MATERIALS AND METHODS

Diarrhea stool sample collection. Stool samples were obtained from a previous study (17, 18) that identified inflammatory markers and viral copathogens during *C. difficile* infection. Stool samples from patients with diarrhea were collected from inpatient, outpatient, and emergency department visits at St. Louis Children's Hospital (SLCH) between July 2011 and July 2012. Patients were ≤18 years old with a variety of underlying diseases. Patients with stool residuals of <1 ml were excluded from the study. This study was approved by the Institutional Review Board of the Washington University School of Medicine.

Pathogen detection by the diagnostic laboratory and research laboratory. In the clinical diagnostic laboratory, a glutamate dehydrogenase enzyme immunoassay (EIA) (Wampole C. diff Quik Chek; Alere, Orlando, FL) was used to screen for *C. difficile*. Positive samples were confirmed by GeneXpert *C. difficile* PCR (Cepheid, Sunnyvale, CA). The two assays were performed according to the manufacturers' instructions. We also performed quantitative PCR (qPCR) to evaluate the abundance of *C. difficile* in the samples in our research laboratory, as described previously (19). In brief, SYBR green-based real-time PCR was performed using the 7500 Fast real-time PCR system (Applied Biosystems) by including 10 μl of fast SYBR green master mix (Applied Biosystems), 0.5 μM primers, and 50 to 100 ng of nucleic acid in a 20 μl PCR. Monoplex TaqMan reverse transcription-PCR (RT-PCR) was also performed to detect the presence of norovirus and sapovirus as described previously (17, 20). In brief, primer/probe sets, reaction buffers, and a 100 ng template were mixed in a final 25-μl reaction volume. The RT-PCR was performed for 10 min at 45°C (reverse transcription temperature), 10 min at 95°C (*Taq* polymerase activation), 45 cycles of 15 s at 95°C, and 1 min at 60°C. Twenty-two *C. difficile* PCR-positive samples in addition to 5 samples from patients with diarrhea whose *C. difficile* PCRs were negative in the clinical diagnostic laboratory were selected for 16S rRNA gene sequencing and whole-genome shotgun sequencing.

Metagenomic sequencing and analysis. Total nucleic acid (DNA and RNA) was extracted using the NucliSens easyMag automated system (bioMérieux, Marcy l'Etoile, France) according to the manufacturer's instructions. In brief, samples were placed in the sample vessel and were followed by lysis incubation. Magnetic silica was added to the samples followed by the automatic extraction. 16S rRNA gene sequencing and MSS were performed in the McDonnell Genome Institute at the Washington University School of Medicine. Preparation of 16S rRNA gene libraries, sequencing, and data processing followed the standard operational protocols of the Human Microbiome Project (HMP) consortium (21). Briefly, the V3 to V5 region of the 16S rRNA gene was amplified using primers 357F (5′-CCTACGGGAGGCAGCAG-3′) and 926R (5′-CCGTCAATTCMTTTRAGT-3′). PCR was performed with the following conditions: 30 cycles of 95°C for 2 min, 50°C for 0.5 min, and 72°C for 5 min. Amplicons were purified, pooled at equimolar concentrations, and pyrosequenced on the Roche 454 Titanium platform. Samples were binned by allowing one mismatch in the barcode. Low-quality reads (average quality of <35 for a read), short reads (<200 bp), and reads with chimeric 16S rRNA gene sequences were removed. High-quality sequences were classified from the phylum to genus levels by the Ribosomal Database Project Naive Bayesian Classifier version 2.5 using training set 9. As *C. difficile* is distinct from other *Clostridia* based on the 16S rRNA gene (<97% identity), we further classified *Clostridium* reads to *C. difficile* by blasting them against a clostridial database that we constructed by incorporating all *Clostridium* species in the RDP (https://rdp.cme.msu.edu/classifier/classifier.jsp) and Silva (http://www.arb-silva.de/) databases. The top hit with at least 97% identity and 97% coverage to the reference was designated the *Clostridium* species for a 16S rRNA gene sequence. If a read had the same bit score for more than one *Clostridium* species, it was designated an unclassified *Clostridium* spp. To avoid read depth biasing the detection of *C. difficile*, all samples were subsampled to 3,000 reads/sample.

For MSS, single-indexed sequencing libraries were constructed from total nucleic acid with insert sizes of 300 to 500 bp. In brief, total nucleic acid was subjected to reverse transcription and second strand synthesis to convert the RNA to DNA using random primers (22). The DNA was then sheared using the Covaris instrument, and library construction was performed using standard methods for end repair, A-tailing, adaptor ligation, and amplification using the Phusion enzyme (NEB). Libraries were pooled (7 to 8 samples per lane). MSS was performed on the Illumina HiSeq platform, and 100 base-paired end reads were generated. MSS reads were subjected to quality trimming, host contamination removal, and low-complexity region masking. The subsequent sequences were aligned to microbial databases using RTG mapping (Real Time Genomics) against >5,000 reference genomes (23) with the following parameters: –repeat-freq 97% -e 10% -T 4 –w 15 –n 255. Alignments against bacterial and fungal genomes were performed with the unique mapping mode of RTG, in which only the reads uniquely aligned to a reference genome were used for bacterial and fungal species identification. The species relative abundances were normalized by taking into account the number of reads and the length of the reference genomes that the reads hit. For virus identification, alignments were performed as described previously (12). Briefly, a nucleotide sequence alignment was performed with RTG (–repeat-freq 97% -e 10% -T 4 –w 15 –n 255 –top-random). Unaligned sequences were further interrogated for viruses. Translated alignments were carried out using MBLASTX software (MulticoreWare) (24) against a database of translated sequences from all of the viral reference genomes with the following parameters: -m 32 – e 1e-02 –I 50. Virus sequences were confirmed to be unambiguously viral by realignment to larger nucleotide (NT) and nonredundant (NR) databases using RTG mapping and MBLASTX with the same parameters described above. Sequences were counted as viral only if there were no similar alignments to other taxonomic divisions. Because the single-index sequencing libraries were pooled, some incorrect binning of sequences was expected (25). In order to address this conservatively, we disregarded relatively low virus counts from samples in the same pool with a sample that had a relatively high number of reads for the same virus.

To determine the presence of resistance genes in the metagenomic samples, human-free and high-quality WGS reads were mapped to the Antibiotic Resistance Genes Database (ARDB) (http://ardb.cbcb.umd.edu/). The resistance gene was defined as present when the reads had 100% identity to the reference gene, and the reference gene was covered 100% in length by the reads mapped to the gene.

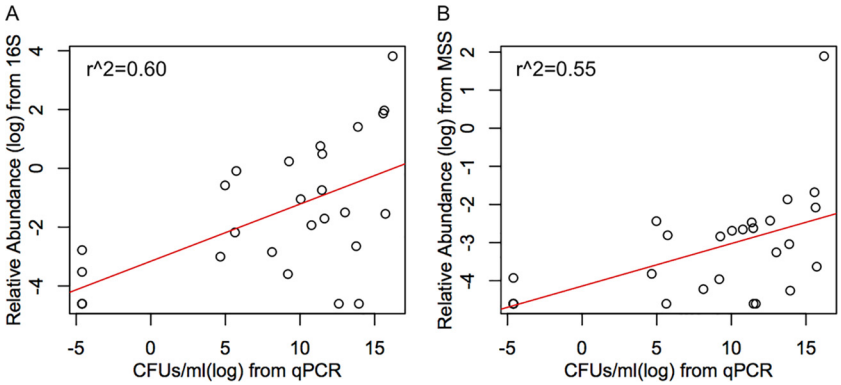

FIG 1 Correlation of qPCR with metagenomic sequencing in detection of *C. difficile* in the diarrhea samples. CFU derived from qPCR were positively correlated with the relative abundances of *C. difficile* detected by 16S rRNA gene sequencing (A) and MSS (B).

Molecular validation of pathogens identified by sequencing. PCR primers Cdiff16s-F (5′-AGCTCTTGAAACTGGGAGACTTGAG-3′) and Cdiff16s-R (5′-AGGGAACTCTCCGATTAAGGAGATGTC-3′), designed to amplify the 16S rRNA gene of *C. difficile* (26), were used to confirm the presence of *C. difficile* in samples that were *C. difficile* negative by qPCR (detected the *tcdB* gene) but positive by sequencing. Real-time PCR was performed (27) to detect *Salmonella enterica* in samples that were *S. enterica* negative in the diagnostic laboratory (by culture) but positive by sequencing. Parechovirus and anellovirus, which were discovered by MSS, were further validated by PCR as described previously (28, 29).

Nucleotide sequence accession number. All reads were deposited in the Sequence Read Archive database at NCBI under accession number PRJNA293986.

RESULTS AND DISCUSSION

Comparison of *C. difficile* detection by metagenomic sequencing and qPCR. To determine the concordance between sequencing and molecular-based techniques in the detection of *C. difficile*, 22 *C. difficile*-positive stool samples from patients with diarrhea detected by PCR in the diagnostic laboratory and qPCR in our research laboratory were selected for sequencing with 16S rRNA gene sequencing and MSS. We also sequenced five *C. difficile*-negative stool samples (by EIA and PCR) from the patients who had diarrhea. These samples served as symptomatic controls for diarrhea caused by *C. difficile*. The potential causes, based on cultures and medical records, for the diarrhea in symptomatic controls were *Campylobacter* and *Salmonella* infections, drug side effect, inflammatory bowel disease (IBD), and unknown, respectively.

The relative abundances of *C. difficile* ranged from 0.02% to 45.4% as measured by 16S rRNA gene sequencing in *C. difficile*-positive samples. CFU (range, 106 to 10,957,641/ml) calculated from qPCR (17) were positively correlated with the relative abundances of *C. difficile* from 16S rRNA gene sequencing (Pearson correlation, $r^2 = -0.60$; $P = 0.001$) (Fig. 1A), which corroborated that the two approaches to *C. difficile* quantification produced similar results. Specifically, *C. difficile* was detected by 16S rRNA gene sequencing in 20 (90.9%) of the 22 samples that were qPCR positive (threshold cycle [C_T] value of <46) (Table 1). Two samples in which *C. difficile* was not detected by 16S rRNA gene sequencing (C_T values of 29.7 and 31.5) produced an abundance of 16S rRNA gene reads (4,813 and 11,468, respectively), so sampling depth was not an issue.

Surprisingly, we also detected a sparse *C. difficile* presence by 16S rRNA gene sequencing in two symptomatic control samples. The clinical diagnoses for these two samples were drug side effect and *Salmonella* infection. The *C. difficile* reads were blasted against the NT database to further validate the specificity of the taxon calling. *C. difficile* was the top hit with a high identity (>97%), which suggests that those reads are likely from *C. difficile*. Because the qPCR was negative for the *tcdB* gene and 16S rRNA genes are indistinguishable between toxigenic and nontoxigenic *C. difficile*, we first reasoned that these reads may be from nontoxigenic *C. difficile*. Primers designed to specifically amplify the *C. difficile* 16S rRNA gene were used to validate the presence of *C. difficile* regardless of the toxin genes. PCR assay for the 16S rRNA gene was negative for the two symptomatic control samples. The detection of *C. difficile* by 16S sequencing but the lack of confirmation by PCR from the original samples suggests that the *C. difficile* reads may be from contamination in different steps of the study. Because the PCR of the *C. difficile*-specific 16S rRNA gene is a gel-based assay, the rareness of *C. difficile* in the samples (only 5 and 7 reads were detected in the 16S rRNA gene sequencing) can also lead to the negative observation from the gel. The main goal of the study is to assess the general concordance of pathogen identification by sequencing and laboratory testing. The discordance in the above samples prompts us to further investigate the factors (sequencing depth and source of contamination) in greater detail in future study. In addition, because 16S rRNA gene sequencing does not differentiate toxigenic and nontoxigenic *C. difficile*, 16S rRNA gene sequencing used for the detection of *C. difficile* may have a similar utility as the EIA in the diagnostic laboratory.

We detected *Campylobacter* and *Salmonella* by 16S rRNA gene sequencing in two symptomatic *C. difficile*-negative but *Campylobacter*- and *Salmonella*-positive samples.

As shown in Fig. 1B, the abundances of *C. difficile* from MSS agreed with the qPCR results (Pearson correlation, −0.55) and showed the same trend as 16S rRNA gene sequencing (Pearson correlation, 0.98). MSS successfully detected *C. difficile* in all samples with C_T values of <20, 86.7% of samples with C_T values of 20 to 35, and 75% of samples with C_T values of 35 to 46. Three samples that were qPCR positive were negative by MSS (Table 1), but these samples had the lowest MSS read depth, which suggested that the inability to detect *C. difficile* by MSS in these cases may have been due to insufficient read depth. We also detected a low

TABLE 1 Detection of the copresence of bacteria and viruses in the diarrhea samples

Sample identification	*C. difficile* detection by:			*C. perfringens* detection by 16S + MSS	Norovirus detection by:		Sapovirus detection by:	
	qPCR	16S	MSS		qRT-PCR	MSS	qRT-PCR	MSS
CDAF.131.131	+[a]	+	+	−[b]	−	−	−	−
CDAF.136.136	+	+	+	−	−	−	+	−
CDAF.137.137	+	+	−	−	−	−	−	+
CDAF.139.139	+	−	+	−	−	−	−	−
CDAF.142.142	+	+	+	−	−	−	−	−
CDAF.143.143	+	+	+	−	+	+	−	−
CDAF.178.178	+	+	+	−	+	+	−	−
CDAF.180.180	+	+	+	−	−	−	−	−
CDAF.193.193	+	+	−	−	−	−	−	−
CDAF.198.198	+	−	+	−	−	+	−	−
CDAF.218.218	+	+	+	−	−	−	−	−
CDAF.224.224	+	+	+	−	−	−	−	−
CDAF.230.230	+	+	+	−	−	−	−	−
CDAF.231.231	+	+	+	−	+	−	−	−
CDAF.243.243	+	+	+	−	+	−	−	−
CDAF.245.245	+	+	+	−	−	−	−	−
CDAF.267.267	+	+	+	−	−	−	−	−
CDAF.41949.A	+	+	+	+	+	+	−	−
CDAF.41951.C	+	+	+	−	−	−	+	+
CDAF.41953.E	+	+	+	−	−	−	−	−
CDAF.41955.G	+	+	+	−	−	−	−	−
CDAF.41958.J - *C. difficile* + *Salmonella*	+	+	−	−	−	−	−	−
CDAF.41950.B -NC[c] (medicine side effect)	−	+[d]	−	+	−	−	−	−
CDAF.41952.D-NC (inflammatory bowel disease)	−	−	−	+	−	−	−	−
CDAF.41954.F-NC (*Campylobacter*)	−	−	+[d]	−	−	−	+	+
CDAF.41956.H-NC (unknow cause)	−	−	−	−	−	−	−	−
CDAF.41957.I-NC (Salmonella)	−	+[d]	−	+	−	−	−	−

[a] +, Present in the sample.
[b] −, Not present in the sample.
[c] NC, negative control.
[d] Detected by sequencing but not confirmed by 16S rRNA gene PCR.

abundance of *C. difficile* by MSS in the sample from the *Campylobacter* control in which *C. difficile* was not detected by PCR. Alignment of the sequences to the NT database confirmed the specificity of the reads to *C. difficile*. However, a gel-based PCR with amplification for the 16S rRNA gene from the original samples failed to support the presence of *C. difficile*. This may be due to the same artifact noted above. MSS successfully detected *Campylobacter* and *Salmonella* in two controls that were known to contain these agents by PCR and were also detected by 16S rRNA gene analysis.

We also performed reverse transcription-quantitative PCR (qRT-PCR) to determine if norovirus and sapovirus were present in these samples and compared the detection sensitivity with MSS. Five samples were norovirus positive and 3 samples were sapovirus positive by qRT-PCR (Table 1), four samples were norovirus positive and 2 samples were sapovirus positive by MSS, and norovirus and sapovirus were detected in 3 and 2 sample by qRT-PCR and MSS, respectively. The correlation between MSS and qRT-PCR in viral detection was low. This is probably because viral genomes are small and, therefore, viral nucleic acid often accounts for a relatively small proportion of the total nucleic acid from a sample if the virus is not abundant and because the MSS procedure in this study did not include the viral enrichment step that is sometimes used for viral discovery. Our previous work showed that sequencing depth affects the sensitivity of viral detection in clinical samples. Increased sequence depth (i.e., 20 million reads/sample) strengthens vial signals and allows for novel viral detection (8).

In summary, the targeted 16S rRNA gene sequencing and the MSS showed moderate correlation in *C. difficile* identification compared to that of diagnostic laboratory and research laboratory testing. The consistency of the MSS and qRT-PCR was lower for the detection of low-abundance organisms, such as viruses. One limitation of this study is its small sample size, especially because it included relatively few virus-positive samples. Future studies with larger sample sizes will provide more insights into the sensitivity of PCR and MSS in the detection of viral pathogens. Discordance between sequence-positive and PCR-negative samples deserves further investigation.

Whole microbiome community revealed by MSS. Figure 2 illustrates the microbial community compositions and abundances from the 27 diarrhea samples using MSS. *C. difficile* and any organisms present in greater abundance than *C. difficile* were included in the heatmap. First, the relative abundance of *C. difficile* in the bacterial communities from MSS varied widely, ranging from 0.005% to 6.7% of total reads in the *C. difficile*-positive samples. It is not clear what level of *C. difficile* can cause diarrhea, but our recent study showed that the load of *C. difficile* was not associated with clinical outcome (19). Second, the microbial communities were quite distinct in the *C. difficile*-positive samples

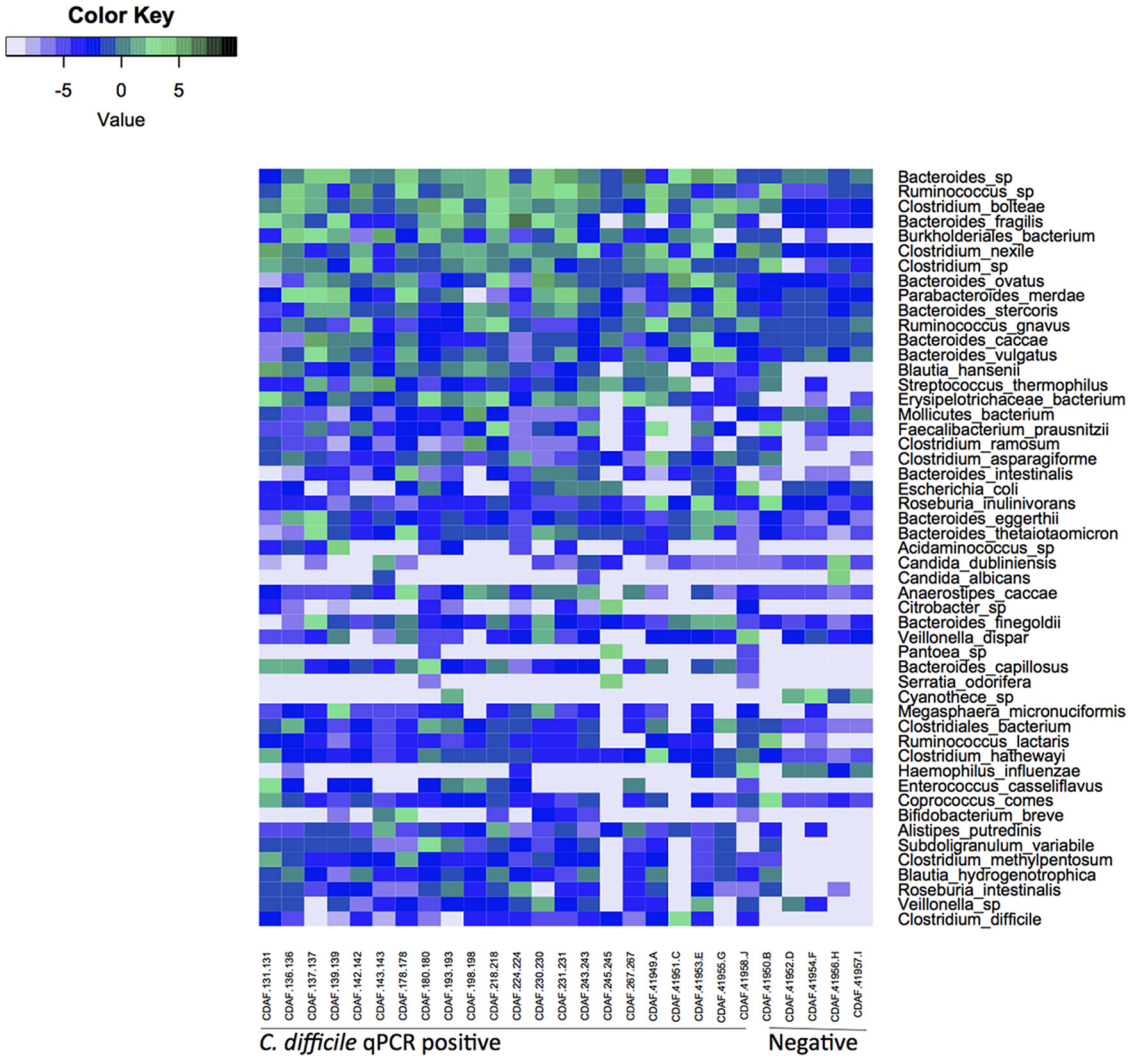

FIG 2 Microbial community profile of the diarrhea samples revealed by MSS. The distribution of *C. difficile* and the taxa whose relative abundances are higher than that of *C. difficile* are illustrated by heatmap. Each row represents a taxon, and each column represents a sample. The samples are in the same order as Table 1. Relative abundances with log10 transformation are used in the heatmap.

(Fig. 2). The dominant species in the majority of the samples were commensal gut flora, including *Bacteroides* spp. and *Ruminococcus* spp., which are the major enterotypes identified in healthy human stool (30). We also found that one patient sample was dominated by *Candida* spp. (35.5% of relative abundance). Interestingly, this patient was a symptomatic *C. difficile*-negative control patient without another clear cause of diarrhea. We further found this patient was treated with several antibiotics, including gentamicin, nafcillin, rifampin, trimethoprim-sulfamethoxazole, and vancomycin in the 3 months before diarrhea occurred. It is unclear whether fecal domination with *Candida* is a cause of diarrhea or simply a consequence of antibiotic therapy (31), but either observation has clinical relevance and would not have been identified by the cultures or PCR-based diagnostic studies typically performed in the clinical laboratory on stool samples.

Diverse microbial communities from patients with the same clinical symptoms are not surprising, as the microbiota are highly variable even between healthy subjects (32). Age, geographical location, diet, and environmental factors all potentially affect microbial community structure. The high intersubject variation of the bacterial communities in a diarrheal condition may reflect the inherent variation of gut microbiota before the patients had diarrhea. Antibiotic usage, long-term diet, and the underlying diseases in those patients may also contribute to the microbial variation between patients in the disease status.

Detection of pathogen copresence in diarrhea samples by MSS. A major advantage of metagenomic sequencing for pathogen identification is its potential to detect simultaneous coinfection with multiple pathogens, including bacteria and viruses. Few studies have reported the frequency of pathogenic bacterial coinfection with *C. difficile* infection. In this study, we focused on *Clostridium perfringens* to determine its copresence with *C. difficile* because it is a common clinically diagnosed bacterial pathogen that causes diarrhea. In addition, 16S rRNA gene sequencing is capable of identifying *C. perfringens* at the species level (33). Considering the difficulty of detecting low-abundance organisms using the metagenomic approach, the presence of *C. perfringens* was designated only when the organism was identified by both the 16S

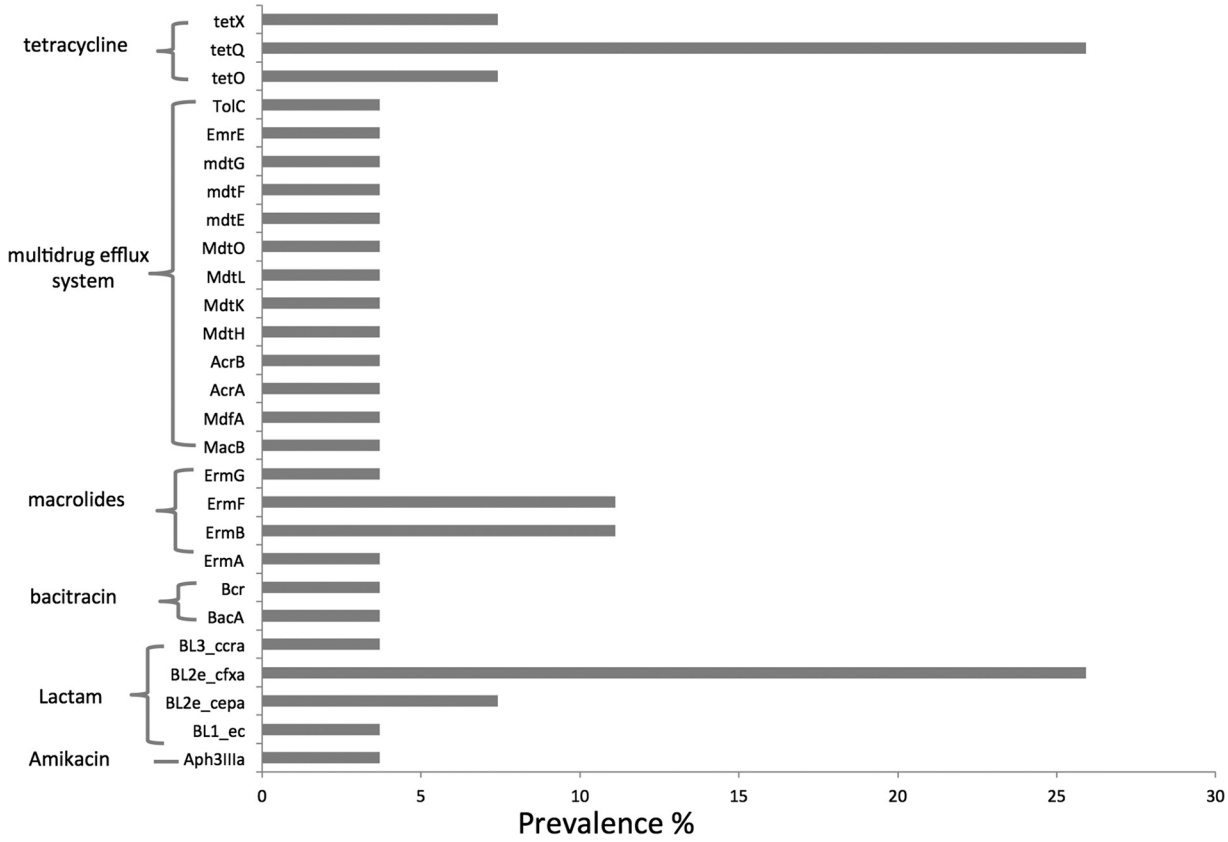

FIG 3 Prevalence of antibiotic resistance genes. The prevalence of antibiotic resistance genes is illustrated by a bar plot. The antibiotic categories are listed on the left side of the bar plot.

rRNA gene approach and MSS. *C. perfringens* was found to be copresent with *C. difficile* in one *C. difficile*-positive sample (Table 1). We also found *C. perfringens* in a symptomatic control patient whose diarrhea was thought to be caused by medications. The detection of *C. perfringens* raises another etiologic possibility. *C. perfringens* was also detected in IBD- and *S. enterica*-symptomatic control samples. The presence of *C. perfringens* was further validated by aligning the reads to the NT database.

Viral pathogens were also detected in *C. difficile*-positive samples by MSS. In addition to norovirus and sapovirus detected by qRT-PCR assays, we also detected anellovirus and parechovirus using MSS. These two viruses were not tested by our diagnostic and research laboratories before sequencing. We later confirmed the presence of the two viruses by PCR assay, as described in the Materials and Methods. The four viral genera were detected in 27.3% (6/22) of *C. difficile*-positive samples and 1 symptomatic control. Norovirus was the most prevalent virus in these samples, as it was detected in 18.2% (4/22) of the *C. difficile*-positive samples. We also found a copresence of norovirus, *C. difficile*, and *C. perfringens* in 1 sample. Sapovirus was found in 1 *C. difficile*-positive sample and 1 symptomatic control sample with an unknown cause of diarrhea from the clinical lab. As described above, we found that *Candida* was predominant in this symptomatic control. Of the above viruses, only norovirus and sapovirus are associated with diarrhea (21). It is unclear whether they may be the primary or secondary cause of the symptoms observed in these patients. These viruses are also sometimes detected in asymptomatic individuals. Viral detection by multiplex PCR is widely used in clinical diagnostic laboratories. Because viral detection using MSS can detect unexpected and novel viruses, it should be considered an alternative tool for viral discovery, especially when antigen detection and PCR fail to detect such agents.

Of note, the accuracy of microbial identification from MSS depends on the completeness of the reference database and the relatedness of clinical query strains to the reference strains in the database. Furthermore, the sequencing depth is likely to affect the robustness of the metagenomic approach. Because of the difficulties in recovering the whole genome of a bacterium or virus from a complex metagenomic sample, the species identification is based on read depth and the coverage of the reference genome. Therefore, MSS data should be interpreted with caution, especially given the low abundances of the pathogens we found in some of the specimens. Finally, the interpretation of simultaneous detection of *C. difficile* along with other pathogenic bacteria and viruses in the same patient requires further study. The current analytical approach only supports their concomitant presence in the gut environment but does not indicate which of the agents is responsible for disease manifestations. Using approaches including multiplex PCR and sequencing to facilitate the diagnosis of infectious diseases provides greater understanding of the diseases while also raising the question of which is the real causative agent.

Antibiotic resistance prediction from metagenomic sequences. Using strict criteria to define the presence of antibiotic resistance genes, we identified 27 antibiotic resistance genes in our

samples, and 55.6% of the samples contained at least one such locus. The most prevalent antibiotic resistance genes were *Bl2e_cfxa* (25.9%) and *tetQ* (25.9%) (Fig. 3), encoding a class A beta-lactamase that confers resistance to cephalosporin and tetracycline resistance, respectively. *ermA*, *ermB*, *ermF*, and *ermG* genes, which are responsible for resistance to macrolide antibiotics, were also identified in 3.7% to 11.1% of the samples. *tet* genes are the most common resistance genes identified in stool samples from healthy adults (34, 35). Indeed, a recent study indicated that tetracycline, beta-lactamases, and multiple drug resistance genes were commonly found in the stool of children <12 months of age (36). We also identified genes encoding multidrug efflux system proteins in one sample. Whole-genome shotgun sequencing of cultured bacteria revealed antibiotic resistance phenotypes with high accuracy. MSS has the capability to identify the resistance genes in the whole bacterial community. To pin down the bacterial origin of the resistance, deep sequencing and subsequent assembly of the bacterial genome or other alternative approaches are needed.

Conclusion. In summary, MSS correlates well with standard clinical diagnostic laboratory testing and qPCR in a research laboratory in its ability to identify *C. difficile*. It enables detection of multiple potential pathogens without *a priori* knowledge in clinical samples. Future amplicon-based sequencing targeting full-length 16S rRNA genes and rRNA internal transcribed spacers (ITS) (37) is likely to increase the resolving power of the taxonomic classification of bacteria. This ever-evolving sequencing technology aims to lower sequence cost, increase throughput, and decrease turnaround time. These developments will expedite the implementation of sequencing technology in diagnostic testing in the clinic.

ACKNOWLEDGMENTS

We thank Phillip Tarr and Carey-Ann Burnham for their careful and critical reading. We thank Sheila Mason and Richard Buller for their work on the PCR validation of sequencing results.

FUNDING INFORMATION

NIH provided funding to George Weinstock under grant number U54HG004968.

REFERENCES

1. **Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P.** 2013. Metagenomics for pathogen detection in public health. Genome Med **5:**81. http://dx.doi.org/10.1186/gm485.
2. **Capobianchi MR, Giombini E, Rozera G.** 2013. Next-generation sequencing technology in clinical virology. Clin Microbiol Infect **19:**15–22. http://dx.doi.org/10.1111/1469-0691.12056.
3. **Fournier PE, Drancourt M, Colson P, Rolain JM, La Scola B, Raoult D.** 2013. Modern clinical microbiology: new challenges and solutions. Nat Rev Microbiol **11:**574–585. http://dx.doi.org/10.1038/nrmicro3068.
4. **Koser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ.** 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog **8:**e1002824. http://dx.doi.org/10.1371/journal.ppat.1002824.
5. **Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ.** 2011. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A **108**(Suppl)**:**S4680–S4687.
6. **Wylie KM, Truty RM, Sharpton TJ, Mihindukulasuriya KA, Zhou Y, Gao H, Sodergren E, Weinstock GM, Pollard KS.** 2012. Novel bacterial taxa in the human microbiome. PLoS One **7**(6)**:**e35294. http://dx.doi.org/10.1371/journal.pone.0035294.
7. **Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, Wang D.** 2008. Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathog **4:**e1000011. http://dx.doi.org/10.1371/journal.ppat.1000011.
8. **Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA.** 2012. Sequence analysis of the human virome in febrile and afebrile children. PLoS One **7**(6)**:**e27735. http://dx.doi.org/10.1371/journal.pone.0027735.
9. **Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, Sokolic R, Garabedian E, Candotti F, Buckley RH, Reed KD, Meyer TL, Seroogy CM, Galloway R, Henderson SL, Gern JE, DeRisi JL, Chiu CY.** 2014. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. N Engl J Med **370:**2408–2417. http://dx.doi.org/10.1056/NEJMoa1401268.
10. **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.
11. **Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI.** 2011. Human nutrition, the gut microbiome and the immune system. Nature **474:**327–336. http://dx.doi.org/10.1038/nature10213.
12. **Wylie KM, Mihindukulasuriya KA, Zhou Y, Sodergren E, Storch GA, Weinstock GM.** 2014. Metagenomic Analysis Of Double-Stranded DNA Viruses in Healthy Adults. BMC Biol **12:**71. http://dx.doi.org/10.1186/s12915-014-0071-7.
13. **Cho I, Blaser MJ.** 2012. The human microbiome: at the interface of health and disease. Nat Rev Genet **13:**260–270.
14. **Madupu R, Szpakowski S, Nelson KE.** 2013. Microbiome in human health and disease. Sci Prog **96:**153–170. http://dx.doi.org/10.3184/003685013X13683759820813.
15. **Pflughoeft KJ, Versalovic J.** 2012. Human microbiome in health and disease. Annu Rev Pathol **7:**99–122. http://dx.doi.org/10.1146/annurev-pathol-011811-132421.
16. **Antharam VC, Li EC, Ishmael A, Sharma A, Mai V, Rand KH, Wang GP.** 2013. Intestinal dysbiosis and depletion of butyrogenic bacteria in *Clostridium difficile* infection and nosocomial diarrhea. J Clin Microbiol **51:**2884–2892. http://dx.doi.org/10.1128/JCM.00845-13.
17. **El Feghaly RE, Stauber JL, Tarr PI, Haslam DB.** 2013. Viral co-infections are common and are associated with higher bacterial burden in children with *Clostridium difficile* infection. J Pediatr Gastroenterol Nutr **57:**813–816. http://dx.doi.org/10.1097/MPG.0b013e3182a3202f.
18. **El Feghaly RE, Stauber JL, Tarr PI, Haslam DB.** 2013. Intestinal inflammatory biomarkers and outcome in pediatric *Clostridium difficile* infections. J Pediatr **163:**1697–1704. http://dx.doi.org/10.1016/j.jpeds.2013.07.029.
19. **El Feghaly RE, Stauber JL, Deych E, Gonzalez C, Tarr PI, Haslam DB.** 2013. Markers of intestinal inflammation, not bacterial burden, correlate with clinical outcomes in *Clostridium difficile* infection. Clin Infect Dis **56:**1713–1721. http://dx.doi.org/10.1093/cid/cit147.
20. **Grant L, Vinje J, Parashar U, Watt J, Reid R, Weatherholtz R, Santosham M, Gentsch J, O'Brien K.** 2012. Epidemiologic and clinical features of other enteric viruses associated with acute gastroenteritis in American Indian infants. J Pediatr **161:**110–115. http://dx.doi.org/10.1016/j.jpeds.2011.12.046.
21. **Human Microbiome Project Consortium.** 2012. A framework for human microbiome research. Nature **486:**215–221. http://dx.doi.org/10.1038/nature11209.
22. **Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J, Latreille JP, Wilson RK, Ganem D, DeRisi JL.** 2003. Viral discovery and sequence recovery using DNA microarrays. PLoS Biol **1:**E2. http://dx.doi.org/10.1371/journal.pbio.0000002.
23. **Martin J, Sykes S, Young S, Kota K, Sanka R, Sheth N, Orvis J, Sodergren E, Wang Z, Weinstock GM, Mitreva M.** 2012. Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. PLoS One **7:**e36427. http://dx.doi.org/10.1371/journal.pone.0036427.
24. **Davis CKK, Baldhandapani V, Gong W, Abubucker S, Becker E, Martin J, Wylie K, Khetani R, Hudson M, Weinstock G, Mitreva M.** 2013. mBLAST: keeping up with the sequencing explosion for (meta)genome analysis. J Data Mining Genomics Proteomics **4:**135.
25. **Kircher M, Sawyer S, Meyer M.** 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res **40:**e3. http://dx.doi.org/10.1093/nar/gkr771.

26. **Goncalves C, Decre D, Barbut F, Burghoffer B, Petit JC.** 2004. Prevalence and characterization of a binary toxin (actin-specific ADP-ribosyltransferase) from *Clostridium difficile*. J Clin Microbiol **42:**1933–1939. http://dx.doi.org/10.1128/JCM.42.5.1933-1939.2004.
27. **Chen J, Zhang L, Paoli GC, Shi C, Tu SI, Shi X.** 2010. A real-time PCR method for the detection of *Salmonella enterica* from food using a target sequence identified by comparative genomic analysis. Int J Food Microbiol **137:**168–174. http://dx.doi.org/10.1016/j.ijfoodmicro.2009.12.004.
28. **McElvania TeKippe E, Wylie KM, Deych E, Sodergren E, Weinstock G, Storch GA.** 2012. Increased prevalence of anellovirus in pediatric patients with fever. PLoS One **7:**e50937. http://dx.doi.org/10.1371/journal.pone.0050937.
29. **Nix WA, Maher K, Johansson ES, Niklasson B, Lindberg AM, Pallansch MA, Oberste MS.** 2008. Detection of all known parechoviruses by real-time PCR. J Clin Microbiol **46:**2519–2524. http://dx.doi.org/10.1128/JCM.00277-08.
30. **Zhou Y, Mihindukulasuriya KA, Gao H, La Rosa PS, Wylie KM, Martin JC, Kota K, Shannon WD, Mitreva M, Sodergren E, Weinstock GM.** 2014. Exploration of bacterial community classes in major human habitats. Genome Biol **15:**R66. http://dx.doi.org/10.1186/gb-2014-15-5-r66.
31. **Krause R, Schwab E, Bachhiesl D, Daxbock F, Wenisch C, Krejs GJ, Reisinger EC.** 2001. Role of *Candida* in antibiotic-associated diarrhea. J Infect Dis **184:**1065–1069. http://dx.doi.org/10.1086/323550.
32. **Zhou Y, Gao H, Mihindukulasuriya KA, La Rosa PS, Wylie KM, Vishnivetskaya T, Podar M, Warner B, Tarr PI, Nelson DE, Fortenberry JD, Holland MJ, Burr SE, Shannon WD, Sodergren E, Weinstock GM.** 2013. Biogeography of the ecosystems of the healthy human body. Genome Biol **14:**R1. http://dx.doi.org/10.1186/gb-2013-14-1-r1.
33. **Woo PC, Lau SK, Chan KM, Fung AM, Tang BS, Yuen KY.** 2005. *Clostridium* bacteraemia characterised by 16S ribosomal RNA gene sequencing. J Clin Pathol **58:**301–307. http://dx.doi.org/10.1136/jcp.2004.022830.
34. **Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, Pan Y, Li J, Zhu L, Wang X, Meng Z, Zhao F, Liu D, Ma J, Qin N, Xiang C, Xiao Y, Li L, Yang H, Wang J, Yang R, Gao GF, Wang J, Zhu B.** 2013. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nat Commun **4:**2151.
35. **Forslund K, Sunagawa S, Kultima JR, Mende DR, Arumugam M, Typas A, Bork P.** 2013. Country-specific antibiotic use practices impact the human gut resistome. Genome Res **23:**1163–1169. http://dx.doi.org/10.1101/gr.155465.113.
36. **Moore AM, Patel S, Forsberg KJ, Wang B, Bentley G, Razia Y, Qin X, Tarr PI, Dantas G.** 2013. Pediatric fecal microbiota harbor diverse and novel antibiotic resistance genes. PLoS One **8:**e78822. http://dx.doi.org/10.1371/journal.pone.0078822.
37. **Ruegger PM, Clark RT, Weger JR, Braun J, Borneman J.** 2014. Improved resolution of bacteria by high throughput sequence analysis of the rRNA internal transcribed spacer. J Microbiol Methods **105:**82–87. http://dx.doi.org/10.1016/j.mimet.2014.07.001.