August 2019

Development of Indicators for Human Fecal Pollution Using Deep-Sequencing of Microbial Communities

Shuchen Feng
University of Wisconsin-Milwaukee

DEVELOPMENT OF INDICATORS FOR HUMAN FECAL POLLUTION USING DEEP-SEQUENCING OF MICROBIAL COMMUNITIES

by

Shuchen Feng

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Freshwater Sciences

at

The University of Wisconsin-Milwaukee

August 2019

ABSTRACT

DEVELOPMENT OF INDICATORS FOR HUMAN FECAL POLLUTION USING DEEP-SEQUENCING OF MICROBIAL COMMUNITIES

by

Shuchen Feng

The University of Wisconsin-Milwaukee, 2019
Under the Supervision of Professor Sandra L. McLellan

The gut microbiome is shaped by both host physiology and environmental factors, which results in unique communities that contain certain members specific to a host. Microbial source tracking (MST) methods that rely on host-specific fecal microorganisms have been applied to detect human fecal pollution over the past two decades. The most common approach uses quantitative polymerase chain reaction (qPCR) to amplify certain sequences of these microorganisms, or human fecal markers. To date, most bacterial human fecal markers have focused on the HF183 cluster within the genus *Bacteroides*. However, cross-reactions with animals or variable *Bacteroides* abundance in humans have been found. In addition, the traditional clone library method to identify fecal markers targets order *Bacteroidales*, thereby excluding other taxonomic groups that might also contain host-specific members. Here we employed deep 16S ribosomal RNA (rRNA) gene next-generation sequencing (NGS) of sewage and animal fecal samples (n=469) to explore human-specific microorganisms. Multiple marker candidates were identified from the family *Lachnospiraceae* and non-HF183 cluster of *Bacteroides.* Assays were developed for two human-associated *Lachnospiraceae* (i.e., Lachno3 and Lachno12) and two sewer pipe-derived *Bacteroides* (i.e., BacV4V5-1 and BacV6-21). Validation studies of these

qPCR assays in host and non-host samples demonstrated their specificity to human fecal source. Low-level animal cross-reactions have been reported for all bacterial human fecal markers, including our newly identified human- and sewage-associated markers; however, the mechanism is poorly understood. We examined cross-reactivity in 180 animal fecal samples using NGS and qPCR assays (i.e., Lachno3, multiplexed *Escherichia coli* and human *Bacteroides*, and multiplexed *Enterococcus* spp. and BacV6-21). All three human fecal markers showed over 90% specificity in both NGS and qPCR results. Human marker cross-reactions could correlate with certain composition of its corresponding genus and could putatively correlate with environmental factors. In particular, discrepancies between NGS and qPCR marker positives could primarily be explained by amplification of the marker's closely-related organisms. Overall, this work provided a new generation of reliable human fecal markers, identified mechanisms for their cross-reactions both ecologically and technically and highlighted the utility of deep sequencing of microbial communities for MST method development.

To my husband and our son.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CN	Copy Numbers
CSO	Combined Sewer Overflow
Ct	Cycle threshold
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
ENT	Enterococci/*Enterococcus* spp.
FIB	Fecal Indicator Bacteria
LLOQ	Lower Limit of Quantification
MB	Method Blank
MST	Microbial Source Tracking
NGS	Next-generation Sequencing
PCR	Polymerase Chain Reaction
qPCR	Quantitative Polymerase Chain Reaction
SPC	Sample Processing Control
rRNA	Ribosomal Ribonucleic Acid
SS	Salmon Sperm
SSO	Sanitary Sewer Overflow
USEPA	United States Environmental Protection Agency
WWTP	Wastewater Treatment Plant

ACKNOWLEDGEMENTS

I would like to express my most sincere gratitude to my advisor, Professor Sandra L. McLellan, for her strong support and constant guidance for this work. Over the past years, I benefited hugely from her excellence as a scientist, mentor and working mother. She is always one of the role models of my career and family life.

I would like to thank the rest of my committee: Professor Ryan J. Newton, Professor Charles F. Wimpee, Dr. Orin C. Shanks and Dr. Matthew C. Smith. They made each committee meeting happen from different geographical locations in the U.S. as well as outside of the U.S.. Without their insightful suggestions, this work would not have been accomplished.

I thank all my current and former lab mates for their help and feedbacks. I would like to thank our lab manager, Deborah K. Dila, for proofreading this thesis and for the immense help she has provided to my daily work over the past years. I specially thank Dr. Adélaïde Roguet for her generous help with my bioinformatics work, my writing and presentations. I thank Dr. Patricia A. Bower and Melinda J. Bootsma, who have been taught me experimental skills since my first day in the lab and have always been available to answer my questions. I also thank Dr. Jill McClary in the Newton lab for her brilliant comments on my manuscripts and presentations.

Finally, I would like to thank my family: my grandparents, my parents, my elder brother, my parents-in-law, my husband and my son. Their continuous and unparalleled love, encouragement and support have motivated me all the time. I am lucky to have them in my life and I dedicate this milestone to them.

Chapter 1 Introduction

Waterborne diseases.

Waterborne diseases are usually caused by pathogenic microorganisms transmitted in water sources (1). Fecal pollution is one of the main sources for these waterborne pathogens (1, 2), such as pathogenic *Escherichia coli* (*E. coli*), *Salmonella, Cryptosporidium*, *Giardia* and norovirus (1, 3, 4). Some symptoms of waterborne diseases include gastroenteritis, respiratory infections, conjunctivitis and skin rash (2, 5–8). Among all populations, young children, the elderly, and those with weakened immune systems are the most affected (5, 9, 10). These diseases are not always self-limited; some infections cause high morbidity or even death (3, 5, 10).

In the United States (U.S.), waterborne pathogens caused 4.3 to 19.5 million cases of acute gastrointestinal illness annually through drinking water sources (9–11). The largest documented waterborne disease outbreak in the U.S. happened in 1993 in Milwaukee, Wisconsin. Caused by the human fecal pathogen *Cryptosporidium* in drinking water from Lake Michigan, the outbreak affected about 25% of Milwaukee residents and led to economic losses of over $96 million. The source of the pathogen has never been determined (12, 13). Recreational water is another main exposure route for waterborne pathogens (14, 15). During 2000 to 2014, there were 363 reported pathogen-related waterborne disease outbreaks in treated recreational waters (e.g., pools, hot tubs and water playgrounds) in the U.S.. Fifty-eight percent of these outbreaks were caused by *Cryptosporidium*, resulting in more than 21,600 cases (16). In untreated recreational waters (e.g., rivers, lakes and oceans), 95 outbreaks were reported, among which 84% were caused by enteric pathogens and led to more than 2,700 cases (17). However, reported numbers greatly underestimate the real incidence of illness cases, as surveillance is voluntary and sporadic cases or small

outbreaks may be unrecognized or unreported (5, 9, 18). In fact, it is estimated that 90 million recreational waterborne illnesses occurred annually in the U.S. with costs of 2.2 to 3.7 billion dollars (5).

Understanding the full scope of the frequency, prevalence and pathogenic agent of waterborne diseases is critical for development of public health risk assessments and preventive measures. Identifying fecal pollution presence is a key step in the process of accurately interpreting the source and distribution of waterborne pathogens in environmental waters.

Human fecal pollution as a major cause of waterborne diseases.

Urban watersheds often have multiple fecal pollution sources present (e.g., sewage, pets, wildlife and agricultural runoff) (19). It is generally agreed that human fecal pollution usually poses more health risk to the public than domestic and wild animal feces (20–22). This is assumed as a result of the "species barrier", where the types of pathogens that pose a health risk to human are fewer in animal feces than in human feces (20, 23). Pathogens that are derived from human fecal pollution enter water environments via various pathways. Some main pathways include combined sewer overflows (CSOs) and sanitary sewer overflows (SSOs), both of which discharge untreated sewage to surface water directly (24–26). It was reported that CSO and SSO events introduced more than 850 million gallons of untreated sewage into the U.S. waterways annually (24). Other pathways also deliver human pathogens into water environments, such as illicit cross connections between stormwater and sewer systems, and leaking sewer pipes that infiltrate to groundwater and stormwater systems (19, 24, 27). It was estimated that 23% of the nation's river and stream

miles and 31% of the nation's bays and estuaries are impaired, with pathogens from fecal pollution as one of the main causes (28).

Human fecal pollution in receiving water is a persistent issue in the U.S. and is ubiquitous in urbanized areas (29–31). This situation could be much worse for future generations, as the population and urbanization is increasing (32) while investment for new sewer infrastructure is insufficient (33). At the same time, climate change has been expected to add to the burden of waterborne diseases by increasing pathogen delivery to surface water via higher storm frequency and severity in certain regions (34–36). Reliable identification of human fecal pollution in waters is particularly important for microbial water quality assessment and public health protection, as well as reduction in economic losses.

General fecal indicator bacteria.

Direct monitoring for waterborne pathogens is challenging because it is difficult to identify the causative agent from the great variety of waterborne pathogens that are present in human fecal pollution (37, 38). Furthermore, waterborne pathogens have an uneven distribution and are usually in low concentrations in water environments, making it problematic to detect these organisms (37, 38). Over the past 100 years, the standard approach for microbial water quality assessment has been to monitor the concentrations of nonpathogenic general fecal indicator bacteria (FIB), which are very abundant in human feces and sewage (2, 37). These FIB include fecal coliforms, *E. coli* and enterococci, and have been used worldwide for microbial water quality assessment for recreational waters (2, 39–41). It has been reported that certain FIB levels are positively correlated with pathogen presence in freshwater (42–44) and marine water (45–47). However, there are

also many studies that have failed to establish direct or significant correlations between FIB levels and human pathogen levels or human health outcomes (8, 48–50). The significant positive relationship between FIB and pathogens can occur when the fecal pollution is dominated by the human source since humans contribute both FIB and pathogens (51). In urban water environments where multiple pollution sources are often present (e.g., stormwater runoff), FIB levels can be unrelated to pathogen concentrations, as fecal pollution from non-human sources, such as animal feces, also contribute to the FIB levels but do not introduce human pathogens (52). The health risks caused by these human fecal pathogens are usually much higher compared to animal sources (22, 53). Therefore, the inability of FIB to provide host source information can lead to inaccurate fecal pollution source identification and false public health risk assessment (31, 37). To solve this problem, host-specific alternative fecal indicators have been developed and were used to assess microbial water quality.

Microbial source tracking method and the application of 16S ribosomal RNA gene.

Microbial source tracking (MST) has been largely focused on determining fecal pollution sources in water environments (37, 54) and employs chemical (e.g., fecal steroids and artificial sweeteners) (55, 56), viral (e.g., F-specific RNA bacteriophages, human adenovirus) or bacterial indicators that distinguish the source of fecal pollution (e.g., host-specific members of the genus *Bacteroides*) (37, 54). An ideal fecal marker for MST should meet the following criteria: 1) the marker should be highly specific to its host source and be ubiquitously present in individuals of its host source; 2) the marker should be of high concentration in its host source to be easily detectable; 3) the marker should be of similar

or better persistence in the environment compared to FIB; and 4) the presence of the marker should be correlated with human pathogen in the same environment (54).

The 16S ribosomal RNA (rRNA) gene is approximately 1,500 base pairs long, with a composition of both conserved (i.e., a consistent sequence type within certain bacterial phylogenetic lineages) and hypervariable regions (i.e., different sequence types among taxa) (57). This gene has been applied as a target for MST fecal marker assays, which were developed from certain host-associated microorganisms. Some reasons include: 1) the 16S rRNA gene is a "gold standard" for reconstructing bacterial phylogenies due to its slow evolution rate (i.e., high degree of conservation) in bacterial cells; 2) it is universally present in bacterial genomes, usually with multiple copies in a single bacterium, making it more sensitive to detection than single copy genes; and 3) the V1- V9 hypervariable regions make it possible to use the 16S rRNA gene to characterize and cluster organisms of lower taxonomic levels (e.g., genus and species) (31). In particular, the degree of variability of hypervariable regions varies between different taxonomic lineages (58). This provides useful information for correlating organisms (e.g., species level or lower) with host niches.

Established *Bacteroides* fecal marker assays.

In as early as the 1980s, the genus *Bacteroides*, which is one of the most predominant genera in the human gut, was suggested as a potential indicator for human fecal pollution (59, 60). In 2000, one of the first *Bacteroides* marker assays for tracking human fecal pollution was developed targeting a specific sequence (designated as the HF183 marker) within this human-specific organism (61). The HF183 marker is located in the V2 hypervariable region of the 16S rRNA gene of the HF183 cluster of organisms, which was identified to include *Bacteroides dorei* (61). To date, the genus *Bacteroides* has

become one of the most characterized human-associated fecal genera, with many PCR/quantitative PCR (qPCR) assays developed. Most of the *Bacteroides* assays target the 16S rRNA gene within the same phylogeny as the HF183 cluster (26, 62–69), such as the widely-used HF183/BacR287 (69) and BacHum-UCD (65) assays. Some assays also target *Bacteroides* outside of the HF183 cluster, such as the 16S rRNA gene and genomic sequence of *Bacteroides thetaiotomicron* (67, 70, 71). A primer map that includes most of the established *Bacteroides* 16S rRNA gene marker assays is shown in Figure 1.1.

Many efforts have been made to validate and assess the performance of established MST fecal marker assays (Table 1.1). The main strategy is to test these human marker assays in their host samples, such as human feces and sewage samples, and other non-host samples, such as animal fecal samples. The two major criteria in assay performance are specificity and sensitivity (37, 54, 61, 72). Specificity refers the proportion of true negative samples in marker assay's tested non-host samples. Some assays and their reported average specificities are listed as follows: HF183/SSHBac-R (91.1%) (62, 65, 73–75), HF183/BFDrev (76.8%) (67, 69), HB (90.9%) (26, 52), HF183/BacR287 (91.2%) (52, 69, 75), BacHum-UCD (77.9%) (65, 74–77), BacH (92.6%) (64, 75, 77), HuBac (54.5%) (63, 65, 77, 78), Human-Bac1 (44.4%) (66, 77) and BacHuman (81.5%) (68) (Table 1.1). Interestingly, assays that use the HF183 marker as the forward primer directly (i.e., HF183/SSHBac_R, HF183/BFDrev, HB, and HF183/BacR287) and assays that use primers or probes that overlap with the HF183 marker (i.e., BacHum-UCD and BacH) reported lower-level animal cross-reactions compared to the other assays, further demonstrating the human specificity of the HF183 marker (Table 1.1).

Sensitivity refers to the prevalence, or the true positive rate, of a marker assay in its tested host samples (i.e., human feces and sewage). A sensitivity of 100% indicates the marker assay is always present in the host source. It was reported that not all individual human fecal samples were positive for these human fecal marker assays (62, 65, 73, 74, 78–80). However, this should not be an issue affecting these assays' sensitivity since most fecal pollution is derived from multiple human inputs (i.e. septic systems, household or neighborhood leaking sanitary sewer pipes. Most studies included sewage samples for sensitivity testing; sewage represents a comprehensive fecal microbial community of the population from a large geographical scale (81) and is the main targeted pollution source of human fecal marker assays.

To date, there is no strict benchmark criteria for host specificity and sensitivity of human fecal marker assays. However, it was recommended that a good marker should have a host specificity value of > 0.90 and a sensitivity value of > 0.80 (54, 82, 83). Despite the numbers of human fecal marker assays that have been developed, there is no single marker assay that is exclusively specific to human and sewage sources. Cross-reaction with animal sources such as cat, dog, pig, chicken, turkey, cow, and deer have been reported for these previously described marker assays (Table 1.1).

Clone library method for fecal marker assay development.

The HF183 marker was identified based on 16S rRNA gene clone sequences from *Bacteroides* with primers Bac32F/Bac708R (84). Subsequently, many *Bacteroides* 16S rRNA gene marker assays were developed based on clone sequences amplified using the same primers (62, 63, 65, 79). These assays were limited to the V2 - V4 hypervariable regions due to the amplicon length of Bac32F/Bac708R (Figure 1.1).

One advantage of the clone library method is that it is feasible to get large piece of DNA (e.g., near full-length 16S rRNA gene), which provides an approach for examining host specificity of the targeted microorganism across different regions of 16S rRNA gene. However, clone library method is time-consuming and complex. For example, one picked colony represents one sample for sequencing (i.e., Sanger sequencing), and only one sequence can be obtained from it. Also, clone sequences cannot represent all members that the targeted organism (e.g., genus *Bacteroides*) contains, as usually only dominant members are captured (85). This can cause problems when designing assays based on clone libraries that are not of enough depth. For example, when comparing host sequences to non-host sequences, some sequences that appear to be exclusive to the host source could still exist in non-host sources. Assays designed based on such sequences would likely to have low host specificity; this could be at least one of the reasons for low specificities of some clone library-based *Bacteroides* assays (63, 66). Using the clone library method, bacterial fecal marker assays were developed in only a few microorganisms, such as *Bacteroidales* and *Bifodobacterium* (31, 37), leaving a large population of fecal microorganisms untouched.

Next-generation sequencing application in fecal marker assay development.

DNA sequencing technology has been applied to analyze 16S rRNA gene sequences for application of MST methods in the last decade (86–89). Next-generation sequencing (NGS) technology, which has the ability to yield as many as millions of reads per sample, provides the opportunity to gain an in-depth inventory of the microbial community in a sample and makes it possible for similar or identical sequencing reads to be mapped to different hosts even if they are in low abundance (31, 90, 91). Beginning

with 454 pyrosequencing technology and was later replaced by the current MiSeq, HiSeq and NextSeq Illumina sequencing platforms, sequencing performance metrics such as depth (i.e., the number and the length of reads sequenced and aligned to a reference sequence) and detection sensitivity have been improved (91, 92). At the same time, the cost has been reduced greatly (92). Taking this advantage, researchers are able to use 16S rRNA gene NGS data to characterize the taxonomic composition of microbial communities in environmental samples and apply this approach to MST (93). For example, some studies successfully identified human fecal pollution in surface waters by tracking distribution patterns of human fecal bacteria in microbial communities of environmental water samples (89, 94).

The common approaches for 16S rRNA gene sequence analysis after raw read processing includes reference database-dependent taxa classification and *de novo* clustering (31, 95). The reference database-dependent method assigns the reads taxonomic information through direct sequence comparison with a reference database, which is composed of comprehensive 16S rRNA gene sequences from known (e.g., cultured) organisms (e.g., the SILVA database) (96). The *de novo* method is to cluster reads based on their similarities to each other and therefore has no requirement for a taxonomic reference database. Clustering of sequences, which are aggregated to operational taxonomic units (OTUs), usually obeys a standard threshold of 97% sequence similarity, or 3% sequence dissimilarity. This threshold, however, was determined based on full-length 16S rRNA gene sequence, and is not sensitive enough to differentiate similar organisms in NGS data, which are partial sequences of 16S rRNA gene (31). Studies have demonstrated that 16S rRNA gene sequences (i.e., V6 region NGS data) of more than 99%

similarity could correspond to different ecological niches (95). Therefore, clustering 16S NGS data based on an arbitrary similarity criteria (i.e., 97% similarity) could result in failure to discern potential host-specific markers (31). Approaches that are sensitive and accurate enough to correlate sequence types with host niches are required.

The bacterial family *Lachnospiraceae* as a reservoir for alternative human-specific fecal markers.

The bacterial orders *Bacteroidales* and *Clostridiales* are abundant and consistently present in sewage microbial communities (86). Members of *Bacteroidales* have been demonstrated to be human-associated, such as the HF183 cluster within the genus *Bacteroides* (65, 70, 79, 84). However, less has been reported about *Clostridiales*. McLellan et al. (86) first suggested the family *Lachnospiraceae* within the order *Clostridiales* as a promising group for alternative human fecal marker because of its abundance and high diversity in microbial communities of untreated sewage samples. Later the first human-associated *Lachnospiraceae* 16S rRNA gene fecal marker assay (i.e., Lachno2) was developed (30). This assay showed high sensitivity in tracking sewage pollution in freshwater and had a significant correlation with human *Bacteroides* (30). Since then, more investigations into *Clostridiales* and *Lachnospiraceae* has been done (90, 97, 98). By analyzing *Clostridiales* V6 region sequences in untreated sewage, *Lachnospiraceae* was proved to be the most abundant group, with genera *Roseburia* and *Blautia* identified as the two most abundant genera (98). The same study also examined members within *Clostridiales* across sewage, human, cow and chicken feces, and demonstrated the human specificity of *Lachnospiraceae* (98). Further analysis of *Blautia* showed distinguishing distribution patterns in human groups versus other animal sources

(e.g., pig, cat, dog, deer, cow and chicken), strongly suggesting the potential of family *Lachnospiraceae* members as marker candidates for tracking human fecal pollution (90, 97).

The scope of this thesis work.

The genus *Bacteroides* has been extensively used as a target for human fecal marker assays with the HF183 cluster most widely employed. However, the detection of human fecal pollution based on a single organism (e.g., the HF183 marker) could be biased by animal cross-reactions (i.e., assay false positives) (62) and a lack of the targeted organism (i.e., assay false negatives) (99). Therefore, it is necessary to re-examine the potential for additional highly specific human-associated fecal marker assays.

Most *Bacteroides* fecal marker studies have been developed based on V2-V4 regions 16S rRNA gene clone libraries (61, 63–66, 68). This method is unable to access the full composition of microbial communities and may neglect the presence of host-associated members in non-host samples, resulting in low host specificity of the chosen "host-specific" organisms. In fact, even for the HF183 marker that has been considered the most human-specific, non-host amplifications have always been reported (67, 73–75, 78, 100). The mechanism behind these non-host cross-reactions is still poorly understood (54).

The NGS technology has been proven to be an appropriate approach for discovering additional human-associated fecal microorganisms (86, 90, 97). Studies have suggested that *Bacteroidales* and *Clostridiales* are dominant fecal microbiome members in sewage influent samples (86). Further analysis of NGS data from sewage and animal fecal samples indicate that members of *Lachnospiraceae* (e.g., genus *Blautia*) have enormous potential of being human-specific (90, 97). In this work, NGS data from V4V5 and V6 regions

isolated from a wide variety of sewage and animal samples (n = 469) offered the opportunity to examine the potential of *Lachnospiraceae* and *Bacteroides* for human-specific fecal markers across different regions of 16S rRNA gene and to explore mechanisms for human fecal marker cross-reactions.

Based on the all the ongoing efforts for human fecal marker assay development and the innovation in advancing NGS technology and bioinformatic tools, this thesis work aims to: 1) mine data for alternative highly specific human fecal marker candidates from certain human-associated fecal microorganisms (i.e., family *Lachnospiraceae* and genus *Bacteroides*); 2) develop reliable highly specific and sensitive human-associated fecal marker assays; and 3) explore mechanisms for marker cross-reactions from the perspectives of microbial community composition and qPCR assay amplification.

Chapter 2 describes the development of two human *Lachnospiraceae* fecal marker assays from the V6 region of *Lachnospiraceae* 16S rRNA gene. This work advanced the application of human-associated *Lachnospiraceae* organisms and demonstrated the usage of NGS data in human fecal marker assay development. This work was published in *Applied and Environmental Microbiology* in 2018 (52).

The key results include:

1. Assessment with 97% sequence similarity criteria did not resolve *Lachnospiraceae* members into host groups.
2. The V6 region of *Lachnospiraceae* 16S rRNA gene was more variable than the V4V5 region and was more ideal for developing *Lachnospiraceae* marker assays.

3. A list of 40 V6 *Lachnospiraceae* marker candidates were identified. The two most abundant candidates in sewage, Lachno3 and Lachno12, were developed to TaqMan qPCR assays. The Lachno3 assay was considered highly human-specific (specificity = 96.4%). The Lachno12 assay amplified in cow and pig fecal sources at low levels.
4. The applications of the Lachno3 and Lachno12 assays, together with a dog fecal marker assay, resolved presumptive fecal pollution source(s) in Milwaukee urban water samples that previously demonstrated inconsistent results in the HB and the Lachno2 assays (i.e., high CN in one assay but low in another, or results were not at the same order of magnitude).

Chapter 3 reports the development of two sewage specific *Bacteroides* marker assays from the V4V5 and V6 regions of 16S rRNA gene, respectively. These assays targeted a sewer pipe-derived *Bacteroides* (i.e., *Bacteroides graminisolvens*) and were independent of human or animal gut microbiota. This work provided evidence for *Bacteroides* host specificity and explored this genus across different regions of 16S rRNA gene for human fecal marker identification. This work also proved the feasibility to use resident organisms of sewer pipe system for sewage tracking in addition to fecal anaerobes. This work was published in *Applied and Environmental Microbiology* in 2019 (101).

The key results include:

1. Genus *Bacteroides* showed consistent oligotype patterns in sewage samples and were dissimilar from patterns in animal fecal samples.
2. The HF183 cluster comprised ~ 3% of the sewage *Bacteroides* clone library and its downstream V4V5 and V6 regions were not human-specific.

3. Multiple sewage-specific *Bacteroides* markers within the V4V5 (n=7) and V6 regions (n=21) were identified. Two HF183-independent sewage-specific *Bacteroides* TaqMan qPCR assays were developed from the V4V5 (i.e., BacV4V5-1 assay) and V6 region (i.e., BacV6-21 assay). The BacV4V5-1 assay showed 98.7% specificity with a very low signal in one pig in animal validations (n=76). The BacV6-21 assay showed 100% specificity.
4. The *Bacteroides* assay validations for the BacV4V5-1 and BacV6-21 assays in sewage and environmental water samples demonstrated they were the same organism. The human *Bacteroides* (HB) and HF183/BacR287 assays targeted a different organism. The results of the sewage *Bacteroides* assays and the HF183 assays were overall correlated in environmental water samples.

Chapter 4 focuses on exploring cross-reaction mechanisms for the Lachno3, HB and BacV6-21 assays. This work compared the presence of human fecal markers in V6 NGS data (n=271) and qPCR results (n=180) and identified human fecal marker cross-reaction mechanisms from the distribution patterns of fecal microorganisms and technical details of qPCR amplification. This work is in preparation for publication.

The key results include:

1. Host physiology and environmental factors such as diet and habitat both showed influences on the compositions of animal fecal microbial communities, including family *Lachnospiraceae* and genus *Bacteroides*.
2. Cross-reaction of human marker co-varied with the presence/absence of its closely-related organisms. High level marker cross-reactions (e.g., 10^4 copy numbers per ng of DNA) correlated with changes in marker's corresponding

genus (e.g., genus *Bacteroides*). Also, marker cross-reaction could be correlated with certain environmental factors such as diet and habitat.

3. Specificities of human and sewage fecal markers in NGS data (n = 271) included Lachno3 (97.0%) and BacV6-21 (99.6%). Specificities of human and sewage fecal marker assays in qPCR results (n = 180) were BacV6-21 (95.6%) > Lachno3 (92.8%) > HB (91.7%).
4. Discrepancies were observed for human markers in animal fecal samples between NGS and qPCR results. Amplification of organisms closely-related to the marker could be responsible for qPCR positive-only cross-reactions.

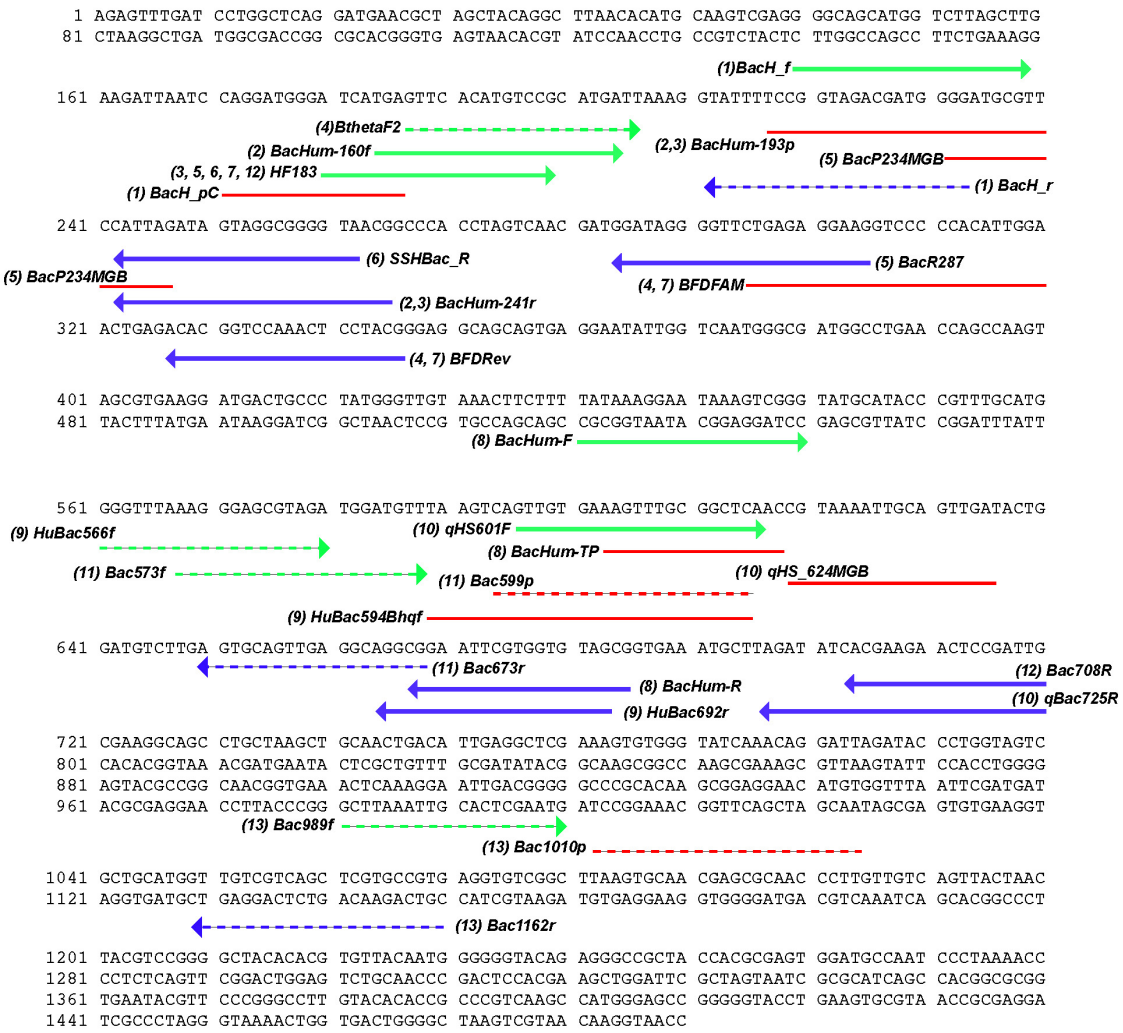


Figure 1.1 Alignment of established 16S rRNA gene *Bacteroides* assays. Primer and probe sequences are aligned to a reference sequence (GenBank accession number AB242143). Forward primers are shown in green arrows, probes are shown in red lines, and reverse primers are shown in blue arrows. Sequences that are not 100% matched with the reference are shown in dashed line. Each primer/probe name is labeled at the start or end of the sequence. Numbers in parentheses represent the following assays: (1) BacH (Reischer et al. 2006), (2) BacHum-UCD (Kildare et al. 2007), (3) HB (Templar et al. 2016), (4) BthetaF2 (Haugland et al. 2010), (5) HF183/BacR287 (Green et al. 2014), (6) HF183/SSHBac_R (Seurinck et al. 2005), (7) HF183/BFDrev (Haugland et al. 2010), (8) BacHuman (Lee et al. 2010), (9) HuBac (Layton et al. 2006), (10) HumanBac-1 (Okabe et al. 2007), (11) BacV4V5-1, developed in this work, (12) HF183/Bac708R (Bernhard and Field 2000), (13) BacV6-21, developed in this work.

Table 1.1 Established *Bacteroides* marker assays with their reported animal cross-reactions and specificities. References in bold are the developers of assays.

Target	Marker assays	Type	Tested animals	Positive animals	Specificity [a]	Reference
16S rRNA gene/*B. dorei*	HF183/Bac708R	PCR	Cow, cat, deer, dog, duck, elk, goat, llama, pig, seagull, sheep (n= 46)	None	100.0%	**Bernhard and Field (2000)** (61)
		SYBR qPCR	Cow, pig, sheep, goat, horse, chicken, dog, duck, pelican, kangaroo (n=136)	Sheep	99.3%	Ahmed et al (2009) (13)
		PCR	Pronghorn, moose, deer, duck, pelican, raccoon, gull, elk, cattle, goat, pig, turkey, sheep, chicken, dog, cat, dog and 3 marine animals (animals were tested as 22 pools of composite DNA for qPCR and as individuals for PCR) (n=158)	Dog	99.4%	Shanks et al (2010) (14)
	HF183/SSHBac-R[b]	SYBR qPCR	Chicken, horse, cow, dog and pig (n=19)	Chicken	94.7%[c]	**Seurinck et al (2005)** (62)
		SYBR qPCR	Cow, horse, dog, cat and seagull (n=41)	Dog and cat	92.7%	Kildare et al (2007) (65)
		SYBR qPCR	Cows, cat, dog and chicken (n=30)	Cat and dog	93.3%	Ahmed et al (2010) (17)
		SYBR qPCR	Cat, dog, gull, rat and raccoon (n=47)	Cat	97.2%	Van De Werfhorst et al (2011) (74)
		SYBR qPCR	Monkey, wild boar, bird, chicken, rabbit, cat and dog (n=220)	Chicken, rabbit and dog	88.6% 80.0% [d]	Nshimyimana et al (2017) (75)
	HF183/BFDrev	TaqMan qPCR	Cow, pig, chicken, dog, cat (each animal as one group of composite DNA) (n=50)	Chicken and dog	60.0%	**Haugland et al (2010)** (67)
		TaqMan qPCR	Chicken, turkey, dog, cat, deer, pronghorn, pig and cow (n=123)	Chicken and turkey	93.5%	Green et al (2014) (69)
	HF183/BacR287	TaqMan qPCR	As described above	Chicken and turkey	93.5%	**Green et al (2014)** (21)
		TaqMan qPCR	As described above	Chicken and rabbit	90.0% 86.7% [d]	Nshimyimana et al (2017) (75)

		TaqMan qPCR	Cat, dog, pig, cow, deer and gull (n=55)	Deer	94.5%	Feng et al (2018) (52)
	HB	TaqMan qPCR	As described above	Deer and dog	90.9%	**Templar et al (2016) (31)** Feng et al (2018) (52)
	HF183/BthetaF2	TaqMan qPCR	As described above	Chicken and dog	90.9%	Shanks et al (2010) (78)
16S rRNA gene	BacHum-UCD	TaqMan qPCR	As described above	Dog	97.6%	**Kildare et al (2007)** (65)
		TaqMan qPCR	Dog, cow, horse and Canadian goose (as groups of composite DNA) (n=41)	Dog, cow and horse	70.7%	Silkie and Nelson (2009) (76)
		PCR	As described above	Pig, sheep, horse and dog	95.6%	Ahmed et al (2009) (77)
		TaqMan qPCR	As described above	Cat, dog, gull and raccoon	38.9%	Van De Werfhorst et al (2011) (74)
		TaqMan qPCR	As described above	Chicken, rabbit and dog	91.4% 73.3% [d]	Nshimyimana et al (2017) (75)
16S rRNA gene	BacH	TaqMan qPCR	Cow, deer, chamois, roe deer, sheep, goat, horse, fox, dog, cat, pig, chicken, turkey, swan, duck and black grouse (n=302)	Cat	99.7%	**Reischer et al (2007)** (64)
		PCR	As described above	Sheep, goat and dog	94.1%	Ahmed et al (2009) (77)
		TaqMan qPCR	As described above	Chicken and rabbit	90.0% 86.7% [d]	Nshimyimana et al (2017) (75)
16S rRNA gene	HuBac	TaqMan qPCR	Cow, pig, horse and dog (n=18)	Cow, pig and dog	67.9%	**Layton et al (2006)** (63)
		TaqMan qPCR	As described above	Cow, horse, dog and cat	61.0%	Kildare et al (2007) (65)
		PCR	As described above	Cow, pig, sheep, horse, dog and ducks	63.2%	Ahmed et al (2009) (77)
		TaqMan qPCR	As described above	Deer, Canadian goose, duck, raccoon, elk, cow,	22.7%	Shanks et al (2010) (78)

				pig, turkey, sheep, chicken, dog, cat and dog		
16S rRNA gene/ *B. fragilis*	Human-Bac1	TaqMan qPCR	Cow and pig	Cow and pig	10.0% (37)	**Okabe et al (2007)** (66)
		PCR	As described above	Cow, sheep, horse, dog and kangaroo	78.7%	Ahmed et al (2009) (77)
16S rRNA gene	BacHuman	TaqMan qPCR	Cow, pig, deer, horse, dog, cat, gull, goose and raccoon (n=54)	Pig, dog and cat	81.5% [c]	**Lee et al (2010)** (68)
Genomic sequence/ *B. thetaiotomicron*	B. theta	PCR	Dog, cow, chicken, turkey, horse, pig and goose (n=241)	Dog	97.9%	**Carson et al (2005)** (71)
		PCR	As described above	Dog	98.7%	Shanks et al (2010) (78)
16S rRNA gene/ *B. thetaiotomicron*	BthetaF2	TaqMan qPCR	As described above	Pig, chicken, dog and cat	20.0%	**Haugland et al (2010)** (67)
		TaqMan qPCR	As described above	Pronghorn, moose, goose, duck, raccoon, gull, elk, dairy cow, pig, sheep, chicken, dog, cat, sea lion and elephant seal	31.8%	Shanks et al (2010) (78)
Genomic sequence/ *B. thetaiotomicron α-mannanase*	B. theta α[b]	TaqMan qPCR	Dog, cow, horse, pig, chicken, turkey and goose (n=160)	None	100%	**Yampara-Iquise et al (2008)** (70)
		TaqMan qPCR	As described above	Cat	98.6% 93.3% [d]	Nshimyimana et al (2017) (75)

a. Specificity is calculated as the percentage of negative animal fecal samples.
b. This marked assay was named by Harwood et al. 2014 (37).
c. Animal false positives were reported in the reference publications.
d. The upper percentage represents specificity from individual animals, and the lower percentage represents the specificity from pooled animals.

Chapter 2 Development of human-associated fecal marker assays from family *Lachnospiraceae*

Abstract

Assessing urban water microbial quality is challenging since water can be impacted by many fecal sources such as sewage, pet waste and urban wildlife. How to track the human source fecal pollution (i.e., sewage) has been an important issue since it is the source that most likely carries human pathogen. The human gut microbiome contains many organisms that could potentially be used as indicators of human fecal pollution. In this study we developed two next-generation sequencing (NGS) data-based human-associated fecal marker assays from certain organisms in bacterial family *Lachnospiraceae*. V6 hypervariable region sequences of the 16S rRNA gene from sewage and animal fecal samples were used, and 40 human-associated marker candidates with a robust signal in sewage and low or no occurrence in animal hosts were identified. Two of them were chosen for quantitative PCR (qPCR) assay development by mapping to V2 to V9 region sequences generated from sewage and animal clone libraries; the developed qPCR assays were designated Lachno3 and Lachno12. Assay validations were performed for fecal samples (n=55) from cat, dog, pig, cow, deer, and gull sources, and compared with established human fecal marker assays (Lachno2, and two human *Bacteroides* assays; HB and HF183/BacR287). Each of the established assays cross-reacted with at least one other animal, including animals common in urban areas. The Lachno3 and Lachno12 assays were primarily human-associated; Lachno12 demonstrated low levels of cross-reactivity with select cows, and non-specific amplification in pigs. However, this limitation may not be problematic when testing urban waters. These markers resolved ambiguous results from previous investigations in stormwater-impacted waters, demonstrating their utility. Combined marker assays will provide the highest resolution and specificity for assessing fecal pollution sources in urban waters.

Introduction

Human fecal pollution enters urban waters via ways such as combined sewage overflows (CSOs), sanitary sewage overflows (SSOs), illicit connections, or failing sanitary sewers that infiltrate stormwater systems (19, 27, 30). Pathogenic microorganisms from fecal pollution, including bacteria, viruses, and protozoans, pose a risk of waterborne disease for those exposed to the polluted surface waters (6, 7, 102). General fecal indicator bacteria (FIB) such as fecal coliforms, *Escherichia coli* (*E. coli*) and enterococci have historically been used to assess the microbial water quality because of their high abundance in sewage and feces (2, 37). However, many studies have also failed to establish direct correlations between FIB concentrations and pathogen presence or human health outcomes (8, 48–50). In urban surface water, this is most likely due to the presence of nonhuman source fecal pollutions from non-point sources such as stormwater runoff, which contribute to FIB concentrations but not introduce human pathogens. Since FIB are common in all warm-blooded animals intestines and do not distinguish human source from animal source fecal pollution (37, 50), there is a need to develop alternative fecal indicators to assess water quality in complex environments where multiple fecal pollution sources contribute.

It has been demonstrated that human fecal anaerobes are useful for tracking fecal pollution sources because they are abundant in the human intestinal tract with some taxa specifically associated with host physiology (37, 54, 90). The emerging of next-generation sequencing (NGS) technology, which provides the opportunity to gain taxonomy and relative abundance information of taxa community wide, makes it possible for identifying host-associated even host-specific microorganisms (31, 90, 91). Fecal anaerobes within *Bacteroidales*, in particular members of the genera *Bacteroides* and *Prevotella*, have been well studied and successfully applied for fecal

pollution identification (61, 62, 103). However, a large portion of the human microbiome remains untapped for host-associated indicator development, including members of *Clostridiales*, which can comprise more than half of the human source fecal microbial community (104). Additional indicators from such microorganisms could be very useful in cases where the mostly used *Bacteroidales* markers are not abundant enough in populations within in geographical regions due to diet, culture, or other environmental factor impacts (54, 99, 105, 106). In addition, it has been widely reported animal cross-reactions, such as cat, dog, chicken, turkey and raccoon, for established human-associated *Bacteroides* assays (67, 69, 73, 74). Fecal markers from a different microorganism could add a layer of verification to source tracking studies that are being used to guide mitigation efforts, which are often of high cost and require strong stakeholder support.

Previous studies which used NGS technology to create an inventory of potential new indicators found that about 97% of the human fecal community oligotypes were present in sewage with the most abundant ones matched (81), thus demonstrating that sewage comprehensively represents human fecal microbial community composition. Members of the family *Lachnospiraceae* are promising candidates for host-associated genetic marker because of their high abundance and diversity in sewage (86). In particular, the genus *Blautia* within family *Lachnospiraceae* has been demonstrated of specificity and preference pattern among sewage, human and animal hosts (90, 97, 99). The human-associated *Lachnospiraceae* genetic marker Lachno2 (30) had been identified based on presence in sewage but not cows, although the Lachno2 V6 marker sequence was subsequently found in cats and dog fecal samples (90). Despite noted cross reactivity, in sewage-contaminated water, the Lachno2 assay and the human *Bacteroides* (HB) assay, which is a hybrid of the HF183 marker (26) and the BacHum-UCD marker assay (65), are strongly correlated and improved accuracy of sewage detection (19, 26).

Here we examined the population structure of the family *Lachnospiraceae* in sewage and animal hosts using near full-length sequences of 16S rRNA gene and identified 40 human-associated (i.e. preferred for the human host and only found sporadically in other animals) *Lachnospiraceae* genetic marker candidates. Two genetic markers, designated Lachno3 and Lachno12, were chosen for quantitative PCR (qPCR) assay development. These assays host specificities were validated using animal fecal samples (n=55) across six hosts from multiple locations. Results were compared with established human fecal marker assays, including the Lachno2, HB, and the HF183/BacR287 assays (26, 30, 69). Further testes of urban environmental water samples derived from non-point source pollution demonstrated the applicability of Lachno3 and Lachno12 marker assays via comparison of the Lachno2, HB, HF183/BacR287, and DogBact assays (107).

Material and Methods

Samples collection and DNA extraction.

Two sets of animal fecal samples were used in this study; Set 1 was used for clone library construction, including five cats, five dogs and ten pig samples. Set 2 was used for qPCR assay validation, including 11 cats, ten dogs, nine pigs, 11 deer, ten cows and four gulls. Samples were different in Set 1 from Set 2. The majority of samples were also sequenced (n=44), except in cases where there was not enough material available. Detailed metadata information of these animal fecal samples is in Supplemental Data Set 2.1. Animal fecal samples were transported to University of Wisconsin-Milwaukee (Milwaukee, WI, USA) on ice within 24 hours of collection and stored at -80°C upon arrival until DNA extraction. Fecal samples preparation and DNA extraction used QIAamp DNA Stool Mini Kit (Qiagen Inc., Valencia, CA), following protocol for pathogen detection, which increases the yield of non-host fecal genomic DNA. In some cases,

extracted DNA was sent directly from the originating laboratory (annotated in Data Set 2.1). All DNA samples were stored at -20°C.

Cone libraries of fecal samples.

Clone libraries were generated from Set 1 animal fecal samples using *Clostridium coccoides* (*C. coccoides*) cluster targeted forward primer and a universal 16S rRNA gene reverse primer (Ccoc-F/1492R) to amplify a portion of the 16S rRNA gene from *Lachnospiraceae* (30, 98, 108). Amplicons were sequenced by Sanger sequencing (ABI Prism 3700xi genetic analyzer, Applied Biosystems, Foster City, CA). Clone sequences from two other published studies were also used: (i) *C. coccoides* sewage clone libraries (GenBank Access Numbers JX228967 - JX230954) (98), and (ii) whole community cow fecal clone libraries (GenBank Access Number FJ672948-FJ674268 and FJ675665-FJ685516) (109). Only *Lachnospiraceae* sequences from the cow libraries were used. Both libraries were subsampled to 200 sequences. Cloning and sequencing methods were as previously described, including steps of PCR, ligation, transformation, plasmid preparation, and sequencing reactions (30).

Sequence processing and analysis.

For animal clone library Sanger sequencing, three primers (Ccoc-F, 331F and 1492R) were used. Sequences were assembled using SeqMan Pro program (Lasergene v12, DNASTAR, Madison, WI), and these less than 900 bp were discarded, with chimeras subsequently removed using Chimera Vsearch (110) in mothur (111). A total of 718 sequences were analyzed, including 200 sewage, 80 cat, 85 dog, 153 pig, and 200 cow sequences. Operational Taxonomic Units (OTUs) were created using the nearest neighbor method at 97% similarity level in mothur based on SILVA 119 taxonomic reference database (96).

A phylogenetic tree of OTU representative sequences was constructed to examine the phylogenetic relationships of *Lachnospiraceae* organisms from different hosts. Host source was annotated for each OTU (i.e. human only, animal only, or human/animal). The OTU representative sequences were aligned using MUSCLE (112) and trimmed to the same length with MEGA7 (113), along with two *E. coli* 16S rRNA gene sequences as an outgroup (GenBank Access Numbers HF584706 and LT745986). The tree was constructed using maximum-likelihood method in Kimura 2-parameter (K2) model with gamma-distribute rates and invariant sites (G+I), bootstrapped for 1000 replicates and visualized in Interactive Tree Of Life (114) (iTOL, http://itol.embl.de). Representative sequences and their host annotation for each OTU were shown in Supplemental Data Set 2.2. A heatmap was generated to display the *Lachnospiraceae* relative abundance in different hosts based on clone libraries (Appendix A Figure 1). To better visualize the distribution of clones in different hosts, the relative abundance was normalized to 100% for each of the most abundant 70 OTUs.

NGS datasets.

Sequences were generated using Illumina MiSeq and HiSeq platforms at the Marine Biological Laboratory (MBL), University of Chicago. Whole community datasets of partial 16S rRNA gene sequences were generated from the V4V5 (518F/926R) (115) and V6 (967F/1064R) (116) regions, and stored in the Visualization and Analysis of Microbial Populations (VAMPs) platform (https://vamps2.mbl.edu) (117). Sequence counts were normalized to the median count of all samples' total bacterial sequences, and singleton sequences were removed. *Lachnospiraceae* sequences were extracted using taxonomy assignments in GAST (118). Sequences from the newly described family *Christensenellaceae* (119), which were previously designated as *Lachnospiraceae* and were very likely also host-adapted, were added into the datasets. In all, the

dataset included 20,587 unique V4V5 sequences with 741,927 reads, and 100,242 unique V6 sequences with 17,143,353 reads. The *Lachnospiraceae* sequences were enumerated according to their rank abundance in the composite dataset of sewage samples. The second and third most abundant *Lachnospiraceae* in this dataset had appeared in the inverse order in previous analysis and had been designated Lachno3 and Lachno2, respectively. Likewise, the tenth most abundant *Lachnospiraceae* in the dataset had previously been designated Lachno12. Since the exact order is somewhat dependent on the sewage samples used in analysis, we chose to keep the original designation for these two instances. Therefore, Lachno3 in this study is the second most abundant *Lachnospiraceae* in this dataset, and Lachno2 is the third most abundant. Lachno12 and Lachno10 designations correspond to the tenth and twelfth most abundant *Lachnospiraceae*, respectively. The 100 most abundant sequences for each animal and sewage sources are detailed in supplemental Data Set 2.3.

Design of human-specific molecular assays.

Animal and sewage samples that were both sequenced for V4V5 and V6 regions were compared using R package "indicspecies" (120) with 999 permutation tests to identify the region that would provide the most specific and sensitive *Lachnospiraceae* marker candidates. The human-associated marker candidates were first chosen by the criteria that they were above 90% sensitivity and specificity, and among the top 95% abundant *Lachnospiraceae* in sewage. Candidates were retained if they were present at lower levels in two or less other animal hosts (i.e. cat, dog, pig, cow, deer, chicken and raccoon). We chose the V6 region as the most promising marker regions and then compared the V6 NGS dataset sequences to the sewage clone library using BLAST+ (121) to find clones that represented longer sequences and contained each V6 marker sequence for primer design.

Two qPCR assays were developed to target the V6 region Lachno3 and Lachno12 markers (Table 2.1). Primers and probes were designed base on alignments of animal and sewage clone sequences in MegAlign Pro program in DNASTAR software (Lasergene v12, DNASTAR, Madison, WI). Alignments included each respective V6 marker sequence, a *Lachnospiraceae* full-length 16S rRNA gene reference sequence (GenBank Access Number EF036467), and the marker's exact matches of the sewage clone library sequences. Animal clones that had >97% similarity with the markers, and representative sequences from the top 10 OTUs of all animal sequences, were also added into the alignment.

The Lachno3 and Lachno12 markers were also mapped into longer sequence reads that included the V4V5 region to search for their correlated V4V5 sequences, which were then identified in our V4V5 NGS dataset to look for their host specificity information.

Quantitative PCR analysis.

All qPCR experiments were performed on an Applied Biosystems StepOne Plus™ Real-Time PCR System Thermal Cycling Block (Applied Biosystems; Foster City, CA). To validate Lachno3 and Lachno12 marker assays, animal fecal samples from Set 2 were tested using at least six individual samples and one pool (made from two individuals), except gulls that were run as single individuals. Concentration of each animal fecal DNA was measured by Nanodrop spectrophotometer. Each sample was then diluted to 1 ng μL^{-1}, 0.1 ng μL^{-1} and 0.01 ng μL^{-1} with 5 µL used in each qPCR reaction. Sewage samples were diluted 1:100 volume to volume, and environmental samples were run without dilution. All standard curves were run in triplicate with DNA from sewage clones that match the Lachno3 and Lachno12 assays and were serially diluted from 1.5×10^6 to 1.5 copies per reaction. For each validation run, positive control using sewage DNA and blank control using DNA-grade sterile water were used. The qPCR reaction setting was

as described by Templar *et al.* (26). To optimize annealing temperature for these two assays, we tested diluted sewage DNA samples (diluted at the ratios of 1:100, 1:500, 1:1000, 1:2000, 1:4000 and 1:8000) from 60°C to 64°C to determine if any amplification efficiency was lost. The amplification program included one cycle at 50°C for 2 min, followed by one cycle at 95°C for 10 min, then 40 cycles of 95°C for 15 s followed by 1 min at 64°C for Lachno3 or 61°C for Lachno12; the Lachno2 assay was run at 61°C in this study. The qPCR assays slopes, y-intercepts, and efficiencies were shown in Appendix A Table 1. For validation result output, each animal's qPCR reaction copy number (CN) was converted to CN per ng of DNA, CN per 0.1 ng of DNA and CN per 0.01 ng of DNA, and each sewage sample result was also converted to CN per ng of DNA.

Nucleotide sequence accession numbers.

The partial 16S rRNA gene clone libraries sequences were deposited in the GenBank database under accession numbers MG702648-MG702965. A portion of the NGS data used in this study was from National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) SRP041262 (V6 sequences) and BioProject PRJNA261344 (V4V5 sequences). NGS data generated for this study was stored in the SRA projects under accession numbers SRP132402 (V6 region sequences) and SRP132403 (V4V5 region sequences).

Results

Population structure of *Lachnospiraceae* in human and animal hosts.

We examined 718 sequences in *C. coccoides* libraries from sewage, cat, dog, and pig, as well as previously published *Lachnospiraceae* sequences from a near full-length library comprised of cows (109). In total, there were 200 OTUs clustered at 97% sequence similarity, within which 70 OTUs contained multiple sequences. A phylogenetic tree was constructed using OTU

representative sequences annotated with host information (Figure 2.1), demonstrating that phylogenetically-related *Lachnospiraceae* OTUs did not correspond to host sources. Two OTUs (OTU 185 and 198) were classified as family *Defluviitaleaceae,* which was included in the family *Lachnospiraceae* in earlier version of reference taxonomy and select members were able to be amplified with the *C. coccoides* primer. Overall, 31 out of the 70 OTUs with multiple sequences contained sequences from both animals and sewage, suggesting that assessment with 97% sequence similarity criteria does not resolve these organisms into host groups. Appendix A Figure 1 shows the *Lachnospiraceae* OTUs distributions in different hosts.

Comparison of V4V5 and V6 regions as reservoirs for human-associated markers.

We analyzed a subset of samples that were sequenced for both the V4V5 and V6 regions to determine the more useful region to identify markers for organisms found in sewage (i.e. human fecal pollution sources). Fifty-two animal samples and 16 sewage samples were utilized for permutation tests in "indicspecies". The results demonstrated that the V6 region had more human markers of specificity and sensitivity over 90% (n=193) than V4V5 (n=22), and the V6 region showed 49 indicators of 100% specificity and sensitivity while the V4V5 region showed none (Figure 2.2). In this analysis, a larger number of specific indicators were identified for the V6 region because fewer animals were included in the dataset. Overall, these results suggest that the V6 region is more ideal as a marker region for host-associated organisms and potentially an ideal target region for *Lachnospiraceae* assays to discriminate sources of fecal pollution.

V6 region *Lachnospiraceae* markers identification.

We examined of *Lachnospiraceae* host distribution patterns using V6 region unique NGS sequences to identify organisms uniquely found in human source (i.e., sewage). The NGS dataset was more extensive than the clone libraries, with 198 samples, ten host types, and 100,242 unique

V6 region sequences recovered without sequence singletons. In total, 88 indicators were selected with both sensitivity and specificity above 90% (p values = 0.001); seven out of 88 were identified with 100% specificity and sensitivity (Appendix A Figure 2). The final list of *Lachnospiraceae* V6 markers candidates that meet our criteria contained 40 candidates, including ten exclusively in sewage and 30 presented at low relative abundance in one or two animal hosts (Appendix A Figure 3).

Continuity of V4V5 region host specificity for V6 region *Lachnospiraceae* markers.

Two V6 region marker candidates, designated as Lachno3 and Lachno12, were chosen for this study. The clone libraries allowed us to identify the V4V5 region that matched the Lachno3 and Lachno12 in the organisms contain these markers. We were then able to use V4V5 region NGS dataset to examine if these matching V4V5 sequence types were unique to sewage samples or also found in animals. In the clones that contained the Lachno3 marker (n=79), 18 were matched with NGS V4V5 dataset with eight unique sequence types (Figure 2.3). All of the V4V5 types showed dominance in sewage but several V4V5 types were also found in animals with lower abundance, suggesting that the Lachno3 organism-correlated V4V5 region is not as specific as its V6 region. For the Lachno12 marker, only one V4V5 type (i.e. V4V5_15) was found in clone sequences, which was exclusive to sewage. However, it is possible that there are other Lachno12 related V4V5 sequence types that occur in animal hosts as the depth of the sewage clone library limits the identification of more Lachno12 related V4V5 sequence types.

Development of qPCR assays for Lachno3 and Lachno12.

Two human-associated *Lachnospiraceae* assays were developed based on Lachno3 and Lachno12 markers by mapping these markers onto *C. coccoides* sewage and animal clone libraries to find regions of specificity. The Lachno3 and Lachno12 clones with the V4V5 region that did

not cross over into animals were considered targets, and the animal clone library sequences were used for exclusion of non-targets. The forward and reverse Lachno3 and Lachno12 primers had at least 1 mismatch with animal sequences, and the probes had several mismatches. We purposefully chose one assay (i.e. Lachno3) that was strictly specific, based on our NGS animal dataset, and one that was more abundant in sewage, but had low levels in other hosts (i.e. Lachno12) as a means to benchmark performance of these assays when these organisms may occur at very low levels in non-target animal sources.

Lachno3 and Lachno12 assay validation.

To validate Lachno3 and Lachno12 assay sensitivities, these marker assays were applied in sewage sample tests. Relative abundance levels of Lachno3 and Lachno12 in sequenced sewage samples (n=28) indicated that the Lachno3 marker was generally about 3.0 ± 1.3 folds of the Lachno12 marker. The qPCR results of Lachno3 and Lachno12 assays in untreated Milwaukee sewage influent samples (n=8) indicated that we could expect the Lachno3 marker CN ($1.2 \times 10^5 \pm 7.7 \times 10^4$) to be about two-fold of the Lachno12 marker CN ($6.6 \times 10^4 \pm 4.5 \times 10^4$).

For specificity validation, we tested Lachno3 and Lachno12 in 55 animal fecal samples across six hosts (Figure 2.4). The Lachno3 assay demonstrated a overall specificity of 98.4%, with very low level of amplification in two cat samples in 1 ng μL^{-1} ng DNA template level; however, the Lachno3 sequence was not found in their NGS dataset, suggesting it may be non-specific amplification. All other samples were negative for Lachno3. Lachno12 cross-reacted with four cows (25% of tested cows) (Figure 2.4A, Appendix A Table 2) with average copy number of 2.2×10^2, which is equivalent to 1: 300 of sewage DNA. The Lachno12 also showed positive in the qPCR results of three pigs (33.3% of pigs) with average CN of 12; however, the NGS data of these pig samples showed no presence of the Lachno12 marker, indicating non-specific amplifications.

In addition, a low occurrence of Lachno12 (1: 280 relative abundance compared to sewage) was observed in one dog in the NGS dataset, while this marker was not detected in qPCR of that dog sample, or any other dogs tested. This demonstrated that sequencing data may be more sensitive than what can practically be amplified in a sample. Lachno12 was considered human associated with cow cross-reaction, while Lachno3 was considered as human-specific in our results.

Animal validations were also carried out using established assays designed for human *Bacteroides*. The HB assay was positive in one dog (10% of dogs) and two deer (18.2% of deer); the HF183/BacR287 assay was positive in two deer (18.2% of deer) but not in any dog. The Lachno2 assay was run at an annealing temperature of 61°C rather than the previously reported 60°C and showed cross-reactions with cat (82% of cats), dog (70% of dogs) and pig (100% of pigs) samples at the highest concentration of fecal material. Some cats and pigs were also positive at lower concentrations of fecal material. The Lachno2 qPCR results also showed low levels of amplification in three deer (27% of deer) and seven cows (70% of cows). Lachno2 marker was not found in six out of seven cow samples NGS data, indicating cow Lachno2 signals were mostly from qPCR amplifications of non-target sequences; the three positive deer were not sequenced, but all had decreased CN with the increased temperature from 60°C to 61°C, indicating a possibility for optimization of the Lachno2 assay. The gull samples were negative for all five assays.

Lachno3 and Lachno12 assay applications in non-point source polluted urban water samples.

We tested several environmental water samples that demonstrated inconsistent results between the HB and Lachno2 assays (i.e. high CN in one human marker assay but low in another, or results were not at the same magnitude). The HB assay targets on the HF183 cluster of organisms that have been also found in dogs and deer. Lachno2 is sensitive for detecting sewage

but also sporadically cross-reacts with dogs and cats, and other non-urban animals. Samples were tested with Lachno3 and Lachno12 and a combination of other available assays (Table 2.2). Samples that were non-detects (or below detection limit) in the DogBact assay could be potentially excluded from dog source. Samples with positive Lachno3 CN were considered contaminated with human fecal pollution. Because Lachno12 is less specific and sensitive than Lachno3, ratios of these markers that were not typical of what was found in sewage, or the presence of only Lachno12 was considered suspicious for non-human sources. In addition, because the BacHum-UCD assay showed cross-reacts with raccoon (74), dog (65) and deer, urban water samples that only showed positive in HB assay were interpreted as containing contamination from raccoon when the DogBact assay was negative (e.g. sample FT15268) as deer is not expected to be in this highly urbanized area. Human contamination from a limited number of individuals that had atypical microbiome compositions could not be ruled out as an explanation for inconstancies in human-associated marker results.

Discussion

Host-associated organisms offer an opportunity to discover new indicators of fecal pollution.

The gut microbiome of human and animals is largely shaped by diet and host physiology (104, 122), and organisms specifically adapted to fill a niche within a host are promising candidates for developing new indicators for fecal pollution sources. The gut microbiome of humans and animals have a limited number of bacterial families and genera, but have extensive species and strain diversity that could indicate diversification among heterogeneous hosts (123). Our work to examine the population structure within the family of *Lachnospiraceae* found OTU clustering at 97% similarity is not sufficient to distinguish patterns of host specificity (Figure 2.1) (124). This

suggests that genetic traits that determine host association do not map to overall phylogeny within a group.

While there were not overall phylogenetic patterns, we found finer scale methods could track host-associated organisms within the family *Lachnospiraceae*. Our previous work within the genus *Blautia* showed that using a 60 bp region within the 16S rRNA gene as a marker region was sufficient to reveal ecologically relevant distribution patterns among hosts (97). Here, we expand this work to include all *Lachnospiraceae*, and demonstrate that this family is rich in potential indicators, with 88 V6 sequences identified by the biomarker identification program "indicspecies" (120). Analysis of the V4V5 regions in clone libraries demonstrated that organisms tracked by a particular V6 can be further discriminated into subpopulations by their V4V5 sequences (i.e. one V6 region could have multiple associated V4V5 sequence types), with some of the V4V5 sequences for the Lachno3 organisms found in other animals. The "indicspecies" analysis (125) identified far fewer markers in the V4V5 region than the V6, demonstrating V6 was more discriminatory of host patterns. Analysis of two regions at the same sequences depth verified that sequencing depth could not account for these results. Overall, the V4V5 region, while longer in length, offered less resolution for tracking host-associated populations. These results are consistent with the V6 region showing the highest variability (126). Our findings support the hypothesis that marker gene distribution patterns may reflect differences in the genome that accounts for presence in different host niches, but reiterates that only a portion of 16S rRNA gene cannot represent the exact organism that it comes from, and mapping the genetic markers to longer sequence reads could improve the tracking of specific organisms that are uniquely adapted to a host.

Environmental factors may affect the presence of fecal genetic markers. We found the most abundant *Lachnospiraceae* V6 sequence in sewage in our NGS dataset (designated Lachno1)

was not found in dairy cows in this study, but we have recovered this marker in the steer population in previous studies (30). Lachno1 was also found in the "cow" clone library, which was from beef cattle's feces. This could be attributed to the different diets of these cow populations, as it was reported that beef cattle fecal microbial communities are very likely to be shaped by feeding operations (127). In addition, beef cattle and dairy cows have been found to have different abundance patterns of major and minor gut bacterial groups (109). Our qPCR results demonstrated three out of the four cows positive for Lachno12 were from the same farm in Racine, WI, and all six of the negative cows came from different farms but in the same city of Brodhead, WI. Considering the possibility of different diets in cattle populations, there may be tradeoffs in sensitivity and specificity when choosing markers, and it might be necessary to develop markers that are directed toward certain types of animal operations or feeding regimens.

The most abundant markers are stable in sewage.

The ranks of marker abundance differed slightly across various sewage samples; but within most of our sewage samples or sewage contaminated water samples (n=38), the Lachno1, Lachno2, and Lachno3 markers were within the top four most abundant *Lachnospiraceae* sequences. Stability of these markers have also been found over a three-year period at two WWTPS in a single city (124). The initial taxonomy of the NGS dataset was based on SILVA 102 and was later updated to SILVA 119, and previously annotated *Lachnospiraceae* were annotated to *Christensenellaceae* and *Defluviitaleaceae* within order *Clostridiales*. We included sequences annotated as *Christensenellaceae*, recently described in a human fecal microbial community study (119) as it appears to be found preferentially in humans. We exclude the *Defluviitaleaceae* sequences because this organism appears to be non-fecal within the sewer systems (128). However, these sequences might be good candidates for tracing sewage release into the environment, since

they were not found in any of the animals tested; they may ultimately demonstrate the presence of sewage more specifically than any of the human derived markers that are found to crossover with other non-target hosts.

Lachno3 is highly human-specific.

Deep sequencing has revealed that only on rare occasion do marker sequences appear exclusive to a host, and even in these cases, further sequencing may reveal it is shared between two or more hosts. Rather, fecal community members appear to be host preferred more so than strictly human-specific (31). For human-associated marker assays, including Lachno2 (30), Lachno12, and the previously published human *Bacteroides* assays (61, 62, 69), cross reactivity was found, but usually for a low number of animals (Figure 2.4). The use of these assays synergistically could ultimately improve specificity. Current fecal identification is often based on usage of single human-associated alternative fecal indicator, however, there are several factors that can influence sharing of human and animal microbiome organisms (e.g. similar diets or cohabitation), and the use of a combination of human-associated assays can exclude false positive detection of human sources (65).

True animal cross-reaction needs to be differentiated from non-specific amplification by assay primers. In the qPCR validation portion of this study, Lachno3 qPCR results showed very low copy numbers in two cats, but the V6 marker sequences were absent in these cats' sequencing results. The Lachno2 assay validation results also included non-specific amplification. These signals could be caused by primers amplifying targets that are very close to the markers V6 sequences. High levels of similar but non-target DNA could account for non-specific amplification in other studies (78, 129). Increasing the temperature could reduce non-specific amplification but may negatively impact assay efficiency. This was observed in the case of optimizing the

temperature for the Lachno3 assay, as well as validating the Lachno2 assay; the slight increasing of temperature eliminated Lachno2 false positives in six deer, one cat and one pig in the highest fecal material level. This complication further highlights the usefulness of using two unrelated assays to detect human fecal pollution.

We also developed the Lachno12 assay that was primarily human-associated, but found in low levels in dogs, and sporadically present in certain cows. We found that despite a very low occurrence in one dog sample's sequencing result, the marker was not detected by qPCR. This finding illustrates that while sequencing may reveal low level of an organism, it may not be relevant in practical applications such as detection in water samples, where fecal material is already diluted. Further, these results helped confirm that low levels of amplification in cat samples by the Lachno3 assay was most likely non-specific as the cat samples were sequenced to a similar depth and the Lachno3 sequence was absent.

Given the high diversity of the microbiome of animals, mechanisms like co-habitation that give rise to shared gut microbiome, and diet and geographic differences among individuals within a host type, assessments for host specificity and sensitivity of markers should be ongoing. For example, Lachno2 was originally chosen for its high sensitively in sewage, and absence in cows (30). With the inclusion of dairy cows in this study, we observed cross over with this target. Additionally, we found sporadic presence in cats, dogs, and pigs, demonstrating the high sensitivity but low specificity of this marker. Similarly, the HF183 assay was later found to amplify signals in cat and dog samples, however redesign of the reverse primer and probe improved specificity in subsequent work (69).

Future application of *Lachnospiraceae* assays to fecal source detection in urban waters.

Humans and animals in urban areas contribute fecal pollution to waterways, including recreational beaches. It is not practical or perhaps even feasible to develop assays for every possible source in a complex watershed comprised of urban land use, however, use of multiple assays and interpretation of results in a tiered approach may provide insight into possible sources. For example, use of Lachno2 with highly specific assays like Lachno3 and HF183/BacR287 could help identify when nonhuman sources are present, without running separate assays for dogs, cats, or raccoons. Further, human-associated indicators target that is generally present and the most abundant in the human population (81), but when fecal pollution is derived from a smaller number of individuals, such as a broken lateral from a home, or a cross connection, results may be atypical. Multiple assays may be necessary when investigating small-scale contamination, like locating failures in sanitary sewer systems. Future work to increase the types and number of animals tested and increase the geographic coverage would provide more comprehensive assessments of specificity. Stormwater with fecal contamination from urban wildlife in particular lacks characterization and is difficult to distinguish from contributions from a limited number of humans. Shared resources such as fecal sample banks may be useful for researchers to validate use of assays in their watershed and so that they may compare with other areas. Overall a use of a combination of human-associated fecal marker assays with known cross-reaction potentials, as well as animal marker assays, will improve the resolution of fecal pollution source identification. This information is crucial for assessing possible risk from co-occurring pathogens, and for remediation of pollution sources in urban water environments.

Acknowledgements

We thank the Marine Biological Laboratory (MBL), University of Chicago and the Great Lakes Genomics Center (GLGC), University of Wisconsin- Milwaukee for providing expertise in NGS and Sanger sequencing. We thank Dr. Adélaïde Roguet for lending bioinformatics expertise to the project and Dr. Keri A. Lydon for her insightful discussion, and Katherine Halmo for assistance with DNA extractions. Funding for this work was provided by National Institutes of Health (NIH), grant number R01 AI091829.

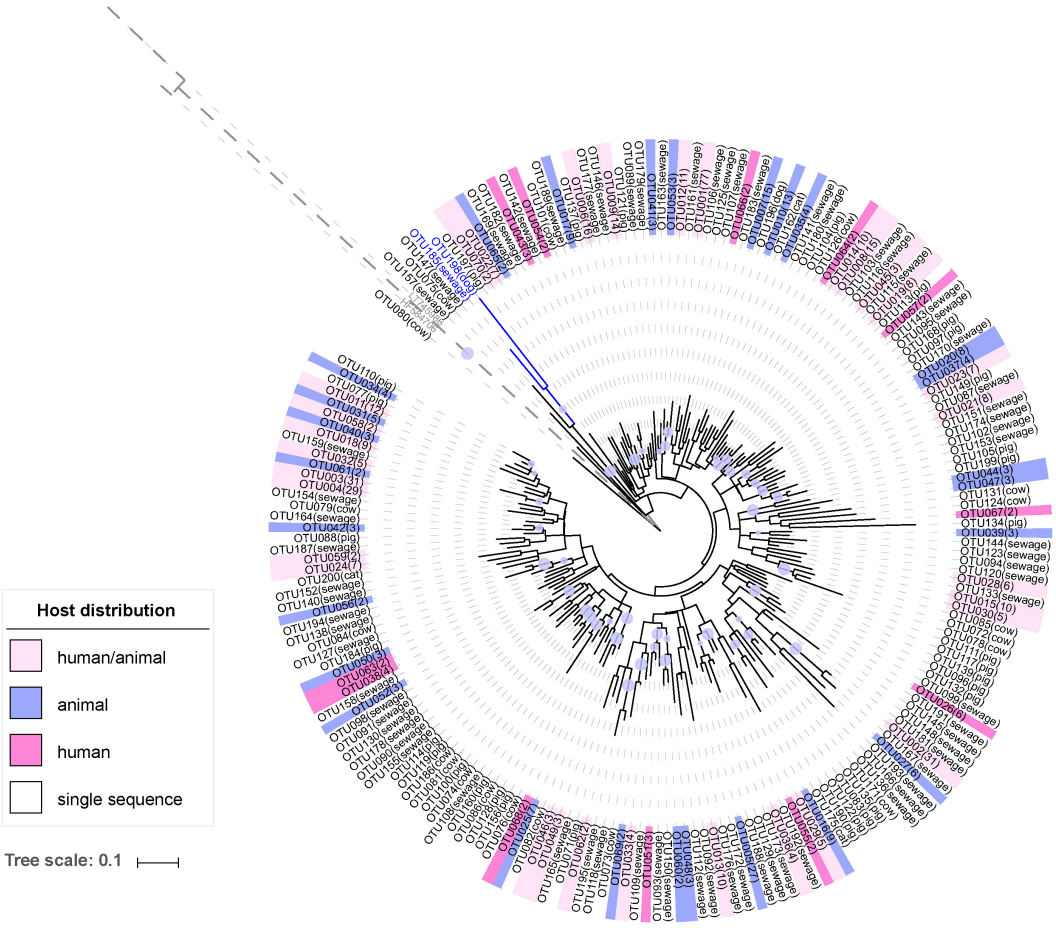


Figure 2.1 Phylogenetic tree comprised of the 200 representative OTU sequences from *Lachnospiraceae* clone libraries. The color range represents OTU host types (i.e., human only, animal only, or human/animal). The number of sequences found in each OTU is in parentheses. The family *Defluviitaleaceae* clade is in blue color. The *E. coli* outgroup clade is in dashed lines in gray color. Bootstrap values larger than 0.7 are indicated by lavender circles, and the values are proportional to the circle sizes.

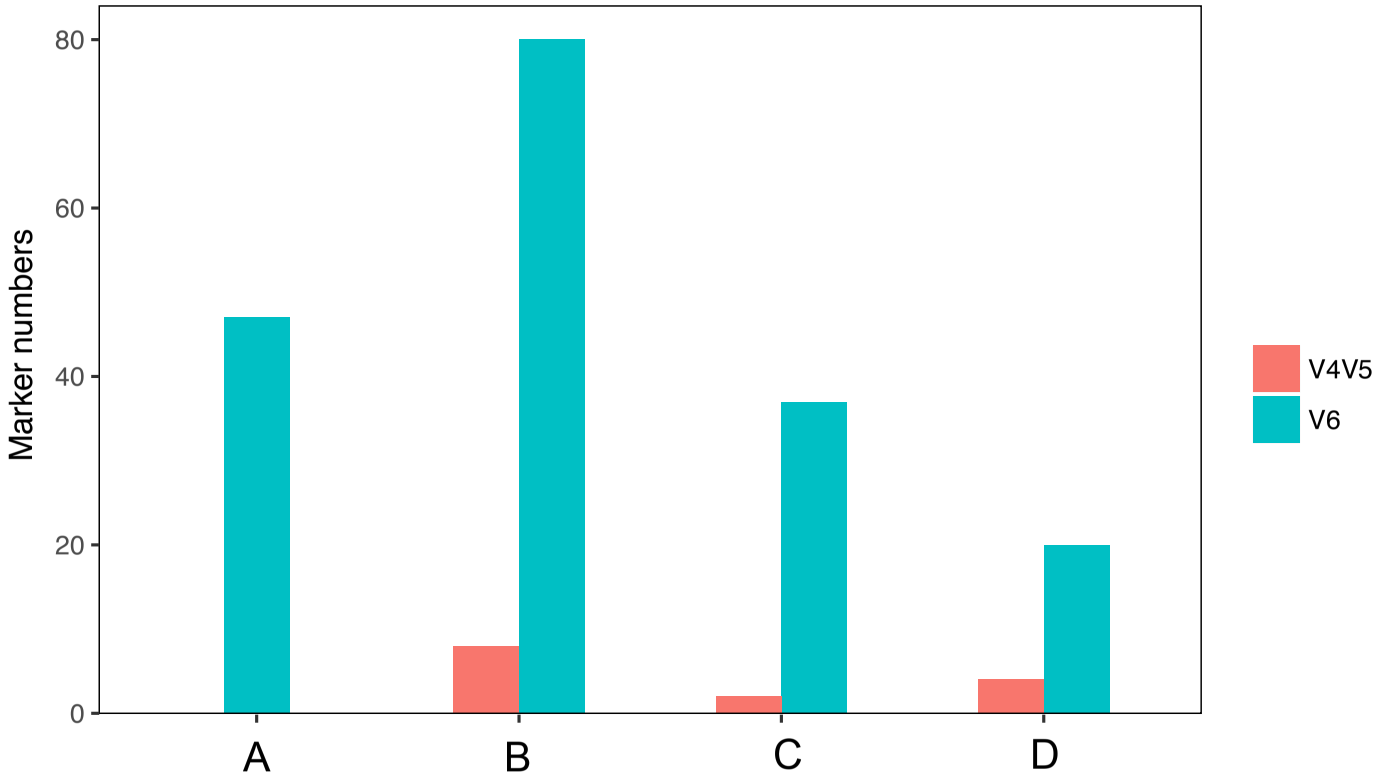


Figure 2.2 Comparison of *Lachnospiraceae* marker candidate numbers in V4V5 and V6 regions using a subset of same samples. Sequence regions are in different colors. Group A shows numbers of marker candidates that have 100% specificity and sensitivity. Group B shows numbers of marker candidates that have 100% sensitivity but 90% - 100% specificity. Group C shows numbers of marker candidates that have 100% specificity but 90% - 100% sensitivity. Group D shows numbers of candidates that only have 90% - 100% specificity and sensitivity.

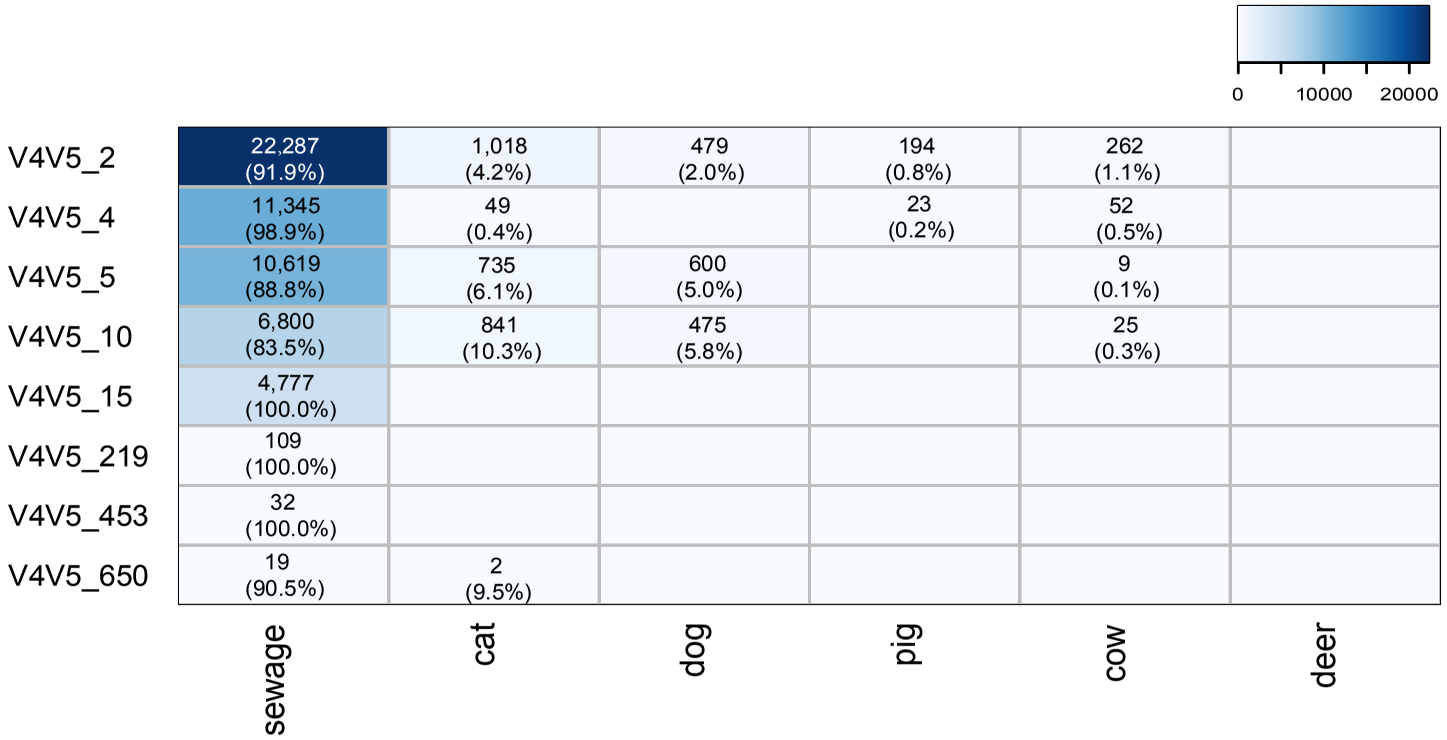


Figure 2.3 Abundances of Lachno3-associated V4V5 sequence types in sewage and five animal hosts. The abundance shown in each cell is normalized to the median sequence count for all samples and is converted to a percentage according to each V4V5 type's total abundance. The values increase from white to blue.

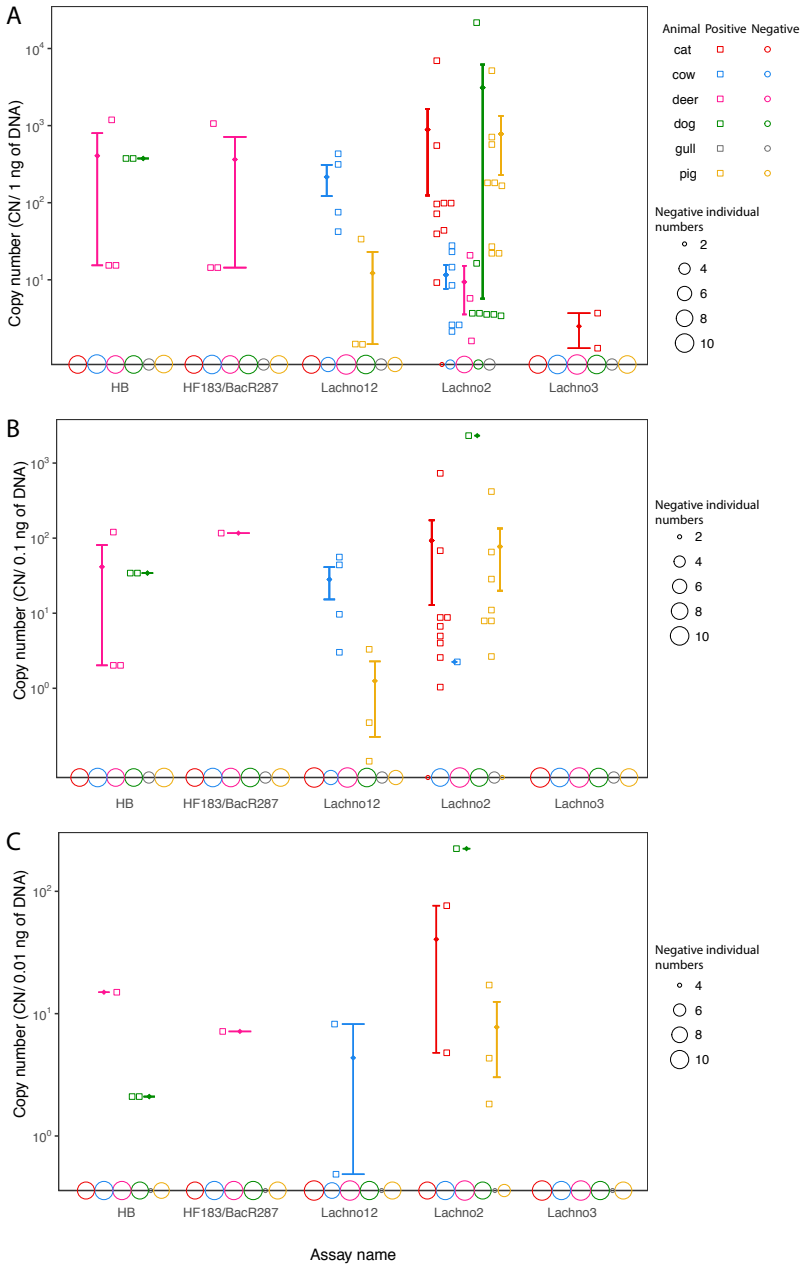


Figure 2.4 qPCR results of the Lachno3, Lachno12, Lachno2, HB, and HF183/BacR287 assays in animal fecal samples. Y-axis indicates the copy numbers, and X-axis shows the assays. The results shown are from (A) 1 ng μL^{-1}, (B) 0.1 ng μL^{-1}, and (C) 0.01 ng μL^{-1} DNA templates. Animals are shown in different colors. The result for each positive sample is shown as a hollow square, with those pooled shown as two adjacent squares. The results for all negative samples are shown as open circles on X-axis, with the sizes of the circles being directly proportional to the number of negative individuals. The error bars represent the mean CN (shown as a rhombus) with the standard error

Table 2.1 Primer and probe sequences of the Lachno3 and Lachno12 marker assays.

Assay name	Forward primer	Probe	Reverse Primer
Lachno3	5'-CAACGCGAAGAACCTTACCAAA -3'	FAM 5'-CTCTGACCGGTCTTTAATCGGA -3' NFQ-MGB	5'-CCCAGAGTGCCCACCTTAAAT-3'
Lachno12	5'-ATCTTGACATCCCTCTGACCGGGA -3'	FAM 5'-CGTCCCTTTCCTTCGGGACAGG -3' NFQ-MGB	5'-CTCAGAGTGCCCACCACTACGT -3'

Table 2.2 Applications of the Lachno3 and Lachno12 assays to environmental samples that had inconsistent results in HB and Lachno2 assays.

Sample name	Type	Site	Sample date	HB	Lachno2	Lachno3	Lachno12	Dog-Bact	Interpretation of presumptive sources*
				CN /100ml					
FT21217	Rivers	Kinnickinnic River grab	5/3/16	801	27,300	6,510	4,450	0	Human
FT21380	Stormwater	Kinnickinnic River grab	6/7/16	7,500	548,000	173,000	40,500	0	Human
FT20574	Rivers	Kinnickinnic River autosampler	9/8/15	39,700	188,000	75,400	37,300	15,800	Human/Dog
FT21332	Stormwater	Kinnickinnic River Manhole	5/10/16	0	1,350	0	170	19,200	Dog
FT12198	Stormwater	Wilson Park Creek Outfall 25	6/21/12	566	0	0	132	0	Raccoon
FT12431	Stormwater	Honey Creek 05	7/24/12	672	318	151	162	0	Human
FT14569	Beaches	South shore old beach 001	7/9/13	BLD	1,760	394	391	276	Human/ Dog
FT14570	Beaches	South shore old beach 002	7/9/13	166	3,460	1,000	1,430	1,060	Human/Dog
FT14571	Beaches	South shore old beach 003	7/9/13	0	18,100	985	765	27,900	Human/Dog
FT15268	Stormwater	Kinnickinnic River Outfall 47	10/31/13	3,540	0	0	0	0	Raccoon
FT15280	Stormwater	Kinnickinnic River Outfall New	11/6/13	225	33,700	249	196	8,710	Human/Dog
FT17167	Rivers	Kinnickinnic River	7/22/14	1,381	34,000	6,730	8,900	944	Human/Dog
FT17171	Rivers	Kinnickinnic River	7/22/14	375	6,450	821	1,610	0	Human/Cow
FT17708	Stormwater	Wilson Park Creek Outfall 07	8/25/14	BLD**	9,020	1,630	466	839	Human/Dog
FT17713	Stormwater	Wilson Park Creek Outfall 15	8/25/14	0	6,150	107	193	408	Human/Dog
FT18040	Stormwater	Wilson Park Creek Outfall 18	10/14/14	BLD**	9,620	1,890	185	0	Human
FT19920	Rivers	Menomonee River	7/9/15	0	675	265	161	0	Human
FT20193	Beaches	South Shore Old Beach 001	8/10/15	0	1,320	132	0	320	Human/ Dog
FT20724	Stormwater	Russell Avenue Manhole	10/28/15	8,560	0	45	256	0	Raccoon

* Fecal sources of cow, pig, and deer are not expected in these urban water samples
** BLD: Below the limit of detection.

Chapter 3 Highly specific sewage *Bacteroides* fecal marker assays

Abstract

The identification of sewage contamination in water has primarily relied on detection of the human *Bacteroides* using markers within the V2 region of the 16S ribosomal RNA (rRNA) gene. Despite establishment of multiple assays that target the HF183 cluster (e.g., *Bacteroides dorei*) and other *Bacteroides* organisms (e.g., *Bacteroides thetaiotomicron*), the potential for more human-associated markers in this genus has not been explored in depth. Here we examined genus *Bacteroides* population structure in sewage and animal hosts across the V6 hypervariable region and demonstrated the its specificity in sewage. Using near full-length clone sequences, we identified the sequences in the V4V5 and V6 hypervariable regions that are linked to the HF183 marker in the V2 region and found these sequences were present in multiple animals, demonstrating that regions downstream of the HF183 marker are not human-specific. In addition, the V4V5 and V6 regions contained human fecal marker sequences for organisms that were independent of HF183 cluster. The most abundant *Bacteroides* in untreated sewage was free-living, not human-associated but pipe derived. Two TaqMan qPCR assays were developed targeting the V4V5 and V6 regions of this organism. Validation studies using fecal samples from seven animal hosts (n=76) and uncontaminated water samples (n=30) demonstrated their high specificity for sewage. Freshwater *Bacteroides* were also identified in uncontaminated water samples, demonstrating that measures of total *Bacteroides* do not reflect fecal pollution. Comparison of two previously described human *Bacteroides* assays (HB and HF183/BacR287) in municipal wastewater influent and sewage contaminated urban water samples produced identical results, illustrating they target the same organism. While it is widely known that *Bacteroides* are major members

of the gut microbiota and host-specific, organisms within this genus have been used extensively to gain information on pollution sources. The detection of *Bacteroides* organisms that are specific to sewer pipe environment offers measures that are independent of the human microbiome for identifying sewage pollution in water.

Introduction

Human fecal pollution in urban waters from untreated sewage contains pathogenic bacteria, virus, and protozoa that cause gastrointestinal diseases through the ingestion of polluted water (6, 7), or skin, eye and respiratory infections through direct contact (6). Human source fecal pollution is considered a higher health risk to the public than animal sources (22, 53). Detection of traditional fecal indicator bacteria (FIB), such as fecal coliforms, *Escherichia coli* (*E. coli*) and enterococci (2) do not distinguish human source from animal sources of fecal pollution because they commonly occur in all mammalian intestines (31). Studies have demonstrated a lack of correlation between FIB levels and pathogen occurrence or adverse human health outcomes (8, 48–50) because some sources of fecal pollution do not carry human pathogens.

Microbial source tracking (MST) methods, which rely on quantification of levels of certain fecal microorganisms that are specific to a host (37), have been used for fecal source identification for a number of years (84). To date, the most characterized microorganisms used in MST belong to the genus *Bacteroides*, one of the most predominant genera in the human gut. The best-studied human *Bacteroides* marker to date is the HF183 marker, which is found in *Bacteroides dorei* (*B. dorei*) and its closely related taxa (54) and located in the V2 hypervariable region of the 16S ribosomal RNA (rRNA)

gene. This marker was first reported by Bernhard and Field (2000) as a PCR assay (i.e., HF183F/Bac708R) (61, 77, 78).

Because most human-associated *Bacteroides* markers have been developed using clone libraries that target order *Bacteroidales* using the Bac708R primer (54), established quantitative PCR (qPCR) assays have been limited to the V2-V4 hypervariable regions. Assays and their average specificities include HF183/SSHBac-R (91.1%) (62, 65, 73–75), HF183/BFDrev (76.8%) (67, 69), HB (90.9%) (26, 52), HF183/BacR287 (91.2%) (52, 69, 75), BacHum-UCD (77.9%) (65, 74–77), BacH (92.6%) (64, 75, 77), HuBac (54.5%) (63, 65, 77, 78), Human-Bac1 (44.4%) (66, 77) and BacHuman (81.5%) (68) (see Chapter 1, Table 1.1).

In addition to the assays that used the HF183 marker directly as a forward primer (i.e., HF183/ SSHBac_R, HF183/BFDrev, HB, and HF183/BacR287), assays that use primers or probes that overlap with the HF183 marker, such as the BacHum-UCD and BacH assays, were also reported low level animal cross-reactivity, further demonstrating human specificity of the HF183 marker. Human *Bacteroides* fecal marker PCR/qPCR assays have also been developed within 16S rRNA gene and genomic sequences of *B. thetaiotaomicron* (67, 70, 71), another predominant species in human feces that usually shows up more often in human feces than animals sources (71, 130). Overall, there is no bacterial fecal marker assay exclusively human-specific, and animal source cross-reactions were reported for all the PCR/qPCR assays mentioned above (Table 1.1).

The goal of this study was to explore the potential of genus *Bacteroides* for MST markers, in addition to the widely applied HF183 marker, to expand methods for sewage detection and quantification. By characterizing *Bacteroides* population structure in other

hypervariable regions other than the V2 region and delineating the association patterns among markers across V2, V4V5 and V6 regions, it may be revealed additional host-preferred and/or host-specific *Bacteroides* organisms, as well as help couple community sequencing data to marker assays. In this study, we compared the population structure of *Bacteroides* in 27 sewage and 151 animal fecal samples using next-generation sequencing (NGS) data in V6 region to explore its host specificity. We also explored the human *Bacteroides* V2, V4V5, and V6 regions sequence linkages and specificities by analyzing V2-V9 region sewage *Bacteroides* clone libraries. Multiple sewage-specific *Bacteroides* markers not related to the HF183 marker, including one from V4V5 region and one from V6 region, were identified from NGS data. We also identified a sewage-associated *Bacteroides* species that appears to be specifically propagated in urban sewer systems and developed two TaqMan qPCR assays targeting the V4V5 region and the V6 region, respectively.

Material and Methods

Sample collection and DNA extraction.

Influent sewage samples used for qPCR in this study were from Jones Island (JI) and South Shore (SS) wastewater treatment plants (WWTPs) in Milwaukee, WI (n=20), along with ten other U.S. cities representing geographical regions of the U.S. that were sampled in two different seasons over a year (n=20) (81). Sewage-contaminated river water samples (n=20) were collected during a 2016 Milwaukee combined sewer overflow (CSO) event. Agricultural-contaminated water samples were collected from the Milwaukee River (n=13) after rain in spring and early summer of 2014 and 2015; these samples also had evidence of sewage contamination, but at three to four orders of magnitude lower than

ruminant contamination as determined using a ruminant marker (131). Freshwater samples that had no evidence of human fecal contamination (i.e., had zero or extremely low colony counts of FIB and were negative in HB and Human *Lachnospiraceae* qPCR assays) (52) were collected from Lake Michigan (n=20) and Milwaukee area beaches (n=10).

A total number of 76 animal fecal samples, including 22 pigs, 13 dogs, 12 cats, 11 deer, 10 cows, four gulls, and four chickens, were collected for qPCR assay validation. Among these animal fecal samples, 46 were extracted in a previous study (52) but re-diluted for qPCR experiment in this study. Fecal sample processing and storage were as described previously (52).

All sewage, animal fecal and environmental water sample details, including their associated studies and qPCR results, are listed in supplemental Data Set 3.1.

NGS data used for oligotyping, clone comparisons and marker identification.

To examine overall population structure of *Bacteroides* populations, V6 region sequence data generated from two previous studies (52, 90) from 27 sewage samples and 151 animal fecal samples, including hosts of cat, dog, pig, cow, deer, raccoon and chicken, were analyzed using oligotyping (95). All raw sequences were trimmed using "cutadapt" software (132) and assembled using PEAR (133) software. Sequences were then classified using GAST (118) with comparison to SILVA reference database version 132 to parse out *Bacteroides* sequences. Oligotyping was run with parameters *-s* (the minimum number of samples where an oligotype present) equal to 9 (5% of total sample), *-M* (the minimum substantive abundance) equal to 85 and *-c* (number of base locations) equal to 33. The output of the oligotype count matrix was plotted using "ggplot2" package (134) in R

(version 3.5.1) (135). The statistical analysis of sewage and animal oligotypes was performed using the *adonis* function in the "vegan" package (136) in R.

For clone comparisons and marker identification (described below), V4V5 and V6 sequence datasets from previous studies (52, 81, 90) were obtained from the Visualization and Analysis of Microbial Population Structures platform (VAMPS, https://vamps2.mbl.edu) (117) with reference to SILVA database version 119. A "taxbyseq" file, which described whole community unique sequences, taxonomy, and abundance in each sample, was used. The total number of sequences for each sample was normalized to the median count for all samples sequence counts (V4V5 NGS dataset = 89341, V6 NGS dataset = 741189. Singletons were removed to form the whole community NGS datasets. The genus *Bacteroides* data was then extracted. The samples, their usage in this study, the associated studies, and SRA studies' accession numbers are listed in Dataset 3.2 Tab 1.

Sewage clone libraries.

Two sewage clone libraries were generated using four sewage influent samples from different U.S. cities (Milwaukee, Palo Alto, Laramie, and Key West) collected in August 2012 (81). The first clone library (library 1) was constructed using a human *Bacteroides* group forward primer (BacH_f) (64) and a universal 16S rRNA gene reverse primer (1492R); the BacH_f primer was chosen to form human *Bacteroides* amplicons that were long enough to cover the V2 region. The second clone library (library 2) was generated using the universal 8F primer and a new reverse primer, designated as 1030R (5'- CCACCTTCCTCACATCTTACGA -3'), which was designed to target *Bacteroides* broadly. The Probe Match function in the Ribosomal Database Project (RDP) (137) demonstrated that the 1030R primer matched 34,100 of 35,602 *Bacteroides*. The PCR

products were cloned into the pCR2.1 vector using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA), and plasmid were extracted as previously detailed (87). Sanger sequencing was performed with M13F, 331F and M13R primers using the ABI Big Dye Terminator Kit (Applied Biosystems, Foster City, CA) on an ABI 3730xl DNA Analyzer (Applied Biosystems, Foster City, CA) (87). In all, 332 sequences were generated in library 1 and 375 sequences were generated in library 2.

Linkage of the HF183 marker representing V2 region with the V4V5 region and the V6 region of *Bacteroides*.

The sewage clone libraries were compared with the HF183 marker sequence to identify clones containing this marker, and then with the unique NGS V4V5 and V6 sequences types to identify clones containing corresponding types; both comparisons were performed using BLAST+ (121). The V4V5 and V6 sequence types for each HF183 clone were compiled in Excel using the "VLOOKUP" function.

Freshwater *Bacteroides* population identification.

We used freshwater samples with low or absent levels of fecal pollution (n=35) that were previously sequenced for the V6 region to identify environmental *Bacteroides* (Data Set 3.2 Tab 1). These were compared to the sewage and animal fecal samples used in oligotyping. The "uncontaminated" samples were collected under baseflow conditions (i.e. no rain in the previous 48 hours) from Lake Michigan nearshore and offshore surface water (n=6) as grab samples (138), and the Milwaukee River, Kinnickinnic River and Menomonee River (n=29) using automated Teledyne ISCO 3700 full-size, portable, sequential samplers (131). To identify freshwater group preferred *Bacteroides* sequences,

the R package "indicspecies" (120) was applied with a setting of 999 permutations. These V6 region *Bacteroides* sequences are listed in Data Set 3.2 Tab 2.

***Bacteroides* marker identification.**

We used previously sequenced V4V5 region (52, 81) and V6 region (52, 90) sewage, sewage-contaminated water and animal fecal samples for marker identification (Data Set 3.2 Tab 1). For the V6 region, 22,006 unique *Bacteroides* sequences were present from 40 sewage and sewage-contaminated water samples, and 156 animal fecal samples; for the V4V5 region, 22,104 unique *Bacteroides* sequences were present from 195 sewage and 60 animal samples (see Data Set 3.2, Tab 3 and Tab 4 for the 100 most abundant *Bacteroides* V4V5 and V6 region sequences in sewage). These unique *Bacteroides* sequences were named according to their abundance ranks in sewage samples in the dataset (e.g. V4V5-1 and V6-1 are the most abundant V4V5 and V6 unique *Bacteroides* sequences in sewage NGS data, respectively).

To identify sewage-associated *Bacteroides* markers in the V4V5 and V6 regions, we used a subset of 16 sewage and 51 animal fecal samples from the NGS datasets, all of which had V4V5 and V6 regions sequenced. These data were analyzed using R package "indicspecies" (120) and the number of indicators from each region that had over 90% specificity and sensitivity were compared. To identify sewage-associated *Bacteroides* marker candidates, criteria that they must be over 90% sensitive and 100% specific from "indicspecies" results was applied. To identify the probable source of these marker candidates (i.e., whether they are human derived or likely residents of the sewer pipes), we compared sequences of these marker candidates with the National Center for Biotechnology Information (NCBI) nucleotide database and published V3V5 (139–141),

V4V6 (142), V6 (143) and V6V8 (144) region human stool sequences using BLAST+. The most abundant *Bacteroides* in sewage was specific to sewage but did not appear to be of fecal origin. This organism was chosen for qPCR assay development, with the corresponding markers identified as V4V5-1 (V4V5 region) and V6-21 (V6 region). Candidate markers, their specificities, the probable source and the sequences are shown in Appendix B Table 1 and Appendix B Table 2, respectively.

Phylogenetic placement of sewer pipe-associated markers.

Near full-length *Bacteroides* clones containing matched V4V5 and/or the V6 marker and marker candidates identified by "indicspecies" were used to construct a maximum likelihood tree in MEGA7 (113), based on Kimura 2 parameters (145) with Gamma distribution and invariant sites (K2 + G + I) and bootstrapping of 1000 replications.

Design of sewage-specific *Bacteroides* 16S rRNA gene fecal marker assays.

Primers and probes were designed based on 16S rRNA gene sequences alignment of animal fecal and sewage samples and visualized in MegAlign Pro program in DNASTAR software (version Lasergene 12). The marker sequences, a *B. dorei* 16S rRNA gene reference sequence (GenBank Accession Number AB242142) (146) and sewage clone library sequences containing the V4V5-1 and V6-21 marker sequences were included in the alignment. In addition, published near-full-length animal fecal *Bacteroides* clone sequences were also included in the sequences alignment from pig (147), dog (148), cow (109), chicken (149, 150) and mice (151), to discriminate from possible animal sources in the assay design. Primers and probes were named according to their base pair locations aligned to an *E. coli* reference sequence (GenBank Accession Number J01859) with comparison of universal 16S rRNA gene primers. Details are shown in Table 3.1. The

amplicon of the two assays and their reference clone GenBank accession numbers are listed in Appendix B Table 3.

QPCR experiments.

The qPCR reaction conditions, volumes, methods for establishing the standard curve and testing inhibitions were described in Chapter 2 as well as in a previous study (26). Each run included a sewage positive control and a no DNA control. The annealing temperatures were optimized by running a gradient qPCR using 1:100 volume to volume diluted sewage DNA (n=4) from 59°C to 64°C. Using the optimized annealing temperature, assays were applied to these sewage samples with different dilution ratios as described previously (52) to make sure no amplification efficiency was lost. The amplification program included one cycle at 50°C for 2 min, followed by one cycle at 95°C for 10 min, then 40 cycles of 95°C for 15 s followed by 1 min at 64°C for the BacV4V5-1 assay and 60°C for the BacV6-21 assay, respectively.

For assay validation to test for cross-reactivity, cat, dog, pig, cow, and deer fecal samples were tested in the format of individual samples (i.e., from a single animal) and pooled samples (i.e., from two single animals of the same type). Pooled samples were tested individually unless there was insufficient material. Gull and chicken fecal samples were tested as only individuals. Each animal fecal sample was tested at DNA template concentrations of 1 ng μL^{-1}, 0.1 ng μL^{-1} and 0.01 ng μL^{-1}, and the animal qPCR results were converted to the units of copy number (CN) per ng of input DNA, CN per 0.1 ng of input DNA and CN per 0.01ng of input DNA. For sewage samples, DNA templates were diluted 1:100 volume to volume. For environment water samples, DNA templates were tested without dilution. All the sewage and environmental water results were expressed in CN per

100 mL filtrated sample. A subset of 40 samples, including sewage, animal feces and environmental water samples, were tested for inhibition using salmon sperm DNA (~1,000 copies per reaction) as internal control as previously described (26). No inhibition was observed in these samples. Statistical analysis of qPCR assays correlations was performed using *cor* and *cor.test* functions in R. The qPCR assays slopes, y-intercepts, R^2, and efficiency values are shown in Appendix B Table 4.

Nucleotide sequence accession numbers. The partial 16S rRNA gene sequences of the sewage clone libraries were deposited in the NCBI GenBank database. The library 1 sequences were deposited under accession numbers MH515295 - MH515584 and MH515940 - MH515981, and library 2 sequences were under accession numbers MH515585 - MH515939 and MH515982 -MH516001. All V4V5 region NGS sequences of sewage and animal samples were from BioProject PRJNA261344 (81) and BioProject PRJNA433408 (52). The V6 region NGS sequences of sewage and animal were from NCBI Sequence Read Archive (SRA) SRP041262 (90) and BioProject PRJNA433407 (52); V6 region NGS sequences of baseflow lake samples were from SRA SRP056973 (138), and the baseflow river samples sequences are deposited to NCBI SRA SRP168560.

Results

Bacteroides population structures in sewage, animal hosts and freshwater samples.

We applied oligotyping to V6 region sequences of *Bacteroides* from 27 sewage influent samples and 151 animal fecal samples. In total, 1.48×10^7 *Bacteroides* reads (97.66% of total reads) were analyzed, including 1.96×10^6 reads from sewage samples and 1.29×10^7 reads from animal fecal samples. The oligotypes (n=1,730) distribution pattern in each sample is shown in Figure 3.1. There were 82 oligotypes exclusively found

in sewage, and of these, 30 were among the 100 most abundant sewage oligotypes. The sewage oligotype patterns were consistent between U.S. and Spain sewage samples and were distinguishable from animal hosts (Figure 3.1). The animal and sewage oligotype profiles were dissimilar in individual host groups (n=8, R^2 = 0.419, P = 0.001), and in sewage compared to a pooled animal group (R^2 = 0.119, P = 0.001). Bray-Cutis-dissimilarity-based hierarchical cluster analysis of *Bacteroides* oligotypes demonstrated animal and sewage samples clustered by source, and sewage was the most distant sample group compared to all other animal groups (Appendix B Figure 1). In addition, certain oligotypes were associated with specific hosts (Figure 3.1, Appendix B Figure 1). For example, 80 oligotypes were found only in cows, 11 were only in deer, five were only in dogs and four were only in pigs, indicating that organisms within genus *Bacteroides* could also be good targets for some animal fecal markers.

We also examined freshwater samples using "indicspecies" package to identify potential freshwater *Bacteroides* sequences based on the relative abundances of unique V6 sequences in freshwater samples (n=35), compared with sewage and animal samples (n=178). Three unique *Bacteroides* V6 sequences were found only in freshwater samples (relative abundance 4.7% ± 9.3% of all *Bacteroides*, mean ± SD), and 27 unique sequences were found in freshwater with comparatively low occurrence in sewage (relative abundance 37.4% ± 32.1% of all *Bacteroides*, compared with 1.5% ± 0.74% in sewage) and with no occurrence in animal samples. BLAST results against NCBI nucleotide database showed no identical match from human fecal source with the three freshwater specific sequences; for the 27 "freshwater-preferred" sequences, only two were found to have identical matches with human stool source, and another two matched with bioreactors

using farm animal waste (e.g. cow and pig). This indicates that there is a potential for *Bacteroides* populations to occur in the freshwater environment in the absence of fecal contamination.

Identification of V4V5 and V6 regions downstream of the HF183 human *Bacteroides* marker.

We utilized our sewage clone libraries to examine the specific marker sequences in the V4V5 and V6 regions of 16S rRNA gene that were downstream of the HF183 marker. A total of 136 clones matching the HF183 marker (41% of sequences) were found in library 1, which used the BacH_f primer to amplify human *Bacteroides* from the locus ahead of the HF183 marker. The HF183 organisms were associated with one primary V4V5 and one primary V6 sequence, designated V4V5-4 and V6-4 according to their rank of abundance in sewage samples in corresponding NGS datasets, respectively (Figure 3.2 A1). Only 3% of sequences in library 2, which represented total *Bacteroides* from sewage, had the HF183 marker (Figure 3.2 B1), indicating that the HF183 marker cluster is a small fraction of *Bacteroides* in sewage. In addition, all HF183 positive clones in library 2 had the BacH_f primer site, supporting that library 1 was inclusive of HF183 cluster of organisms.

We used the V4V5 and V6 NGS datasets to examine the host specificity of the primary sequences downstream of the HF183 marker and found they occurred in multiple animals. The V4V5-4 sequence occurred in 40% of the samples including cat, dog, cow, and deer, and the V6-4 sequence was found in 16.8% of the samples including cat, dog, cow, pig, chicken, deer, raccoon and rabbit, indicating the regions downstream of the HF183 marker are not specific to humans (Appendix B Figure 2). We tested a subset of these samples for the HF183 marker by qPCR in cases where DNA material was available. Overall, two of 13 samples containing the V4V5-4 sequence were positive for the

HF183/BacR287 assay. For available samples containing the V6-4 sequences, only one of three was positive with HF183/BacR287 assay. These results support that the downstream region is not specific, as opposed to these animals carrying a HF183 positive organism.

Potential human and sewage markers in *Bacteroides* V4V5 and V6 regions that are not associated with the HF183 cluster.

We aimed to identify additional human or sewage-associated *Bacteroides* markers in the V4V5 and V6 regions so that they could be used in PCR applications, but more importantly also be used as markers in sequencing datasets since these regions are commonly sequenced. We applied the "indicspecies" permutation test and identified markers from the V4V5 region and V6 region that were over 90% sewage specific and sensitive for sewage. Within these, there were nearly 20-fold more V6 region markers than V4V5 markers that were 100% specific and sensitive to sewage. These results may be due to the higher variability in the V6 region, which provides more resolution and therefore more unique human- or sewage-associated sequences than the V4V5 region. Although the V4V5 region had fewer markers, there were two that had 100% specificity and sensitivity (V4V5-1 and V4V5-7), both of which did not appear in human gut microbiome datasets, suggesting they were organisms from non-fecal fraction of the sewage microbial community. There were seven markers identified in the V6 region, with only one of these associated with human feces. The most abundant sewer pipe-associated markers fell within a clade of *B. graminisolvens* (Appendix B Figure 3). Human-associated and sewer pipe-associated markers and their specificities for the V4V5 region are in Appendix B Table 1 and markers in the V6 region are in Appendix B Table 2.

Assays development and sensitivity for sewage detection.

Two TaqMan qPCR assays were developed targeting the most abundant sewage-specific *Bacteroides*, designated as the BacV4V5-1 assay and BacV6-21 assay, respectively (Table 3.1). We tested these two assays in 40 U.S. sewage influent samples, including 20 from Milwaukee and 20 from ten other U.S. cities, and compared with the HB and HF183/BacR287 assays (Figure 3.2). All four assays showed 100% sensitivity in sewage. In Milwaukee sewage samples, the BacV6-21 assay showed about the same magnitude of CN ($4.9 \times 10^7 \pm 5.8 \times 10^7$ CN / 100 mL, mean ± SD) as the HB assay ($5.8 \times 10^7 \pm 3.3 \times 10^7$ CN / 100 mL) and HF183/BacR287R assay ($5.1 \times 10^7 \pm 2.6 \times 10^7$ CN / 100 mL), but was found to have greater fluctuation. The BacV4V5-1 marker was about four-fold higher than the V6-21 marker, with CN equal to $1.9 \times 10^8 \pm 2.2 \times 10^8$ per 100 mL sewage. In other U.S. cities, similar sewage sensitivities were detected for BacV4V5-1 and BacV6-21 assays, with the BacV6-21 assay of $5.5 \times 10^7 \pm 6.7 \times 10^7$ CN / 100 mL and the BacV4V5-1 assay of $1.9 \times 10^8 \pm 2.3 \times 10^8$ CN / 100 mL. The BacV4V5-1 assay mirrored the BacV6-21 assay fluctuation, suggesting they target the same organism (Pearson' r = 0.931, $P < 2.2 \times 10^{-16}$). Likewise, the HB assay and the HF183/BacR287 assay were tightly coupled (Pearson's r = 0.990, $P < 2.2 \times 10^{-16}$). The BacV5V5-1 and BacV6-21 assays were not correlated to either the HB or the HF183/BacR287 assays, with the Pearson' r ranging from -0.083 to -0.061.

The four *Bacteroides* assays were also tested in freshwater samples that had no known evidence of human fecal pollution (n=30). The HB, HF183/BacR287 and the BacV6-21 assay all showed negative results. The BacV4V5-1 assay, however, showed low CN in two lake/harbor samples (CN = 200 ± 6 per 100 mL, mean ± SD) and six beach

samples (CN = 180 ± 111 per 100 mL). All qPCR results for the four assays are shown Data Set 3.1.

***Bacteroides* assay validations in animal fecal samples.**

We validated the two sewage *Bacteroides* assays in 76 animals in the formats of individuals and pooled samples. Animal hosts included cat, dog, pig, cow, deer, gull, and chicken (Table 3.2). The BacV4V5-1 and BacV6-1 assays showed higher specificity than the two HF183 assays. The BacV4V5-1 assay gave a very low signal (5.2 CN per ng DNA) in one pig (pig pool 3) at 1 ng μL^{-1} and was negative at 0.1 ng μL^{-1} and 0.01 ng μL^{-1} DNA template levels. The BacV6-21 assay was negative in all animals at all three dilutions of DNA template. In contrast, the HB and HF183/BacR287 assays showed sporadic cross-reactivity with animals. The HB assay cross-reacted with one dog pool sample, whereas the HF183/BacR287 assay was negative for this sample (52). Results for all three dilutions of DNA are detailed in Data Set 3.1.

Sensitivity of *Bacteroides* assays in environmental water samples.

We tested the BacV4V5-1, BacV6-21, HB, and HF183/BacR287 assays in 20 sewage-contaminated local river water samples and 13 known-agricultural-contaminated local river water samples. Overall, the four assays were significantly correlated in these environmental water samples (Table 3.3). The BacV4V5-1 assay and BacV6-21 assay showed very similar fluctuation patterns and were more highly correlated to each other than the HB or HF183/BacR287 assay (Table 3.3). The BacV4V5-1 marker CN was at 4.0 ± 1.4 (mean ± SD) fold higher concentrations than the BacV6-21 marker, which corresponded to levels in the U.S. sewage samples that were tested. The HB and HF183/BacR287 assays showed identical CN fluctuation patterns in sewage-contaminated

river samples (Appendix B Figure 4A) and agricultural contaminated samples containing low levels of sewage (Appendix B Figure 4B) and were highly correlated to each other, indicating the equivalency of these two HF183 marker-based assays. In addition, all four assays clearly distinguished human contamination from ruminant in agricultural-contaminated river samples. Detailed CN data is shown in Data Set 3.1.

Discussion

Genus *Bacteroides* is a potential reservoir of sewage marker and certain animal host marker.

The identification of human fecal pollution provides evidence to assess public health risks caused by waterborne diseases. The fecal anaerobic microorganism *Bacteroides* has been utilized as a target for human fecal source detection since the specificity of the HF183 cluster was identified (84). Our study further explores the host specificity patterns of this genus among 27 sewage and 151 animal fecal samples across seven hosts using deep sequencing data. We demonstrated the host-specific nature of *Bacteroides* populations, consistent with previous studies using the V4V5 and V6 variable regions (90) and the V2 region (63, 65). The oligotyping results from previous studies and this study suggests that V6 region from genus *Bacteroides* could be used for marker identification for certain animals, such as cows and deer, since specific patterns were evident within these hosts. For example, the dairy cow and beef cattle oligotype patterns in our results was dissimilar, which may be caused by dietary differences (109, 127) (Figure 3.1, Appendix B Figure 1). With high variability among cattle, development of more restrictive host animal fecal markers could be useful; for example, specific markers

targeting dairy cows, or cattle on forage diets common to certain regions may be more feasible than employing a "universal" cattle marker.

***Bacteroides* organisms could be sewer system and freshwater derived.**

Among our identified sewage-specific NGS marker candidates, many did not match human microbiome organisms but appeared to be associated with the sewer pipe environment, and these organisms were among the most abundant *Bacteroides* in sewage. *Bacteroides* in mammalian guts is responsible for breakdown of complex polysaccharides (152–155). In addition, studies have been focused on free living *Bacteroides* species, which also have the ability to degrade complex organic matter, such as polysaccharides (156–158). The sewer pipe-derived *Bacteroides* organism represented by the V4V5-1 and V6-2 (and V6-21) markers closely matched *B. graminisolvens* based on near-full-length clone sequences (Appendix B Figure 4). This organism was isolated from a methanogenic reactor at a cattle farm where it was implicated in breakdown of hemecellulose (156), and has been detected in sequences generated from a microbial fuel cell reactor where it perform similar functions (i.e., degrading carbohydrates) (159). Just as *Bacteroides* have co-evolved and been selected for in the human gut (160), it appears that *Bacteroides* with urban sewer infrastructure may have been selected for or evolved in sewer pipes as a result of the available nutrition from sewer system inputs, where they provide further breakdown of material not completely utilized in the gut. Most notable is the ubiquitous occurrence of identical V4V5 and V6 marker sequences in all of the cities studied. It is unknown if these organisms were originally deposited as minor members of the human gut microbiome, or if they arose from environmental source. Given the short transit time in some of these systems studied (i.e. 6-24 hours) (87), coupled with the high abundance patterns in relation

to what is found in human fecal material, the sewer pipe-derived *Bacteroides* appear to be residents with the system.

We designed assays targeting the most abundant *Bacteroides* represented by the V4V5-1 marker. The most common marker for this organism in the V6 regions was V6-2 (Figure 3.2). However, this sequence was also found in animals. Therefore, we targeted a smaller subpopulation for qPCR assays represented by the V6-21 marker for the V6 region assay. The sewer pipe-associated marker assays strengthen confidence of sewage pollution detection because the targeted organisms are sewer pipe derived and have essentially no cross-reactivity with either human or animal feces, unlike gut derived organisms, where distinguishing members of the community are more often human-preferred organisms with lower occurrence in animal guts rather than strictly human-specific (31). Further, sewer pipe-associated markers may not be subjected to differences in the human microbiome in different regions, as is observed with some of the human gut derived markers (99, 161). Further testing of urban sewer systems worldwide is needed to determine their applicability in areas where the HF183 marker is low or absent. Importantly, since this organism appears to be free living rather than host associated, further validation studies of uncontaminated water are needed to determine if this organism is exclusively found in sewer systems and similar environments (manure detention ponds, anaerobic digesters, etc.).

In addition, we demonstrated the presence of *Bacteroides* in freshwater environment, which differed from *Bacteroides* in sewage and the seven animal fecal sources. *Bacteroides* in freshwater has previously been reported on *Cladophora* mats (162, 163), which is consistent with their ability to breakdown complex polysaccharides. In general, there was a single dominant sequence type in an apparently uncontaminated

sample, and different samples each had a different sequence type (Data Set 3.2 Tab 2), indicating the freshwater *Bacteroides* population may be very diverse and specific to location. Freshwater *Bacteroides* may be detected when using universal *Bacteroides* marker assays (65), causing false positive results for fecal pollution detection. In addition, high levels of these organisms may also interfere with *Bacteroides* assays that employ closely related primer sequences.

HF183 assays and sewage *Bacteroides* assays target on two separate organisms.

The HF183 assays and the sewer-associated *Bacteroides* assays target two independent *Bacteroides* organisms but are overall correlated. The high correlation between the two HF183-based assays (Table 3.3) indicated that they amplify the same *Bacteroides* organism. In our animal validation results, the HF183/BacR287 assay showed better specificity (93.2%) than the HB assay (90.5%) because of cross-reactivity with a certain dog sample in the latter assay. Overall, these two assays showed near identical sensitivity patterns among sewage, and sewage- and agricultural- contaminated environmental water samples, demonstrating they are interchangeable for the purpose of human fecal source detection. However, their application needs to be considered cautiously when employing the HB marker if dog waste is suspected, and specific testing using a canine marker or verification using a second human marker should be considered.

The BacV4V5-1 and BacV6-21 markers (targeting a sewer pipe-derived *Bacteroides*) had consistent ratios in sewage and sewage-contaminated environmental water samples (i.e., the BacV4V5-1 assay was about 4.0 ± 1.0 folds CN of the BacV6-21 assay) and were highly correlated in environmental waters. The linkage of the V4V5-1 and V6-21 markers in clone libraries (Figure 3.2 B2) verified that these two assays target the

same Bacteroides organism. In water samples where sewage was present, all four assays were correlated, demonstrating that they all detect sewage similarly.

NGS could reveal potential human fecal marker cross-reactions with animals.

Having access to a large V4V5 NGS dataset allowed us to examine other established human-associated *Bacteroides* assays targeting the V4 region. We compared the HumanBac-1 (66) and HuBac (63) assays, which are both located in the V4 region of the Bacteroides 16S rRNA gene, with our V4V5 NGS dataset; exact primer and probe matches of both assays were found in cat, dog, pig, cow, deer and rabbit, suggesting true animal cross-reactions occur, which explains the comparatively low human specificity of these assays (Table 1.1).

Cross-reaction of human fecal markers with animals can be influenced by multiple complex factors, such as similarities in gut microbial community as a result of dietary factors (52, 104, 127) and possible animal ingestion of human waste (164). We have previously noted that employing markers from two different bacterial groups such as *Bacteroides* and *Lachnospiraceae* can increase confidence in results where cross reactivity is suspected (52). For a well-designed fecal marker qPCR assay (e.g., optimized for avoiding dimers, hairpin structures, annealing temperature etc.), NGS could not only be used to verify the assay's host specificity, but also identify closely related sequences that might interfere. Deep sequencing has also been proven to be valuable for identification of host-associated markers on the scale of whole microbial community without the effort of constructing sequence clone library. However, linking different regions to the same organism is difficult without continuous, more extended sequence data since some variable regions appear to be less discriminatory and found in multiple host types. Therefore,

sequencing databases in common regions of 16S rRNA gene of sewage and animal fecal samples from a wide geographical range (e.g., across the U.S.) with key information of host (e.g., animal diet, cohabitation) could verify the applicability of markers and interpret site specific data. Data like this could be shared between research laboratories and would be extremely useful in assay validation in silico, therefore providing substantial evidence of specificity and sensitivity (52).

Combining NGS and qPCR for water quality assessments.

qPCR is indispensable for rapidly quantifying sources of fecal pollution such as human or cattle waste. However, most contamination scenarios are complex, especially in urban environment where there may be sewage contamination mixed with non-point sources from stormwater that add a significant fecal indicator bacteria burden (26). In addition, there are known sensitivity and specificity issues with each single fecal bacteria marker. Without annotating the microbiota composition in animal sources, human fecal marker cross-reactivity has not been completely characterized (165). NGS data creates a high-resolution inventory of presented organisms. With falling sequencing costs, NGS may be feasible for directly characterizing fecal pollution sources in the future. Further, fecal bacteria sequences within these datasets that do not match a characterized source could be used to indicate extraneous sources that may be contributing fecal indicator bacteria but are not considered as a significant human health risk (i.e., bird or pet waste, urban wildlife). Anchoring the relative abundance derived from sequencing with qPCR for host-associated markers will provide quantification. As the complexity of fecal pollution signals is unraveled, combining NGS with qPCR methods for source tracking may become common metrics for assessing microbial water quality.

Acknowledgements

We thank the Marine Biological Laboratory (MBL), University of Chicago, and the Great Lakes Genomics Center (GLGC), University of Wisconsin - Milwaukee, for offering expertise in NGS and Sanger sequencing, respectively. We thank Dr. Adélaïde Roguet for help with raw sequence processing, and Melinda J. Bootsma for assistance with qPCR in this study. Funding for this study was provided by National Institutes of Health (NIH), grant number R01 AI091829.

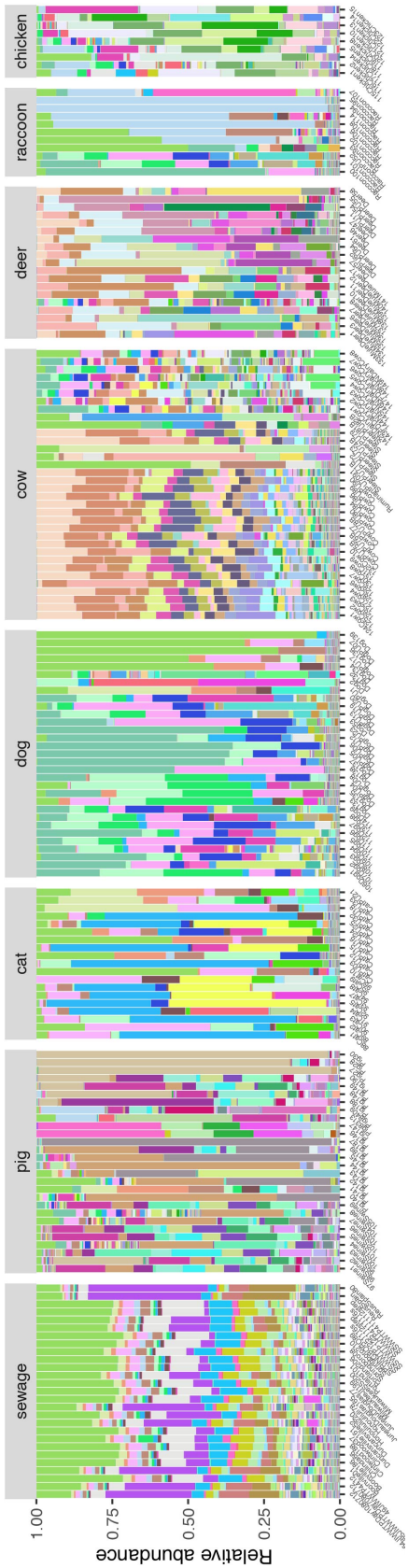


Figure 3.1 Oligotype patterns of the V6 region sequences of *Bacteroides* 16S rRNA gene in sewage and seven animal hosts. Samples are grouped by host types. Oligotypes are represented in different colors, with bar height of each color representing the relative abundance.

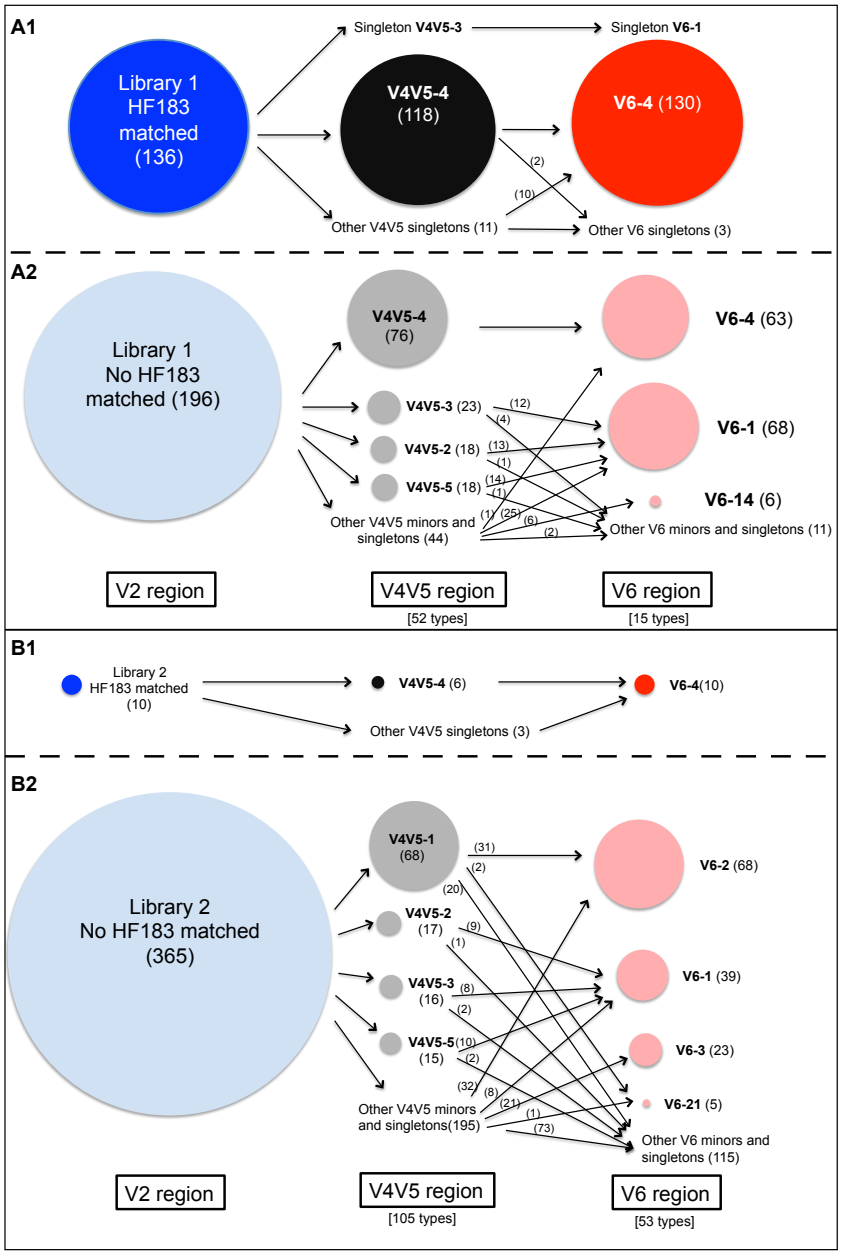


Figure 3.2 Associations of the V2, V4V5 and V6 regions of sewage *Bacteroides* clone sequences. A1. Associations in HF183 clones in library 1; A2. Associations in non-HF183 clones in library 1; B1. Associations in HF183 clones in library 2; B2. Associations in non-HF183 clones in library 2. The deep/light blue circles represent V2 region, black/gray circles represent V4V5 region, and the red/pink circles represent V6 region. Circle sizes are proportional to the sequence reads numbers except the non-HF183 matched V2 region in library 2, which is smaller than the actual proportional area for a better visualization. The unique sequence numbers in V4V5 and V6 regions identified in the clone libraries are annotated at the bottom of A and B. Numbers within parentheses indicated clone sequence numbers. Clone sequences that have no NGS matches are not included.

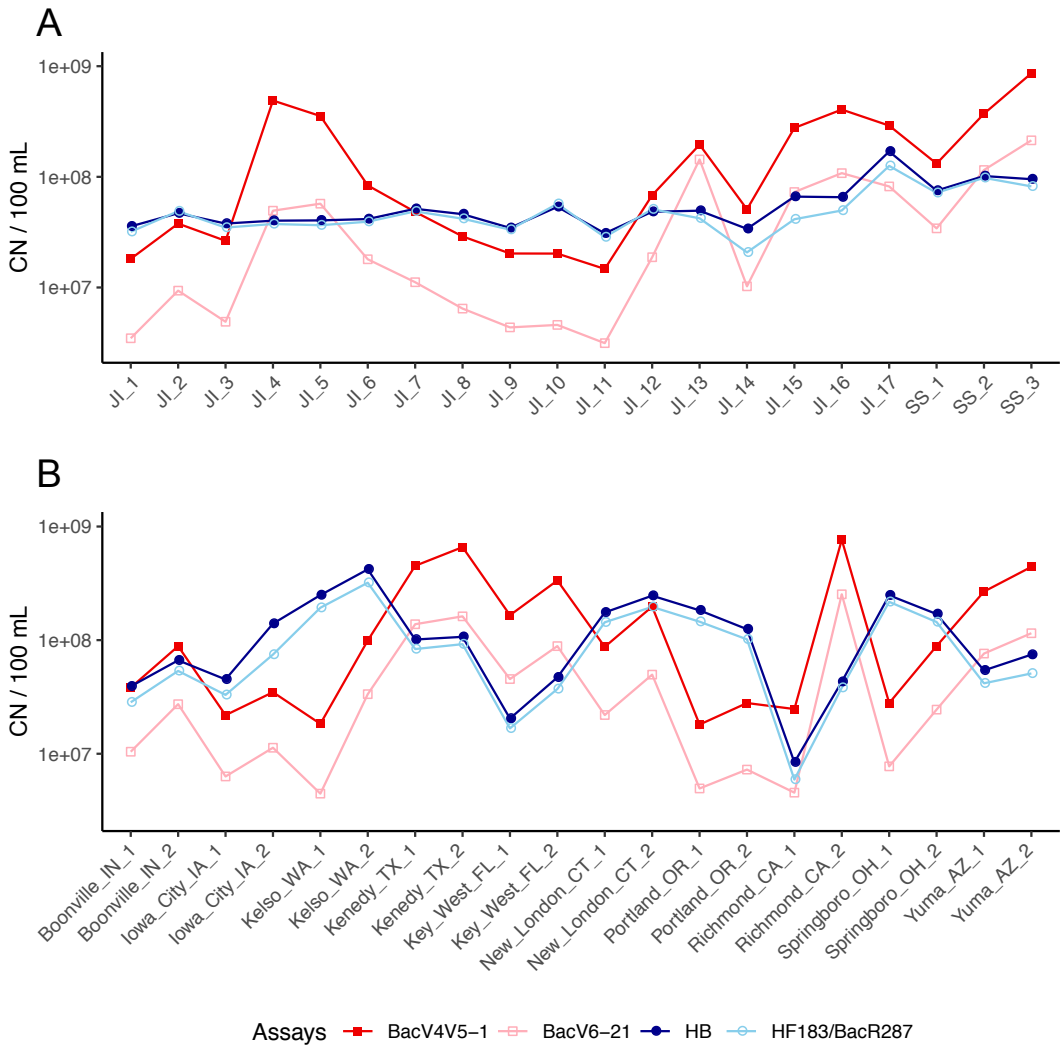


Figure 3.3 Comparison of the BacV4V5-1, BacV6-21, HB, and HF183/BacR287 assays copy numbers (CNs) in sewage samples. Line graph is used to show the fluctuation patterns of assay results, not correlations of samples. A shows the four assay CNs in 20 local sewage samples from Jones Island (JI) and South Shore (SS) WWTPs, Milwaukee. B shows the four assays CNs in 20 sewage samples from ten other U.S. cities, each tested at two different time points.

Table 3.1 The BacV4V5-1 and BacV6-21 marker assays.

Assay name	Forward primer (5’ – 3’)	Probe (5’ – 3’)	Reverse Primer (5’ – 3’)
	(Primer/probe name) sequence		
BacV4V5-1	(Bac573f) AAGGGAGCGTAGGTTGACATA	(Bac599p) FAM-CAGCTGTGAAAGTTTACGGCTC-NFQ-MGB	(Bac673r) CGCCCACCTCTTGTACACT
BacV6-21	(Bac989f) GCTTGAATTGCAGAGGAATA	(Bac1010p) FAM-AGTTGAAAGATTATGGCCGCA-NFQ-MGB	(Bac1162r) GCAGTCTCACTAGAGTCCTCAG

Table 3.2 Animal validation results of the *Bacteroides* assays.

Animal	Total number (number of pools containing two samples)	BacV4V5-1		BacV6-21		HB[a]		HF183/BacR287[a]	
		Positive numbers	Average CN per ng DNA Average CN per gram of fece	Positive numbers	Average CN per ng DNA Average CN per gram of fece	Positive numbers	Average CN per ng DNA Average CN per gram of fece	Positive numbers	Average CN per ng DNA Average CN per gram of fece
Cat	13(1)	0	0	0	0	1	8 115,000	1	5 77,000
Cow	10(2)	0	0	0	0	0	0	0	0
Deer	11(1)	0	0	0	0	3[b,c]	406 404,000	3[b,c]	364 362,000
Dog	13(2)	0	0	0	0	2[b]	375 3,330,000	0	0
Pig	22(2)	2[b,c]	5 84,700	0	0	0	0	0	0
Chicken	4	0	0	0	0	0	0	0	0
Gull	4	0	0	0	0	0	0	0	0

[a] Partial results of the HB and HF183/BacR287 assay validations were generated in Chapter 2.
[b] A pool was positive in each of these animal groups and was counted as two positive animals.
[c] The positive pooled samples were also tested in individuals at 1 ng μL^{-1} DNA level, see Data Set 3.1 for details.

Table 3.3 Pearson's correlation of the four *Bacteroides* assays in 20 sewage-contaminated and 13 agricultural contaminated water samples.

Assays	BacV4V5-1	BacV6-21	HB	HF183/BacR287
	Pearson' r (*P* value)			
BacV4V5-1	1.000	-	-	-
BacV6-21	0.995 ($< 2.2\times10^{-16}$)	1.000	-	-
HB	0.842 (7.9×10^{-10})	0.817 (6.7×10^{-9})	1.000	-
HF183/BacR287	0.824 (3.8×10^{-9})	0.792 (3.9×10^{-8})	0.995($<2.2\times10^{-16}$)	1.000

Chapter 4 Exploring mechanisms for cross-reaction of human fecal markers using animal fecal microbial communities

Abstract

Quantitative polymerase chain reaction (qPCR) assay for human-associated fecal marker has been one of the main approaches for human fecal pollution detection in water. Human-associated fecal marker assays (i.e., Lachno3, HF183 and BacV6-21) have demonstrated sporadic positive results in animal sources despite their high specificities to the human fecal source. It is unclear whether these amplifications are caused by low or sporadic levels of a marker, or the presence of a closely-related organism with a highly similar sequence in the regions of the marker. In addition, the distribution patterns of recently-described human-associated markers in animal fecal microbial communities have not been explored in depth, which is crucial for evaluating if detection of a marker in animal sources is a true or false reaction. Here we analyzed V6 region 16S ribosomal RNA (rRNA) gene NGS data from 271 animal fecal samples, among which 180 were also tested using human marker qPCR assays. The Lachno3 assay, a multiplexed *Escherichia coli* (*E. coli*) and human *Bacteroides* (HB) assay, and a multiplexed *Enterococcus* spp. and BacV6-21 assay were performed. The two multiplexed assays were validated in this study. Our results suggest that compositions of animal fecal microbial communities were influenced by both host physiology and environmental factors on whole community, single family (i.e., *Lachnospiraceae*) and single genus (i.e., *Blautia* and *Bacteroides*) levels. Cross-reaction of human markers with animal fecal samples were associated with certain compositions of *Blautia* or *Bacteroides* at the unique sequence level. In addition, in certain cases (e.g., domestic rabbit versus wild rabbit), factors such as diet and habitat correlated to high-level amplification of human markers. Overall specificities of human markers in NGS data were 99.6% for BacV6-21 and 97.0% for Lachno3. Specificity of human marker assays in qPCR

results were BacV6-21 (95.6%) > Lachno3 (92.8%) > HB (91.7%). Most human/sewage marker positive qPCR results in animals were in low signals. The Lachno3, HB and BacV6-21 assays' copy numbers (CNs) in animals were on average two to three orders of magnitude lower than the average CNs of general fecal indicators, indicating such cross-reactions may not affect markers' ability to indicate human fecal source. We found that the presence of organisms that have sequences highly similar to the marker sequence was the main reason for false positive reactions in animal fecal samples. The qPCR cross-reaction mechanisms can be used for guiding improvement of corresponding assays such as assay modification or result interpretation. Our finding also supports previous findings that human fecal markers are less likely "host-specific" but rather "host-preferred". A combination of marker assays from different targeted microorganisms should be used for increasing the confidence of human fecal pollution detection.

Introduction

The microbial source tracking (MST) method has been applied to track human fecal pollution in water environments for about two decades (61, 84, 166, 167). The basic theory of MST method is to detect the presence of host-specific fecal microorganisms to indicate the source of fecal pollution (37, 54). Bacterial human fecal marker assays were initially developed as polymerase chain reaction (PCR) assays (61) and subsequently quantitative PCR (qPCR) assays (62, 65, 67, 79). Many of these assays target the V2-V4 hypervariable regions of 16S ribosomal RNA (rRNA) gene of *Bacteroides*, including one of the best studied markers, HF183, from a human-specific *Bacteroides*. More recently, assays have been developed from non-fecal *Bacteroides* in sewage (101). These assays, designated the BacV4V5-1 and BacV6 -21 assays, provided another measurement for sewage detection independent of human and animal fecal microorganisms. In addition, human fecal marker assays have also been developed from bacterial family *Lachnospiraceae*, within which the Lachno3 assay demonstrated high human specificity (52). The sewage *Bacteroides* assays and the human *Lachnospiraceae* assay were developed based on next generation sequencing (NGS) data, which allowed for comparison of animal and sewage microbial communities for marker identification.

Perhaps the most important performance characteristic of human fecal marker assay is host specificity, which refers to the marker's ability to accurately detect targeted fecal source (54, 72). Despite the large number of bacterial human fecal marker assays that have been developed, cross-reaction with animal fecal sources have been reported for all. Reduced specificities of previously developed human marker assays (63, 66, 68) could be attributed to the limitation of clone library methods, where the representation of targeted

host organism clones for a host source could be inadequate (54, 101). In the case of NGS-based assays such as Lachno3 and BacV6-21, the host specificities have been validated *in silico* using an in-depth sequence inventory. A recent study tested the Lachno3 and BacV6-21 assays and a HF183 marker assay in a total of 360 animal fecal samples across 14 hosts (100). Although Lachno3 demonstrated high specificity (95%), animal cross-reactions were observed for both Lachno3 and BacV6-21 assays (100). The reason why these two marker assays were positive in animal sources is poorly understood, especially for the BacV6-21 marker that targets a sewer pipe-derived *Bacteroides* rather than human or animal fecal organism. Some hypotheses for the positive results of these human marker assays in animal sources include: 1) the qPCR assay amplified sequence that is highly similar but not identical to with the marker gene; 2) the qPCR amplified a marker that is commonly present in an animal host, but this host was not included in the NGS dataset used for marker identification; or 3) the qPCR amplified the marker in an animal individual that has an atypical gut microbial community composition compared to marker negative individuals, due to random environmental factors that could shape gut microbial community composition.

To date, most studies use qPCR to validate marker cross-reactions. When sequencing data was available, the presence of a marker in an animal fecal sample could also be identified *in silico* (52, 91). Here we analyzed V6 NGS data from 271 individual and pooled animal fecal samples collected from the United States (U.S.) and Australia. For a subset of 180 samples, qPCR experiments for three human marker assays were performed, including the Lachno3 assay, a multiplexed *Escherichia coli* (168)/ human *Bacteroides* (HB) assay (26, 61, 65) and a multiplexed *Enterococcus* spp. (169)/BacV6-21 assay. By

comparing a marker's presence in NGS data and the corresponding qPCR assay, we identified animal samples that were positive for both approaches or only positive for one approach. Mechanisms for human fecal marker cross-reactions in these animals were then explored by examining the microbial community compositions of the animal fecal samples for atypical patterns, which suggested shifts in a certain genus (i.e., genus *Bacteroides*) towards a human pattern. Also, sequences closely-related to markers that amplified by qPCR is another possible reason for positive results of human marker assays in animal fecal samples.

Material and methods

Sample collection and processing.

A total number of 379 single animal fecal samples across 22 hosts were collected from the U.S. and Australia and sent in as raw samples or extracted DNA (See Data Set 4.1 for detailed information). For raw sample collection, up to 5 mL or gram sample were collected into a 50 mL conical sterile centrifuge tube with 2 mL ASL Buffer (i.e., stool lysis buffer, Qiagen Inc., Valencia, CA) added. Sample tubes were stored at -80°C within 4 hours of collection, and later sent on ice to University of Wisconsin – Milwaukee (Milwaukee, WI). DNA samples were sent freeze-dried. For raw samples, DNA was extracted in the formats of individual samples (n=244) and pooled samples (n=27 pools). For individual sample, DNA was extracted using the standard protocol of QIAamp DNA Stool Mini Kit (Qiagen Inc., Valencia, CA). For each pooled sample, five individual samples that belong to the same host type and from the same sampling event (i.e., from the same location and the same sender) were combined. DNA extraction of pooled samples combined five lytic stool samples in step 2 of the standard protocol, with each of the lytic

sample at 1/5 volume of that used in an individual sample extraction. A final number of 271 animal fecal samples were prepared for sequencing. Details of fecal sample processing, storage and DNA extraction were also described previously (52, 101).

NGS data analysis.

All collected animal fecal samples were sequenced for V6 region using Illumina Hiseq sequencing platform at the Marine Biological Laboratory at the University of Chicago. The paired-end, short reads sequencing method was described in a previous publication (170). NGS data of healthy human fecal samples (n=6) and U.S. sewage samples (n=8) were obtained from a public data set (171) and previous studies (52, 81), respectively. Animal hosts were purposely grouped for sequence analysis, including cat and dog as the "pet" group; antelope, cow, deer, goat and sheep as the "ruminant" group; chicken, duck, gull, goose and parrot as the "bird" group; and bear and raccoon as the "wildlife" group. Other animal hosts, including horse, pig, rabbit, kangaroo, flying fox and alligator were not grouped. All sequence data was stored and managed on the Visualization and Analysis of Microbial Population Structures platform (VAMPS, https://vamps2.mbl.edu) (117). The total number of sequences in each sample was normalized to the median total sequence count of all samples (median = 513,566). Singletons of sequence count were then removed. In all, 319,418 unique V6 sequences from 137,193,309 reads were analyzed. Analysis of the dataset was performed in R (version 3.5.1) (135).

Statistical analysis.

Statistical analysis was performed in the "vegan" package (136) in R. Non-metric multi-dimensional scaling (NMDS) analysis was performed based on Bray-Curtis dissimilarities of samples, which were calculated based on counts of unique sequences.

QPCR experiment.

The NGS animal samples that had enough DNA extract for three sets of assays (i.e., total amount of DNA > 40 ng) were validated for human fecal marker assays using qPCR (n=180). Sixty-eight Australia animal samples and 11 U.S. animal samples did not have enough DNA for qPCR. Three sets of TaqMan qPCR assays were adopted in this study, including the Lachno3 assay (52), multiplexed *E.coli* (168)/human *Bacteroides* assay (26) (*E. coli*/HB), and multiplexed *Enterococcus* spp. (169)/BacV6-21 assay (101) (ENT/BacV6-21). Each sample's DNA was tested in duplicate. Standard curves were tested in triplicates using plasmids at each concentration, ranging from of 1.5×10^6 to 1.5 copy numbers (CNs). For all three assays, animal fecal sample were tested at DNA template concentrations of 1 ng μL^{-1}, 0.1 ng μL^{-1}, and 0.01 ng μL^{-1}. Each run included a positive control using sewage sample and a blank control using sterile DNA-grade water. Lachno3 assay qPCR validation experiments were performed as previously described (52).

For multiplexed assays, VIC reporter dye was used for general indicator assay probes and FAM reporter dye was used for human marker assay probes. For both multiplexed assays, a 25 μL qPCR reaction system was used, including TaqMan Gene Expression Master Mix (Applied Biosystems; Foster City, CA), 1 μM each primer, 80 nM for each VIC reporter dye probe and 80 nM for each FAM reporter dye probe; DNA input volume was 5 μL. The amplification program for both multiplexed assays included one

cycle at 50°C for 2 min, followed by one cycle at 95°C for 10 min, and then 40 cycles of 95°C for 15 s followed by 1 min at 60 °C. Validations of multiplexed assays were performed for standard curves and samples. Standard curve validation was carried out using each target's plasmid or genomic DNA under single assay conditions and under multiplexed assay conditions (Appendix C Table 1). Sewage samples (n=16) and animal fecal samples that were known to have no cross-reaction with Lachno3, HB and BacV6-21 assays (n=4) were used for comparing single and multiplexed conditions. Student's t-test was performed to test statistical difference between cycle threshold (Ct) values of single and multiplexed runs for both sets of multiplexed assays.

Method blanks (MB) were extracted with no extraneous DNA added (n=3). Sample processing controls (SPC, n=3) were extracted with 0.2 ng μL^{-1} salmon sperm (SS) genomic DNA spiked in MB extractions. A subset of six animal fecal samples were re-extracted with 0.2 ng μL^{-1} SS genomic DNA spiked in to test for extraction efficiency. MB, SPC, SS DNA-added animal fecal samples and non-SS DNA added animal fecal samples were then amplified using the Sketa22 assay with data acceptance criteria as described (169, 172). Extraction efficiency was then determined for each of the six animal fecal samples based on recovery of DNA amount (ng) using the standard curve method (i.e., Log_{10}[ng of DNA] versus Ct). An inhibition test was performed by spiking in about 0.03 ng SS genomic DNA in each reaction in fecal samples of 10 ng μL^{-1} DNA with four no template controls. A subset of 46 animal fecal samples that encompassed all animal hosts were tested; four no template control samples were also included. Lower limit of quantification (LLOQ) for each assay was defined as the 95% prediction of the upper limit of the 15 copies DNA

standard dilution based on corresponding standard curve. The qPCR assay slopes, y-intercepts, R^2 values, efficiencies and LLOQ values are reported in Appendix C Table 2.

Results

Distribution patterns of *Lachnospiraceae*, *Blautia* and *Bacteroides* in human, sewage and animal groups.

A total number of 271 bacterial families were classified from all human feces, sewage and animal fecal samples (n = 271). Human samples (n = 6) contained 124 families, sewage samples (n = 8) contained 249 families, and all animal samples combined comprised 256 families. Figure 4.1 shows the distribution patterns of families of more than 1% abundance of the whole community (n = 24) across all the different sample types. Microbial communities of human, sewage and animal host groups (n = 12) had a correlation of $R^2 = 0.361$ (p value = 0.001). Appendix C Figure 1 shows the distribution patterns of these 24 families in sewage, human and animal host groups. Seven out of the 24 families were within the top 20 abundant families in sewage, human and animal host groups simultaneously, including *Lachnospiraceae* and *Bacteroidaceae*. Distribution patterns of *Lachnospiraceae*, *Blautia* and *Bacteroides* were then explored in depth. These three taxa were present in sewage, human and all animal hosts, suggesting that they are common taxa in human and animal fecal sources.

Lachnospiraceae, Blautia and *Bacteroides* were shaped by host physiology and diet.

Distribution patterns of *Lachnospiraceae*, *Blautia* and *Bacteroides* were explored based on Bray-Curtis dissimilarities of human, sewage and animal samples within each taxonomic group (Appendix C Figure 2). For all taxa, similarities were observed within most host groups, including human, sewage, pet (i.e., dog and cat), ruminant, horse and

pig. The *Lachnospiraceae* pattern as a family was mostly consistent with the genus *Blautia*, without obvious similarity between human and animal sources observed (Appendix C Figure 2A, B). *Bacteroides* had a stronger in-group similarity in cows compared to the other two taxa (Appendix C Figure 2C). Moderate similarities between human and some animal hosts were observed in some cases, such as human and pet/pig for *Lachnospiraceae* and *Blautia*, and human and ruminant/wildlife/rabbit for *Bacteroides*. In all, these taxa showed distribution patterns that mostly corresponded to the host group.

To identify whether the distributions of *Lachnospiraceae*, *Blautia* and *Bacteroides* in different hosts was impacted by potential environmental factors (e.g., geographical region and diet), we performed non-metric multi-dimensional scaling (NMDS) analysis for all mammal samples (n = 219) based on Bray-Curtis dissimilarity (Figure 4.2). For whole community, *Lachnospiraceae*, *Blautia* and *Bacteroides*, herbivores (e.g., ruminant, horse and rabbit) were well separated from carnivores (i.e., dog and cat) on X-axis, indicating that diet is an important factor shaping the fecal microbial communities in these animals. In particular, the pig samples, which had plant- and/or grain-based diet, were clustered closer to the herbivore group than the omnivore group, further supporting that diet impacts the fecal microbial communities of different animal hosts.

We also explored the impact of geographical region on these taxa's distributions among different host groups (Figure 4.3). Three regions that all had cow, deer, dog, horse and pig samples collected (n = 108) were chosen, including Australia (AUS), Texas (TX) and Wisconsin (WI). Cow and deer samples were combined as the ruminant group. For all taxa, samples were more closely grouped by host type rather than geographical region,

demonstrating host physiology and/or diet rather than geographical region was the main factor(s) shaping fecal microbial communities of different animal hosts.

Multiplexed qPCR assay validations and sample processing control results.

All single and multiplexed standard curves of *E. coli*/HB assay and ENT/BacV6-21 assay showed R^2 values of > 0.990. The amplification efficiencies ranged from 92.7% (singe BacV6-21 assay) to 100.1% (singe HB assay). Parameters of all standard curves are shown in Appendix C Table 1. For sample processing control, five out of six animal fecal samples had a Ct value of 17.55 ± 0.37 (mean ± SD) and were within the SPC acceptance threshold (i.e., Sketa22 MB Ct mean + 3 × standard deviations, Ct = 20.90). One sample had a Ct of 23.82 and failed SPC acceptance threshold, but was eligible for Ct adjustment (i.e., sample Sketa22 mean Ct – Sketa22 MB mean Ct ≤ 3.3). Extraction efficiency was calculated as 22.4% ± 12.2% (mean ± SD). No inhibition was observed for the 46 animal fecal samples that were tested.

Discrepancies of human marker positives in NGS and qPCR.

A total of 180 out of 271 animal fecal samples were validated for the Lachno3, HB and BacV6-21 marker presence in qPCR assays and NGS data (Figure 4.4, Figure 4.5). Animals that were positive in marker presence in both NGS and qPCR were considered true cross-reactions. Animals that were only positive in a qPCR assay were considered false amplifications. The Lachno3 qPCR assay had a specificity of 92.8%, with positive reactions observed in one dog (CN = 107 per ng of DNA), two kangaroos (CN = 4 ± 2 per ng of DNA, mean ± SD) and ten horses (CN = 90 ± 163 per ng of DNA) (Figure 4.4). Lachno3 marker showed a specificity of 97.0% in NGS data of 271 animals. The Lachno3

NGS positives included the same dog and kangaroos that were positive in qPCR, and another three kangaroos, one cat and one raccoon that were only positive in NGS (Figure 4.5A). NGS positive-only samples had low levels of the Lachno3 marker sequence in these samples (count = 14.2 ± 22.4, mean ± SD), indicating that either sequencing is sensitive in detecting organism of low abundance, or sequencing error occurred for these samples. In addition, the qPCR positive horses were negative for the Lachno3 marker in their NGS data.

The BacV6-21 assay showed a qPCR specificity of 95.6%, including positives from three alligators (CN = 46 ± 14 per ng of DNA), three geese (CN = 11 ± 1 per ng of DNA), one horse (CN = 8 per ng of DNA) and one pig (CN = 2 per ng of DNA) (Figure 4.4). However, BacV6-21 NGS marker was negative in all 180 animal samples (Figure 4.5C).

The HB assay had a qPCR specificity of 91.7%, showing positives in two cats, three deer, three dogs, two rabbits, two raccoons, one alligator, one chicken and one sheep (Figure 4.4). Among these positive animals, two rabbits and one deer showed high level qPCR concentrations (CNs of 9,930 ± 7,630 and 1,190 per ng of DNA, respectively). The rest of the positive samples had much lower level signals with CNs of 5 ± 9 per ng of DNA. One of the positive rabbits was a pooled sample (i.e., Rabbit Pool1) and was subsequently tested in individuals. Four out of the five individuals were positive for the HB assay with CNs of 20,900 ± 15,100 per ng of DNA. The HB positive in NGS data was indexed by its main related V6 region sequence (BacV6-4) (145), which had an NGS specificity of 96.3%. BacV6-4 was positive in one chicken, two deer and two rabbits, all of which were also positive for the HB qPCR and were considered as true cross-reactions of the HF183 marker.

The true cross-reacted animal samples for HB were also confirmed by the HF183/BacR287 assay. BacV6-4 was also positive in one raccoon that was negative for the HB qPCR (Figure 4.4C).

Although both true and false amplifications were observed, specificities of these marker assays were all over 90%. In addition, most of the cross-reacted samples were low in concentrations (i.e., 67% of the positive animals were lower than 15 CNs per ng of fecal DNA) (Figure 4.4). NGS results and qPCR results of all three DNA concentrations are detailed in Data Set 4.1.

Mechanisms for qPCR positive-only human marker cross-reactions.

The Lachno3 and BacV6-21 markers both showed qPCR-only positives in animal fecal samples. BLAST+ analysis of the ten horses NGS data against the Lachno3 assay showed that these horses all had sequences matched with the Lachno3 probe (i.e., unique sequences, n = 64) (Figure 4.6A), which is located within the V6 region. Most of the probe-matched sequences that have high similarity (e.g., 90%) to Lachno3 were overall human- and sewage- preferred with very low occurrences in animal hosts, indicating amplification of these sequences should not impact Lachno3 assay's specificity. However, certain members of these sequences (i.e., LC4 and LC8) showed presence in horse and kangaroo. The sequence type LC4 showed up only in these two animals, including all horse and kangaroo individuals that were positive in the Lachno3 qPCR. This suggested that the amplification of such sequences was responsible for the qPCR-only positives of Lachno3 in horse.

Similar cases were also observed in the BacV6-21 marker assay. The qPCR positive-only samples had no identical match with the BacV6-21 marker in their NGS data

but showed matches with sequences highly similar to the BacV6-21 marker (Figure 4.6B). Some of the highly similar sequences of BacV6-21 (i.e., BC3, BC4 and BC5) only had one base difference with the BacV6-21 marker, either in the forward primer or probe. These sequences showed up in alligator, goose, horse and pig, corresponding to the qPCR-only amplifications of the BacV6-21 assay. In addition, the BacV6-21 qPCR assay was performed on these animals with one-degree higher annealing temperature (i.e., 61℃). Two out of eight positives were eliminated. This indicated that organisms that have sequences highly similar to the marker is one possible cause of the BacV6-21 qPCR assay amplification in animal fecal samples.

Mechanisms for human marker cross-reactions that were positive in both NGS and qPCR.

For the animal samples that showed human marker positives in both NGS and qPCR (i.e., Lachno3 and HB), the marker's presence was compared to the sample's microbial community composition. For Lachno3, one dog and two kangaroos were considered as true positives. The microbial community composition of the positive dog was similar to negative dogs. Kangaroos were more similar within Lachno3 positive and negative groups than between the two groups. For the HF183 marker, one chicken, two deer and two rabbits were considered as true positives. The chicken sample and one deer (PU259) were similar to negative chicken and deer samples, respectively. The other deer (PU123) and the two rabbits (PU27 and Pool1) that were positive with high signals in both NGS and qPCR showed higher similarity with human and sewage within genus *Bacteroides* compared to negative deer and rabbit samples. The HB/BacV6-4 positive rabbits were artificially fed (e.g., pet and domestic rabbits) and collected from different geographical regions (i.e., WI and TX), while the negative rabbits were all wild from the

same sender in TX. This suggests that the high-level human marker cross-reaction in rabbit could be correlated with domestic raising that may cause an atypical fecal microbial community composition. However, there was also one marker negative deer (i.e., Deer PU121) that showed a similar Bray-Curtis pattern with the marker positive deer (PU123). Both of these two deer showed different pattern compared with the other deer, suggesting that an atypical animal fecal microbial community does not always lead to change of a single marker organism.

To further explore marker cross-reactions in animal fecal microbial communities, we compared *Blautia*/*Bacteroides* unique sequences in Lachno3/HB positive and negative animals. The BacV6-21 marker was not included in this analysis as there was no NGS positive samples. Lachno3 positive dog and kangaroo had multiple unique *Blautia* sequences of over 98% sequence identity to the Lachno3 marker. Lachno3 negative dog did not have any sequences of over 90% identity to Lachno3, and Lachno3 negative kangaroo had *Blautia* sequences with the highest identity of 91.7%. In addition, even with the unequal Lachno3 positive and negative dog sample sizes (i.e., one and 18, respectively), 19.2% of all the *Blautia* unique sequences in dog samples only existed in the positive dog and 12.9% were shared. In kangaroo, five Lachno3 positive individuals and five Lachno3 negative individuals were compared. Seventy-four percent of the *Blautia* unique sequences were only in Lachno3 positive individuals, 11.5% were only in negative individuals and 14.5% were shared. In addition, 45% of the unique *Bacteroides* sequences in BacV6-4/HB positive rabbits were >98% similar with the BacV6-4 sequence, while the negative rabbit group only showed a highest identity of 87% with the BacV6-4 sequence.

From these results, cross-reaction of human markers in animal fecal samples is more likely the detection of a cluster of phylogenetically closely-related organisms. Atypical microbiomes also could explain some results; this shift of microbial community could be reflected in composition change of the corresponding genus and may be correlated with environmental factors.

QPCR results for general fecal indicator assays.

Both the *E. coli* and ENT assays had higher CN levels and prevalence in hosts such as alligator, birds and raccoon, and were found in lower levels and had lower prevalence in hosts such as horse, cow and deer. Appendix C Figure 4 shows positive results of the two FIB assays in animal host groups. *E. coli* was positive in 108 out of 180 animal samples (60%), including all hosts but duck (n=1), at 1 ng μL^{-1} template level (CN = 818 ± 2,000 per ng of DNA, mean ± SD). ENT was positive in 94 out of 180 samples (52%) in all hosts at 1 ng μL^{-1} DNA template level (CN = 1,840 ± 5,900 per ng of DNA). *E. coli* assay CN mean was 109 ± 126-fold higher than Lachno3 CNs in Lachno3 positive samples, 540 ± 484-fold higher than HB CNs in HB positive samples and 113 ± 164-fold higher than BacV6-21 CNs in BacV6-21 positive samples. ENT assay CN was 245 ± 282-fold higher than Lachno3 CNs in Lachno3 positive samples, 1,210 ± 1,090-fold higher than HB CNs in HB positive samples and 253 ± 368-fold higher than BacV6-21 CNs in BacV6-21 positive samples. Overall these results demonstrated that human marker CNs were two to three orders of magnitude lower than the general indicator qPCR assays.

Discussion

Host physiology and environmental factors both affect microorganism distribution patterns in animal hosts.

The gut microbial community is an acquired system in vertebrate animals. The composition of gut microbiota can be affected by many factors, including host physiology and environmental factors such as cohabitation and diet (i.e., herbivore, omnivore and carnivore) (160). Impacts of host physiology and/or an animal's general diet were consistently observed in our data; for example, microbial communities of ruminant, horse, pig and dog were grouped by host species, regardless that these samples were collected from different continents (Figure 4.3). Specific diet and cohabitation also showed impacts; for example, the microbial community of mammal herbivores were separated from carnivores, and the human fecal microbial community was close to the pet group (Figure 4.2). These observations were consistent with other mammal gut microbiome studies where host species and diet both showed significant impacts in network- and UniFrac-based microbial community composition analysis (160, 173). In addition, our qPCR and NGS validations of the HB assay in certain rabbits showed that high cross-reaction could correlate with cohabitation and/or feeding operation, which also indicates the impacts of environmental factors on compositions of animal fecal microbial communities. Overall, our results support the conclusion that human fecal marker organisms could be affected by diverse environmental factors that contribute to sporadic cross-reactions in animal individuals.

Exploring qPCR-only amplifications of human fecal markers to improve assay performance.

Marker identification based on NGS data requires resolution of the unique sequence level within the sequenced region to distinguish organisms that are specific to a host niche. When designing qPCR assays, sequences of similar organisms of the NGS marker may still be amplified due to the assay's inadequate representativeness for the full-length marker sequence. It is possible that some low-abundance sequences in our NGS data that are very similar to the marker could be derived from sequencing errors of Illumina's sequencing-by-synthesis technology, which are usually caused by single nucleotide substitutions (174). However, it is also possible that these single base changes are derived from single-nucleotide polymorphism changes in bacterial genomes. Considering the deep, paired-end sequencing method we used, and the fact that these organisms could have relatively high abundance (e.g., the BC3 sequence type in sewage, alligators and geese, Figure 4.6B) and could present independent of the marker organism, it is reasonable that organisms with these sequences were truly present in animal fecal samples and were amplified by the qPCR assay as templates of low concentrations.

For development of more specific qPCR assays using NGS data, or for re-optimization of these established qPCR assays, it is necessary to investigate host specificities of the sequences of marker's closely-related organisms, and optimization could include these sequences in assay designing to avoid amplification of potential organisms with low host specificities. Continuous efforts for expanding NGS dataset, such as V2 region NGS for HF183-positive animal samples, can help identify the true or false presence of the HF183 marker. Also, validation and optimization of qPCR assays using NGS data are needed to improve assay performance.

General fecal indicator qPCR assays may not reflect total fecal pollution.

The general fecal indicator assays only showed up in about 50% to 60% of the tested animal fecal samples at 1 ng μL^{-1} DNA template. The patterns of general indicator assays in animal hosts were similar to another study (175), where *E. coli* and enterococci were at about the same order of magnitude as this study for chicken and racoon, and were below the limit of quantification for deer and cow. This indicated that the levels of DNA used in these studies could not detect general fecal indicators in animals such as cow and deer. In animals positive for human or sewage markers, the human fecal marker assays were on average two to three orders of magnitude lower than FIB assays, indicating that such level cross-reactions should not impact human marker assays' ability to detect human fecal source.

Bacterial 16S human fecal marker assays are host preferred.

Our analysis of NGS data at unique sequence level in cross-reacted animals revealed that the existence of organisms, which have sequences closely-related to the Lachno3 and BacV6-21 markers, could cause qPCR amplification of these marker assays. These organisms could co-occur with the marker organism. It was further demonstrated that the host specificities of these organisms are not always consistent with the markers. Cross-reaction cause by qPCR amplification of these organisms was usually of much lower signals compared to marker presence in human/sewage source. Such cross-reaction should not affect the marker assay's performance in urban waters, where human source usually overwhelms others. However, in a few cases, human marker signals that even surpassed their signals in sewage were observed (i.e. HB assay results for one deer and two rabbits). Considering environmental factor impacts (e.g., diet and habitat) that can change animal

fecal microbial communities, as well as other random events such as ingestion of human waste by animal hosts (164), it seems unavoidable for human fecal marker organisms to be complete absent in other sources. This supports reported finding that it is more appropriate to treat these human fecal marker assays as "human-preferred" or "human-associated" (100). Combined use of marker assays that are derived from different microorganisms is an approach to eliminate potential animal cross-reactions. For example, in our validation results of 180 samples, a combination of Lachno3 and HB assays resulted in only one dog that was positive for both assays (i.e., presumptive specificity = 99.4%), a combination of Lachno3 and BacV6-21 would not have any sample that was positive for both assays (presumptive specificity = 100%), and a combination of HB and BacV6-21 would only have one alligator that was positive for both assays (presumptive specificity = 99.4%). Therefore, using such assay combinations will improve the confidence for human fecal source detection by reducing possible influence from animal sources.

Acknowledgements

We thank the following researchers for assistance with animal fecal sample collection or contribution of animal fecal DNA for this work: Dr. Scott E. Henke (Texas A&M University- Kingsville), Dr. Garret Suen (University of Wisconsin – Madison), Dr. Terrance Arthur (U.S. Meat Animal Research Center), Dr. Jason Gill (Texas A&M University), Dr. Jill Stewart (University of North Carolina), Dr. Valerie Harwood (University of South Florida), Dr. Karl Miller (University of Georgia) and Melony Wilson (University of Georgia Extension). We thank Hillary Morrison at the Bay Paul Center, Marine Biological Laboratory (MBL), University of Chicago for offering expertise in NGS

sequencing. Funding for this study was provided by National Institutes of Health (NIH), grant number R01 AI091829.

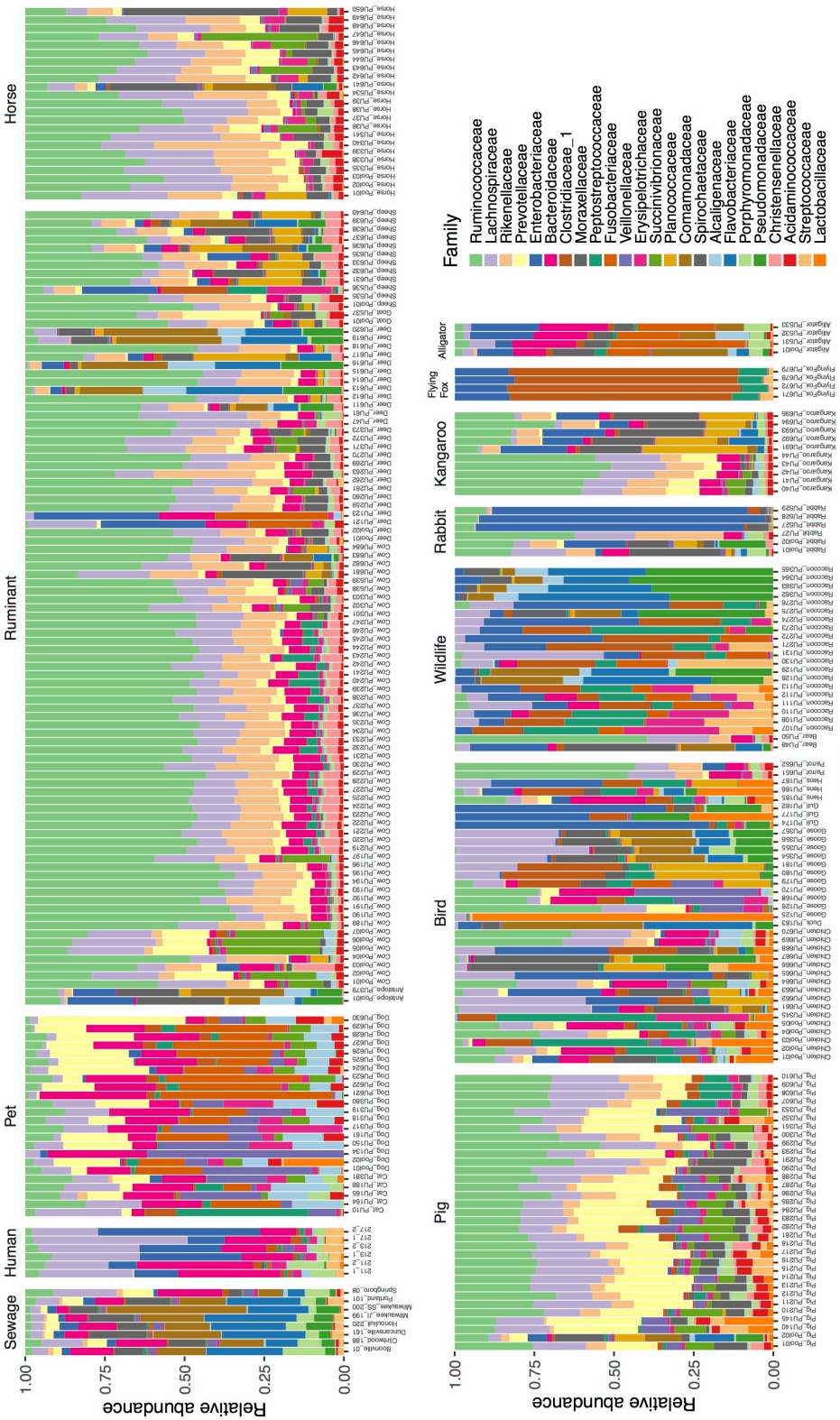


Figure 4.1 Whole community compositions of sewage, human and animal fecal samples examined on family level. Families of more than 1% relative abundance of the whole community are plotted and visualized in host groups. The Y- axis represents relative abundance, and the X- axis shows individual samples in host groups.

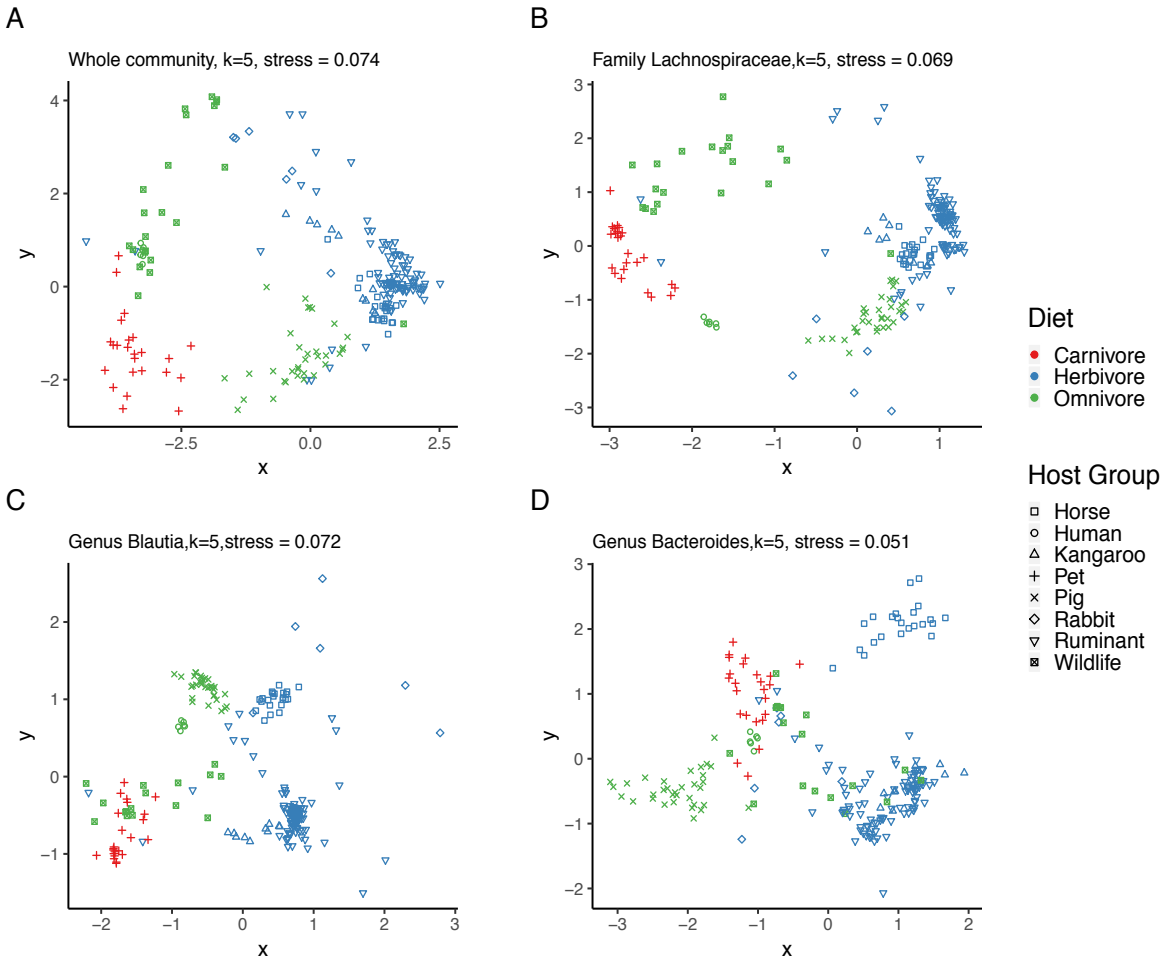


Figure 4.2 Non-metric multidimensional scaling (NMDS) analysis of microbial communities of all mammal samples (n=219). Analysis is performed for A. Whole community, B. Family *Lachnospiraceae*, C. Genus *Blautia* and D. Genus *Bacteroides*. Animal host groups are shown in different shapes. Diet groups are indicated in different colors: carnivore is in red, omnivore is in green, and herbivore is in blue.

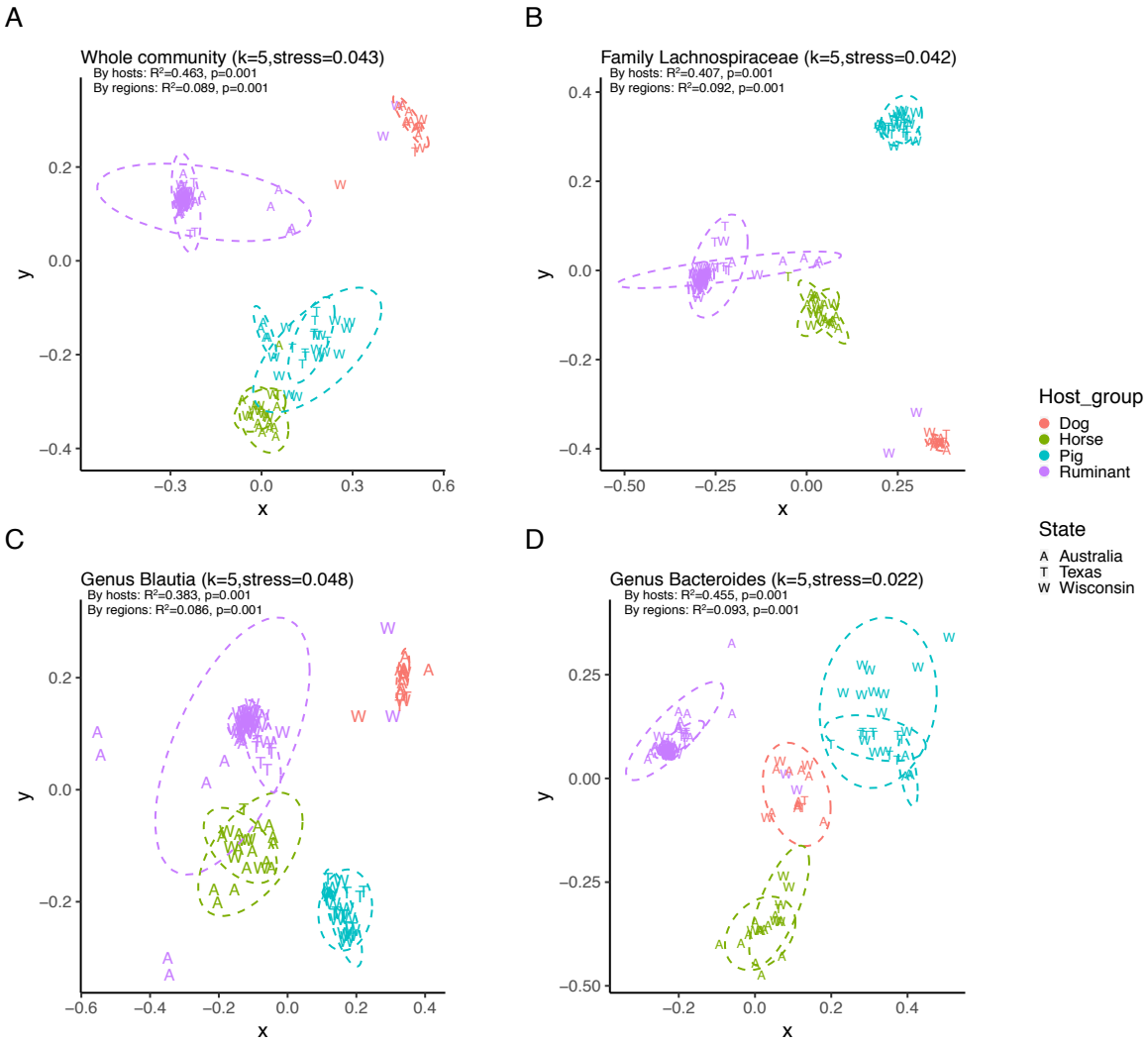


Figure 4.3 Nonmetric multidimensional scaling (NMDS) analysis of fecal microbial communities of animals that have samples collected from Australia, Texas and Wisconsin. Analysis is performed for A. whole community, B. family *Lachnospiraceae*, C. genus *Blautia* and D. genus *Bacteroides*. Geographical regions are in different point shapes. Ellipses represent 95% confidence interval with colors corresponding to host groups.

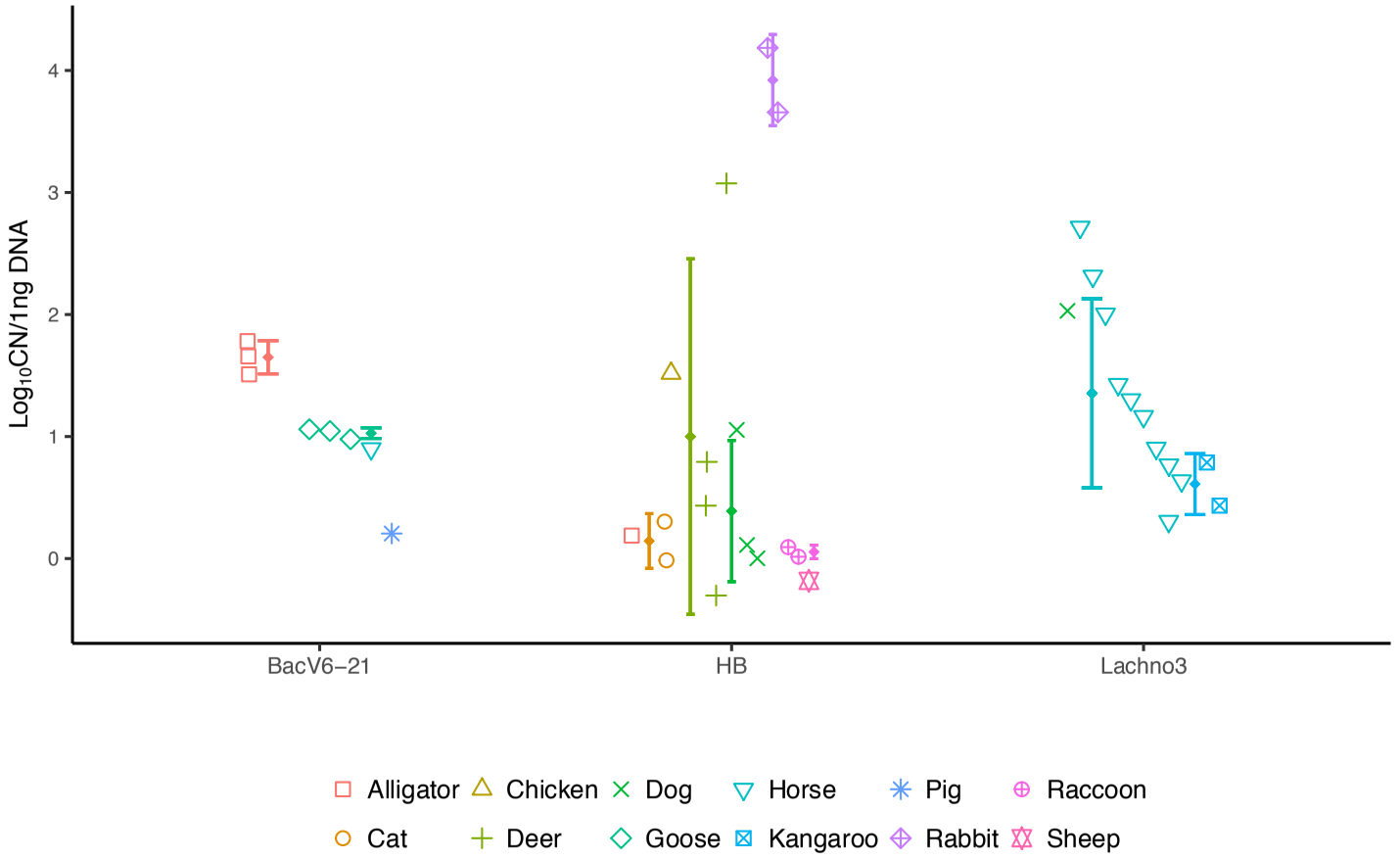


Figure 4.4 Human fecal marker qPCR assay positive results. The Y- axis shows log10-transformed CN per ng of DNA. The X- axis shows assays that were tested. Each animal host is shown in different color and shape. Error bar stands for mean ± SD, with the rhombus in the middle representing mean value.

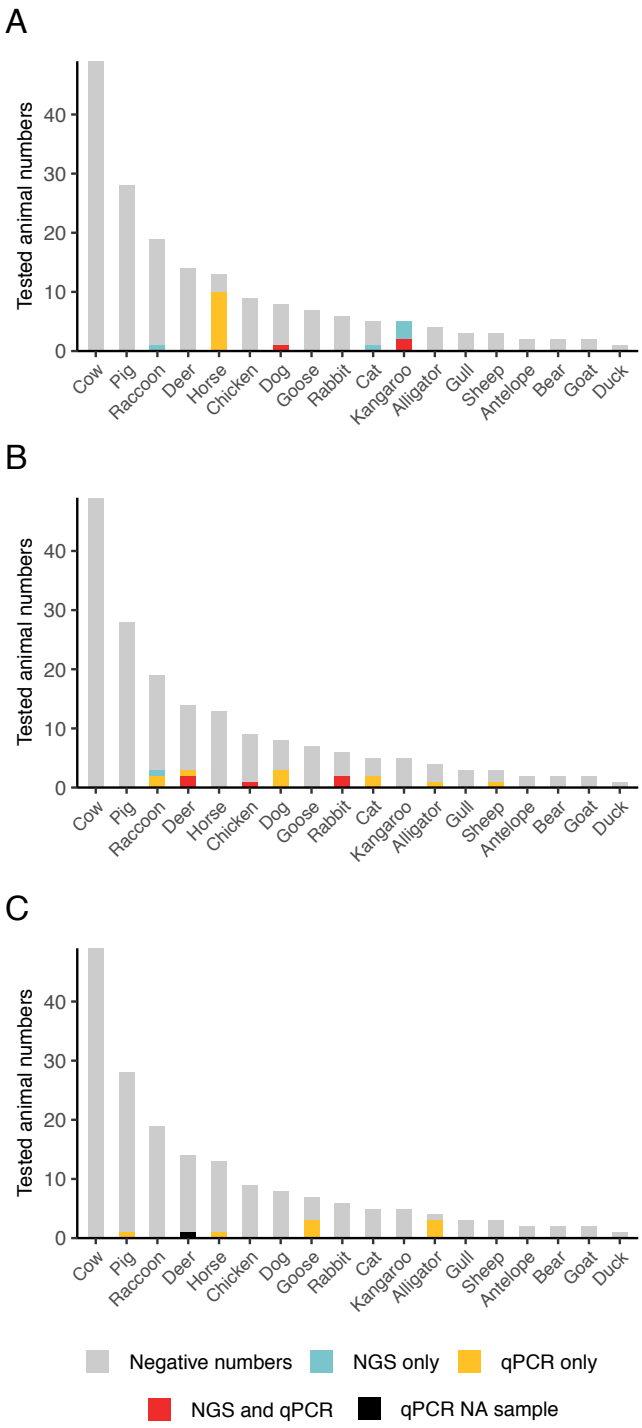


Figure 4.5 NGS and qPCR validation results in 180 animal samples. Markers included are A. Lachno3, B. HB (BacV6-4), C. BacV6-21. X-axis shows animal hosts. The Y-axis shows total tested animal numbers. Gray bars represent number of samples that are negative for both NGS and qPCR, green bars represent number of samples that are only positive in NGS data, yellow bars represent number of samples that are only positive in qPCR, red bars represent number of samples that are both positive in NGS and qPCR and the black bar represents one deer sample (PU123) that was not tested for BacV6-21 qPCR assay due to the lack of DNA.

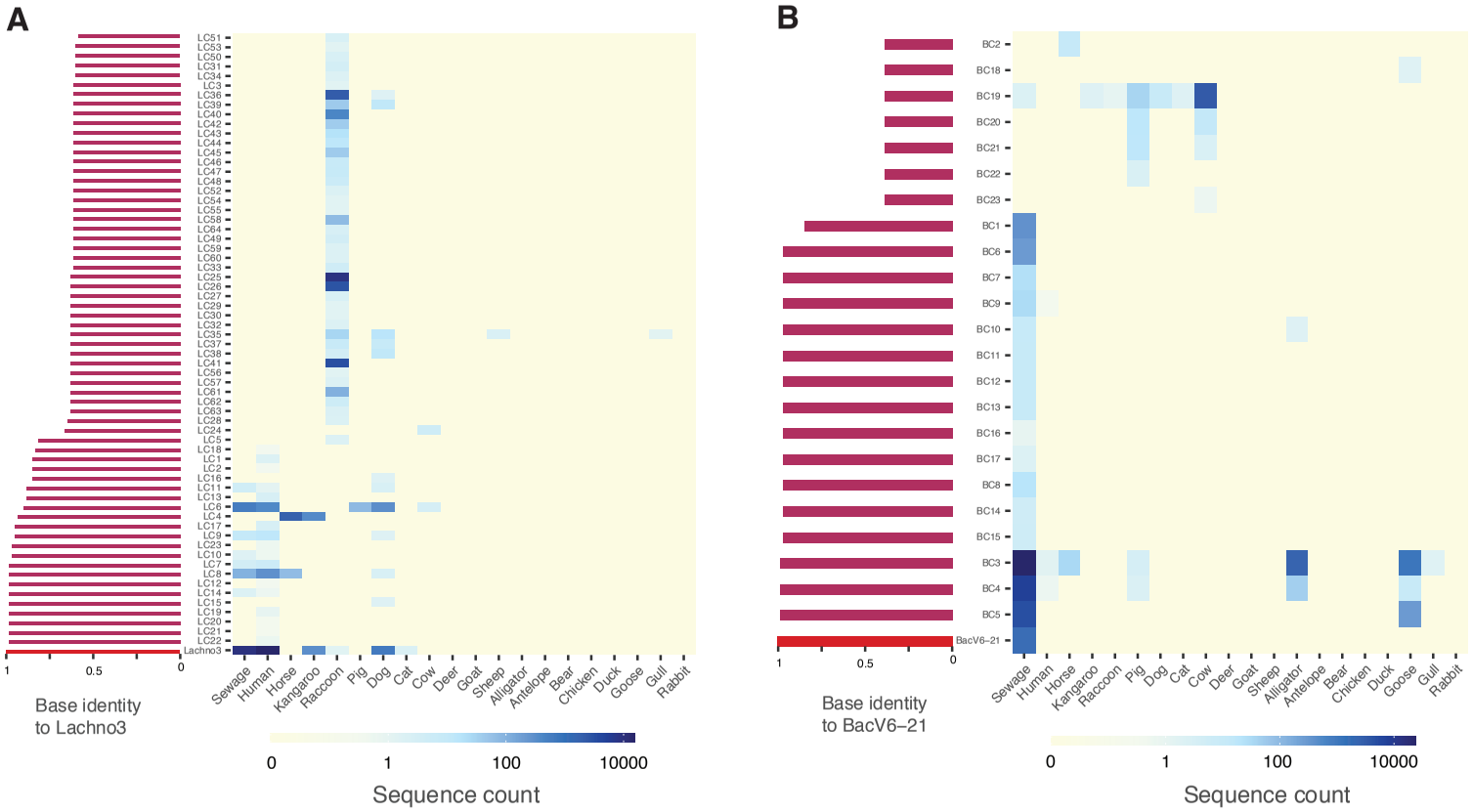


Figure 4.6 Distribution patterns of human fecal markers' closely-realted organisms. A. Distribution patterns of animal fecal sequence types that 100% match the probe of Lachno3 assay. B. Distribution patterns of animal fecal sequence types that are highly similar to the forward primer or probe of BacV6-21 assay. The maroon horizontal bars on the left side represent sequences' base identities to the marker. Marker sequences are shown in the bottom as references of 100% identity (in red bars). The heatmaps on the right show normalized counts of these sequences in sewage, human and animal fecal samples. Sequence count increases from light to dark blue color.

Chapter 5 General discussion

Summary of this work.

The purpose of this work was to expand the scope of microorganisms identified as specific to the human fecal source and apply these new bacterial indicators to track human fecal pollution (i.e., sewage) in water environments. The HF183 cluster of organisms has been used extensively as a target for detecting human fecal pollution over the past two decades. However, identification of human fecal source based on one single cluster of organisms may be insufficient when fecal pollution from other animal sources is present, or the concentrations of these organisms are low in the human population. In addition, MST applications have largely focused on the HF183 cluster, leaving other taxa within sewage and human fecal microbial communities unexplored.

By using next-generation sequencing (NGS) data from a large inventory of sewage and animal fecal samples, we expanded the inventory of organisms that are shown to be specific to sewage or human fecal sources. We explored distribution patterns of organisms within the family *Lachnospiraceae* in sewage and animals and identified human-specific fecal markers (i.e., Lachno3 and Lachno12) from the genus *Blautia*. Also, by exploring population structure of *Bacteroides* in sewage, we identified highly sewage-specific markers from a sewer pipe-derived, HF183 independent *Bacteroides* (i.e., BacV4V5-1 and BacV6-21). Quantitative PCR (qPCR) assays were developed for all these markers for subsequent use in sewage pollution detection. When validation studies of Lachno3, HF183 and BacV6-21 marker assays all showed sporadic amplifications in animal fecal samples, we further explained these cross-reactions. Our exploration of animal fecal sample microbial communities shed light on the "human-preferred" pattern of marker organisms

by identifying correlations between marker presence, composition of animal fecal microbial communities and environmental factors. We also identified potential mechanism for qPCR positive-only cross-reactions of these markers, which could be attributed to the amplification of the markers' closely related organisms that may co-occur with the markers in lower abundance.

Overall, this work highlighted the usage of sequencing data as a reservoir for host-specific fecal marker organisms, as well as a reference for human specificity of potential marker sequences. The development of NGS-based markers in this work provided a new generation of highly specific indicators for tracking human fecal pollution. The exploration of marker cross-reaction mechanisms could be applied to explain cross-reactions of human fecal marker assays observed in field tests. Using a combination of markers from different organisms (e.g., *Lachnospiraceae* and *Bacteroides*), the identification of human fecal pollution will be more accurate by discriminating marker false positives (e.g., presence of cross-reacted fecal source) or false negatives (e.g., the marker organism has low abundance).

Human fecal marker specificity and sensitivity are impacted by host physiology and environmental factors.

The acquisition of gut microbial community, which starts from birth, is inherited from parents vertically and is shaped by environmental factors horizontally (e.g., diet) (173, 176, 177). It has been shown that the diet-induced colonization of species shaped the structure of gut microbial communities in larger phylogenetic lineages (i.e., closer to the root of bacterial tree), and co-speciation of gut microbiota with the host correlated with

community composition changes in smaller phylogenetic lineages (i.e., closer to the leaf of bacteria tree) (177). *Bacteroidetes* and *Firmicutes* are the main phyla commonly present in most mammalian gut microbial communities, whereas their sub-taxa have various distribution patterns in different host niches (160, 178, 179). Such differentiation caused by host physiology and environmental factors result in stable, resilient and different compositions of intra-host gut microbial communities, and even more variable inter-host communities (178). Gut microbial community composition is therefore "host-adapted" due to the long-term symbiosis between gut microbiota and their hosts (159, 173, 176). This supports the premise of microbial source tracking methods, where lower taxonomic level microorganisms such as a single genus (e.g., *Bacteroides*) shows host specificity. However, the difference of gut microbiota between individuals (e.g., marker presence or absence) cannot be simply ignored. Instead of exploring using individual human samples to identify human marker candidates, we identified human-associated markers from sewage influent, which is a comprehensive representation of human fecal microbial communities and also the main format of human fecal pollution entering surface water (30, 52, 81, 101).

There are multiple mechanisms for certain taxa to adapt to the host gut environment, which may contribute to the presence/absence of a marker in its host source. For example, it has been observed that members of the family *Bacteroidaceae* in human and great-ape species both showed a vertical transmission that passed from generation to generation as seeded members of the gut microbiota (180). *Bacteroides* has been shown as one of the gut microorganisms that have retained "hallmarks of co-diversification", as the phylogenetic relationship among *Bacteroides* members mirrors their hosts (181). Members of *Lachnospiraceae* are on the opposite side. They are acquired from other sources and

transfer horizontally between hosts, thereby could be completely different members in gut microbial communities of mothers and children (180–182). This may be correlated with the spore-forming feature of *Lachnospiraceae*, which could facilitate their dispersal in different hosts (180, 181). The adaptation patterns of *Bacteroides* and *Lachnospiraceae* may further result in various distributions of their members in different host individuals, contributing to the observation that a single marker organism, such as the HF183 or Lachno3, are not 100% present in all tested human individuals (62, 100).

Human marker specificity is usually evaluated by its presence/absence in animal fecal sources (i.e., the proportion of total marker negatives in tested animal samples). Increasing human activities are disturbing animal gut microbiota through habitat degradation and transition to captive programs (183). This is also well supported by our observations that the fecal communities of pigs on plant- and grain-based diets were clustered with the herbivore group more than the omnivore group, and that domestic rabbits were positive for HF183 at high concentrations while the tested wild rabbits were negative (see Chapter 4). In addition, the human gut microbiota has also been changing. It has been observed that humans with modern living styles (e.g., urbanization, western style diet that is low in microbiota accessible carbohydrates) are forming overall more similar gut microbiomes with shrinking diversity despite the geographical differences (160, 178, 184–186). Human and animal gut microbiota composition is "dynamic" under the influence of environmental factors, which impacts the presence of human fecal marker organisms in fecal microbial communities in both sources. This may impact the human marker specificities (i.e., occurrence in animal sources) and sensitivities (i.e., occurrence in the human source) in the long run. This also supports the suggestion from multiple studies that

a combination of marker assays from different microorganisms should be adopted for confident detection of human fecal source, and that studies for human fecal marker validations are always needed (31, 100).

Application and limitation of NGS in identification of human fecal pollution.

NGS studies shed light on its large potential for exploring microorganisms in different sample categories, such as sewage, human and animal fecal microbiomes. Using the Illumina sequencing platform (187, 188), deep sequencing profiles offer the opportunity to investigate relative abundances of organisms community-wide to explore their host specificities *in silico*. This was successfully demonstrated by establishing the host specificities of the Lachno3 and BacV6-21 assays, which were validated in this work and (100). In addition to host-associated fecal marker identification, NGS data also revealed the potential to identify MST markers from sewage infrastructure-associated organisms (e.g., sewage *Bacteroides*, see Chapter 3) (101). Assays for such microorganisms can capture untreated sewage pollution in surface water using criteria that are independent of human and animal gut microbiomes, offering additional measures to increase confidence in sewage pollution identification (101). Moreover, NGS applied to computational or machine learning approaches such as Bayesian approach (165) or random forest (179) can use sequence abundance patterns of the whole community, or of taxonomic groups, rather than a single organism to identify pollution signals within a water sample. These NGS-based methods rely on a signature of sequences that include their relative abundance patterns within the community, and sequences shared between sources generally will not also share the overall relative abundance pattern within the signature (179). NGS-based studies have also been applied to characterize human pathogen

distributions in environmental waters, and found pathogen-like sequences (e.g., genera *Acinetobacter, Arcobacter* and *Clostridium*) in human activity-impacted area with or without statistical correlation with FIB (189–191). All these studies indicate the appropriate applicability of NGS to fecal source tracking and even public health risk assessment.

However, limitations of NGS application in marker assay development do exist. Markers identified in this study are unique sequences of hypervariable region(s). It has been observed in our study that markers such as Lachno3 and BacV6-21 all have closely-related organisms, which have sequences highly similar to the marker but may not share the host specificity pattern with the marker. Amplification of such organisms could cause "cross-reaction" of a marker assay when they are present in animal fecal sources. Although these closely-related sequences can cause qPCR false positives, markers based on the full length of hypervariable region(s) are very useful in identifying human fecal source *in silico*. For example, identification of Lachno3 marker sequence in an environmental water sample's V6 NGS data would suggest the presence of human fecal pollution. Another limitation is that it is still challenging to use NGS data to determine the actual levels of taxa of interest. Therefore, the quantification of marker organisms is performed by subsequent qPCR amplification, which also verifies the specificity of human markers as observed in the NGS data. In addition, with emerging commercially-available controls that can be added to samples (e.g., microbes completely unrelated to the human microbiome), the quantification of taxa in relation to relative abundance data may facilitate the use of NGS for quantitative usage in fecal source tracking in the future.

QPCR technical details are critical for successful fecal marker assay performance.

The PCR-based method has been favored by research laboratories for fecal source detection due to its high levels of specificity and sensitivity to the amplification target (37, 54, 83, 172). Design and optimization for an ideal assay performance are critical steps for successful assay application. In our study, the design of a marker assay was based on a set of considerations, including: 1) sequence variability between host and non-host sources; 2) melting temperatures of primers and probe, which could be affected by factors such as the length of the chosen sequence and the percentage of Guanine-Cytosine (i.e., GC content); 3) a suggested amplicon length of between 50 to 150 bases; and 4) the degree of primer and probe sequences being matched with targeted species (e.g., genus *Blautia* and *Bacteroides*), which can be verified *in silico* using tools such as BLAST and the probe match function in RDP (https://rdp.cme.msu.edu/). In addition, qPCR program annealing temperature should be well optimized. During the assay development process, we tested assays based on criteria obtained from the standard curve analysis, no template and positive control results and sample inhibition test results. For validation efforts, we added more control criteria, such as a sample processing control and extraction efficiency test. When multiplexing assays, both standard curves and sample tests were compared between multiplexed and non-multiplexed settings to make sure that no bias was introduced. Throughout our study, sample processing, qPCR reagent preparation and qPCR amplification were carried out in separate laboratory areas to avoid false positives caused by extraneous source DNA contamination. DNA extractions of sewage samples, environmental water samples and animal fecal samples were also performed in separate lab areas.

The qPCR experiment procedures of NGS-based assays were developed based on proficient and validated protocols that have been used in the lab with consistent reagents and instruments. The recently established EPA Method 1696 for the HF183/BacR287 qPCR assay offers the best example for human marker qPCR assay standardization, and should be followed for successful and reliable performance of human marker assays (192). It is recommended to future users of assays developed in this work that qPCR parameters should always be reported, such as standard curve parameters, primer/probe concentrations, amplification program, reagents and instruments. Standardization of MST assay is a long-term effort that needs assay performance validation from numerous research laboratories (172). The human fecal marker assays developed in this study have been realizing their applications. More field application of these assays can further demonstrate and propel their usage in source tracking and load estimations of human fecal pollution in environmental waters, especially in complex cases where multiple fecal sources present.

Guidance and recommendations for usages of marker assays developed in this work.

We encourage applications of the fecal marker assays developed in this study with guidance and recommendations listed below:

1. For fecal and environmental water samples that have sequence data available for V4V5 and V6 regions of the 16S rRNA gene, NGS marker (i.e., Lachno3, Lachno12, BacV4V5-1, or BacV6-21) presence should be identified *in silico* to indicate the true presence/absence of the marker. But this step should not be a replacement for the qPCR assay because sensitivity of marker detection will be

diminished as fecal pollution is mixed with environmental microbial communities.

2. For field application, selection of marker should be based on validated host specificity results and local area knowledge. For example, all four assays are recommended for sewage pollution identification in urban waters; while in rural area where livestock sources may appear, the Lachno3 and sewage *Bacteroides* assays are recommended rather than the Lachno12 assay, which has demonstrated cross-reactions with cow and pig. Also, sewer pipe-derived *Bacteroides* assays could potentially be used for distinguishing sewer system-independent human fecal pollution sources, such as cesspools, from untreated sewage released from wastewater conveyance systems. This application requires additional study.
3. A combination of marker assays from different organisms is recommended for field applications, especially in geographical regions where no fecal source tracking study has been employed before. This is to avoid false positive interpretation of results caused by cross-reaction with animal sources and false negatives caused by insufficient abundance of an assay's targeted organism in tested samples due to the influence of some environmental factors (e.g. diet in the human population).
4. Ongoing validation efforts are needed for specificity and sensitivity of NGS-based marker assays developed in this study. For assay validations, technical details such as standard curve parameters, lower limit of quantification and inhibition test results should be reported to facilitate transfer of technology to

other labs. Also, quantitative values should be reported in units of copy number per nanogram of fecal DNA or copy number per volume of water sample. Pooled samples can be used to reduce the cost of screening large numbers of samples. However, for positive pooled samples, testing of individual samples should be performed.

5. Additional studies are needed for these marker assays developed from NGS data, including determining their decay rates and correlations with pathogen presence. This can be done with the ongoing host specificity validation efforts to form a comprehensive understanding of these markers for their application in field studies.

REFERENCES

1. Ferguson C, De Roda Husman AM, Altavilla N, Deere D, Ashbolt N. 2003. Fate and transport of surface water pathogens in watersheds. Crit Rev Environ Sci Technol 33:299–361.
2. U. S. Environmental Protection Agency (EPA). 2012. Recreational Water Quality Criteria. Washington, DC. EPA 820-F-12-061.
3. Gargano JW, Adam EA, Collier SA, Fullerton KE, Feinman SJ, Beach MJ. 2017. Mortality from selected diseases that can be transmitted by water - United States, 2003-2009. J Water Health 15:438–450.
4. Adam EA, Collier SA, Fullerton KE, Gargano JW, Beach MJ. 2017. Prevalence and direct costs of emergency department visits and hospitalizations for selected diseases that can be transmitted by water, United States. J Water Health 15:673–683.
5. 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 17:3.
6. Griffin DW, Donaldson KA, Paul JH, Rose JB. 2003. Pathogenic human viruses in coastal waters. Clin Microbiol Rev 16:129–143.
7. Donovan E, Unice K, Roberts JD, Harris M, Finley B. 2008. Risk of gastrointestinal disease associated with exposure to pathogens in the water of the Lower Passaic River. Appl Environ Microbiol 74:994–1003.
8. Colford JM, Wade TJ, Schiff KC, Wright CC, Griffith JF, Sandhu SK, Burns S, Sobsey M, Lovelace G, Weisberg SB. 2007. Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. Epidemiology 18:27–35.
9. Colford JM, Roy SL, Beach MJ, Hightower A, Shaw SE, Wade TJ. 2006. A review of household drinking water intervention trials and an approach to the estimation of endemic waterborne gastroenteritis in the United States. J Water Health 4:71–88.
10. Reynolds KA, Mena KD, Gerba CP. 2008. Risk of waterborne illness via drinking water in the United States. Rev Environ Contam Toxicol 192:117–158.
11. Messner M, Shaw S, Regli S, Rotert K, Blank V, Soller J. 2006. An approach for developing a national estimate of waterborne disease due to drinking water and a national estimate model application. J Water Heal 4 Suppl 2:201–240.
12. Hoxie NJ, Davis JP, Vergeront JM, Nashold RD, Blair KA. 1997. Cryptosporidiosis-associated mortality following a massive waterborne outbreak in Milwaukee, Wisconsin. Am J Public Health 87:2032–2035.
13. Corso PS, Kramer MH, Blair KA, Addiss DG, Davis JP, Haddix AC. 2003. Cost of illness in the 1993 waterborne *Cryptosporidium* outbreak, Milwaukee, Wisconsin. Emerg Infect Dis 9:426–431.
14. Fleisher JM, Kay D, Wyer MD, Godfree AF. 1998. Estimates of the severity of illnesses associated with bathing in marine recreational waters contaminated with domestic sewage. Int J Epidemiol 27:722–726.
15. Wade TJ, Calderon RL, Brenner KP, Sams E, Beach M, Haugland R, Wymer L,

Dufour AP. 2008. High sensitivity of children to swimming-associated gastrointestinal illness: Results using a rapid assay of recreational water quality. Epidemiology 19:375–383.
16. Hlavsa MC, Cikesh BL, Roberts VA, Kahler AM, Vigar M, Hilborn ED, Wade TJ, Roellig DM, Murphy JL, Xiao L, Yates KM, Kunz JM, Arduino MJ, Reddy SC, Fullerton KE, Cooley LA, Beach MJ, Hill VR, Yoder JS. 2018. Outbreaks associated with treated recreational water — United States, 2000–2014. MMWR Morb Mortal Wkly Rep 67:547–551.
17. Graciaa DS, Cope JR, Roberts VA, Cikesh BL, Kahler AM, Vigar M, Hilborn ED, Wade TJ, Backer LC, Montgomery SP, Secor WE, Hill VR, Beach MJ, Fullerton KE, Yoder JS, Hlavsa MC. 2018. Outbreaks associated with untreated recreational water — United States, 2000–2014. MMWR Morb Mortal Wkly Rep 67:701-706.
18. Craun MF, Craun GF, Calderon RL, Beach MJ. 2006. Waterborne outbreaks reported in the United States. J Water Health 4:19–30.
19. Sauer EP, VandeWalle JL, Bootsma MJ, McLellan SL. 2011. Detection of the human specific *Bacteroides* genetic marker provides evidence of widespread sewage contamination of stormwater in the urban environment. Water Res 45:4081–4091.
20. Bartram J, Rees G. 2000. Monitoring bathing waters – a practical guide to the design and implementation of assessments and monitoring programmes. E & FN Spon, London.
21. U.S. Environmental Protection Agency (EPA). 2009. Review of published studies to characterize relative risks from different sources of fecal contamination in recreational water. Washington DC. EPA 822-R-09-002.
22. 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.
23. World Health Organization. 2015. Animal waste, water quality and human healthiwa publishing. IWA Publishing, London.
24. U.S. Environmental Protection Agency (EPA). 2004. Report to Congress on impacts and control of combined sewer overflows and sanitary sewer overflows. Washington DC. EPA 833-R-04-001.
25. McLellan SL, Hollis EJ, Depas MM, Van Dyke M, Harris J, Scopel CO. 2008. Distribution and fate of *Escherichia coli* in lake michigan following contamination with urban stormwater and combined sewer overflows. J Great Lakes Res 33:566–580.
26. Templar HA, Dila DK, Bootsma MJ, Corsi SR, McLellan SL. 2016. Quantification of human-associated fecal indicators reveal sewage from urban watersheds as a source of pollution to Lake Michigan. Water Res 100:556–567.
27. Marsalek J, Rochfort Q. 2004. Urban wet-weather flows: sources of fecal contamination impacting on recreational waters and threatening drinking-water sources. J Toxicol Environ Health A 67:1765–1777.
28. U.S. Environmental Protection Agency (EPA). 2017. National water quality inventory: Report to Congress. Washington DC. EPA 841-R-16-011.
29. Arnone RD, Walling JP. 2007. Waterborne pathogens in urban watersheds. J Water Health 5:149–162.

30. Newton RJ, VandeWalle JL, Borchardt MA, Gorelick MH, McLellan SL. 2011. *Lachnospiraceae* and *Bacteroidales* alternative fecal indicators reveal chronic human sewage contamination in an Urban harbor. Appl Environ Microbiol 77:6972–6981.
31. McLellan SL, Eren AM. 2014. Discovering new indicators of fecal pollution. Trends Microbiol 22:697–706.
32. Domingo JWS, Ashbolt NJ. 2008. Fecal pollution of water. *In* Cutler J. Cleveland (ed.), Encyclopedia of Earth. National Council for Science and the Environment, Washington, DC.
33. Tibbetts J. 2005. Combined sewer systems: Down, dirty, and out of date. Environ Health Perspect 113:464–467.
34. National Research Council. 2001. Under the weather: Climate, ecosystems, and infectious Disease. The National Academies Pres, Washington, DC.
35. Patz JA, Vavrus SJ, Uejio CK, McLellan SL. 2008. Climate change and waterborne disease risk in the Great Lakes region of the U.S. Am J Prev Med 35:451–458.
36. Drayna P, McLellan SL, Simpson P, Li SH, Gorelick MH. 2010. Association between rainfall and pediatric emergency department visits for acute gastrointestinal illness. Environ Health Perspect 118:1439–1443.
37. 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.
38. Field KG, Samadpour M. 2007. Fecal source tracking, the indicator paradigm, and managing water quality. Water Res 41:3517–3538.
39. World Health Organization (WHO). 2003. Faecal pollution and water quality, p. 51–101. *In* Guidelines for safe recreational water environments. World Health Organization, Geneva.
40. Environmental Protection Department. 2005. Water Quality Criteria / Standards Adopted in the Asia Pacific Region. The Government of Hong Kong Special Administrative Region.
41. European Environment Agency. 2018. European Bathing Water Quality in 2017. Denmark.
42. Dorevitch S, Doi M, Hsu FC, Lin KT, Roberts JD, Liu LC, Gladding R, Vannoy E, Li H, Javor M, Scheff PA. 2011. A comparison of rapid and conventional measures of indicator bacteria as predictors of waterborne protozoan pathogen presence and density. J Environ Monit 13:2427–2435.
43. Duris JW, Reif AG, Krouse DA, Isaacs NM. 2013. Factors related to occurrence and distribution of selected bacterial and protozoan pathogens in Pennsylvania streams. Water Res 47:300–314.
44. Oster RJ, Wijesinghe RU, Haack SK, Fogarty LR, Tucker TR, Riley SC. 2014. Bacterial pathogen gene abundance and relation to recreational water quality at seven Great Lakes beaches. Environ Sci Technol 48:14148–14157.
45. McQuaig SM, Scott TM, Harwood VJ, Farrah SR, Lukasik JO. 2006. Detection of human-derived fecal pollution in environmental waters by use of a PCR-based human polyomavirus assay. Appl Environ Microbiol 72:7567–7574.

46. Graczyk TK, Sunderland D, Awantang GN, Mashinski Y, Lucy FE, Graczyk Z, Chomicz L, Breysse PN. 2010. Relationships among bather density, levels of human waterborne pathogens, and fecal coliform counts in marine recreational beach water. Parasitol Res 106:1103–1108.
47. Viau EJ, Goodwin KD, Yamahara KM, Layton BA, Sassoubre LM, Burns SL, Tong HI, Wong SHC, Lu Y, Boehm AB. 2011. Bacterial pathogens in Hawaiian coastal streams-associations with fecal indicators, land cover, and water quality. Water Res 45:3279–3290.
48. Harwood VJ, Levine AD, Scott TM, Chivukula V, Lukasik J, Farrah SR, Rose JB. 2005. Validity of the indicator organism paradigm for pathogen reduction in reclaimed water and public health protection. Appl Environ Microbiol 71:3163–3170.
49. Lemarchand K, Lebaron P. 2003. Occurrence of *Salmonella* spp. and *Cryptosporidium* spp. in a French coastal watershed: relationship with fecal indicators. FEMS Microbiol Lett 218:203–209.
50. Ercumen A, Pickering AJ, Kwong LH, Arnold B, Parvez SM, Alam M, Sen D, Islam S, Kullmann C, Chase C, Ahmed R, Unicomb L, Luby S, Colford JM. 2017. Animal feces contribute to domestic fecal contamination: Evidence from *E. coli* measured in water, hands, food, flies and soil in Bangladesh. Environ Sci Technol 51:8725-8734.
51. Korajkic A, McMinn BR, Harwood VJ. 2018. Relationships between microbial indicators and pathogens in recreational water settings. Int J Environ Res Public Health 15:2842.
52. Feng S, Bootsma M, McLellan SL. 2018. Human-associated *Lachnospiraceae* genetic markers improve detection of fecal pollution sources in urban waters. Appl Environ Microbiol 84:e00309-18.
53. Sinton LW, Finlay RK, Hannah DJ. 1998. Distinguishing human from animal faecal contamination in water: A review. New Zeal J Mar Freshw Res 32:323–348.
54. Ahmed W, Hughes B, Harwood VJ. 2016. Current status of marker genes of bacteroides and related taxa for identifying sewage pollution in environmental waters. Water 8:231.
55. Gourmelon M, Caprais MP, Mieszkin S, Marti R, Wéry N, Jardé E, Derrien M, Jadas-Hécart A, Communal PY, Jaffrezic A, Pourcher AM. 2010. Development of microbial and chemical MST tools to identify the origin of the faecal pollution in bathing and shellfish harvesting waters in France. Water Res 44:4812–4824.
56. Lim FY, Ong SL, Hu J. 2017. Recent advances in the use of chemical markers for tracing wastewater contamination in aquatic environment: A review. Water (Switzerland) 9:143.
57. Clarridge JE. 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev 17:840–862.
58. Yu Z, Morrison M. 2004. Comparisons of different hypervariable regions of rrs genes for use in fingerprinting of microbial communities by PCR-denaturing gradient gel electrophoresis. Appl Environ Microbiol 70:4800–4806.
59. Allsop K, Stickler DJ. 1985. An assessment of *Bacteroides fragilis* group organisms as indicators of human faecal pollution. J Appl Bacteriol 58:95–99.

60. Fiksdal L, Maki JS, LaCroix SJ, Staley JT. 1985. Survival and detection of *Bacteroides* spp., prospective indicator bacteria. Appl Environ Microbiol 49:148–150.
61. Bernhard AE, Field KG. 2000. A PCR assay to discriminate human and ruminant feces on the basis of host differences in *Bacteroides-Prevotella* genes encoding 16S rRNA. Appl Environ Microbiol 66:4571–4574.
62. Seurinck S, Defoirdt T, Verstraete W, Siciliano SD. 2005. Detection and quantification of the human-specific HF183 *Bacteroides* 16S rRNA genetic marker with real-time PCR for assessment of human faecal pollution in freshwater. Environ Microbiol 7:249–259.
63. Layton A, McKay L, Williams D, Garrett V, Gentry R, Sayler G. 2006. Development of *Bacteroides* 16S rRNA gene taqman-based real-time PCR assays for estimation of total, human, and bovine fecal pollution in water. Appl Environ Microbiol 72:4214–4224.
64. 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.
65. Kildare BJ, Leutenegger CM, McSwain BS, Bambic DG, Rajal VB, Wuertz S. 2007. 16S rRNA-based assays for quantitative detection of universal, human-, cow-, and dog-specific fecal *Bacteroidales*: A Bayesian approach. Water Res 41:3701–3715.
66. Okabe S, Okayama N, Savichtcheva O, Ito T. 2007. Quantification of host-specific *Bacteroides-Prevotella* 16S rRNA genetic markers for assessment of fecal pollution in freshwater. Appl Microbiol Biotechnol 74:890–901.
67. Haugland RA, Varma M, Sivaganesan M, Kelty C, Peed L, Shanks OC. 2010. Evaluation of genetic markers from the 16S rRNA gene V2 region for use in quantitative detection of selected *Bacteroidales* species and human fecal waste by qPCR. Syst Appl Microbiol 33:348–357.
68. Lee DY, Weir SC, Lee H, Trevors JT. 2010. Quantitative identification of fecal water pollution sources by TaqMan real-time PCR assays using *Bacteroidales* 16S rRNA genetic markers. Appl Microbiol Biotechnol 88:1373–1383.
69. Green HC, Haugland RA, Varma M, Millen HT, Borchardt MA, Field KG, Walters WA, Knight R, Sivaganesan M, Kelty CA, Shanks OC. 2014. Improved HF183 quantitative real-time PCR assay for characterization of human fecal pollution in ambient surface water samples. Appl Environ Microbiol 80:3086–3094.
70. Yampara-Iquise H, Zheng G, Jones JE, Carson CA. 2008. Use of a *Bacteroides thetaiotaomicron*-specific α-1-6, mannanase quantitative PCR to detect human faecal pollution in water. J Appl Microbiol 105:1686–1693.
71. Carson CA, Christiansen JM, Benson VW, Baffaut C, Jerri V, Broz RR, Kurtz WB, Rogers WM, Fales WH, Yampara-iquise H, Davis J V. 2005. Specificity of a *Bacteroides thetaiotaomicron* Marker for Human Feces. Appl Environ Microbiol 71:4945–4949.
72. Stoeckel DM, Harwood VJ. 2007. Performance, design, and analysis in microbial source tracking studies. Appl Environ Microbiol 73:2405–2415.

73. Ahmed W, Yusuf R, Hasan I, Goonetilleke A, Gardner T. 2010. Quantitative PCR assay of sewage-associated *Bacteroides* markers to assess sewage pollution in an urban lake in Dhaka, Bangladesh. Can J Microbiol 56:838–845.
74. Van De Werfhorst LC, Sercu B, Holden PA. 2011. Comparison of the host specificities of two *Bacteroidales* quantitative PCR assays used for tracking human fecal contamination. Appl Environ Microbiol 77:6258–6260.
75. Nshimyimana JP, Cruz MC, Thompson RJ, Wuertz S. 2017. *Bacteroidales* markers for microbial source tracking in Southeast Asia. Water Res 118:239–248.
76. Silkie SS, Nelson KL. 2009. Concentrations of host-specific and generic fecal markers measured by quantitative PCR in raw sewage and fresh animal feces. Water Res 43:4860–4871.
77. Ahmed W, Goonetilleke A, Powell D, Gardner T. 2009. Evaluation of multiple sewage-associated *Bacteroides* PCR markers for sewage pollution tracking. Water Res 43:4872–4877.
78. Shanks OC, White K, Kelty CA, Sivaganesan M, Blannon J, Meckes M, Varma M, Haugland RA. 2010. Performance of PCR-based assays targeting *Bacteroidales* genetic markers of human fecal pollution in sewage and fecal samples. Environ Sci Technol 44:6281–6288.
79. Reischer GH, Kasper DC, Steinborn R, Mach RL, Farnleitner AH. 2006. Quantitative PCR method for sensitive detection of ruminant fecal pollution in freshwater and evaluation of this method in alpine karstic regions. Appl Environ Microbiol 72:5610–5614.
80. Jenkins MW, Tiwari S, Lorente M, Gichaba CM, Wuertz S. 2009. Identifying human and livestock sources of fecal contamination in Kenya with host-specific *Bacteroidales* assays. Water Res 43:4956–4966.
81. Newton RJ, McLellan SL, Dila DK, Vineis JH, Morrison HG, Eren AM, Sogin ML. 2015. Sewage reflects the microbiomes of human populations. MBio 6:1–9.
82. U.S. Environmental Protection Agency (EPA). 2005. Microbial Source Tracking Guide Document. Washington, DC. EPA/600/R-05/064.
83. Boehm AB, Van De Werfhorst LC, Griffith JF, Holden PA, Jay JA, Shanks OC, Wang D, Weisberg SB. 2013. Performance of forty-one microbial source tracking methods: A twenty-seven lab evaluation study. Water Res 47:6812–6828.
84. Bernhard AE, Field KG. 2000. Identification of nonpoint sources of fecal pollution in coastal waters by using host-specific 16S Ribosomal DNA genetic markers from fecal anaerobes. Appl Environ Microbiol 66:1587–1594.
85. Heijs SK, Haese RR, Van Der Wielen PWJJ, Forney LJ, Van Elsas JD. 2007. Use of 16S rRNA gene based clone libraries to assess microbial communities potentially involved in anaerobic methane oxidation in a Mediterranean cold seep. Microb Ecol 53:384–398.
86. McLellan SL, Huse SM, Mueller-Spitz SR, Andreishcheva EN, Sogin ML. 2010. Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent. Environ Microbiol 12:378–392.
87. Vandewalle JL, Goetz GW, Huse SM, Morrison HG, Sogin ML, Hoffmann RG, Yan K, Mclellan SL. 2012. *Acinetobacter*, *Aeromonas* and *Trichococcus* populations dominate the microbial community within urban sewer infrastructure. Environ Microbiol 14:2358–2552.

88. Lee JE, Lee S, Sung J, Ko G. 2011. Analysis of human and animal fecal microbiota for microbial source tracking. ISME J 5:362–365.
89. Unno T, Jang J, Han D, Kim JH, Sadowsky MJ, Kim OS, Chun J, Hur HG. 2010. Use of barcoded pyrosequencing and shared OTUs to determine sources of fecal bacteria in watersheds. Environ Sci Technol 44:7777–7782.
90. Fisher JC, Murat Eren A, 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.
91. Tan B, Ng C, Nshimyimana JP, Loh LL, Gin KYH, Thompson JR. 2015. Next-generation sequencing (NGS) for assessment of microbial water quality: Current progress, challenges, and future opportunities. Front Microbiol 6:Article 1027.
92. Unno T, Staley C, Brown CM, Han D, Sadowsky MJ, Hur HG. 2018. Fecal pollution: new trends and challenges in microbial source tracking using next-generation sequencing. Environ Microbiol 20:3132–3140.
93. Staley C, Sadowsky MJ. 2016. Application of metagenomics to assess microbial communities in water and other environmental matrices. J Mar Biol Assoc United Kingdom 96:121–129.
94. Newton RJ, Bootsma MJ, Morrison HG, Sogin ML, McLellan SL. 2013. A microbial signature approach to identify fecal pollution in the waters off an urbanized coast of Lake Michigan. Environ Microbiol 65:1011–1023.
95. 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.
96. 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.
97. 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.
98. McLellan SL, Newton RJ, Vandewalle JL, Shanks OC, Huse SM, Eren a M, Sogin ML. 2013. Sewage reflects the distribution of human faecal *Lachnospiraceae*. Environ Microbiol 15:2213–27.
99. Koskey AM, Fisher JC, Eren AM, Ponce-Terashima R, Reis MG, Blanton RE, Mclellan SL. 2014. *Blautia* and *Prevotella* sequences distinguish human and animal fecal pollution in Brazil surface waters. Environ Microbiol Rep 6:696–704.
100. Ahmed W, Gyawali P, Feng S, McLellan S. 2019. Host specificity and sensitivity of established and novel sewage-associated marker genes in human and nonhuman fecal samples. Appl Environ Microbiol 85:e0064-19.
101. Feng S, McLellan SL. 2019. Highly specific sewage-derived *Bacteroides* qPCR assays target sewage polluted waters. Appl Environ Microbiol 85:e02696-18.
102. Brokamp C, Beck AF, Muglia L, Ryan P. 2017. Combined sewer overflow events and childhood emergency department visits: A case-crossover study. Sci Total Environ 607–608:1180–1187.
103. Fremaux B, Gritzfeld J, Boa T, Yost CK. 2009. Evaluation of host-specific *Bacteroidales* 16S rRNA gene markers as a complementary tool for detecting fecal

pollution in a prairie watershed. Water Res 43:4838–4849.
104. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA. 2005. Diversity of the human intestinal microbial flora. Science 308:1635–8.
105. Okabe S, Shimazu Y. 2007. Persistence of host-specific *Bacteroides-Prevotella* 16S rRNA genetic markers in environmental waters: Effects of temperature and salinity. Appl Microbiol Biotechnol 76:935–944.
106. Gorvitovskaia A, Holmes SP, Huse SM. 2016. Interpreting *Prevotella* and *Bacteroides* as biomarkers of diet and lifestyle. Microbiome 4:15.
107. Sinigalliano CD, Fleisher JM, Gidley ML, Solo-Gabriele HM, Shibata T, Plano LRW, Elmir SM, Wanless D, Bartkowiak J, Boiteau R, Withum K, Abdelzaher AM, He G, Ortega C, Zhu X, Wright ME, Kish J, Hollenbeck J, Scott T, Backer LC, Fleming LE. 2010. Traditional and molecular analyses for fecal indicator bacteria in non-point source subtropical recreational marine waters. Water Res 44:3763–3772.
108. Matsuki T, Watanabe K, Fujimoto J, Miyamoto Y, Takada T, Matsumoto K, Oyaizu H, Tanaka R. 2002. Development of 16S rRNA-gene-targeted group-specific primers for the detection and identification of predominant bacteria in human feces 68:5445–5451.
109. Durso LM, Harhay GP, Smith TPL, Bono JL, DeSantis TZ, Harhay DM, Andersen GL, Keen JE, Laegreid WW, Clawson ML. 2010. Animal-to-animal variation in fecal microbial diversity among beef cattle. Appl Environ Microbiol 76:4858–4862.
110. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ Prepr 4:e2409v1.
111. 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.
112. Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.
113. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870-1874.
114. 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.
115. Nelson MC, Morrison HG, Benjamino J, Grim SL, Graf J. 2014. Analysis, optimization and verification of Illumina-generated 16s rRNA gene amplicon surveys. PLoS One 9: e94249.
116. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103:12115–12120.
117. Huse SM, Mark Welch DB, Voorhis A, Shipunova A, Morrison HG, Eren AM, Sogin ML. 2014. VAMPS: A website for visualization and analysis of microbial population structures. BMC Bioinformatics 15:41.

118. Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet 4:e1000255.
119. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE. 2014. Human genetics shape the gut microbiome. Cell 159:789–799.
120. De Cáceres M, Legendre P. 2009. Associations between species and groups of sites: Indices and statistical inference. Ecology 90:3566–3574.
121. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
122. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD. 2011. Linking long-term dietary patterns with gut microbial enterotypes. Science 334:105–108.
123. Dethlefsen L, McFall-Ngai M, Relman DA. 2007. An ecological and evolutionary perspective on human–microbe mutualism and disease. Nature 449:811–818.
124. Mclellan SL, Newton RJ, Vandewalle JL, Shanks OC, Susan M, Eren AM, Sogin ML. 2014. Sewage reflects the distribution of human faecal *Lachnospiraceae*. 15:2213–2227.
125. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol 12:R60.
126. Chakravorty S, Helb D, Burday M, Connell N, Alland D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods 69:330–339.
127. Shanks OC, Kelty 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.
128. Jabari L, Gannoun H, Cayol JL, Hamdi M, Fauque G, Ollivier B, Fardeau ML. 2012. Characterization of *Defluviitalea saccharophila* gen. nov., sp. nov., a thermophilic bacterium isolated from an upflow anaerobic filter treating abattoir wastewaters, and proposal of *Defluviitaleaceae* fam. nov. Int J Syst Evol Microbiol 62:550–555.
129. Layton BA, Cao Y, Ebentier DL, Hanley K, Ballesté E, Brandão J, Byappanahalli M, Converse R, Farnleitner AH, Gentry-Shields J, Gidley ML, Gourmelon M, Lee CS, Lee J, Lozach S, Madi T, Meijer WG, Noble R, Peed L, Reischer GH, Rodrigues R, Rose JB, Schriewer A, Sinigalliano C, Srinivasan S, Stewart J, Van De Werfhorst LC, Wang D, Whitman R, Wuertz S, Jay J, Holden PA, Boehm AB, Shanks O, Griffith JF. 2013. Performance of human fecal anaerobe-associated PCR-based assays in a multi-laboratory method evaluation study. Water Res 47:6897–6908.
130. Kreader CA. 1995. Design and evaluation of *Bacteroides* DNA probes for the specific detection of human fecal pollution. Appl Environ Microbiol 61:1171–1179.
131. 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.
132. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.
133. Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620.
134. Wickham H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York.
135. R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
136. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, Mcglinn D, Minchin PR, O 'hara RB, Simpson GL, Solymos P, Henry M, Stevens H, Szoecs E, Wagner H, Oksanen MJ. 2018. vegan: community ecology package. R package version 2.4-5.
137. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res 42.
138. Fisher JC, Newton RJ, Dila DK, McLellan SL. 2015. Urban microbial ecology of a freshwater estuary of Lake Michigan. Elem Sci Anthr 3:p.000064.
139. Labus JS, Hollister EB, Jacobs J, Kirbach K, Oezguen N, Gupta A, Acosta J, Luna RA, Aagaard K, Versalovic J, Savidge T, Hsiao E, Tillisch K, Mayer EA. 2017. Differences in gut microbial composition correlate with regional brain volumes in irritable bowel syndrome. Microbiome 5:49.
140. Strati F, Cavalieri D, Albanese D, De Felice C, Donati C, Hayek J, Jousson O, Leoncini S, Renzi D, Calabrò A, De Filippo C. 2017. New evidences on the altered gut microbiota in autism spectrum disorders. Microbiome 5:24.
141. Zwittink RD, Renes IB, van Lingen RA, van Zoeren-Grobben D, Konstanti P, Norbruis OF, Martin R, Groot Jebbink LJM, Knol J, Belzer C. 2018. Association between duration of intravenous antibiotic administration and early-life microbiota development in late-preterm infants. Eur J Clin Microbiol Infect Dis 37:475–483.
142. Madan JC, Koestle DC, Stanton BA, Davidson L, Moulton LA, Housman ML, Moore JH, Guill MF, Morrison HG, Sogin ML, Hampton TH, Karagas MR, Palumbo PE, Foster JA, Hibberd PL, O'Toole GA. 2012. Serial analysis of the gut and respiratory microbiome in cystic fibrosis in infancy: Interaction between intestinal and respiratory tracts and impact of nutritional exposures. MBio 3:e00251-12.
143. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2009. A core gut microbiome in obese and lean twins. Nature 457:480–484.
144. Smith-Brown P, Morrison M, Krause L, Davies PSW. 2016. Dairy and plant based food intakes are associated with altered faecal microbiota in 2 to 3 year old Australian children. Sci Rep 6: 32385.
145. Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120.

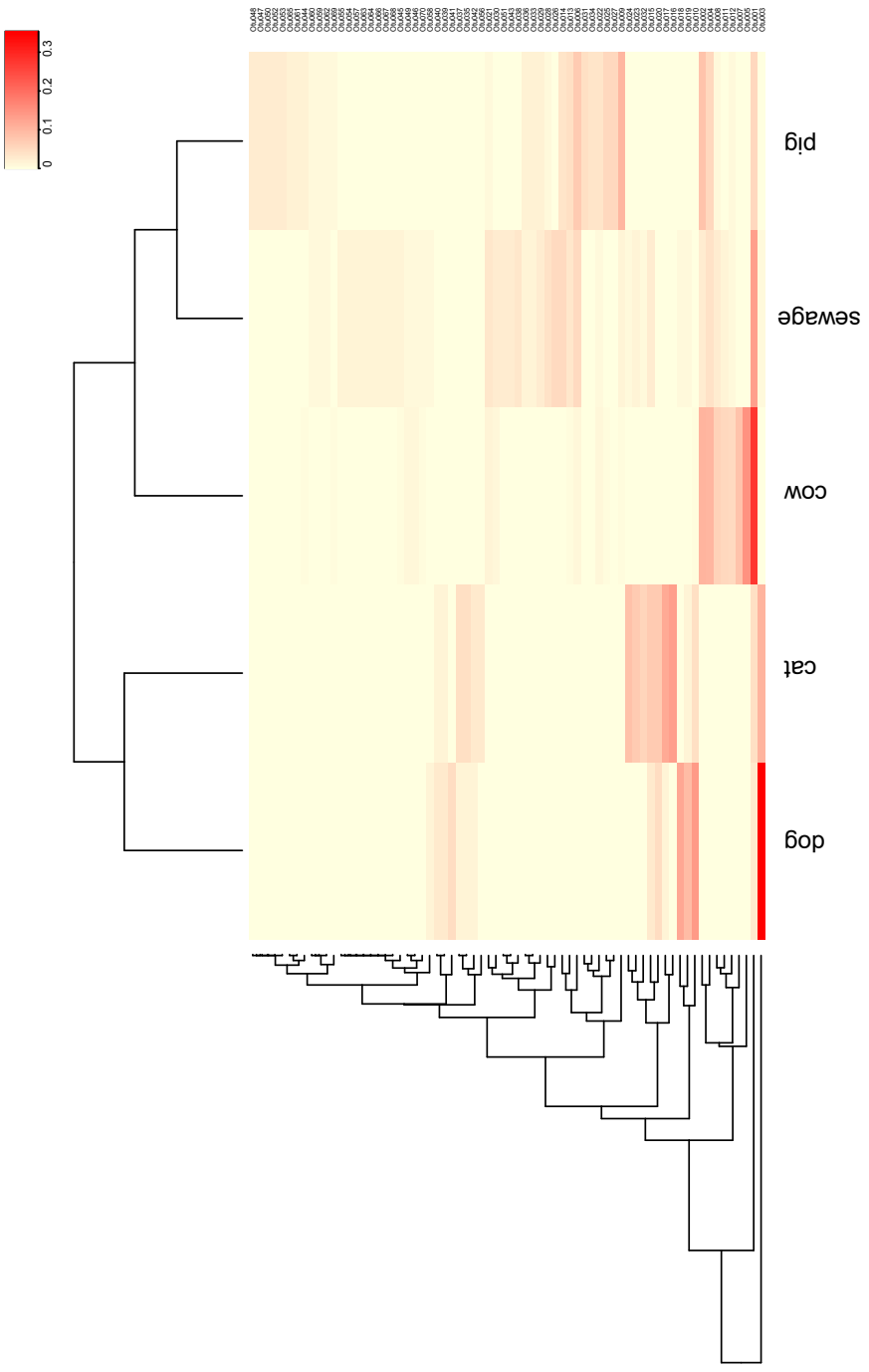
146. Bakir MA, Sakamoto M, Kitahara M, Matsumoto M, Benno Y. 2006. *Bacteriodes dorei* sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 56:1639–1643.
147. Leser TD, Amenuvor JZ, Jensen TK, Lindecrona RH, Boye M, Moøller K. 2002. Culture-independent analysis of gut bacteria: The pig gastrointestinal tract microbiota revisited. Appl Environ Microbiol 68:673–690.
148. Suchodolski JS, Camacho J, Steiner JM. 2008. Analysis of bacterial diversity in the canine duodenum, jejunum, ileum, and colon by comparative 16S rRNA gene analysis. FEMS Microbiol Ecol 66:567–578.
149. Zhu XY, Zhong T, Pandya Y, Joerger RD. 2002. 16S rRNA-based analysis of microbiota from the cecum of broiler chickens. Appl Environ Microbiol 68:124–137.
150. Bjerrum L, Engberg RM, Leser TD, Jensen BB, Finster K, Pedersen K. 2006. Microbial community composition of the ileum and cecum of broiler chickens as revealed by molecular and culture-based techniques. Poult Sci 85:1151–1164.
151. Nozu R, Ueno M, Hayashimoto N. 2016. Composition of fecal microbiota of laboratory mice derived from Japanese commercial breeders using 16S rRNA gene clone libraries. J Vet Med Sci 78:1045–1050.
152. Hespell RB, Whitehead TR. 1990. Physiology and genetics of xylan degradation by gastrointestinal tract bacteria. J Dairy Sci 73:3013–3022.
153. Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper L V., Gordon JI. 2003. A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. Science 299:2074–2076.
154. Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. 2012. Microbial degradation of complex carbohydrates in the gut. Gut Microbes 3:289–306.
155. Wexler AG, Goodman AL. 2017. An insider's perspective: *Bacteroides* as a window into the microbiome. Nat Microbiol 2:17026.
156. Nishiyama T, Ueki A, Kaku N, Watanabe K, Ueki K. 2009. *Bacteroides graminisolvens* sp. nov., a xylanolytic anaerobe isolated from a methanogenic reactor treating cattle waste. Int J Syst Evol Microbiol 59:1901–1907.
157. Hatamoto M, Kaneshige M, Nakamura A, Yamaguchi T. 2014. *Bacteroides luti* sp. nov., an anaerobic, cellulolytic and xylanolytic bacterium isolated from methanogenic sludge. Int J Syst Evol Microbiol 64:1770–1774.
158. Ismaeil M, Yoshida N, Katayama A. 2018. *Bacteroides sedimenti* sp. nov., isolated from a chloroethenes-dechlorinating consortium enriched from river sediment. J Microbiol 56:619–627.
159. Kim JR, Beecroft NJ, Varcoe JR, Dinsdale RM, Guwy AJ, Slade RCT, Thumser A, Avignone-Rossa C, Premier GC. 2011. Spatiotemporal development of the bacterial community in a tubular longitudinal microbial fuel cell. Appl Microbiol Biotechnol 90:1179–1191.
160. 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.
161. Mayer RE, Reischer GH, Ixenmaier SK, Derx J, Blaschke AP, Ebdon JE, Linke R, Egle L, Ahmed W, Blanch AR, Byamukama D, Savill M, Mushi D, Cristóbal HA, Edge TA, Schade MA, Aslan A, Brooks YM, Sommer R, Masago Y, Sato MI, Taylor HD, Rose JB, Wuertz S, Shanks OC, Piringer H, Mach RL, Savio D,

Zessner M, Farnleitner AH. 2018. Global distribution of human-associated fecal genetic markers in reference samples from six continents. Environ Sci Technol 52:5076–5084.
162. Olapade OA, Depas MM, Jensen ET, McLellan SL. 2006. Microbial communities and fecal indicator bacteria associated with *Cladophora* mats on beach sites along Lake Michigan shores. Appl Environ Microbiol 72:1932–1938.
163. Whitman RL, Byappanahalli MN, Spoljaric AM, Przybyla-Kelly K, Shively DA, Nevers MB. 2014. Evidence for free-living *Bacteroides* in *Cladophora* along the shores of the Great Lakes. Aquat Microb Ecol 72:117–126.
164. Alm EW, Daniels-Witt QR, Learman DR, Ryu H, Jordan DW, Gehring TM, Santo Domingo J. 2018. Potential for gulls to transport bacteria from human waste sites to beaches. Sci Total Environ 615:123–130.
165. Brown CM, Staley C, Wang P, Dalzell B, Chun CL, Sadowsky MJ. 2017. A high-throughput DNA-sequencing approach for determining sources of fecal bacteria in a Lake Superior estuary. Environ Sci Technol 51:8263–8271.
166. Wiggins BA. 1996. Discriminant analysis of antibiotic resistance patterns in fecal streptococci, a method to differentiate human and animal sources of fecal pollution in natural waters. Appl Environ Microbiol 62:3997–4002.
167. Parveen S, Portier KM, Robinson K, Edmiston L, Tamplin ML. 1999. Discriminant analysis of ribotype profiles of *Escherichia coli* for differentiating human and nonhuman sources of fecal pollution. Appl Environ Microbiol 65:3142–3147.
168. Li J, McLellan S, Ogawa S. 2006. Accumulation and fate of green fluorescent labeled *Escherichia coli* in laboratory-scale drinking water biofilters. Water Res 40:3023–3028.
169. Haugland RA, Siefring SC, Wymer LJ, Brenner KP, Dufour AP. 2005. Comparison of *Enterococcus* measurements in freshwater at two recreational beaches by quantitative polymerase chain reaction and membrane filter culture analysis. Water Res 39:559–68.
170. 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.
171. Vineis JH, Ringus DL, Morrison HG, Delmont TO, Dalal S, Raffals LH, Antonopoulos DA, Rubin DT, Eren AM, Chang EB, Sogin ML. 2016. Patient-specific *Bacteroides* genome variants in pouchitis. MBio 7:1–11.
172. Shanks OC, Kelty CA, Oshiro R, Haugland RA, Madi T, Brooks L, Field KG, Sivaganesan M. 2016. Data acceptance criteria for standardized human-associated fecal source identification quantitative real-time PCR methods. Appl Environ Microbiol 82:2773–2782.
173. 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 320:1647–1651.
174. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 43:e37.
175. Kelty CA, Varma M, Sivaganesan M, Haugland RA, Shanks OC. 2012.

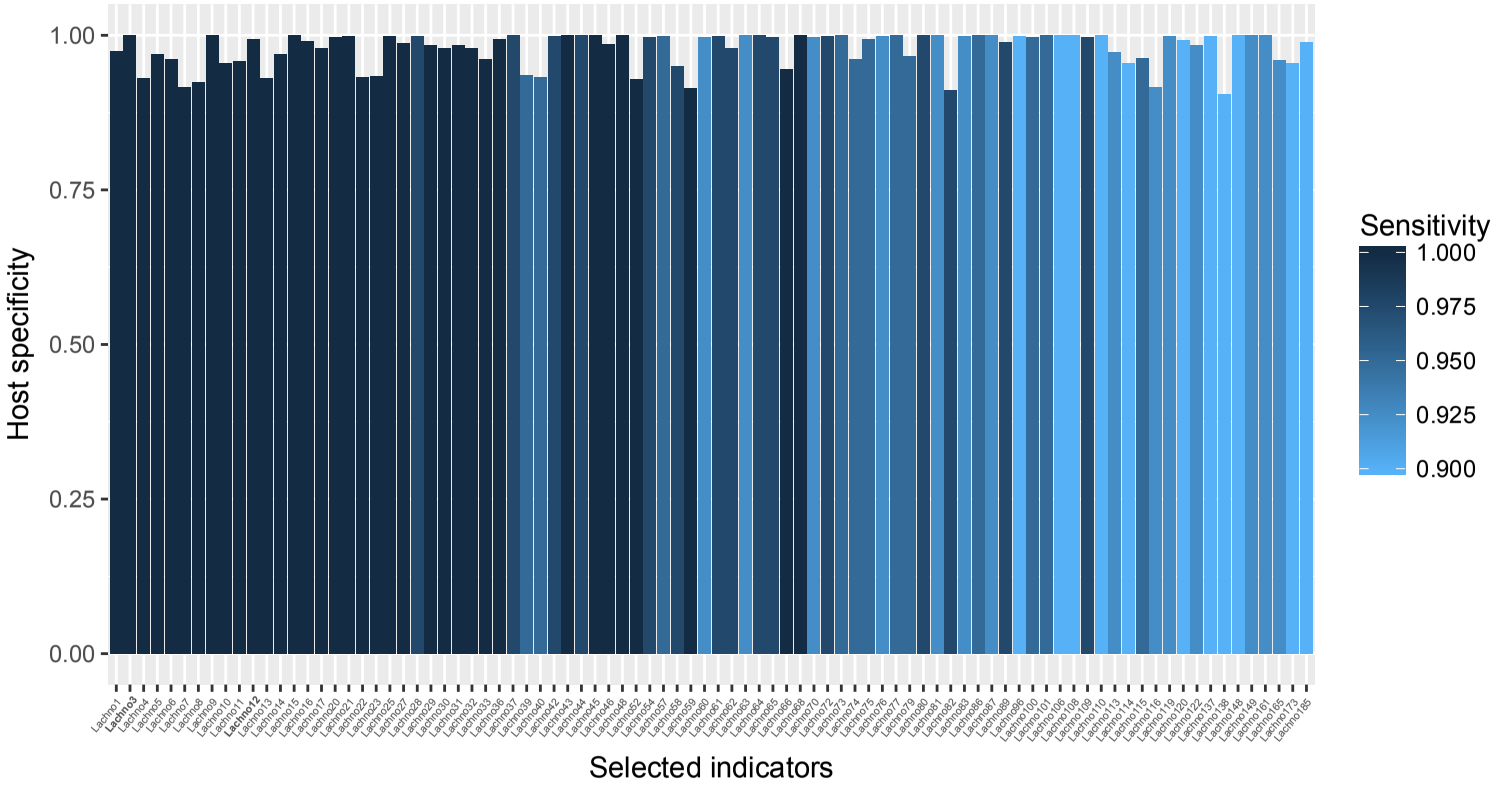
Distribution of genetic marker concentrations for fecal indicator bacteria in sewage and animal feces. Appl Environ Microbiol 78:4225–4232.
176. Ley RE, Peterson DA, Gordon JI. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124:837–848.
177. Groussin M, Mazel F, Sanders JG, Smillie CS, Lavergne S, Thuiller W, Alm EJ. 2017. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nat Commun 8:14319.
178. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. 2012. Diversity, stability and resilience of the human gut microbiota. Nature 489:220–230.
179. Roguet A, Eren AM, Newton RJ, McLellan SL. 2018. Fecal source identification using random forest. Microbiome 6:1–15.
180. Nishida AH, Ochman H. 2019. A great-ape view of the gut microbiome. Nat Rev Genet 20:195–206.
181. Moeller AH, Caro-Quintero A, Mjungu D, Georgiev A V., Lonsdorf E V., Muller MN, Pusey AE, Peeters M, Hahn BH, Ochman H. 2016. Cospeciation of gut microbiota with hominids. Science 353:380–382.
182. Korpela K, Costea P, Coelho LP, Kandels-Lewis S, Willemsen G, Boomsma DI, Segata N, Bork P. 2018. Selective maternal seeding and environment shape the human gut microbiome. Genome Res 28:561–568.
183. West AG, Waite DW, Deines P, Bourne DG, Digby A, McKenzie VJ, Taylor MW. 2019. The microbiome in threatened species conservation. Biol Conserv 229:85–98.
184. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. 2016. Diet-induced extinctions in the gut microbiota compound over generations. Nature 529:212–215.
185. Broussard JL, Devkota S. 2016. The changing microbial landscape of Western society: Diet, dwellings and discordance. Mol Metab 5:737–742.
186. Moeller AH. 2017. The shrinking human gut microbiome. Curr Opin Microbiol 38:30–35.
187. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17:55.
188. Pollock J, Glendinning L, Wisedchanwet T, Watson M. 2018. The madness of microbiome: Attempting to find consensus "best practice" for 16S microbiome studies. Appl Environ Microbiol 84:e02627-17.
189. Ibekwe AM, Leddy M, Murinda SE. 2013. Potential human pathogenic bacteria in a mixed urban watershed as revealed by pyrosequencing. PLoS One 8:e79490.
190. Nshimyimana JP, Freedman AJE, Shanahan P, Chua LCH, Thompson JR. 2017. Variation of bacterial communities with water quality in an urban tropical catchment. Environ Sci Technol 51:5591−5601.
191. Ghaju Shrestha R, Tanaka Y, Malla B, Bhandari D, Tandukar S, Inoue D, Sei K, Sherchand JB, Haramoto E. 2017. Next-generation sequencing identification of pathogenic bacterial genes and their relationship with fecal indicator bacteria in different water sources in the Kathmandu Valley, Nepal. Sci Total Environ 601–602:278–284.

192. U.S. Environmental Protection Agency (EPA). 2019. Method 1696 : Characterization of human fecal pollution in water by polymerase chain reaction (qPCR) assay. Washington, DC. EPA 821-R-19-002.

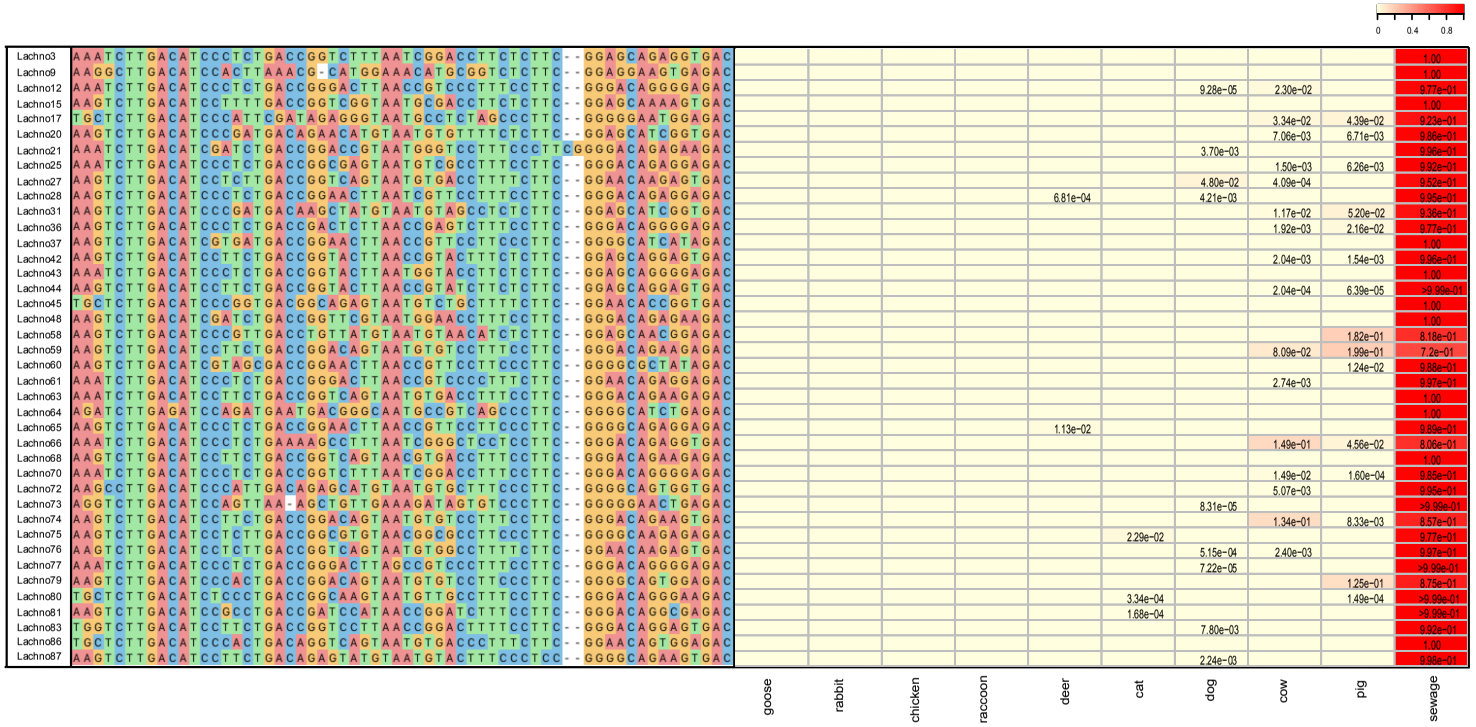
Appendix A. Supplemental Material for Chapter 2



Appendix A Figure 1 The relative abundance of the 70 OTUs in sewage and animal clone libraries. Values of relative abundances increase from yellow to red. Dendrograms represent relationship of abundance patterns (Y-axis) and relationship of hosts using absolute distance between OTUs.



Appendix A Figure 2 Sewage indicator candidates that have over 90% host specificity and sensitivity chosen by "indicspecies". The X-axis lists the chosen indicators and the Y-axis represents their host specificities. Deeper color represents higher indicator sensitivity.



Appendix A Figure 3 List of the top 40 *Lachnospiraceae* fecal markers candidates. Sequences are shown in an alignment. The distribution of each candidate in sewage and nine animal hosts in the V6 NGS dataset are shown in the heatmap. The relative abundances are calculated using total sum of that sequence for all samples and increase from yellow to red. Relative abundance values above zero are annotated on the heatmap.

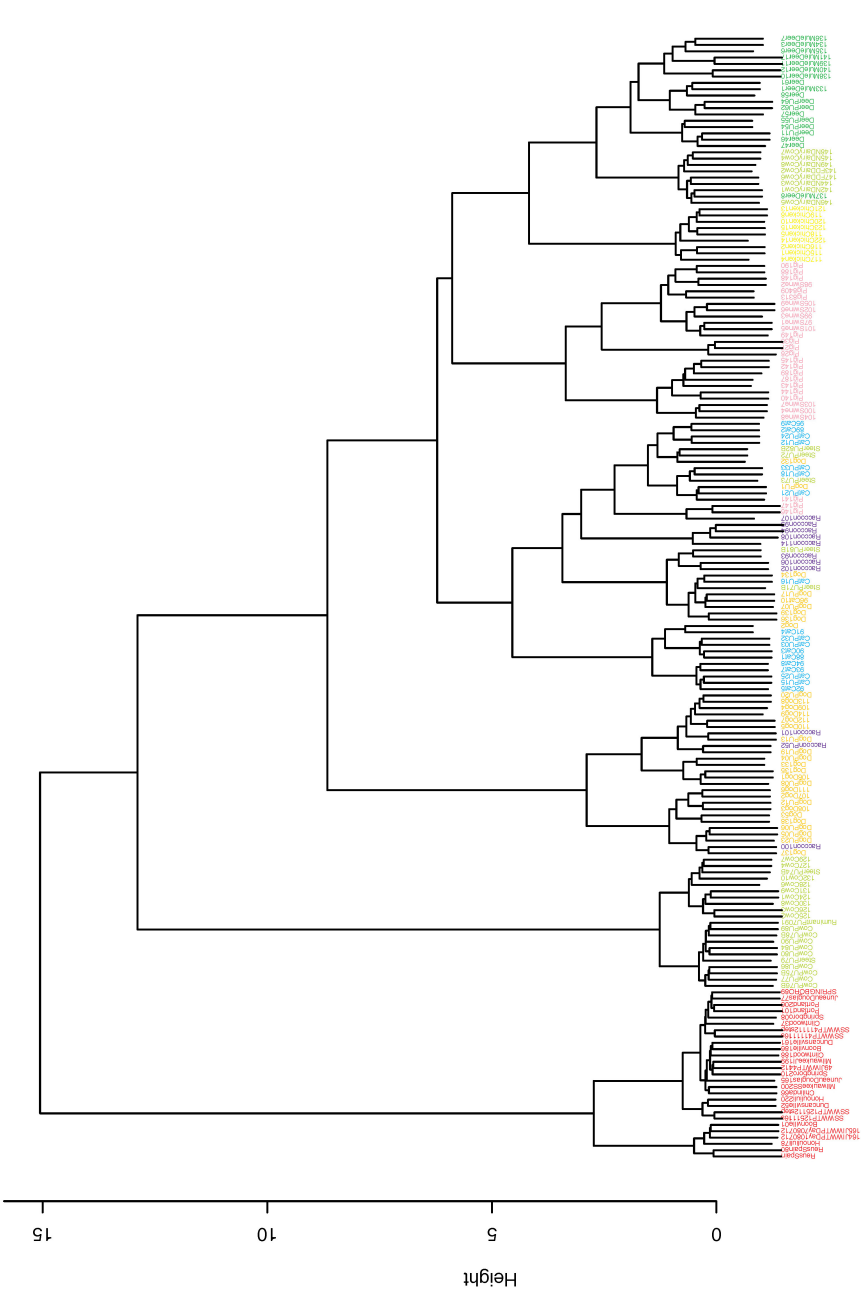
Appendix A Table 1 Standard curve parameters of the qPCR assays used in Chapter 2.

Assay Name	Slope	Y-intercept	R^2	Efficiency (%)
Lachno3	-3.333	38.321	0.999	95.519
Lachno12	-3.827	40.914	0.998	101.483
Lachno2	-3.525	38.182	0.999	92.316
HB	-3.350	37.202	0.999	98.887
HF183/BacR287	-3.515	38.550	0.999	92.515

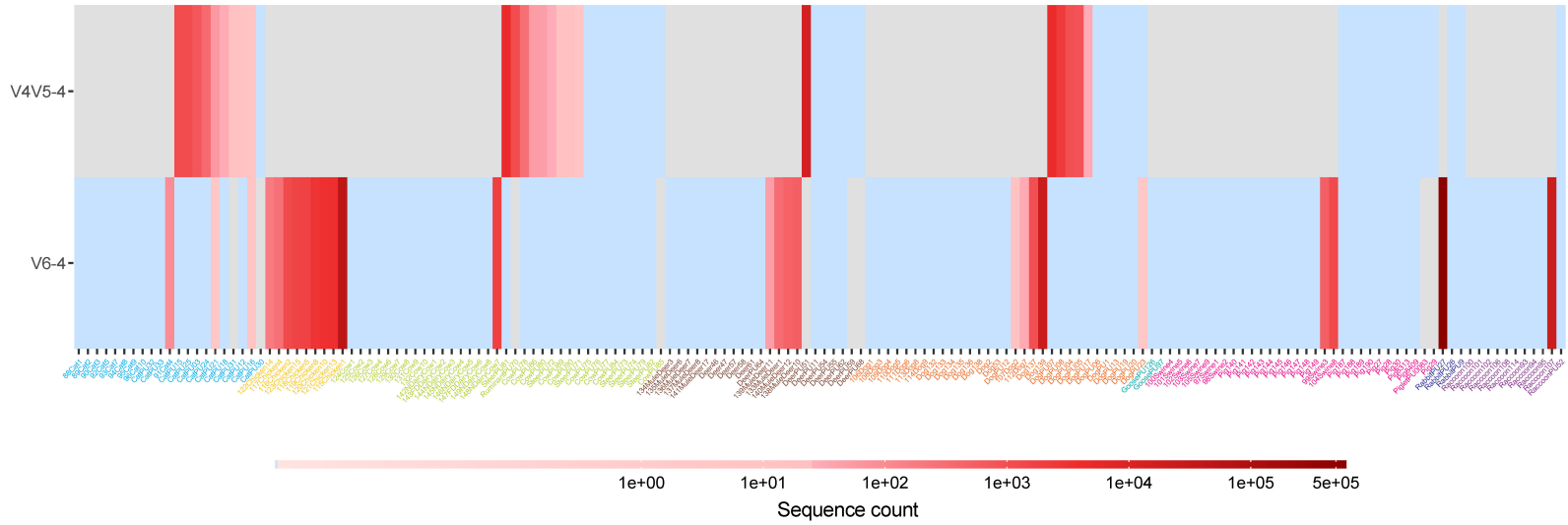
Appendix A Table 2 Validation results of the Lachno3, Lachno12, HB, HF183/BacR287 and Lachno2 assays in animal fecal samples. N is the number of tested animal individuals within a host. Results are displayed as positive individual numbers (n) with the average copy numbers (CN).

Assay name	DNA amt (ng)	Cat (N=11)	Dog (N=10)	Pig (N=9)	Cow (N=10)	Deer (N=11)	Gull (N=4)
	Positive n / CN per amt of DNA						
Lachno3	1	2/2	0	0	0	0	0
	0.1	0	0	0	0	0	0
	0.01	0	0	0	0	0	0
Lachno12	1	0	0	3/12	4/216	0	0
	0.1	0	0	3/1	4/28	0	0
	0.01	0	0	0	2/4	0	0
HB	1	0	2/375	0	0	3/406	0
	0.1	0	2/34	0	0	3/41	0
	0.01	0	2/2	0	0	1/15	0
HF183/BacR287	1	0	0	0	0	3/364	0
	0.1	0	0	0	0	1/117	0
	0.01	0	0	0	0	1/7	0
Lachno2	1	9/884	7/3,100	9/782	7/12	3/9	0
	0.1	9/93	1/2,320	7/77	1/2	0	0
	0.01	2/41	1/223	3/8	0	0	0

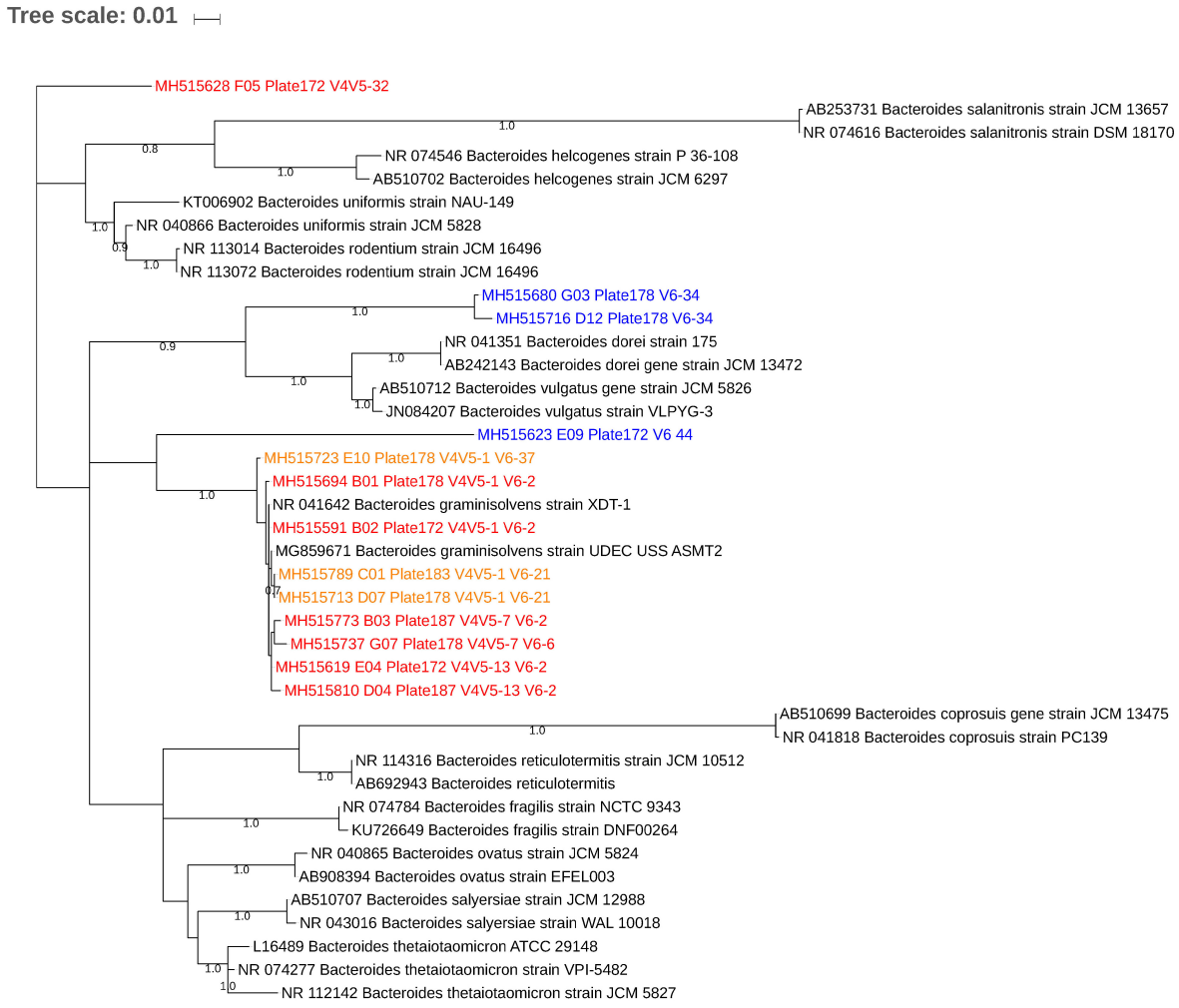
Appendix B. Supplemental Material for Chapter 3



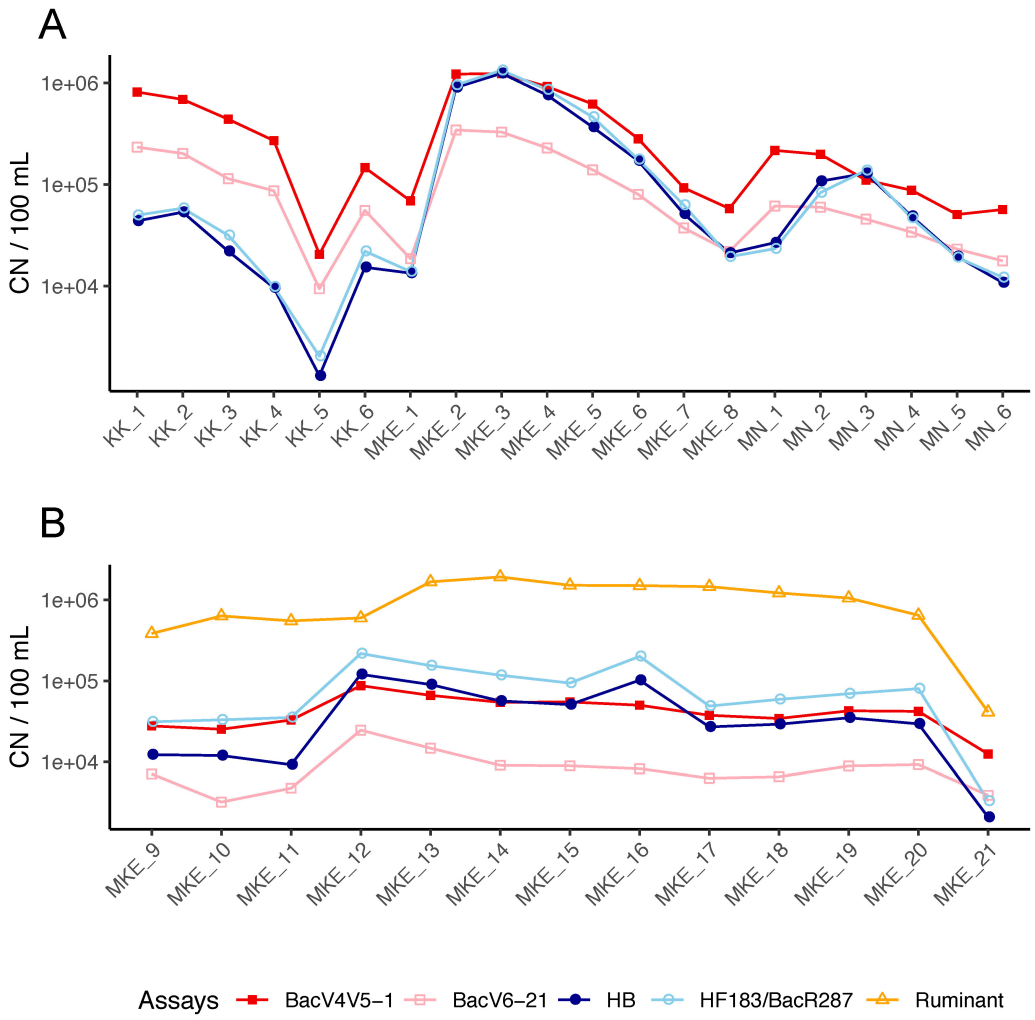
Appendix B Figure 1 *Bacteroides* oligotypes hierarchical cluster analysis based on Bray-Curtis dissimilarity. Sewage samples are labeled in red. Animal hosts are labeled in other different colors.



Appendix B Figure 2 Host specificities of the HF183 downstream regions (V4V5-4 and V6-4). Gray color represents samples that are not sequenced. Light blue color represents samples negative for the sequence type. Positive sequence counts increase from pink color to dark red color. All animals are grouped in host types and labeled in different colors.



Appendix B Figure 3 Maximum likelihood tree constructed from *Bacteroides* reference strains and clones that are found to contain V4V5 and V6 regions marker candidates. The clones contain only the specific V4V5 region marker candidates are labeled in red, only the specific V6 region marker candidates are in blue, and these have both specific marker regions are in orange. Bootstrap values between 0.7 to 1 are shown in the middle position of corresponding branches.



Appendix B Figure 4 Comparison of the four *Bacteroides* assays in environmental water samples. Line graph is used to show the fluctuation patterns of assay results, not correlations of samples. A shows comparison of the BacV4V5-1, BacV6-21, HB and HF183/BacR287 assays in sewage-contaminated water samples from Kinnickinnic River (KK), Milwaukee River (MKE) and Menomonee River (MN) from a 2016 combined sewer overflow (CSO) event. B shows comparison of the four *Bacteroides* assays and one ruminant marker assay in agricultural-contaminated MKE river water samples from rain and post-CSO events.

Appendix B Table 1 The V4V5 marker candidates with their specificities and sensitivities from the permutation test and the NGS dataset.

Marker name	Permutated specificity	Permutated sensitivity	V4V5 NGS dataset specificity	V4V5 NGS dataset sensitivity	Possible Source	Sequence
V4V5-1	1	1	1	1	Sewer pipe	ACGGAGGATCCAAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGTTGACATATAAGTCA GCTGTGAAAGTTTACGGCTCAACCGTGAAATTGCAGTTGATACTGTATGTCTTGAGTGTACAAGAGG TGGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTCCAATTGCGAAGGCA GCTCACTGGGGTACAACTGACACTGAGGCTCGAAAGTGTGGGTATCAAACAGGATTAGATACCCTG GTAGTCCACACAGTAAACGATGAATACTCGCTGTTTGCGATATACAGTAAGCGGCCAAGCGAAAGC ATTAAGTATTCCACCTGGGGAGTACGCCGGCAACGGTGAA
V4V5-7	1	1	1	1	Sewer pipe	ACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGTTGACGTATAAGTCA GCTGTGAAAGTTTACGGCTCAACCGTGAAATTGCAGTTGATACTGTATGTCTTGAGTGTACAAGAGG TGGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTCCAATTGCGAAGGCA GCTCACTGGGGTACAACTGACACTGAGGCTCGAAAGTGTGGGTATCAAACAGGATTAGATACCCTG GTAGTCCACACAGTAAACGATGAATACTCGCTGTTTGCGATATACAGTAAGCGGCCAAGCGAAAGC ATTAAGTATTCCACCTGGGGAGTACGCCGGCAACGGTGAA
V4V5-13	1	0.96	1	0.96	Sewer pipe	ACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGTTGACATATAAGTCA GCTGTGAAAGTTTACGGCTCAACCGTGAAATTGCAGTTGATACTGTATGTCTTGAGTGTACAAGAGG TGGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTCCAATTGCGAAGGCA GCTCACTGGGGTACAACTGACACTGAGGCTCGAAAGTGTGGGTATCAAACAGGATTAGATACCCTG GTAGTCCACACAGTAAACGATGAATACTCGCTGTTTGCGATATACAGTAAGCGGCCAAGCGAAAGC ATTAAGTATTCCACCTGGGGAGTACGCCGGCAACGGTGAA
V4V5-22	1	0.95	1	0.95	Human feces	ACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGCAGACGGGTCGTTAAGTCA GCTGTGAAAGTTTGGGGCTCAACCTTAAAATTGCAGTTGATACTGGCGTCCTTGAGTGCGGTTGAGG TGTGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCA GCACACTAATCCGTAACTGACGTTCATGCTCGAAAGTGTGGGTATCAAACAGGATTAGATACCCTG GTAGTCCACACGGTAAACGATGGATACTCGCTGTTGGCGATATACTGTCAGCGGCTTAGCGAAAGC GTTAAGTATCCCACCTGGGGAGTACGCCGGCAACGGTGAA
V4V5-25	1	0.91	1	0.91	Sewer pipe	ACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGATGTTTAAGTCA GTTGTGAAAGTTTAAGGCTCAACCTTGAAATTGCAGTTGATACTGGATATCTTGAGTACATTGAATG TGGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTCCAATTGCGAAGGCA GCTCACAGTAATGTAACTGACGCTGATGCTCGAAAGTGTGGGTATCAAACAGGATTAGATACCCTG GTAGTCCACACAGTAAACGATGAATACTCGCTGTTTGCGATATACAGTAAGCGGCCAAGCGAAAGC GTTAAGTATTCCACCTGGGGAGTACGCCGGCAACGGTGAA
V4V5-32	1	0.94	1	0.94	Sewer pipe	ACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGGTGCTTAAGTCA GTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGGTACCTTGAGTGCAGCATAGG TAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCA GCTTACTGGACTGTAACTGACGCTGATGCTCGAAAGTGTGGGTATCAAACAGGATTAGATACCCTG GTAGTCCACACAGTAAACGATGAATACTCGCTGTTGGCGATACACAGTCAGCGGCCAAGCGAAAGC ATTAAGTATTCCACCTGGGGAGTACGCCGGCAACGGTGAA
V4V5-37	1	0.93	1	0.93	Sewer pipe	ACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGATTATTAAGTCA GTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGTAGTCTTGAGTGCAGCAGAG GTAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGC AGCTTACTGGACTGTAACTGACGCTGATGCTCGAAAGTGTGGGTATCAAACAGGATTAGATACCCT GGTAGTCCACACAGTAAACGATGAATACTCGCTGTTTGCGATATACAGCAAGCGGCCAAGCGAAAG CATTAAGTATTCCACCTGGGGAGTACGCCGGCAACGGTGAA

Appendix B Table 2 The V6 marker candidates with their specificities and sensitivities from the permutation test and the NGS dataset.

Marker name	Permutated specificity	Permutated sensitivity	V6 NGS dataset specificity	V6 NGS dataset sensitivity	Possible Source	Sequence
V6-21	1	1	1	1	Sewer pipe	CGGGCTTGAATTGCAGAGGAATATAGTTGAAAGATTATGGCCGCAAGGTCTCTGTGA
V6-23	1	1	1	1	Human feces	CGGGCTTAAATTGCAAATGAATTATGGGGAAACCCATAGGCCGTAAGGCATTTGTGA
V6-24	1	1	1	1	Sewer pipe	CAGGCTTAAATTGCAGATGAATATAGTGGAAACATTATAGCCTTCGGGCATCTGTGA
V6-26	1	1	1	1	Sewer pipe	CGGGCTTAAATTGCACAGGAATAATTTGGAAACAGATTAGTCTTCGGACCTGTGTGA
V6-36	1	1	1	1	Sewer pipe	CGGGCTTGAATTGCTAATGAATATATATGAAAGTATATAGCCGCAAGGCATTAGTGA
V6-38	1	1	1	1	Sewer pipe	CGGGCTTGAATTGCTAATGAATGGAGTAGAGATATTTCAGCCGCAAGGCATTAGTGA
V6-44	1	1	1	1	Sewer pipe	CAGGCTTAAATTGCAGATGAATATAGTAGAAATATTATAGCCTTCGGGCATCTGTGA
V6-17	1	0.95	1	1	Sewer pipe	CGGGCTTAAATTGCAAATGAATATAGTGGAAACATTATAGCCAGCAATGGCATTTGTGA
V6-32	1	0.95	1	1	Sewer pipe	CAGGCTTAAATTGCAGATGAATATAGTGGAAACATTATAGTCTTCGGACATCTGTGA
V6-34	1	0.95	1	1	Sewer pipe	CGGGCTTAAATTGCAACTGAATAGCTGAGAGATCAGTTAGCTAGCAATAGCAGTTGTGA
V6-37	1	0.95	1	1	Sewer pipe	CGGGCTTGAATTGCAGAGGAATATAGTTGAAAGATTATAGCCGCAAGGCCTCTGTGA
V6-40	1	0.925	1	1	Sewer pipe	CAGGCTTAAATTGCAGATGAATATGTGGGAAACCATATAGCCAGCAATGGCATCTGTGA
V6-42	1	0.95	1	1	Sewer pipe	CAGGCTTAAATTGCAGATGAATATAGTGGAAACATTATAGCCAGCAATGGCATCTGTGA
V6-45	1	0.9	1	1	Sewer pipe	CAGGCTTAAATTGCAGATGAATATAGTAGAAATATTATAGTCTTCGGACATCTGTGA
V6-50	1	0.975	1	1	Sewer pipe	CGGGCTTGAATTGCAGAGGAATATAGTCGAAAGATTATAGCCGCAAGGTCTCTGTGA
V6-52	1	0.925	1	0.875	Human feces	CGGGCTTAAATTGCAAATGAATATGCCGGAAACGGCATAGCCGCAAGGCATTTGTGA
V6-55	1	0.95	1	0.95	Sewer pipe	CAGGCTTAAATTGCAGATGAATATAGTGGAAACATTATAGCCTTTATGGCATCTGTGA
V6-68	1	0.9	1	1	Sewer pipe	CGGGCTTGAATTGCAGAGGAACATAGTTGAAAGATTATCGCCGCAAGGTCTCTGTGA
V6-73	1	0.925	1	1	Sewer pipe	CGGGCTTAAATTGCAACTGAATAATTGAGAGATCAGTTAGCTAGCAATAGCAGTTGTGA
V6-79	1	0.925	1	1	Sewer pipe	CGGGCTTAAATTGCAACTGAATAACTTAGAGATGAGTTAGCTAGCAATAGCAGTTGTGA
V6-96	1	0.9	1	1	Human feces	CGGGTTTGAACGCATTCGGACCGGAGTGGAAACACTTCTTCTAGCAATAGCCGTTTGCG

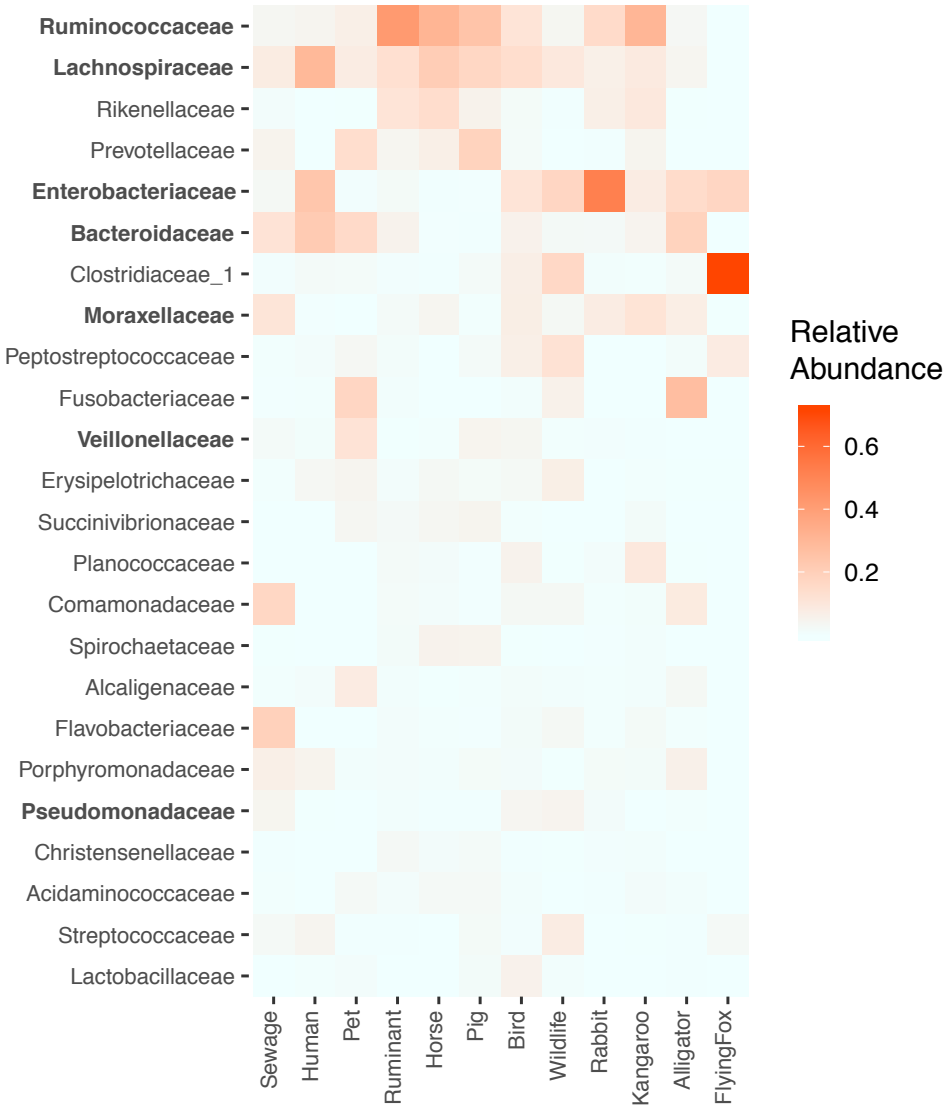
Appendix B Table 3 Amplicon sequences of the BacV4V5-1 and BacV6-21 assays.

Assay name	Amplicon sequence	Reference clone sequences (GenBank Access. No.)
BacV4V5-1	AAGGGAGCGTAGGTTGACATATAAGTCAGCTGTGAAAGTTTACGGCT CAACCGTGAAATTGCAGTTGATACTGTATGTCTTGAGTGTACAAGAG GTGGGCGG	MH515903, MH515911, MH515713
BacV6-21	GCTTGAATTGCAGAGGAATATAGTTGAAAGATTATGGCCGCAAGGTC TCTGTGAAGGTGCTGCATGGTTGTCGTCAGCTCGTGCCGTGAGGTGT CGGCTTAAGTGCCATAACGAGCGCAACCCTTATCATTAGTTACTAAC AGGTCATGCTGAGGACTCTAGTGAGACTGC	MH515733, MH515713

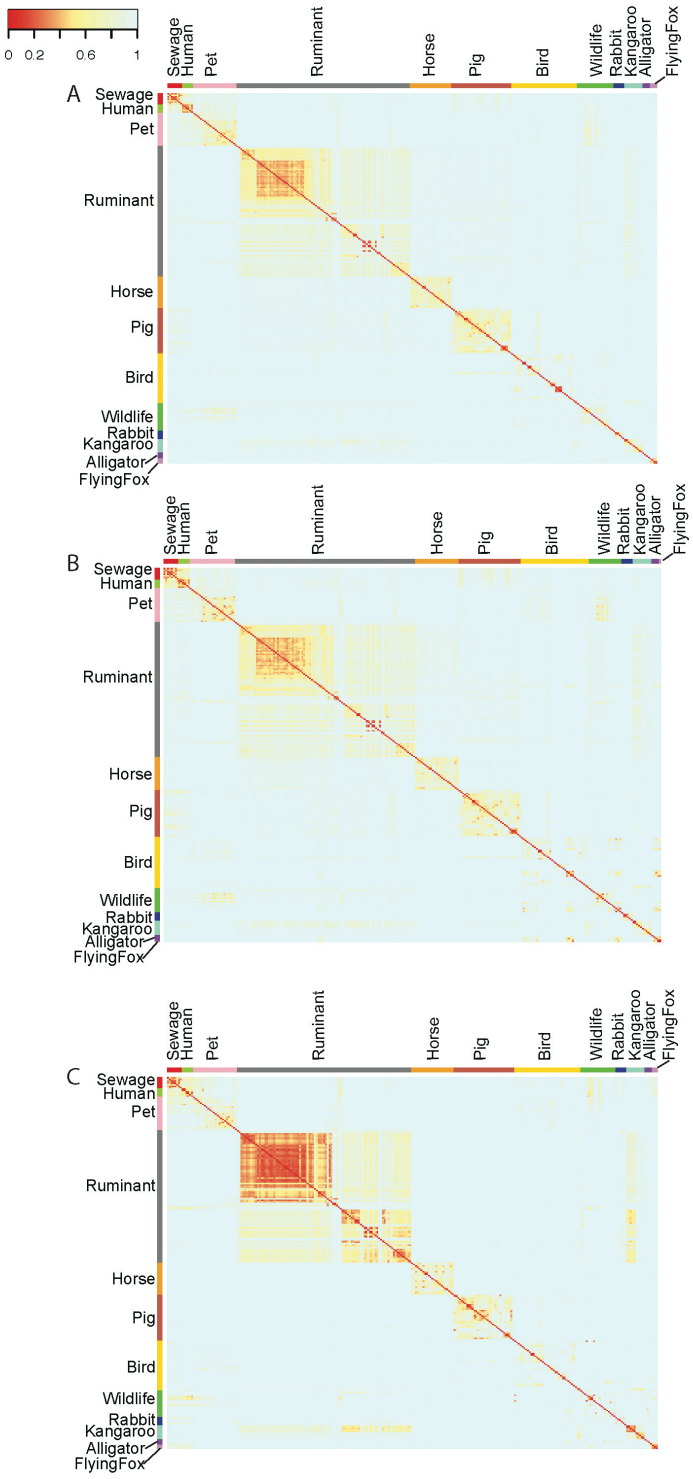
Appendix B Table 4 Slopes, y-intercepts, R^2 and efficiencies of the four *Bacteroides* qPCR assays used in Chapter 3.

Assay name	Slope	Y-intercept	R^2	Efficiency (%)
BacV4V5-1	-3.364	38.056	0.998	98.3235
BacV6-21	-3.399	38.934	0.997	96.869
HB	-3.372	37.468	0.999	98.026
HF183/BacR287	-3.514	38.565	0.999	92.591

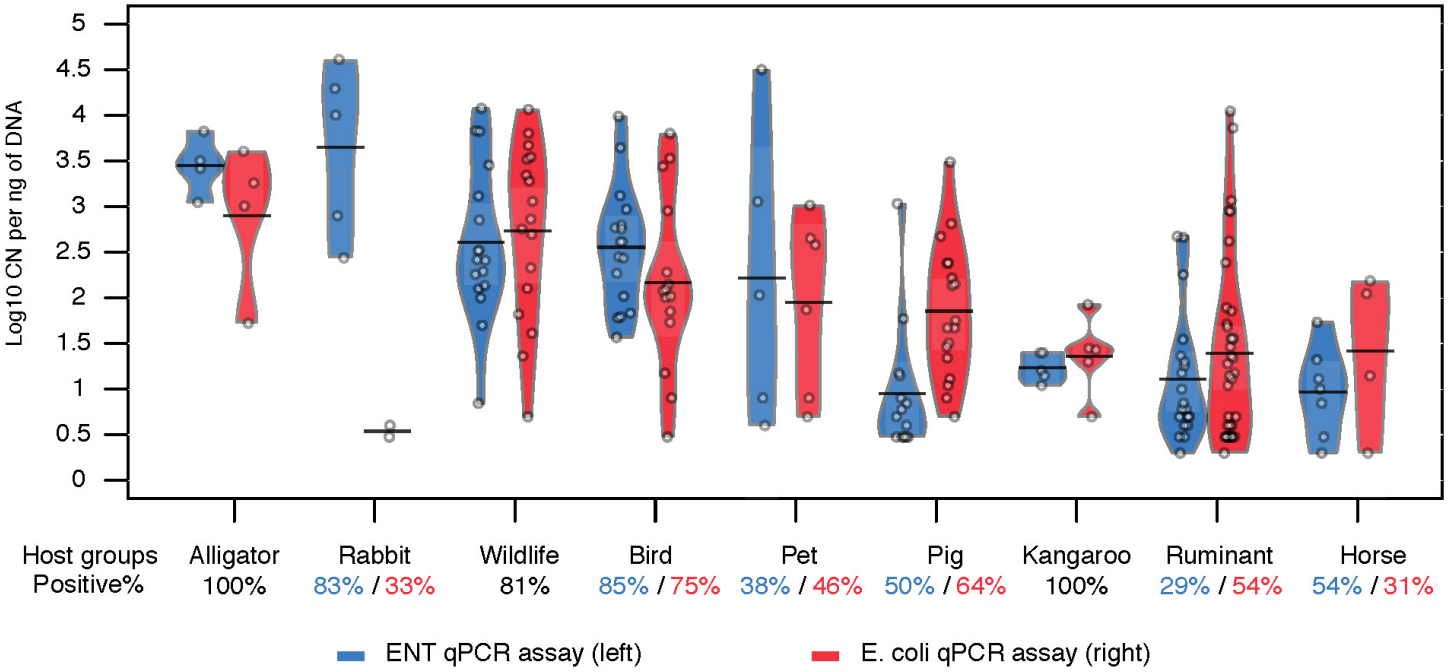
Appendix C. Supplemental Material for Chapter 4



Appendix C Figure 1 Distribution patterns of the 24 most abundant families in all human, sewage and animal fecal samples (n=271). Relative abundance values are normalized to the total sequence counts of the 24 families within each host and increase from light blue color to red color. The X- axis shows groups of sewage, human and animal hosts. The Y-axis shows bacterial families ranked from the most (top) to the least (bottom) abundance. The seven families that all present within the top 20 families of human, sewage and animal groups are in bold font.



Appendix C Figure 2 Bray-Curtis dissimilarity matrix of sewage, human and animal fecal samples. A. Bray-Curtis dissimilarity of family *Lachnospiraceae*, B. Bray-Curtis dissimilarity matrix of genus *Blautia*, C. Bray-Curtis dissimilarity matrix of genus *Bacteroides*. Similarity increases from light blue color to red color. Animal host groups are labeled with different color bars on the top and on the left side with the bar length equals sample numbers.



Appendix C Figure 3 General fecal indicators' qPCR assay positive results at 1 ng · μL^{-1} template level. The Y-axis shows log10-transformed CNs. The X-axis shows animal host groups with ENT results on the left side (blue plot) and *E. coli* results on the right side (red plot). Black line represents mean of copy numbers (CNs). Animal numbers that are positive for each assay are shown in percentages on X-axis below each host group with font color corresponding to the assays. Percentages in black color means both assays show the same percentage (e.g., 100%).

Appendix C Table 1 Standard curve parameters of single and multiplexed assays.

Assay	Standard curve type	Slope	Y-intercept	R^2	Efficiency (%)
E. coli	Multiplexed	-3.441	39.227	1	95.275
	Single	-3.454	39.737	0.999	94.784
HB	Multiplexed	-3.354	36.684	1	98.687
	Single	-3.319	36.739	1	100.122
ENT	Multiplexed	-3.329	38.556	0.998	99.702
	Single	-3.356	38.672	0.999	98.606
BacV6-21	Multiplexed	-3.482	39.155	0.994	93.739
	Single	-3.511	39.303	0.998	92.679

Appendix C Table 2 Standard curve parameters of assays used in Chapter 4.

Assay name	Slope	Y-intercept	R^2	Efficiency (%)	LLOQ
Lachno3	-3.452	38.325	0.999	94.847	34.554
E. coli	-3.376	38.623	0.999	97.814	34.980
HB	-3.347	36.254	0.999	98.999	32.734
ENT	-3.327	38.473	0.999	99.798	34.757
BacV6-21	-3.428	37.698	0.997	95.766	35.093

CURRICULUM VITAE

Shuchen Feng

EDUCATION

University of Wisconsin – Milwaukee, Milwaukee, WI
Ph.D. in Freshwater Sciences
September 2014 – August 2019

Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China
Bachelor of Medicine
Major: Clinical Laboratory Sciences
September 2007 – June 2012

DISSERTATION TITLE

Development of Indicators for Human Fecal Pollution Using Deep-Sequencing of Microbial Communities

MAIN RESEARCH SKILLS

Microbiology and molecular biology	Bacteria culture, phage isolation, nucleic acid extraction, PCR/quantitative PCR, cloning, sequence library preparation
Ecology	Microbial community analysis and indicators identification
Bioinformatics	Sanger sequencing and next-generation sequencing data analysis (raw data processing, clustering analysis, taxonomy assignment and phylogenetic analysis)
Statistical Analysis	Programming in R (large data frame operation, linear and non-linear statistical models)
Language	English and Mandarin

PUBLICATIONS

Feng S, Bootsma M, McLellan SL. 2018. Human-Associated *Lachnospiraceae* Genetic Markers Improve Detection of Fecal Pollution Sources in Urban Waters. Appl Environ Microbiol 84: e00309-18.

Feng S, McLellan SL. 2019. Highly specific sewage-derived *Bacteroides* qPCR assays target sewage polluted waters. Appl Environ Microbiol 85: e02696-18.

Ahmed W, Gyawali P, **Feng S**, McLellan SL. 2019. Host Specificity and Sensitivity of Established and Novel Sewage-Associated Marker Genes in Human and Nonhuman Fecal Samples. Appl Environ Microbiol 85: e00641-19.

PROFESSIONAL ORAL AND POSTER PRESENTATIONS

Feng S, McLellan SL. 2019. Assay development for microbial source tracking by utilizing next-generation sequencing data. American Society of Microbiology Microbe. San Francisco, CA. (Poster)

Feng S, McLellan SL. 2019. Using next-generation sequencing (NGS) and bioinformatics to guide development and validation of new fecal markers. UNC Water Microbiology Conference. Chapel Hill, NC. (Oral)

Feng S, McLellan SL. 2019. Highly Specific Sewage-Derived Bacteroides Quantitative PCR Assays Target Sewage-Polluted Waters. UNC Water Microbiology Conference. Chapel Hill, NC. (Poster)

Feng S, Bootsma MJ, McLellan SL. 2018. Novel human-associated *Lachnospiraceae* genetic markers improve detection of fecal pollution sources in urban waters. UNC Water Microbiology Conference. Chapel Hill, NC. (Poster)

Feng S, McLellan SL. 2016. Candidate Assays for Human Specific Faecal Indicators. American Society of Microbiology Microbe. Boston, MA. (Poster presentation by Dr. Sandra L McLellan)

McLellan SL, Roguet A, **Feng S,** Dila DK. 2019. Molecular approaches for pollution source identification in urban harbors. 3rd Australia New Zealand Marine Biotechnology Society Conference. Australia. (Oral presentation by Dr. Sandra L McLellan)

McLellan SL, Dila DK, **Feng S**, Bootsma MJ. 2017. Climate, Water, and Health: The interface between urban waters and the natural environment. National Water Conference. Milwaukee, WI. (Poster presention by Dr. Sandra L McLellan)

AWARDS

2018-2019 University of Wisconsin - Milwaukee Graduate Student Excellence Fellowship (G.S.E.F.) Award

2019 University of Wisconsin - Milwaukee Graduate Student Travel Award