Evaluation of the Use and Limitations of a Community-Based Microbial Source Tracking Method

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Clairessa Brown

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Michael Sadowsky

July 2018

ACKNOWLEDGMENTS

I would first like to thank my PhD advisor Dr. Michael Sadowsky for providing support and guidance in my professional development. I am thankful for this wonderful opportunity. I would also like to thank Dr. Janet Schottel and Dr. Satoshi Ishii for their valuable suggestions while serving on my committee. I am incredibly grateful to Dr. Daniel Bond and Dr. Michael Smanski for serving on my committee and for all of the many hours of insightful, uplifting, and helpful conversations about science, my career development, and for constantly giving me perspective. I thank you for your investment of energy and time for the betterment of my future.

There are several people who have helped me along the way that I would like to acknowledge. I would like to profusely thank Dr. Chan Lan Chun for helping me in the initial years of my research. Your guidance on life and science was truly invaluable and I thank you for investing in me. I would like to thank Dr. Prince Mathai for the fantastic and often times hilarious hours-long conversations on philosophy, life, and science that helped keep me sane and on track. I will always be grateful of the mentorship you provided. I would also like to thank Dr. Tina Loesekann for always making me laugh ridiculously hard and for her insightful wisdom on everything from science to life. I am also grateful to Julia Beni for her hilarious conversations and unique insights about life, SourceTracker, plotting Dirichlet distributions and how to stay sane in graduate school. I would also like to thank Dr. Chris Staley and Dr. Ping Wang for their

scientific guidance. In addition, I would also like to acknowledge the several undergraduate assistants that helped me along the way.

I would like to thank the 30+ scientist and non-scientist collaborators that made this work possible by helping me collect/donating obscene amounts of animal feces and wastewater treatment plant products.

I dedicate this thesis to some of the people that have shaped who I am. First, I dedicate this thesis to my mom who taught me perseverance and instilled in me a dedication to being the best version of myself. I also dedicate this thesis to my three best friends, Tierney, Julia and Rodelyn, of which without them, I don't know where I would be. I know that I would have not completed this work without all of you supporting me. I can count on all of you to make me laugh and make me feel understood in ways that feed my soul. I love you all.

ABSTRACT

Fecal material from animals and untreated sewage in waterways pose serious risks to human health. Methods that detect and determine sources of aquatic fecal pollution are a part of microbial source tracking (MST). Currently, one of the most widely-used MST methods is susceptible to sensitivity and specificity issues and exhibits variable geographic results which obstruct its widespread use in regulatory capacities. These drawbacks have led to the continued development and evaluation of new MST methodologies. One such new methodology is community-based MST, which uses high-throughput DNA sequencing (HTS) to construct taxonomic profiles of potential fecal sources and environmental samples. SourceTracker, a Bayesian classifier designed to determine sources of contamination in HTS data, is the most common community-based MST tool. Due to the relative novelty of SourceTracker, few studies have evaluated its limitations as a MST tool. This thesis explores the use and limitations of community-based MST with SourceTracker for identifying sources of fecal bacteria in waterways. HTS analysis of microbiota from a diverse collection of fecal samples and environmental samples revealed that the community compositions of freshwater and feces were significantly different, allowing for determination of the presence of fecal inputs. Moreover, the differences in community composition between multiple fecal sources were also statistically significant suggesting that differentiation between fecal sources was possible. When SourceTracker was challenged to identify fecal sources in a freshwater lake with the most diverse and extensive fecal source library used in a

community-based MST field study, SourceTracker was able to identify wastewater effluent from a nearby wastewater treatment plant. SourceTracker also predicted the presence of geese and gull wastes which is in agreement with previous MST studies in that research location. To examine the limitations of community-based MST using SourceTracker, SourceTracker was challenged with identifying known fecal sources in *in situ* mesocosms using different fecal source library configurations. Across nearly all fecal source library configurations, SourceTracker was able to accurately predict most sources in the *in situ* mesocosms. These results were most reliable when the fecal source library contained only the known sources. When fecal sources were missing from the fecal source library, erroneous classifications not seen when all known sources were present became common. Results of this chapter also indicated that the ideal SourceTracker source profile has low-intragroup variability and shares few taxa with other sources. To understand how SourceTracker predictions are influenced by the concentration of feces, geographical variance, diet, and age of source animals, SourceTracker was challenged to identify spiked cow feces from a farm in St. Paul, MN, USA using cow feces from other farms across North America as sources. Analysis of the cow fecal microbiome revealed statistically significant differences in cow fecal taxa at the OTU level when evaluating sample relationships by farm, state, age groups, and diet. Most of the OTU variance was attributed to the individual farms the cows came from, and not to age and diet differences. Source samples from all locations yielded high SourceTracker predictions. On average, higher predictions in source attribution were associated

with more concentrated samples and with cow feces from animals closer to the spiked fecal source location. While SourceTracker was able to detect fecal contamination with source animals that were not from the original location, animals sourced closer to the contamination site provided the most accurate predictions. All of these data demonstrate the ability of SourceTracker to accurately identify sources, making this program a powerful tool for community-based MST.

TABLE OF CONTENTS

ACKNOWLEDGMENTS ..... *i*

ABSTRACT ..... *iii*

TABLE OF CONTENTS ..... *vi*

LIST OF TABLES ..... *viii*

LIST OF FIGURES ..... *ix*

CHAPTER 1 : GENERAL INTRODUCTION *1*

RISKS OF FECAL CONTAMINATION IN WATER TO HUMAN HEALTH 2

FECAL INDICATOR BACTERIA 3

EARLY MICROBIAL SOURCE TRACKING 7

THE IMPACTS OF HIGH-THROUGHPUT DNA SEQUENCING ON MICROBIAL ECOLOGY AND MICROBIAL SOURCE TRACKING 12

COMMUNITY-BASED MICROBIAL SOURCE TRACKING WITH SOURCETRACKER 16

THESIS OBJECTIVES 20

CHAPTER 2 : A HIGH-THROUGHPUT DNA SEQUENCING APPROACH TO DETERMINE SOURCES OF FECAL BACTERIA IN A LAKE SUPERIOR ESTUARY *22*

INTRODUCTION 23

METHODS 25

RESULTS 32

DISCUSSION 42

CHAPTER 3 : EVALUATION OF SOURCETRACKER FOR COMMUNITY-BASED MICROBIAL SOURCE TRACKING IN IN SITU FRESHWATER MESOCOSMS *48*

INTRODUCTION 49

METHODS 51

RESULTS 58

DISCUSSION 69

CHAPTER 4 : INVESTIGATION INTO THE IMPACTS OF AGE, DIET, AND GEOGRAPHY ON SOURCETRACKER PREDICTIONS *74*

INTRODUCTION 75

METHODS 76

RESULTS **79**
DISCUSSION **85**
CHAPTER 5 : CONCLUSIONS AND FUTURE DIRECTIONS ***89***
BIBLIOGRAPHY ***94***

LIST OF TABLES

Table 1.1. Methodology overview of microbial source tracking methods ... 11

Table 1.2. Early community-based microbial source tracking study outcomes that used SourceTracker 19

Table 2.1. GPS coordinates to environmental sites 32

Table 2.2. Diversity indices of all fecal and water samples used in this study 35

Table 2.3. Power analyses of fecal samples to determine appropriate sample amounts for this study 38

Table 2.4. Unique OTUs and taxa used by SourceTracker to assign contamination sources 41

Table 2.5. Relative standard deviation analysis of SourceTracker results after five independent runs of select samples 42

Table 3.1. The shared taxonomic composition between several SourceTracker source profiles 67

Table 3.2. The shared relative abundances of selected source profiles within certain mesocosms 69

Table 4.1. Overview of fecal sample collection 78

LIST OF FIGURES

Figure 2.1. Sampling sites 30

Figure 2.2. Clustering and taxonomic distribution in water and fecal samples at the family level 36

Figure 2.3. Principal coordinate analysis depicting sample relatedness 37

Figure 2.4. SourceTracker results depicting sources of fecal bacteria at seven sites in the Duluth-Superior Harbor and St. Louis River estuary 40

Figure 3.1. Depiction of the shared taxonomic composition and shared relative abundance 58

Figure 3.2. Stacked bar charts analyzing effects of additional sources in the FTL on SourceTracker predictions 60

Figure 3.3. Boxplot depicting intra-group variances within source groups. 61

Figure 3.4. Stacked bar charts analyzing effects of using only the known sources in the FTL on SourceTracker predictions 63

Figure 3.5. Stacked bar charts analyzing effects of missing sources on SourceTracker predictions 65

Figure 4.1. Map of sample locations 79

Figure 4.2. Distribution of relative abundances of the fifteen most abundant family-level taxa 80

Figure 4.3. Averaged taxonomic bar charts of all cow samples 81

Figure 4.4. Influence of age, diet and location on sample relatedness 83

Figure 4.5. SourceTracker analysis of fecal cow spikes from Minnesota ... 84

Figure 4.6. Boxplots of SourceTracker predictions 85

CHAPTER 1 : GENERAL INTRODUCTION

RISKS OF FECAL CONTAMINATION IN WATER TO HUMAN HEALTH

In low-to-middle income nations, over 800,000 deaths are attributed to poor water sanitation and hygiene, accounting for nearly 60% of all diarrheal deaths, which are caused by the ingestion of fecal-associated pathogens (1). In 2012, over 300,000 children under five years of age died from diarrheal diseases attributed to inadequate water sanitation and hygiene (1). An estimated 4% of all deaths can be linked to inadequate water hygiene and sanitation (2). These data demonstrate the importance of determining the presence of feces in water due to the serious health ramifications.

In the United States and Canada, there have been several waterborne outbreaks. In 1993, there was a waterborne outbreak in Wisconsin that caused gastroenteritis in over 400,000 people. This led to 100 deaths and economic and productivity loss estimates of over $90 million (3,4). In 2000, a waterborne outbreak occurred in Walkerton, Ontario, Canada causing gastroenteritis in 2,300 people and killing seven (5). In 2011 to 2012 in the United States, there were 90 reported water-associated outbreaks that resulted in nearly 2000 illnesses, 95 hospitalizations, and 1 death (6). With incidents involving treated water, slightly over half were due to *Crytosporidium* spp. and 33% due to *Escherichia coli* (serotypes 0157:H7 or O111) (6). Avoiding fecal-contaminated water is still an ongoing problem even in highly-developed nations. Human fecal material in untreated sewage poses the greatest health risks due to the presence of human-specific pathogens (7). However, the health risks from animal fecal wastes can

be substantial (8). There are many water-tolerant, fecal-associated pathogens that can cause disease in humans. These pathogens, that can be resistant to a broad range of disinfectants, include protozoa (*Crytospiridium*, *Cyclospora, Entamoeba*, *Giardia duodenalis*), viruses (*Heptatitis A and E, rotavirus, Norwalk and Norwalk-like viruses*), and bacteria (*Clostridium*, *Campylobacter*, *Shigella*, *Salmonella*) (9). These organisms can lead to diseases that range from gastroenteritis to more serious and potentially fatal illnesses that include hemolytic uremic syndrome, severe diarrheal diseases, liver disease, and typhoid fever (9,10).

FECAL INDICATOR BACTERIA

Monitoring all potential pathogens would be difficult due to the number of potential health threats that would require specialized and technically-difficult tests (11). Due to this intractable task, the concept of indicator organism(s) was developed (11,12). These microorganisms are used for detecting the potential presence of pathogenic fecal-associated viruses, protozoa and bacteria. Since bacteria have historically been more readily cultivable and tend to be present at higher environmental concentrations than viruses and protozoa, they are the preferred indicators (9,13). This has led to bacterial indicators of fecal contamination being termed fecal indicator bacteria (FIB). FIB are generally able to be cultivated easily and rapidly (11). To be designated a high-quality FIB, the following criteria (11) must be satisfied:

1. The indicator must be present in high levels in feces

2. The indicator must persist as long as the health risks persists
3. The indicator should not grow outside of the fecal source host species

Some of the first FIB were coliforms, which are a non-taxonomic group of bacteria based mostly on biochemical properties. These bacteria have been found in human and animal feces, sewage, plants, milk, and in soil (14,15). This group includes many *Enterobacteriaceae* members, including *Escherichia* spp., *Klebsiella* spp., *Enterobacter* spp., and *Citrobacter* spp. (16,17). The coliform group includes 80 species across 19 genera which are all defined as gram-negative aerobic and facultative anaerobic, non-spore-forming bacteria that ferment lactose within 48 hours (16).

Several organisms in the coliform group have been identified in extra-intestinal environments, suggesting that many species are not always associated with feces (14,18). These findings led to the designation of the thermotolerant "fecal" coliform group, a subset of the coliform group able to grow at 44.5°C (14). These thermotolerant, fecal coliforms were only supposed to contain bacteria that were identified in feces and therefore would work as more reliable indicators of fecal pollution (14,18). Therefore, the United States Public Health Service started using fecal coliforms to detect fecal pollution in water in the 1960s (19).

Concurrent with the discovery and exploration of coliforms as indicators of fecal pollution, streptococci associated with feces was observed in the early 1900's (12,20). These streptococcus spp. were labeled fecal streptococci and

were found in human and animal feces (20). While these organisms were not found to have naturalized populations in the environment, there were several issues with low recovery numbers in comparison to other FIB, difficulty in quantifying, and a lack of standardized detection methods (21,22).

Unlike other FIB (coliforms and thermotolerant coliforms), *Clostridium perfringens* possesses spore-forming capabilities and is therefore able to persist in the environment longer than most non-spore-formers (23,24). While *C. perfringens* was found to be an indicator capable of predicting the potential presence of viruses and *Cryptosporidium* spp., due to its spore-forming capabilities and its long-term persistence in aquatic sediments, it is generally regarded as not a good indicator of immediate health risks (24,25).

Correlations have been poor with FIB (coliforms, thermotolerant coliforms, *C. perfringens*) and pathogen occurrence (26,27). Since coliforms can be highly susceptible to disinfection, their absence in disinfected water does not correlate with exposure to protozoan (28) or viral (29) pathogens. When coliforms and thermotolerant coliform levels were observed in *in situ* freshwater and saltwater mesocosms spiked with feces (i.e, dog, sewage) and contaminated soil, detection levels of the different FIB varied, being impacted by environmental parameters (saltwater vs. freshwater) (26). When organisms such as *Giardia* cysts and *Cryptosporidium* oocysts were tracked in six waste water treatment plants over the course of a year, none of these organisms had strong correlations with the presence of pathogens (27).

In the late 1970's and through the 1980's, the United States Environmental Protection agency (US EPA) conducted studies that found correlations between the incidences of gastrointestinal illness after swimming and the levels of *E. coli* and *Enterococcus* spp. in fresh and marine waters (19). Therefore, in the late 1980s, the US EPA recommended that *E. coli* and *Enterococcus* spp. be adopted as fecal indicators for waters that had been impacted by feces (19). Both *E. coli* and *Enterococcus* spp. have historically been detected by using culture-based methods partially developed by the US EPA. Within the last few years, however, molecular methods have been developed for *E. coli* and have been approved for *Enterococci* spp. by the US EPA. Both microbes are targeted using a fragment of the 23S rRNA gene in both. These molecular methods allow for more rapid detection and quantification of these fecal indicator bacteria (30,31).

The World Health Organization (WHO) guidelines currently use *E. coli* and thermotolerant coliforms to gauge the safety of water after disinfection (32). However, naturalized populations of *E. coli* and thermotolerant coliforms have been observed to persist outside of the intestinal and fecal environments, suggesting that detection does not necessarily mean fecal contamination (12,32,33). One metareview found that no link existed between diarrhea and thermotolerant coliforms, fecal streptococci, and *E. coli* (34). However, a more recent metareview that looked for links between diarrheal diseases and thermotolerant coliforms and *E. coli* separately found relationships between incidences and *E. coli*, but not for thermotolerant coliforms (35). Additionally,

studies have revealed naturalized *E. coli* populations in beach sand, sediments and soils (36). This data suggests that *E. coli* detection in the environment does not automatically translate to fecal contamination.

EARLY MICROBIAL SOURCE TRACKING

To date, none of the developed FIB methods meet all of the aforementioned required criteria. Additionally, mitigation of fecal pollution in waterbodies demands knowledge of the contaminating sources. The use of FIB failed to reveal sources of fecal bacteria since it targeted organisms that are heavily associated with the feces of most warm-blooded animals. Due to these drawbacks, microbial source tracking (MST) as a science was born.

There have been several MST methodologies developed since the field began (Table 1.1) (37–39). Some of the initial decisions surrounding the development of MST methods involved questions about what would make a high-quality MST marker. The following criteria (38) have been suggested:

1. The marker ideally should be host-specific
2. The marker should be present in all members of the host species with consistent taxonomic levels
3. The detectability and taxonomic levels of the host-specific marker should not change due to geographic or temporal differences
4. The marker ideally should be easily and rapidly quantifiable

5. Ideally, the marker should correlate to all fecal-associated health risks and pathogens and not persist longer than the threat is present

Some of the early MST techniques were library-dependent, phenotypic methods that involved creating characteristic trait libraries (Table 1.1). These methods distinguished between fecal sources based on differences in biochemical or phenotypic profiles. One of the first MST methods to be developed was the carbon source utilization assay where carbon source profiles, or growth patterns, of *Enterococccus* isolates were determined and environmental samples compared against this library (40). While this method is rapid, it suffered from poorer performance relative to other MST methods (40). Other phenotypic methods, such as antibiotic resistance analysis, also performed relatively poorly in comparison to other MST methods (41). Moreover, while these methods could be rapid, easy to perform and cheap, they were demonstrated to be either geographically-specific, lacked source specificity, had many false positives or still lacked extensive limitations testing (Table 1.1).

Ribotyping, which involves hybridizing digested genomic DNA to rRNA probes to differentiate species, was also one of the first molecular MST methods to be used (42). While it generally performs well in comparison to other MST methods, it can be time-consuming with inconclusive results (38). Rep-PCR, a method that uses gel electrophoresis to distinguish species by their palindromic DNA, requires large libraries that are laborious and time-consuming to build (38). Pulse-field gel electrophoresis, another one of the first developed MST methods,

involves analyzing digested genomic DNA with gel electrophoresis (43). While this method performs well when compared to other MST methods with highly discriminatory and conclusive results, it is time-consuming to perform and requires finesse (11,44).

Early library-dependent MST methods consistently required assembling large libraries of bacteria from known animals that were laborious and time-consuming to build. These techniques also suffered from being geographically-specific and therefore these large developed libraries could not be shared between multiple locations or investigators. Additionally, the temporal stability of libraries, which is how effective a library is in allowing the prediction of fecal contamination over time, was either unexplored or found to be poor, hindering the use of these methods for regulatory purposes.

Due to the drawbacks of library-dependent techniques, the field of MST explored using methods that didn't require large databases to build and maintain. These methods generally distinguished fecal sources with libraries based on genetic differences (45). These included library-independent genotypic methods such as host-specific PCR and quantitative PCR (qPCR) (Table 1.1). Host-specific PCR methods targeted different DNA fragments (16S rRNA, toxin genes) in a variety of bacteria (*Bacteroidales*, *Bifodobacterium*, *E. coli*) (38). While a few of the host-specific PCR markers have had superior source discrimination ability when compared to other MST methods, quantification of fecal contamination was not possible (38,41). Additionally, there have been issues with several markers not detecting contamination when it is present (sensitivity), and/or markers that

would detect unintended fecal source groups (specificity) (38). In order to allow semi-quantification of contamination, newer MST methods focused on the use of qPCR, making it currently one of the most widely-used and developed MST techniques. These methods involve quantifying amplified DNA fragments, with most focused on *Bacteroidales* given its prevalence in host microbiomes and its suspected host-specific co-evolution (38). However, the drawbacks to MST with qPCR are similar to the sensitivity and specificity issues of MST with PCR.

Genotypic, library-independent MST methods are culture-independent, rapid and easy to perform. Additionally, some of these methods have been shown to have geographic stability (46). However, there have been many issues with sensitivity and specificity of several markers halting their widespread use in regulatory capacities (47).

Table 1.1. Methodology overview of microbial source tracking methods

Method Classification	Method Examples	Library Type	Advantages	Disadvantages
Phenotypic	Antibiotic resistance analysis (48), carbon source utilization (40), fatty acid methyl esters profiling (49)	Library-dependent	Some methods are rapid, easy, cheap, small libraries needed, highly discriminatory	These methods are either geographically variable or have unknown geographic specificity
Genotypic	Pulse-field gel electrophoresis (50), ribotyping (51), rep-PCR (52)	Library-dependent	Some methods are easy, rapid, and highly reproducible	Some of these methods are complex and time-consuming, large library needed, geographically specific
Genotypic	Host-specific PCR (53), qPCR (54)	Library-independent	Culture-independent, rapid and semi-quantitative, geographically stable	There are documented sensitivity and specificity issues with several molecular targets, some methods aren't able to be quantified
HTS Machine learning	Oligotyping (55), SourceTracker (56)	Library-dependent	Culture-independent, simultaneous community profiling with several sources	Still in early stages of limitations testing, absolute quantification of contamination not possible

HTS is an abbreviation for High-Throughput DNA Sequencing.

THE IMPACTS OF HIGH-THROUGHPUT DNA SEQUENCING ON MICROBIAL ECOLOGY AND MICROBIAL SOURCE TRACKING

Since many bacteria in various environments are unable to be cultured, it is now common to find the use of high-throughput DNA sequencing (HTS) in microbial ecology studies (57). Currently, community analysis studies use massively parallel DNA sequencing of clonally amplified phylogenetic markers, like 16S rRNA genes, to determine the microbial community in an environment. Several environments, from the human body (58) to the phyllosphere (59), have been explored using community analysis techniques with HTS.

A proposal in 1977 calling for phylogenetic classification of all microbial life based on ribosomal RNA genes was a monumental step in allowing for later microbial community analyses in natural systems (60). Months later, an improved sequencing technique (the Sanger sequencing method) was described, allowing for the sequence elucidation of genes (61). Nearly 20 years later, the first microbial community analysis study was done using 16S rRNA, profiled in picoplankton from the Sargasso Sea (62).

The use of the 16S rRNA gene as a molecular chronometer was due to its widespread presence in bacterial (and archaeal) genomes, its resistance to mutations, and its ability to elucidate phylogenetic relatedness (63). The 16S rRNA gene contains regions of conserved DNA sequences where primers for amplification-based molecular techniques can be targeted to (64,65). Inter-spaced in these conserved DNA regions are nine hyper-variable regions, labeled V1-V9, where taxa that share close phylogenetic relationships can have similar

nucleotide sequences (63,66). Different hypervariable regions exhibit varying abilities to distinguish between bacterial taxa, with none being able to distinguish all. There can be drawbacks to using all of the different 16S rRNA variable regions for taxonomic assignment. In a study evaluating which hypervariable regions allowed for the phylogenetic differentiation among pathogenic bacteria, the V2, V3, and V6 regions were found to be most suitable due to their high sequence variability and strong discriminatory power (67).

Early community analysis studies were enhanced by the creation of computational tools like DOTUR (68), the Ribosomal Database Project (RDP) (69), and the NAST alignment algorithm (70). As more community analysis studies for microbial ecology were done, computational approaches to processing these data greatly improved. Commonly used bioinformatic programs that allow quality-control processing of sequence reads today include QIIME (71) and Mothur (72), of which both utilize a number of tools used to align sequences (PyNAST (73)), cluster them into taxonomic units based on sequence similarity (uclust (74)), and ultimately assign sequence reads to taxonomic groups (RDP Classifier (75)).

One of the first realizations upon profiling bacterial communities in feces for early MST methods was that there was a lack of knowledge about the distribution of bacterial markers used in early MST methods in animal hosts of the same species and in varying animal species. This was important because historically, these methods used markers with single indicator organisms or

narrowly-targeted taxonomic groups. The use of single indicator organisms is often based on the presence or absence of detection.

Some of the first community analysis studies with a focus on MST were performed on the feces of animals and sewage, allowing for an understanding of the taxonomic composition in these environments. In a recent HTS study examining the microbial composition of cow feces, the authors used pyro-sequencing on fecal samples from 20 adult cattle from the same herd in the southwestern United States (76). *Bacteroidales* members, including *Bacteroides* spp., *Alistipes* spp., *Prevotella* spp., and unclassified *Bacteroidales* spp., were among the most abundant taxa across the samples (76). *Bacteroides* spp., one of the most abundant genera, ranged from 5-14% of the sequenced bacterial community (76). In qPCR MST methods, *Bacteroides* spp. are some of the most widely-targeted FIB (38). *Clostridium* spp., which includes *C. perfringens*, itself a FIB, were also a substantial makeup of the microbial community (76). *Escherichia* spp. however ranged from less than 1% to approximately 3% (76). More importantly, however, *Escherichia* spp. were not detected in a little over 10% of the study sample size, reflecting an overestimation of the importance placed on cultivable FIB in detecting fecal contamination (76).

Another one of the studies seeking to understand fecal microbiomes specifically for MST was work that profiled the bacterial community of sewage from two wastewater treatment plants in Wisconsin, USA (77). This study looked for a multi-taxa signature that could be used to identify human fecal-associated pollution that had been introduced into freshwater by way of sewage

contamination (77). The authors first profiled the bacterial communities of sewage, freshwater and human feces looking for shared taxa between them (77). While sewage shared several taxa with both human feces and freshwater, there were very few taxa shared between human feces and freshwater (77). The taxa shared between sewage and human feces became the human fecal signature within sewage. The human fecal signature within sewage possessed several members of the *Bacteroidetes* and *Firmicutes* phyla (77).

Unno et. al was one of the first to look for shared taxa between different fecal sources and environmental samples with HTS data (78). This new library-based MST method required bacterial source profiles to be built from 16S rRNA gene HTS data using animal feces and environmental freshwater samples to determine fecal pollution sources in a South Korean river (78). Thirty fecal swabs each were collected from chickens, ducks, swine, beef and dairy cattle from 3 farms in the Jeonnam Province, South Korea (78). Fecal samples from 30 healthy adult humans and wild geese were also collected (78). Since the Yeongsan River runs through land that has multiple uses (agricultural, urban), freshwater samples were taken at multiple sites (78). In this study, major taxonomic differences between animal and human feces and freshwater were detected (78). *Firmicutes* and *Bacteroidetes* dominated the taxa in feces in this study as well as the study in Wisconsin (77,78). To determine the source of fecal bacteria, this study used an algorithm in Mothur to look for taxa shared between the environmental samples and different fecal sources (78). Using this bacterial

community-based MST method, they concluded that human, swine and geese were the main source contributors (78).

COMMUNITY-BASED MICROBIAL SOURCE TRACKING WITH SOURCETRACKER

Machine learning involves using statistical techniques enabling software applications to predict outcomes without explicit programming. One of the first machine learning MST methods using HTS data was the program SourceTracker (56). SourceTracker is a Bayesian classifier designed to determine sources of contamination in HTS samples (56). This program requires source samples that compose the source library and sink samples where it will look for shared taxa from the source library. SourceTracker uses a mixture-modeling approach, assuming that all taxa in a sink sample can be assigned to a source in the source library (56). For taxa that cannot be assigned to any of the user-provided sources, SourceTracker creates an “unknown” source designation (56). Using a Markov Chain Monte Carlo method, SourceTracker predicts the probabilities that certain taxa come from specific sources (56). SourceTracker is then able to assign taxa to specific sources when recognized in contaminated samples (56). From this, it is then able to quantify the relative contribution each source contributes in a contaminated sample (56). SourceTracker also provides the taxonomic composition of the source profiles that were constructed to identify specific sources in each of the contaminated samples.

Several studies have used SourceTracker to detect different kinds of sources. SourceTracker has been used to identify the biogeographical patterns of bacteria in public restrooms (79), to characterize ancient oral microbiota (80), and identify contamination in DNA databases (81). When utilized for MST, SourceTracker can be used to estimate the proportion of fecal contamination that sources contribute to in an environment.

SourceTracker had been used for community-based MST for only a few studies before and during the beginning of this thesis (Table 1.2). The earliest community-based MST studies with SourceTracker used it to characterize the composition of sewage (Table 1.2) (82,83). In these studies, sewage was used as a sink sample while human and non-human feces were evaluated as source samples. This research utilized the fecal signature approaches of community-based MST studies developed before SourceTracker use (82,83).

Other SourceTracker-based studies used “toolbox” approaches where multiple MST methods are utilized to determine fecal pollution sources (Table 1.2). These studies found varying degrees of successful corroboration that were dependent on the methods being compared (84,85). When molecular-based methods, like qPCR, were used, SourceTracker results had varying levels of agreement. This is most likely due to how HTS methods can use different primers to amplify and profile microbial communities than qPCR markers. As mentioned earlier, qPCR markers can suffer from sensitivity, specificity and geographic-specificity issues which could lead to SourceTracker detecting sources qPCR markers could fail to detect. Additionally, specificity, sensitivity

and geographic-specificity in SourceTracker fecal libraries had not been explored extensively prior to the start of this thesis.

Table 1.2. Early community-based microbial source tracking study outcomes that used SourceTracker

Study	Main Objective Related to SourceTracker	Major Finding(s)	Significance
Newton et. al, 2013	Used SourceTracker to identify human fecal taxa signature in sewage samples.	In samples obtained during sewer overflows, there were statistically significant increases in the SourceTracker-identified human fecal taxa signature.	SourceTracker results were logical and consistent with metadata suggesting it may be an accurate tool to detect taxonomic signatures from source samples.
Shanks et. al, 2013	Characterized the source profiles of sewage from 13 different locations across the United States.	The shared taxa between sewage and human feces were consistently found across all samples regardless of location.	SourceTracker results suggests that across the United States, a core fecal community exists in sewage.
Neave et. al, 2014	Using multiple MST methods, determine fecal contamination source(s) at several environmental sites.	SourceTracker findings had variable agreement with other MST methods. When SourceTracker results didn't agree, results were consistent with metadata.	While there was some agreement with other MST methods and metadata, comparisons of SourceTracker with other molecular MST methods can be complicated.
Ahmed et. al, 2015	Using qPCR and SourceTracker, determine sources of fecal bacteria at six sites with varying land use.	SourceTracker and qPCR results were inconsistent about the presence of fecal sources.	While there was some agreement between SourceTracker and qPCR, results comparisons with molecular MST methods can be complicated.

MST is an abbreviation for microbial source tracking.

THESIS OBJECTIVES

This thesis explores the use and limitations of community-based MST using SourceTracker for identifying sources of fecal bacteria in waterways. Due to the relative novelty of SourceTracker, few studies had evaluated its limitations as a MST tool before the beginning of this thesis.

One unexplored aspect of SourceTracker was its use in determining sources of fecal pollution in the coastal areas of Lake Superior in Duluth, Minnesota. Chapter 2 involved the creation of the most diverse and extensive fecal source library used in a SourceTracker field study. SourceTracker was then challenged to identify fecal pollution sources in several coastal locations around this freshwater lake.

Another relatively unexplored, but potential limitation of SourceTracker is understanding how SourceTracker predictions change when the composition of the fecal source library changes. In Chapter 3, the ability of SourceTracker to correctly determine sources of known fecal contamination with *in situ* mesocosms was investigated when different combinations of fecal source library configurations were used. Additionally, the structure and composition of the source profiles SourceTracker used to discern sources was explored.

While early library-dependent MST methods suffered from geographic specificity, SourceTracker was hypothesized to be able to overcome this with its multi-taxa approach in identifying fecal contamination. Additionally, it is unknown how diet, age, and the location of source animals, as well as the concentration of fecal contamination, potentially impact SourceTracker's ability to accurately

predict fecal sources. These questions are explored in Chapter 4 by challenging SourceTracker to identify different dilutions of spiked cow samples in buffer by using different cow populations across North America as the sources in the fecal source library.

CHAPTER 2 : A HIGH-THROUGHPUT DNA SEQUENCING APPROACH TO DETERMINE SOURCES OF FECAL BACTERIA IN A LAKE SUPERIOR ESTUARY

INTRODUCTION

Microbial source tracking methods (MST) rely on the use of specific microorganisms or bacterial profiles, thought to be host-associated (e.g. human, ruminant, bird), to determine sources of fecal bacteria in the environment (86). The intestinal tracts of different animals harbor distinct bacterial communities that vary in the presence and abundance of specific taxa (78). This allows for the determination of different fecal sources by using unique bacterial profiles seen within different animals' fecal microbiomes.

Currently, the most common MST methods involve the use of quantitative PCR (qPCR) and molecular markers (primers) that mainly target the 16S rRNA genes of presumptively host-associated microorganisms. The majority of these markers target members of the genus *Bacteroides*. However, developing qPCR markers that are both sensitive and specific has proven to be challenging in several cases (87–90).

The introduction of high-throughput DNA sequencing (HTS) technology has allowed increased use of culture-independent methods to rapidly assess bacterial community structure in several different environments (78,84,91–93). Community-based MST methods involve the creation of a library of operational taxonomic units (OTUs) present in feces from different animal types, featuring source-specific microbial profiles. These libraries, referred to as fecal taxon libraries (FTL), are compared with bacterial community profiles from environmental samples to find shared taxa or OTUs between the two sample types.

Several community-based MST studies (78,84,91–93). have utilized the SourceTracker program (94) which accepts quality-controlled sequence data, along with a list of samples either identified as a potential fecal source or an environmental sink. SourceTracker uses a Bayesian statistical approach to determine the percentage of the bacterial community, and the probability, that a potential fecal source contributed to an environment (94).

While SourceTracker results include the standard deviation for an estimation of error, higher confidence in the Bayesian model used to predict sources can be achieved by running SourceTracker multiple times, in a bootstrap-like manner, and measuring the uncertainty of the model using the relative standard deviation (RSD) (95). An important consideration with community-based MST methods is determining the appropriate library size of samples for each source. This can be established by using statistical power analysis, which determines whether the sample size is large enough to be able to detect significant differences between sample groups. The Dirichlet multinomial distribution can be used to model data during power analysis, such that Type II error due to overdispersion is alleviated (96). This approach has been successfully used to model microbial community analysis data (97).

Lake Superior, the largest freshwater lake in the world by area, is used for recreational and commercial purposes. Fecal bacterial inputs onto beaches in the Lake Superior harbor area, Park Point, and the St. Louis River estuary, measured by *E. coli* concentrations, have been found to be seasonally impacted by waterfowl (36). Results obtained using qPCR in previous research have also

indicated that the largest point source contributor of fecal-associated bacteria to the Lake Superior-Duluth Harbor was treated wastewater effluent, likely originating from two local wastewater treatment plants (WWTP) (98,99).

In this study, we examined potential sources of fecal bacterial inputs into the Duluth-Superior Harbor and St. Louis River in the Lake Superior watershed employing a community-based MST approach using Illumina DNA sequencing data and SourceTracker analyses. To achieve this goal, water samples were collected over two years, from seven different sites in the Duluth-Superior Harbor and St. Louis River estuary in Duluth, MN, USA along with fecal samples from 11 different types of animal sources and treated wastewater effluent. Samples were sequenced and high-quality HTS data from animal and effluent samples were used to create a FTL for community-based MST. Power analyses, done using a Dirichlet multinomial distribution to model 16S community data, were used to determine whether an appropriate library size had been obtained for the entire FTL. Multiple SourceTracker runs allowed fecal sources to be identified, along with RSD values which allowed for an estimation of confidence to be assigned to the predicted proportions of fecal sources. Results of this study show that a community-based MST approach, with an appropriately-sized FTL, could potentially be used as a tool to allow watershed managers to confidently predict sources of fecal inputs into waterways.

METHODS

Sample collection. Triplicate, two-liter water samples were collected just below the surface from seven sites in the Lake Superior-St. Louis estuary during

summer and fall 2014 and the summer of 2015 (Fig. 2.1). These sites included the: St. Louis River site 1 in 2014 and St. Louis River site 2 in 2015, Western Lake Superior Sanitary District (WLSSD) outfall, Rice's Point, Brewery Creek storm drain, Southworth Marsh storm drain, and Minnesota Point Beach sites in both years (Table 2.1). The St. Louis River site in 2014 changed due to accessibility. Water was sampled on July 23, August 20, and November 5 in 2014. In 2015, water was sampled on June 16 and July 27. All water samples were transported on ice to the lab and stored at 4°C for less than 24 hours prior to filtration.

Fecal samples were collected from 11 different animal types from a variety of sources including farms, via wildlife managers, and from personal pet donations across Minnesota. Several grams of feces were collected from individual chickens, cows, production turkeys, swine, beavers, gulls, Canada geese, wild and domesticated rabbits, deer, cats, and dogs. Triplicate, two-liter samples of treated wastewater effluent were collected from WLSSD during every sampling event.

Sample processing and DNA extraction. Fecal samples were transported at 4°C (on ice) and stored at -20°C until DNA could be extracted. All triplicate 2L environmental water samples were pre-filtered through 5µm nitrocellulose filters (Millipore-Sigma, St. Louis, MO) to remove debris and larger microorganisms and processed as previously described (100). The environmental water was subsequently filtered through 0.45µm, and then 0.22µm, nitrocellulose filters (Millipore-Sigma, St. Louis, MO) to capture all

bacteria. Filters were transferred into 50ml conical tubes containing 2ml of 0.01% sodium pyrophosphate buffer, pH 7.0 containing 0.2% Tween 20 (polyethylene glycol sorbitan monolaurate) and vortexed, twice, for 3 min at room temperature. Each conical tube held up to five filters from the same environmental site, and some of the 0.45μm and 0.22μm filters were placed in the same tube if space allowed. The supernatant, containing re-suspended cells from the filters, were transferred from the conical tubes to 1.5ml Eppendorf tubes and centrifuged for 3 min at 13,300×g. Cell pellets from 0.45μm and 0.22μm filters from the same sample were ultimately combined if the filters from the same sample were placed in different 50ml conical tubes. Samples were stored at -20°C until DNA was extracted. DNA from fecal and environmental water samples was extracted using the DNeasy PowerSoil DNA extraction kit (Qiagen, CA, USA), as per kit directions.

PCR and DNA sequencing. PCR and DNA sequencing were performed at the University of Minnesota Genomics Center (St. Paul, MN, USA) using primers F784 (5' RGGATTAGATACCC 3') (101) and 1046R (5' CGACRRCCATGCANCACCT 3') (102), targeting the V5 and V6 regions of the 16S rRNA gene. Sequencing was done using the dual indexing method as previously described (103). Amplicons were paired-end sequenced on the Illumina HiSeq 2000, HiSeq 2500 (150bp), and MiSeq (300bp) platforms (Illumina, San Diego, CA). Sequencing results can be accessed from GenBank under BioProject PRJNA377760.

Processing of sequence data. All sequencing data obtained from the Illumina MiSeq platform runs were trimmed to 150 bp to match the run length obtained from Illumina HiSeq runs. All sequence processing was performed using QIIME versions 1.8.0 and 1.9.1 software (73). Illumina adapter contamination and low quality base regions were removed using Trimmomatic v. 3.2 (104). Primers, homopolymers >8, and reads smaller than 75% of the amplicon length were removed using Pandaseq (105). This program was also used to concatenate reads using the fastq-join script.(106) Chimeras were removed using UCHIME 6.1 (74). Open reference OTUs were grouped using uclust, at 97% identity, and compared to the SILVA ver. 119 reference database (107,108) using the PyNAST alignment algorithm (73). Taxonomy was assigned utilizing the RDP Classifier, with an 80% bootstrap value (69). Singletons were removed from the dataset and all data that passed quality control were used for statistical analysis. The final dataset was comprised of DNA sequences from 20 beavers, 14 cats, 16 chickens, 32 cows, 19 deer, 17 dogs, 25 geese, 14 gulls, 18 wild and domesticated rabbits (treated as a single source), 18 swine, 18 turkeys, and 22 effluent samples.

Statistical analyses. Statistical analyses were done using QIIME v. 1.8.0 (73), RStudio v. 0.99.896, R v. 3.2.1 (109) and mothur v. 1.34.0 (72). After sequence processing, multiple rarefaction depths were evaluated by random sampling of sequences so that the number of observed taxa were close to the number of expected taxa. A final sequence depth of 25,000 was chosen for all subsequent statistical analysis. Bray Curtis dissimilarity (110) was calculated in

mothur and was used in principal coordinates analysis (PCoA) and analysis of molecular variance (AMOVA) (111). Hierarchical clustering was calculated in the R package pvclust in RStudio (112) using the UPGMA method. Clustering was performed on a Bray Curtis dissimilarity matrix with 1,000 bootstrap iterations containing the averages of family-level taxa abundances from all sample types. Community-based MST was done using SourceTracker v. 1.0 (94) through QIIME v. 1.9.0 (73). While family-level taxa tables rarefied to a sequencing depth of 25,000 were used, SourceTracker was run with default parameters five independent times on the same FTL. Spearman rank correlations were performed in RStudio to determine the correlation between predicted source proportions obtained via SourceTracker and RSD values determined from the five independent runs of the SourceTracker program.

RSD analysis was performed to estimate confidence in SourceTracker proportions and was calculated by using the average standard deviations of a sample across the five independent SourceTracker runs. This value was divided by the average predicted source proportions of that same sample obtained from the five independent SourceTracker runs.

Power analyses were performed on all animal fecal samples to determine whether sufficient samples from each source type were present in the library to avoid increased statistical Type II error. Library size analysis was done by reducing the number of individuals within each source type (ranging from 14 to 32) and then running power analysis using the R HMP package to determine a

minimum size of the library (96). If the source type did not have as many samples as required, all samples were used.

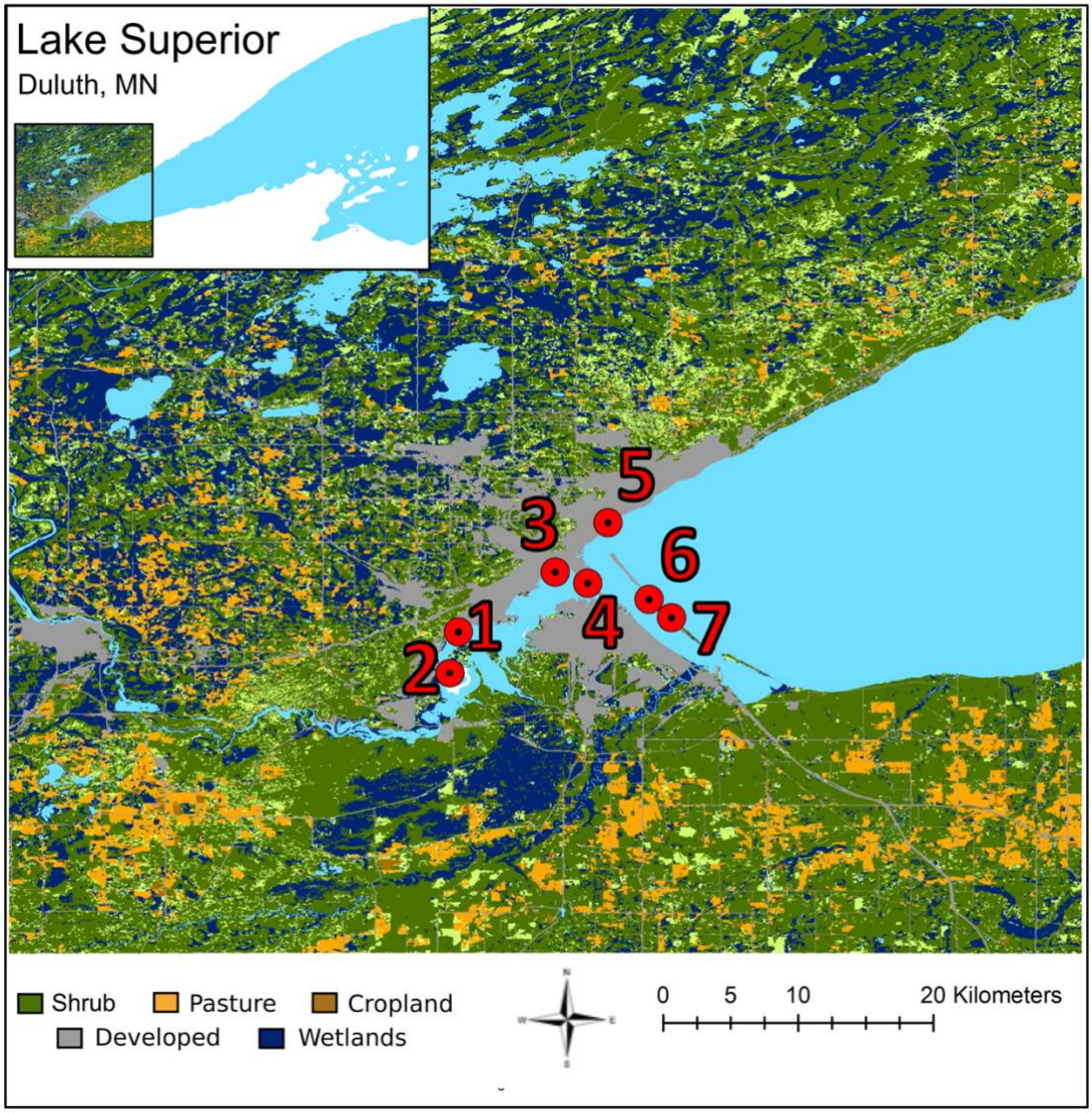

Figure 2.1. Sampling sites

Map showing locations of the seven sites that were sampled in the Lake Superior watershed in 2014 and 2015. The sites are as follows: 1) St. Louis River site 2, 2) St. Louis River site 1, 3) WLSSD treated effluent outfall, 4) Rice's Point, 5) Brewery Creek storm drain, 6) Southworth Marsh storm drain, and 7) Minnesota Point Beach.

Table 2.1. GPS coordinates to environmental sites

Site	GPS Coordinates
St. Louis River site 1	46.719486, 92.189544
St. Louis River site 2	46.700478, 92.207532
Western Lake Superior Sanitary District outfall (WLSSD)	46.758008, 92.120118
Rice's Point	46.751847, 92.103359
Brewery Creek storm drain	46.792084, 92.089915
Southworth Marsh storm drain	46.741114, 92.062519
Minnesota Point Beach	46.728946, 92.047469

RESULTS

Bacterial community structure among water and fecal samples.

Water samples from seven different locations in the Duluth-Superior Harbor and St. Louis River estuary in Duluth, MN and animal fecal samples were collected during 2014 and 2015. The V5 and V6 regions of the 16S rRNA gene were sequenced from 319 environmental lake water, fecal, and treated wastewater effluent samples yielding nearly 60 million reads. The 233 fecal samples were obtained from 12 different domesticated, agricultural, and wild animal types, as well as wastewater effluent, and together the source sequences comprised the FTL used for community-based MST.

The average sequencing coverage for water and feces was 99% and ranged from 97% to 100% (Table 2.2). Feces, on average, had lower diversity

than did lake water (Table 2.2). Feces had, on average, a Shannon index of 4.2 and 1,672 OTUs that clustered at 97% similarity, while freshwater samples had an average Shannon index of 4.8 and 3,272 OTUs clustering at 97% similarity (Table 2.2).

Hierarchical clustering showed that the environmental water and animal samples clustered separately, indicating the microbiota in feces is different from that found in water (Fig. 2.2A). The most common microbiota found in environmental water communities included members of the families *Sporichthyaceae, LD12, and Comamonadaceae* (Fig. 2.2B). In contrast, members of the families *Erysipelotrichaceae*, *Peptostreptococcaceae*, and *Ruminococcaceae* were prevalent in the treated wastewater effluent samples, and animal fecal samples were dominated by members of the families *Ruminococcaceae, Lachnospiraceae*, and *Peptostreptococcaceae* (Fig. 2.2B). Nearly 30% of the sequencing reads from rabbits, swine, deer, cows, and beavers were classified as *Ruminococcaceae* and *Lachnospiraceae*. Cats, dogs, and chickens had nearly equal amounts of those families, as well as several other families, which explains why their samples clustered together (Fig. 2.2A). While geese had nearly equal amounts of *Ruminococcaceae* and *Lachnospiraceae,* which was similar to cats, dogs, and chickens, several of the families that make up the community structure were relatively different (Fig. 2.2A).

Similar to what was found with hierarchical clustering, PCoA revealed that fecal and environmental water samples clustered separately (Fig. 2.3). AMOVA

pairwise-comparisons identified significantly different (p <0.05) bacterial community structures among animal feces, effluent samples and freshwater sites in the Lake Superior-St. Louis River estuary. Fecal samples from individuals in the same source type clustered together (Fig. 2.3). Moreover, AMOVA analyses found that the bacterial community structures were significantly different between animal types.

Table 2.2. Diversity indices of all fecal and water samples used in this study

Sample Type	*n*	Site/Group	Avg. Coverage (%)	Avg. Observed OTUs	Shannon Index
Water	9	St. Louis River 1	100 ± 0	1402 ± 488	3.65 ± 0.7
Water	6	St. Louis River 2	99 ± 0	1914 ± 1067	3.97 ± 1
Water	15	WLSSD Outfall	99 ± 0	2069 ± 1423	4.02 ± 1.2
Water	12	Southworth Marsh	99 ± 0	3084 ± 2895	4.49 ± 1.6
Water	14	MPB	98 ± 0	4281 ± 2976	5.32 ± 1.8
Water	14	Rice's Point	99 ± 0	5297 ± 3212	6.01 ± 1.5
Water	15	Brewery Creek	98 ± 0	4861 ± 3022	6.13 ± 1.5
mean				3272	4.80
Fecal	22	Effluent	99 ± 0.1	2809 ± 797	4.40 ± 0.8
Fecal	20	Beavers	100 ± 0.1	1256 ± 334	4.49 ± 1.0
Fecal	12	Beef Cows	99 ± 0.1	2243 ± 453	5.57 ± 0.2
Fecal	20	Dairy Cows	99 ± 0.1	2920 ± 600	5.85 ± 0.3
Fecal	14	Cats	100 ± 0.1	694 ± 156	3.45 ± 1.0
Fecal	16	Chickens	97 ± 0.1	1251 ± 827	3.39 ± 0.8
Fecal	19	Deer	97 ± 0.1	1748 ± 1223	5.16 ± 0.7
Fecal	25	Geese	100 ± 0	1395 ± 1121	3.61 ± 0.7
Fecal	14	Gulls	99 ± 0	1500 ± 947	3.24 ± 0.8
Fecal	17	Dogs	100 ± 0	624 ± 974	2.35 ± 0.9
Fecal	18	Rabbits	100 ± 0	1330 ± 683	4.83 ± 0.9
Fecal	18	Swine	100 ± 0	2259 ± 767	5.12 ± 0.9
Fecal	18	Turkeys	100 ± 0	536 ± 871	1.99 ± 0.9
mean				1672	4.20

MPB = Minnesota Point Beach.

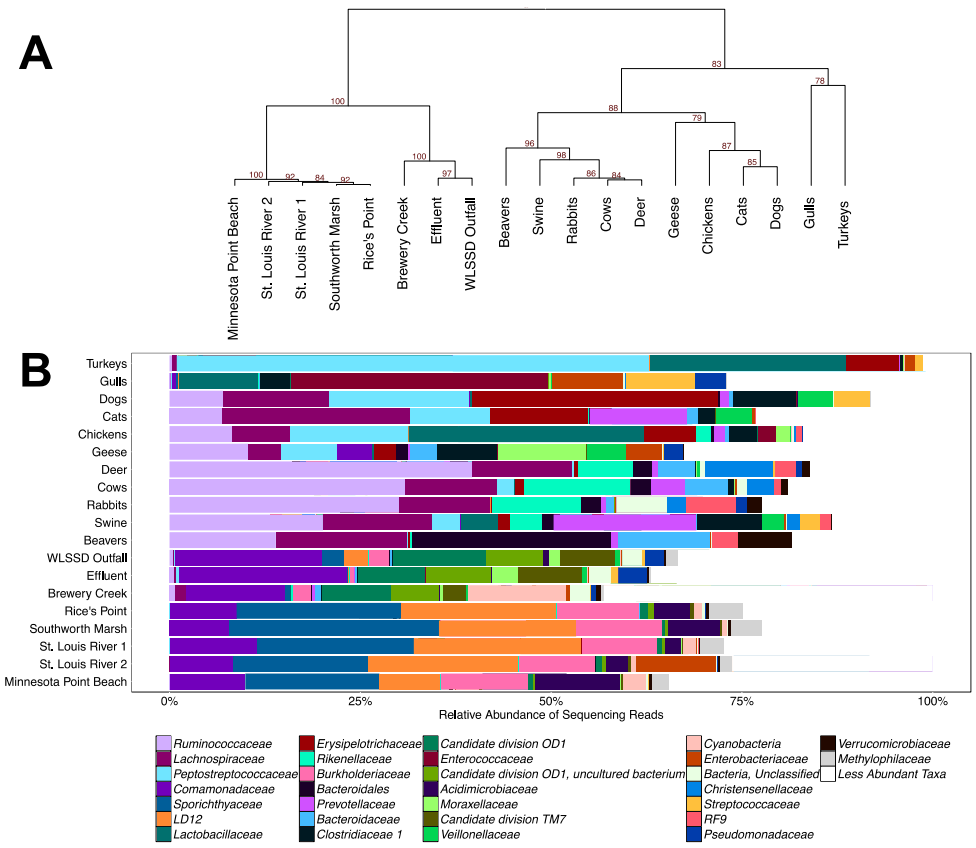

Figure 2.2. Clustering and taxonomic distribution in water and fecal samples at the family level

(A) Hierarchical clustering using the UPGMA method performed on a Bray Curtis distance matrix containing averages of all sample types. Red values at nodes are AU p-values. (B) Stacked taxonomic bar charts depicting the average relative abundances of the thirty most abundant taxa. Hierarchical clustering (Fig. 2A) determined the order of samples for Fig. 2B. The number of animals used in both analyses are as follows: beavers (n=20), cats (n=14), chickens (n=16), cattle (n=32), deer (n=19), dogs (n=17), treated effluent (n=22), geese (n=25), gulls (n=14), rabbits (n=18), swine (n=18), and turkeys (n=18). The number of water samples per site are as follows: St. Louis River 2 (n=6), St. Louis (n=9), WLSSD

outfall (n=15), Rice's Point (n=14), Brewery Creek (n=15), Southworth Marsh (n=12), Minnesota Point Beach (n=14).

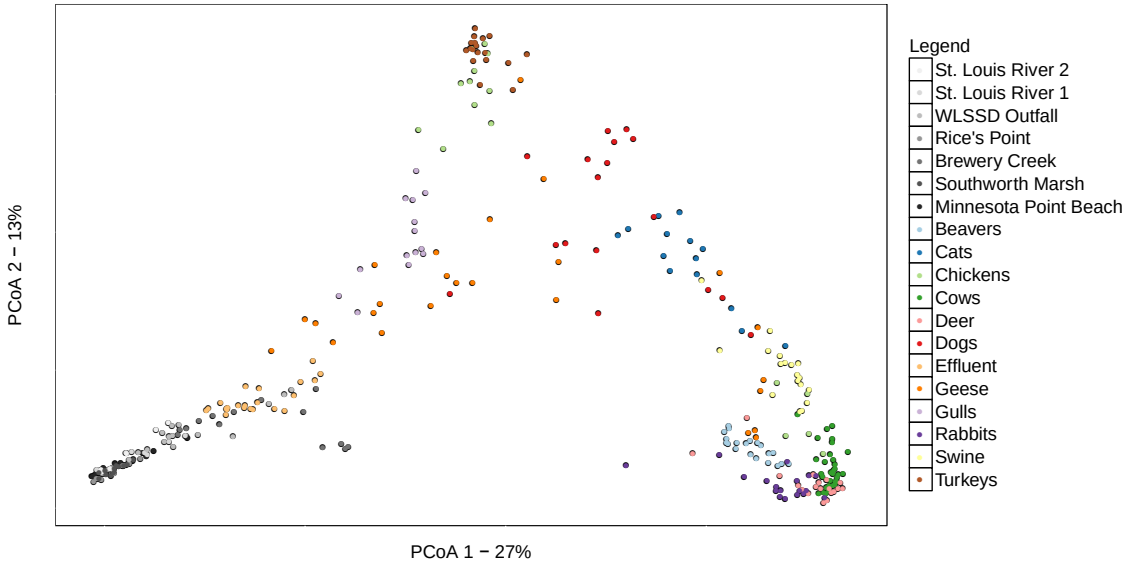

Figure 2.3. Principal coordinate analysis depicting sample relatedness

Principal coordinates analyses were performed using a Bray Curtis dissimilarity distance matrix that described the dissimilarity of all collected fecal and water samples. Water site samples are shown in gray scale and in a variety of shapes while animal fecal samples are depicted in colored circles.

Power analysis to determine appropriate library sizes. Power analyses were used to determine the number of samples in the FTL necessary to show statistical significances between fecal sources. To determine the lowest number of samples needed, library sizes were varied and power analyses re-run on the reformed libraries (Table 2.3). When < 12 animals per type were analyzed the power was 0%, while at 13 animals per type group, the power was ~31% (Table 2.3). Greater than 13 individuals per animal type at 25,000 reads per sample yielded a power of 100% (Table 2.3).

Table 2.3. Power analyses of fecal samples to determine appropriate sample amounts for this study

Library Size	Power (%)
10	0
11	0
12	30.9
13	100
15	100
All	100

Power analyses were performed at the family level on a taxa table containing a sequencing depth of 25,000 reads per sample.

Fecal sources determined by using SourceTracker analyses. Five independent runs of SourceTracker on family-level taxa tables predicted that treated wastewater effluent was likely the most common fecal input source in seven different locations in the Duluth-Superior Harbor and St. Louis River estuary (Fig. 2.4). The most common families used to create the bacterial profile from treated wastewater effluent included the *Comamonadaceae*, *Burkholderiaceae*, and *Candidate division OD1*. SourceTracker used an average of 172 different families, across all seven sites to make bacterial profiles for wastewater effluent (Table 2.4). The number of family-level taxa used to predict fecal inputs at different sites varied.

As expected, the WLSSD outfall that releases treated wastewater effluent into the Lake Superior-Duluth Harbor, had the greatest predicted source contribution associated with effluent (Fig. 2.4). This was followed by St. Louis

River site 2, Brewery Creek storm drain, Minnesota Point Beach, Rice's Point, Southworth Marsh storm drain, and St. Louis River site 1 (Fig. 2.4).

Fecal inputs by waterfowl (geese and gulls) were less widespread, and estimated to account for an average of 10% of the fecal inputs at the Brewery Creek storm drain in June 2015. SourceTracker used a mean of 68 different families to create bacterial profiles for geese at the Brewery Creek site (Table 2.4). The most common families used to determine geese fecal inputs were *Lachnospiraceae, Comamonadaceae, Oxalobacteraceae,* and *Flavobacteriaceae*. In contrast, in June 2015, over half of the total detected contamination at the St. Louis River site 2 was estimated to come from gulls. SourceTracker used 68 different families to create bacterial profiles for gulls including the common orders *Enterobacteriaceae,* LD12 (within the SAR11 clade of the *Alphaproteobacteria*)*, Sporichtyaceae,* and *Burkholderiaceae* (Table 2.4).

Spearman rank correlation (ρ = -0.98, p value < $2e^{-16}$) revealed that the RSD values obtained from the five SourceTracker runs were negatively correlated with the size of the predicted SourceTracker proportions. Larger SourceTracker proportions, like those seen with effluent had lower RSD values, while smaller SourceTracker proportions like geese and gulls had higher RSD values (Table 2.5).

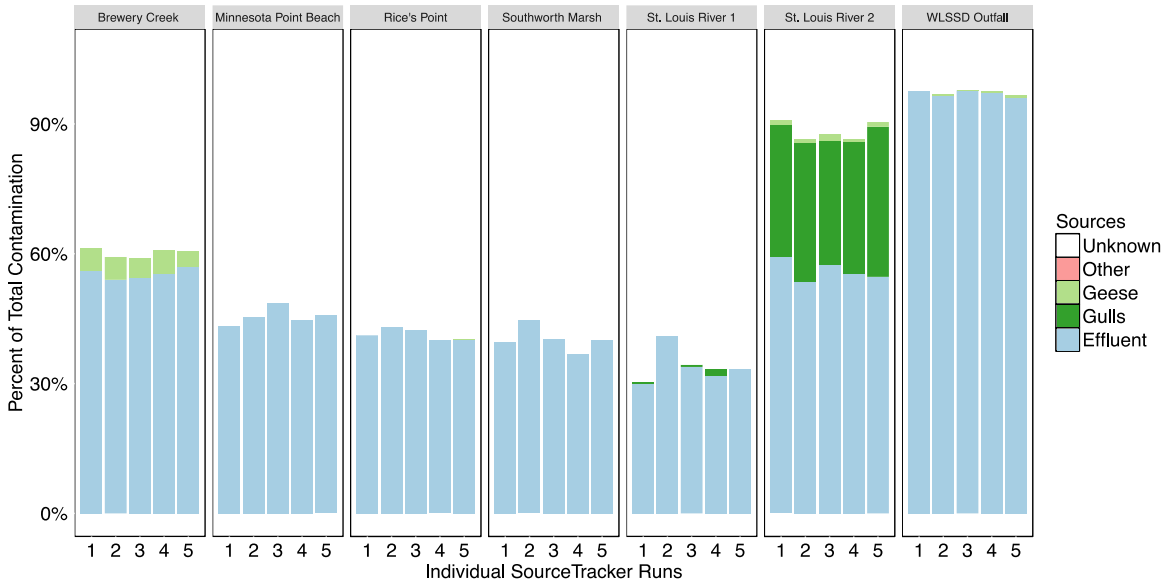

Figure 2.4. SourceTracker results depicting sources of fecal bacteria at seven sites in the Duluth-Superior Harbor and St. Louis River estuary

SourceTracker was run on all collected water samples in five independent runs to perform community-based MST in the Lake Superior watershed. One bar represents the average of predicted fecal sources across all sampling events at one site in one independent SourceTracker run. Only SourceTracker proportions >1%, with RSDs below 100% were used.

Table 2.4. Unique OTUs and taxa used by SourceTracker to assign contamination sources

Site	Source	Most Abundant Taxa	Number of Unique Taxa
SLR 1	Effluent	*Comamonadaceae, Burkholderiaceae, Sporichtyaceae*	151
SLR 2	Effluent	*Comamonadaceae, Burkholderiaceae, LD12*	118
SLR 2	Gulls	*Enterobacteriaceae, LD12, Sporichtyaceae*	79
WLSSD	Effluent	*Comamonadaceae, Candidate division OD1, Candidate TM7*	230
SWM	Effluent	*Burkholderiaceae, Comamonadaceae, Sporichtyaceae*	134
MPB	Effluent	*Comamonadaceae, Burkholderiaceae, Sporichtyaceae*	158
RP	Effluent	*Comamonadaceae, Burkholderiaceae, Sporichtyaceae*	166
BC	Effluent	*Comamonadaceae, Candidate division OD1, Flavobacteriaceae*	245
BC	Geese	*Lachnospiraceae, Comamonadaceae, Oxalobacteraceae*	68

SourceTracker was run on all collected water samples on a taxa table that contained family level taxonomic classifications. The third column contains the average number of family level taxa from five independent SourceTracker runs used for each source at each site. Site abbreviations are as follows: SLR = St. Louis River, WLSSD = the Western Lake Superior Sanitary District outfall, SWM = Southworth Marsh storm drain, MPB = Minnesota Point Beach, RP = Rice's Point, and BC = Brewery Creek storm drain.

Table 2.5. Relative standard deviation analysis of SourceTracker results after five independent runs of select samples

Sites & Dates	Source	Avg. ST Proportion (%)	RSD (%)
WLSSD July 27, 2015	Effluent	94.9	0.71
SLR 2 June 16, 2015	Effluent	68.3	2.48
MPB November 5, 2014	Effluent	33.9	4.07
RP August 20, 2014	Effluent	18.9	7.30
BC June 16, 2015	Geese	6.1	20.71

Select sites and dates were chosen to illustrate the negatively correlated trend of SourceTracker proportion and RSD values. Abbreviations are as follows: ST = SourceTracker; RSD = relative standard deviation, SLR 2 = St. Louis River site 2, WLSSD = Western Lake Superior Sanitary District outfall, RP = Rice's Point, and BC = Brewery Creek storm drain. The RSD was calculated by dividing the average standard deviation over five independent SourceTracker runs by the average source proportion also obtained from five independent SourceTracker runs for each water sample and multiplying it by 100.

DISCUSSION

In this study, the use of high-throughput DNA sequencing to create a FTL was applied as a tool to determine sources of fecal bacteria at seven different sites in the Duluth-Superior Harbor and St. Louis River estuary in Duluth, MN, USA. SourceTracker, a computational tool, was utilized to estimate the sources and the relative contributions of various fecal inputs into the study sites (94). We

performed power and relative standard deviation analyses to determine the appropriate size of the FTL as well as acquire an additional error estimation of the SourceTracker program.

On average, water samples possessed greater microbial diversity than did fecal samples. Shannon indices ranged from 3.65 to 6.13 in water versus 1.99 to 5.85 in feces. This could be due to water communities receiving a variety of microbial inputs from stormwater and soil run-off from rain events. Our results are similar to those from previous research on a wide variety of sample types (113).

The microbial community structures present in animal feces explain the clustering patterns seen in Figs. 2 and 3. All fecal samples from animals had large relative abundances of *Clostridiales,* with *Bacteroidales* as the next most abundant order, except for gulls which was in agreement with previous studies (84,92,114).

Turkeys, geese, chickens and gulls cluster around each other within the fecal cluster. These samples contain families within the order *Lactobacillales* which is consistent with previous literature on these animals (54,88,115–117). Clustered near the avian sample microbiomes are domesticated cats and dogs, which possess bacterial communities that resemble chickens and to a lesser extent, geese microbiomes. Fecal samples from dogs, chickens and cats share several bacterial taxonomic families that drive the close clustering of these groups within our study. Our data also supports the presence of *Erysipelotrichales* in these animals' microbiomes (84,92,117,118).

Agricultural animals such as cattle and swine cluster around wild and domesticated rabbits, deer, and beavers. These animals' microbiome samples possess large amounts of similar bacterial families in this study. In previous work, the fecal microbiomes of beavers and cattle possessed *Ruminococcaceae* and *Lachnospiraceae* which is consistent with our study (76,119). Other research has shown that the fecal microbiomes of rabbits and swine consist of *Firmicutes*, *Bacteroidetes* and *Verrucomicrobia,* which is also consistent with our study (120,121).

The fecal microbiome samples from geese, dogs, cats and chickens exhibited higher levels of variability than other animals. Using the Shannon index as an additional diversity estimator, some of the highly variably groups seen in the principle coordinates analysis possess standard deviations that either exceed the averaged OTU values in an animal type group or are close to it. In addition, the samples within these animal groups are more sparsely clustered relative to other animals in this study. Geese and chicken samples are sparsely clustered throughout the fecal sample cluster in Fig. 3. While cat and dog samples cluster more tightly than geese and chickens, these samples still cluster more disparately than other animal groups. The greater degree of dissimilarity within these sources could potentially confound SourceTracker calculations to identify the presence and source of pollution from animals with high intra-group variability.

SourceTracker predicted the largest source of fecal inputs to the study sites were due to treated wastewater effluent. Two wastewater treatment plants

(WWTP) discharge treated effluent into the Lake Superior-Duluth Harbor. As expected, the greatest proportion of taxa associated with wastewater effluent were at the outfall site in the harbor, while the lowest numbers were found at the St. Louis River site 1. Previous research using qPCR found that fecal inputs in the Lake Superior-Duluth Harbor could be attributed to wastewater effluent from WWTP (98,122). Our results obtained via community-based MST support these previous findings.

Aside from wastewater effluent, the next most common fecal input source predictions included geese and gulls. Similarly, *E. coli* from waterfowl was found to be a significant contributor to fecal inputs at the Duluth Boat Club beach (123). At the Brewery Creek site, geese, on average, were estimated to make up 5% of the fecal contamination sources, while gulls were estimated to have contributed about 30% of the fecal loading at the St. Louis River site 2. While treated wastewater effluent was found here to be the most prevalent source of fecal inputs at the sampled sites, these results do not necessarily mean there are serious health risks associated with these inputs, although there may be potentially unknown risks such as heavy metal contamination associated with effluent discharge (124). The effluent signal is likely stronger than other potential sources given that it is released daily at high concentrations while waterfowl populations are more transient.

SourceTracker estimated that several sites had large contributions from an unknown source. SourceTracker works by modeling the environmental sample as a mixed sink for several source taxa. It then works to assign all the

taxa in an environmental sink sample to a source. If taxa cannot be assigned to a source, it is assigned to an “Unknown” category. These unsourced taxa could be from unknown sources present in the environmental sites or be a part of the indigenous microbiome of freshwater in this estuary.

To increase confidence of the proportion predictions by SourceTracker, the RSD was calculated from five independent SourceTracker runs. Sources with the largest proportions had the lowest RSD values, indicating high confidence. Spearman rank correlation found a strong negative relationship between RSD and the size of predicted source proportions. This trend is in accordance with previous findings, where the program yielded greater variability when quantifying low source contributors (94,95). Therefore, SourceTracker may only be able to predict that low abundance sources are present, rather than accurately quantifying fecal inputs.

The geographic and temporal stability of the multi-taxa markers will need to be further investigated since geographic variability made sharing curated libraries difficult in previous library-dependent MST methods (45,125,126). Another consideration is that the decay rates for different taxa are likely unique and dependent on environmental factors (127). Since community-based MST markers are comprised of several different taxa, their unique decay rates may affect SourceTracker’s ability to accurately assign correct sources to contamination in a watershed.

Additionally, while there were similarities between the animal microbiomes in previous studies and this work, the sequencing of different variable regions,

choice of sequencing technologies and chemistries, as well as the bioinformatic processing could have contributed to the lack of similarities between microbiome community structures. These factors also complicate the use of shared libraries if multi-taxa markers are not spatially and temporally stable and will need to be further researched.

Community-based MST potentially allows for the simultaneous screening of all sources. While this technique is promising, there are still several challenges to be investigated including the geographic variability and temporal stability of FTLs, the cost and time required to build high-throughput sequencing libraries, risk assessments, decay of fecal pollution indicators and the effects of physiochemical parameters on decay, as well as the appropriate amount of each animal/fecal type required for each FTL. However, with a greater understanding of the potential pitfalls of this method, community-based MST could serve as a powerful way of determining fecal pollution sources in waterways.

CHAPTER 3 : EVALUATION OF SOURCETRACKER FOR COMMUNITY-BASED MICROBIAL SOURCE TRACKING IN *IN SITU* FRESHWATER MESOCOSMS

Brown C, Mathai P, Loesekann T, Staley C, Sadowsky M. Manuscript submitted for peer review to Environmental Science & Technology.

INTRODUCTION

Microbial source tracking (MST) is the science of determining sources of fecal contamination in water, food and the environment. Recent MST methods targeted to water have utilized quantitative PCR (qPCR) to amplify host-associated microbial markers to identify potential fecal sources (54,128,129). However, given the knowledge gap on the microbiota of different animal fecal sources, coupled with several qPCR markers suffering from low specificity, sensitivity, and cross-reactivity, a more reliable MST method is needed (47,88–90,129).

In the last two decades, the advancement of culture-independent techniques like high-throughput DNA sequencing have led to large-scale microbial community studies to profile microorganisms present in different environments (77,78,82,84,92,130,131). Several studies have reported distinct microbial communities that are present in the feces of different animals and in environmental samples (77,78,82,84,92,130,131). Leveraging DNA sequence data to acquire the unique microbial community profiles of environmental and fecal sources for MST has been termed community-based MST (84,94,95,130).

Community-based MST approaches have been used in several studies, with some performed in concert with the Bayesian classifier program SourceTracker, which determines the probability that certain taxa are derived from specific fecal sources (84,94,95,130). This program provides an estimated percentage of the sequenced microbial community at a site (the sink) that can be attributed to a specific fecal source. SourceTracker also reveals the composition

of OTUs used to build the source profiles, as well as their relative abundances in each of the community-based source profiles it builds and uses from each of the sources in the fecal taxon library (FTL). To our knowledge, there has been no study to date that has explored the source profiles SourceTracker creates. There is also no study to date that examines how SourceTracker handles closely related sources. More information on the composition of SourceTracker source profiles will yield better insights into how the program uses the provided FTL data to predict sources.

There are still several critical aspects of SourceTracker's ability to perform in community-based MST that have not been explored. The efficacy of SourceTracker to detect known sources is still being established (84,95,130), as well as the consistency of these community-based MST source profiles (83,95). No studies to date have explored how SourceTracker results are impacted if the FTL contains additional sources, or if present sources are missing from the FTL provided. Additionally, to our knowledge, there have been no studies that address whether SourceTracker is able to predict sources that have undergone environmental exposure. Moreover, the impact on SourceTracker predictions when using or excluding autochthonous taxa as a source in these studies has not been investigated. Thus far, the study of SourceTracker's predictive power has been *in vitro* (95). These are all critical points of investigation that may influence the accuracy of SourceTracker results.

Therefore, the objective of this current study was to determine the ability of SourceTracker to detect inputs of known fecal sources that had been exposed

to freshwater environmental conditions and how different combinations of FTLs (configurations) impact determination of input sources. To simulate *in situ* conditions of when fecal contamination occurs, different combinations of fecal sources were spiked into ambient lake water and exposed to a freshwater lake in St. Paul, MN, USA that had an observable but small population of geese. These sources were chosen for their potential impact to human health and to challenge SourceTracker's ability to distinguish closely-related sources. Feces from cows is hypothesized to pose high health risks, second only to the threat from human feces (132). In previous work, the bacterial community structure of secondary wastewater effluent closely resembled the bacterial community structure of freshwater (130). Effluent was therefore used to evaluate whether SourceTracker could discern between two sources (effluent and lake water) that had similar bacterial community structures. The triple-source fecal mixture (horse, cow, effluent) was created to simulate when multiple sources impact a body of water and to test SourceTracker's ability to distinguish between them. We also evaluated the taxonomic composition of the SourceTracker source profiles for insights into how SourceTracker uses the provided data to discern sources.

METHODS

Sample collection, mesocosm creation and environmental placement. To evaluate SourceTracker's predictive power under the influence of environmental exposure, triplicate, 300 ml, lake water samples containing single and triple fecal sources (secondary wastewater effluent, cattle, and horse feces) were contained within dialysis bags and placed in a freshwater lake.

Four types of spiked mesocosms were made for this study: cow-only, secondary wastewater effluent-only, lake water-only and a mixture of cattle and horse feces, and secondary effluent. Triplicate mesocosms of each treatment were prepared by diluting spiked sources into lake water collected from Lake Owasso in St. Paul, Minnesota (MN), USA.

Fecal samples from five cows and five horses were collected from the University of Minnesota Agriculture Extension in St. Paul, MN, USA on two separate days of each experiment in 2016. Pre-treated secondary wastewater effluent was also obtained on the days of the experiments from the Metropolitan Council Environmental Services wastewater plant in St. Paul, MN, USA. These studies that occurred on two different days are referred to as Experiment 1 and Experiment 2.

To construct the cow-only spiked mesocosms, 4 g of feces were collected from five individual cows for each experiment and blended with 1 L sterile phosphate buffered saline (PBS), pH 7.0. The fecal slurry was diluted 1:10 into lake water to make the cow-only spiked mesocosms to a final volume of 300 ml. For the effluent-only spiked mesocosms, a final concentration of 30% (v/v) effluent diluted into ambient lake water was used. For the mixed spiked mesocosms, 12.5 g of cow and horse feces each, along with 300 ml of secondary wastewater effluent were blended with 1 L of sterile PBS. The cow and mix slurries were diluted to 1:10 into 300 ml of a lake water. All spiked mesocosms were transferred to dialysis bags that were 75 mm (flat-width) with a molecular weight cut-off of 12-14 kD (Spectrum, Inc., Rancho Dominguez, CA).

A support frame to hold dialysis bags in water was constructed as previously described (133). The cage, 152.4 cm × 76.2 cm, was made using 1.9 cm PVC pipe, and covered with plastic chicken-mesh wire to allow for water flow-through, while preventing larger debris from puncturing the dialysis bags during strong hydrologic disturbances. Dialysis bags were submerged in Lake Owasso for one hour.

Mesocosm sample processing. The triplicate dialysis bags for each treatment were collected in sterile 2 L bottles and transported to the lab at ambient temperature in approximately 800 mL of freshly collected lake water. The bags were gently massaged and the mesocosm slurries were transferred to sterile 500 ml bottles. All mesocosm samples were processed as previously described (130). Briefly, mesocosm slurries were filtered through 5 µm, 0.45 µm, and 0.22 µm nitrocellulose filters (Millipore-Sigma, St. Louis, MO) (100). The 0.45 µm and 0.22 µm filters were pooled for filter processing and were the only filters used in this study. Filter processing involved adding 2 ml of 0.01% sodium pyrophosphate buffer, pH 7.0, containing 0.2% Tween 20, and vortexing three times. The supernatant was centrifuged at 13,000 × g for 3 min and the cell pellets from the same mesocosm samples were collected, combined, and stored at -20°C until DNA extraction. DNA was extracted using the MoBio Powersoil (now the DNeasy PowerSoil) DNA extraction kit (Qiagen, Hilden, Germany), as per kit directions.

High-throughput DNA sequencing. Amplicon-based DNA sequencing was done as previously described at the University of Minnesota Genomics

Center (St. Paul, MN, USA) (103,130). Primers F784 (GGATTAGATACCC) (101) and 1046R (CGACRRCCATGC ANCACCT) (102) were used to amplify the V5 and V6 regions of the 16S rRNA gene. Amplicons were paired-end sequenced using the dual indexing method on the Illumina HiSeq 2500 and MiSeq platforms (Illumina, San Diego, CA) (103). Sequencing results generated from this study can be accessed from the NCBI Sequence Read Archive under BioProject PRJNA473286.

Quantitative PCR. To evaluate whether waterfowl could be detected by another MST method, qPCR was performed on all mesocosm samples in this study using the GFD marker (54). The forward and reverse primer sequences were: TCGGCTGAGCACTCTAGGG and GCGTCTCTTTGTACATCCCA, respectively (54). Triplicate reactions were performed in 20 µl volumes with iTaq (TM) Universal SYBR® Green Supermix (Biorad, Hercules, CA). Magnesium chloride was used at a 1 mM final concentration (Promega, Madison, WI). The concentration of each primer was 250 nM with 5 µl of template DNA added to each 15 µl reaction. Cycling conditions were as follows: 95°C for 10 minutes, with 40 cycles of 95°C for 15 seconds and 57°C for 30 seconds. Only reactions with amplification efficiencies above 90% and below 110% are reported here.

Bioinformatics. The dataset for FTL construction and SourceTracker analyses included fecal input sources generated from a previous study (130) : 20 beavers, 14 cats, 16 chickens, 32 cows, 19 deer, 17 dogs, 25 geese, 14 gulls, 18 rabbits, 18 swine, 18 turkeys, 22 effluent samples, and 16 horses. All fecal

samples were collected from Minnesota except horses which were collected from Florida.

When all fecal sources were included in the FTL, these configurations were referred to as the All Available Sources (AAS) library. When only the known sources (cows, effluent, horses and occasionally lake water when used as a source) were included in the FTL, the configurations were known as the Only Known Sources (OKS) library. Lake water was used as either a source or a sink to evaluate its effect on SourceTracker predictions in both FTL configurations.

QIIME v. 1.8.0 and 1.9.1 (134) , R v. 3.2.1 (135) and SourceTracker v. 1.0.1 (56) were used for bioinformatic and statistical processing of DNA sequence data. Data, generated from both Illumina HiSeq and MiSeq runs, were processed as previously described (130). Low-quality base calls and Illumina adapter contamination were removed with Trimmomatic v. 3.2 (104), and primers, homopolymers >8, reads smaller than 75% of the amplicon length were removed with Pandaseq (105). UCHIME v. 6.1 (74) was used to remove chimeras, while open-reference OTUs were clustered at 97% with uclust (107). Sequence alignment was performed with the SILVA v. 128 reference database (108) using the PyNAST algorithm (73). Taxonomic assignment, with an 80% bootstrap value, used the SILVA v. 128 reference taxonomy database with the RDP Classifier (69). The remaining high-quality data were used for statistical analyses after removal of singletons from the dataset. After multiple rarefaction depths were analyzed for sample coverage, 30,000 sequences per sample was

chosen. Bray-Curtis dissimilarity (110) was calculated to determine sample relatedness.

The betadisper function (136) of the R vegan package (137) was run to calculate inter-group variation. This test is a multi-variate version of Levene's test for homogeneity of variances and calculates the average distance of group members to the group centroid (138). Significance was determined with permutation tests.

SourceTracker (56) was used to perform community-based MST as previously described (130). Different library configurations were built by filtering the OTU table for only the source and sink samples that were included in the FTL configuration in QIIME. To assess confidence in the SourceTracker predictions, relative standard deviation (RSD) values have been calculated from multiple technical replicates of SourceTracker runs (95,130). These values, however, only address the technical replicates used by SourceTracker, and not the combined error between the biological and technical replicates. To assess the error in biological and technical replicates of SourceTracker sink predictions, five independent runs were performed and all values were averaged. This was done by averaging SourceTracker predictions of the five technical replicates and biological replicates and calculating the standard deviation between all technical and biological replicates. This allowed for calculation of the overall relative standard deviation (RSD), where the standard deviation between the biological and technical replicates was divided by the averaged SourceTracker predictions of each source in each of averaged mesocosm samples.

SourceTracker output includes source-specific sink contribution files for each source. These files contain the relative abundances of OTUs that compose a source profile in every sink sample. While the source profiles of the same source in each mesocosm can be different, this file allows us to view all of the taxa used across all mesocosms to build a specific source profile. Therefore, source-specific OTU lists, which included all of the taxa used to build a particular source across all mesocosms, were constructed from the sink contribution files using taxa that had relative abundances greater than 0%. Intersecting OTUs between two different SourceTracker source profiles were selected to better understand the percentage of taxa shared between them. These intersecting OTUs were termed the shared taxonomic composition (Fig. 3.1). This analysis yielded two numbers for each source combination and was exclusively used on the OKS configurations. Each number was the percentage of the source profile that consisted of shared taxa with the other source profile.

To determine how much of a source profile in a mesocosm consisted of shared taxa, the number of taxa in each source-specific OTU list within a mesocosm was divided by the number of taxa in the shared taxonomic composition that was compared to it (Fig. 3.1). This percentage was termed the shared relative abundance (Fig. 3.1). This analysis yielded shared relative abundances for every source in a sink sample, and was used exclusively on the OKS configurations. Since differences between the two experiments were small, only mean values from the first experiment were reported.

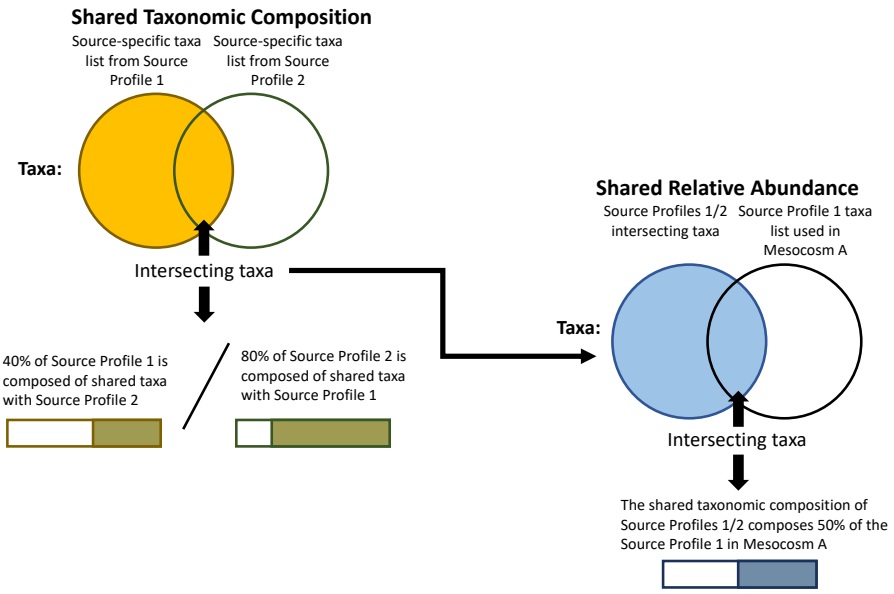

Figure 3.1. Depiction of the shared taxonomic composition and shared relative abundance

The shared taxonomic composition are the percentages of taxa in each marker that are shared between two different markers. The shared relative abundance is the percentage the shared taxonomic composition composes in a marker used to predict a source in a mesocosm.

RESULTS

The effects of using "All Available Sources" on SourceTracker predictions. Two additional FTL configurations were constructed to investigate how additional sources in the FTL, chiefly those sources that were not present in the mesocosms, impacted SourceTracker's ability to discern known and present sources (Fig. 3.2). All available sources (AAS) were analyzed with the lake water mesocosm acting as either a source (Fig. 3.2a) or a sink (Fig. 3.2b).

When lake water was used as a source, SourceTracker predicted that cows, horses, and lake water were in the cow mesocosms, effluent and lake water were in the effluent mesocosms, and cows, horses, and lake water were in the mixed mesocosms (Fig. 3.2a). SourceTracker also predicted a moderate presence of geese in the mixed mesocosms in both experiments and in the cow mesocosms in Experiment 2. Additionally, effluent did not appear in the mixed mesocosms.

When lake water was used as a sink, SourceTracker predicted that the cow mesocosms consisted of a little less than half of cow (Fig. 3.2b). Some of the other sources predicted included waterfowl and effluent (Fig. 3.2b). High levels of effluent, with low levels of waterfowl, were predicted in the lake water mesocosms, while horses, cows, and geese were the main predicted sources in the mixed mesocosms (Fig. 3.2b).

When lake water was used a sink, the lake water mesocosm predictions never summed to around 100% since there were substantial predictions with overall RSD values above 100% that had been removed.

Quantitative-PCR was used to determine whether waterfowl, such as geese, had been present in the lake water. While geese fecal samples had detectable levels of GFD marker, none of the mesocosm samples displayed detectable levels of this marker (data not shown).

Additionally, the intra-group variability of all source profiles was evaluated. Geese had the highest intra-group variability which was found to be significantly different from cows (Fig. 3.3).

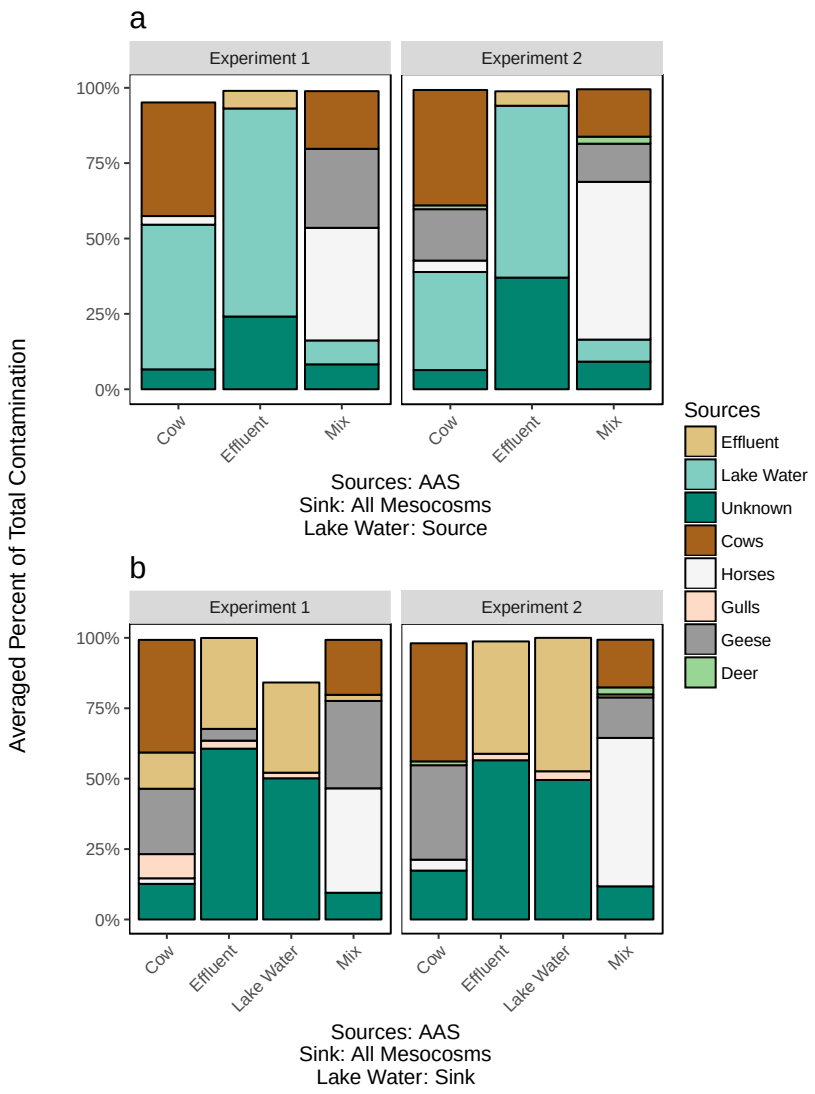

Figure 3.2. Stacked bar charts analyzing effects of additional sources in the FTL on SourceTracker predictions

Only values that are above 1% and have overall RSD values less than 100% were used. The y-axis depicts the averaged percent of total contamination for each source while the x axis are the mesocosms. (a) This figure shows SourceTracker results from the AAS FTL configuration with lake water being utilized as a source while (b) depicts the AAS FTL configuration with lake water as a sink.

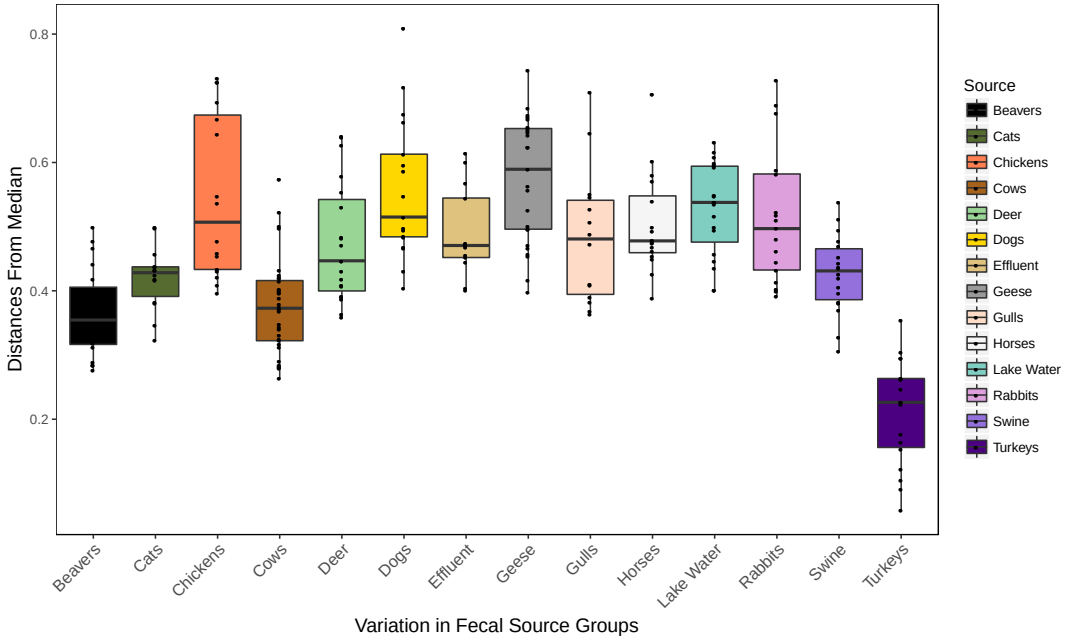

Figure 3.3. Boxplot depicting intra-group variances within source groups. Black dots are all samples within a group. A multivariate version of Levene's test for homogeneity of variances was performed on source group samples. Higher distances from the median indicate higher variation within the group. To have sample numbers that resembled other source groups, the cow source group was reduced to 20 samples instead of 32 for this analysis.

The impact of using "Only Known Sources" on SourceTracker predictions. Two more FTL configurations were constructed to determine how the fecal taxon library (FTL) composition impacted SourceTracker's ability to discern known and present sources (Fig. 3.4). These configurations addressed how SourceTracker predictions were impacted by the presence of only the known sources (OKS) in the FTL.

When lake water was used as a source, SourceTracker predicted that cows and lake water were the major sources in the cow mesocosms, effluent and lake water in the effluent mesocosms, and effluent, cows, horses, and lake water in the mixed mesocosms (Fig. 3.4a). There were low to moderate levels of horse in the cow mesocosms in both experiments (Fig. 3.4a). The effluent signal was not detected in the mixed mesocosms (Fig. 3.4a).

When lake water was used as a sink, SourceTracker predicted mostly cows with varying levels of cows and horses in the cow mesocosms across both experiments. High levels of effluent were predicted in the effluent and lake water mesocosms with horses, cows, and effluent being predicted in the mixed mesocosms (Fig. 3.4b).

Additionally, in Experiment 1 when lake was used as a sink, the lake water mesocosm predictions never summed to around 100%. There were substantial predictions with overall RSD values above 100% that had been removed.

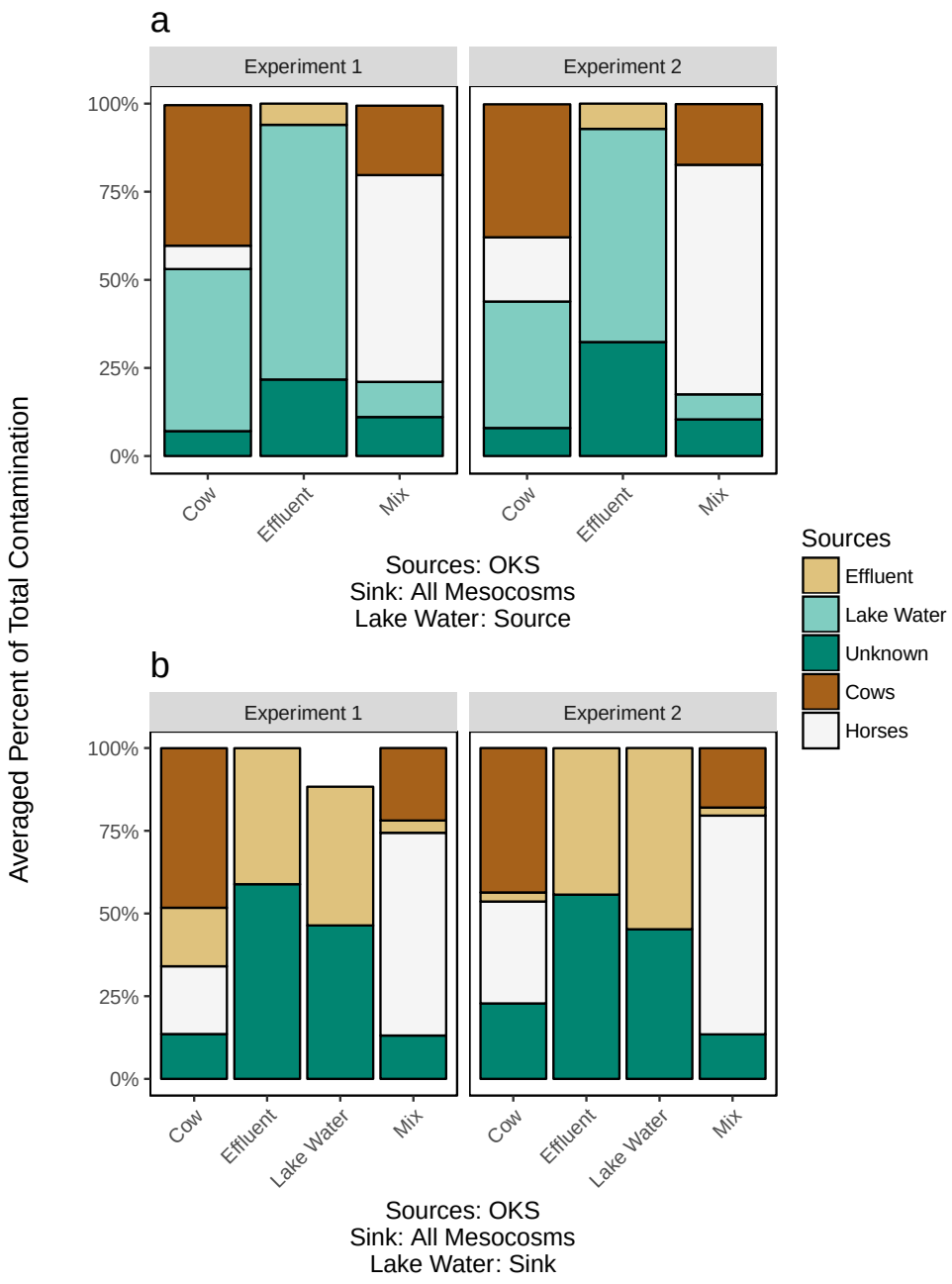

Figure 3.4. Stacked bar charts analyzing effects of using only the known sources in the FTL on SourceTracker predictions

Only values that are above 1% and have overall RSD values less than 100% were used. The y-axis depicts the averaged percent of total contamination for each source while the x axis are the mesocosms. The following configurations

were tested: (a) the OKS FTL configuration with lake water being utilized as a source and (b) depicts the OKS FTL configuration with lake water as a sink.

The effects of FTLs with missing sources on SourceTracker predictions. To address how SourceTracker predictions are impacted when known sources are missing, six additional FTL configurations were constructed (Fig. 3.5). These included examining the effects of when each of the sources were missing from their respective mesocosms with lake water as either a source (Fig. 3.5a-c) or a sink (Fig. 3.5d-e).

When lake water was used as a source, SourceTracker predicted that cow and lake water were in the cow and mixed mesocosms (Fig. 3.5a), that effluent and lake water were in the effluent mesocosms (Fig. 3.5b), and that horses and lake water in the mixed mesocosms (Fig. 3.5c). Horse was also predicted to be present in the cow mesocosm when horse and lake water were an available source in the FTL (Fig. 3.5c). When lake water was used as a sink, SourceTracker predicted that cow and horse were in nearly all mesocosms when they each were used as the only sources available (Fig. 3.5d-f). In addition, effluent was predicted in all mesocosms when it was present as the only source (Fig. 3.5e).

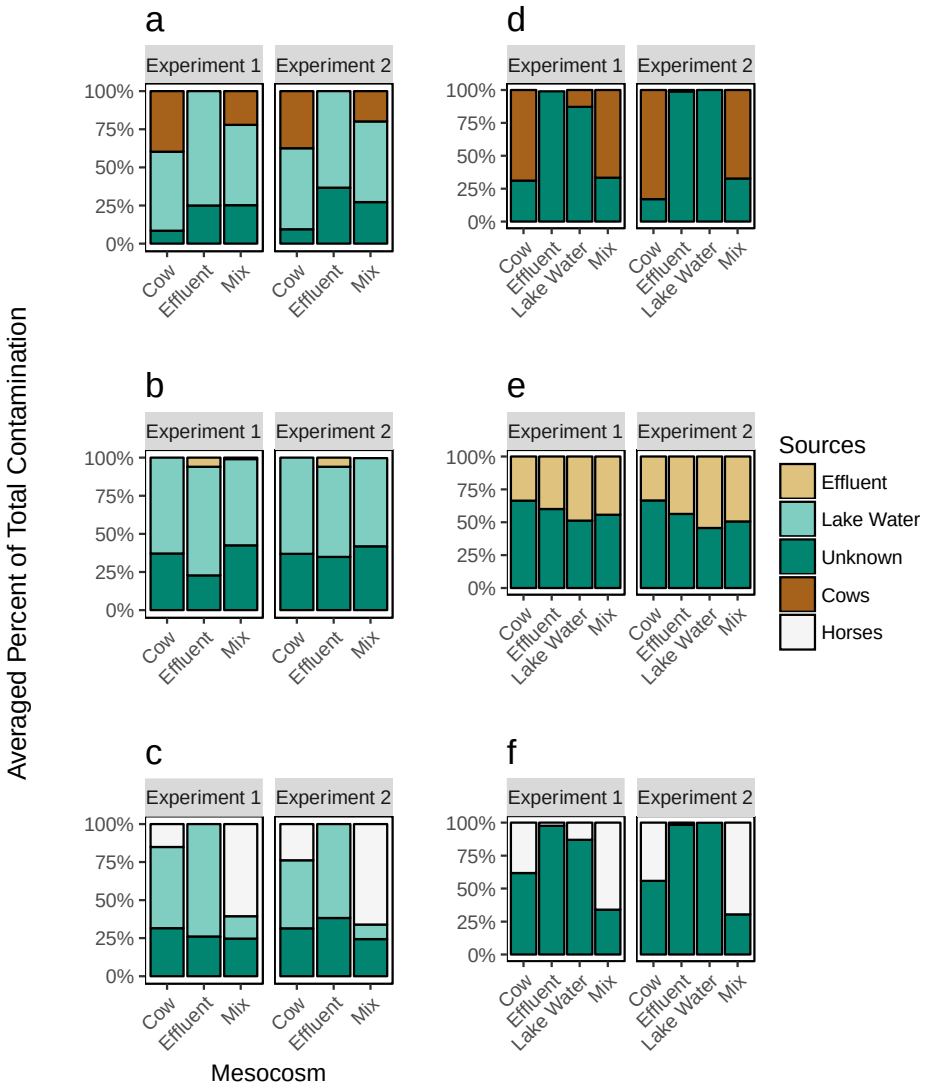

Figure 3.5. Stacked bar charts analyzing effects of missing sources on SourceTracker predictions

Only values that are above 1% and have overall RSD values less than 100% were used. The y-axis depicts the averaged percent of total contamination for each source while the x axis are the mesocosms. The following FTL configurations were tested: (a) when cows and lake water are only present, (b) when only effluent and lake water were present, (c) when only horse and lake water were present, (d) when only cow was present, (e) when only effluent was present, (f) when only horses were present.

Analysis of SourceTracker source profiles. To describe how SourceTracker forms the multi-taxa source profiles, the concept of shared taxonomic composition was created (Fig. 3.1). As defined previously, shared taxonomic composition is the percentage of taxa that are shared between two different source profiles. This analysis exclusively focused on the OKS configurations, with lake water used as a source or a sink due to the few sources available (Fig. 3.3).

A large percentage of the taxa in the effluent source profile was shared between lake water in the configuration where lake water was used as a source configuration, and with unknown in the configuration where lake water was used as a sink configuration (Table 3.1). There were few shared taxa between cow and horse in either configuration (Table 3.1). When comparing across different configurations, a large quantity of the taxa in the lake water source profile, when lake water was used as a source, was shared with the unknown fraction in the configuration where lake water was used as a sink (Table 3.1). There was a relatively moderate percentage of shared taxa between the cow and horse source profiles when comparing the source profiles across different configurations (Table 3.1).

As previously defined, the shared relative abundance determines how much of a source profile in a given mesocosm consists of the shared taxonomic composition between two other source profiles. The shared relative abundances of different SourceTracker source profiles (Fig. 3.1) within various mesocosms

were examined to better understand how SourceTracker uses different sources' taxa to form multi-taxa source profiles (Table 3.2). Nearly all of the taxa shared between effluent and unknown was used to build the effluent source profiles in both the lake water and the effluent mesocosms in the lake water as a sink configuration (Table 3.2). Only a few of the taxa shared between the cow and horse source profiles were used to create the source profiles for cow in the cow and mixed mesocosms (Table 3.2). A lot of the taxa shared between effluent and unknown was used to build the effluent source profiles in the effluent mesocosm in the lake water as a source configuration. Only approximately 40% of the taxa shared between lake water and effluent were used to form the effluent marker in the lake water as a source configuration.

Table 3.1. The shared taxonomic composition between several SourceTracker source profiles

Source Profiles	Shared Taxonomic Composition (%)	FTL Configuration (OKS)
Comparing within the same configuration		
Lake Water/Effluent	9/82	Source/Source
Cow/Horse	18/15	Source/Source
Cow/Horse	24/18	Sink/Sink
Effluent/Unknown	94/17	Sink/Sink
Comparing across different configurations		
Lake Water/Unknown	39/79	Source/Sink
Cow/Cow	56/52	Source/Sink
Horse/Horse	39/34	Source/Sink
Effluent/Effluent	55/23	Source/Sink
Unknown/Unknown	63/22	Source/Sink

This table shows the shared taxonomic composition between two source profiles in the OKS configurations where lake water was used as either a source or sink. This first column shows the two source profiles where the degree that they share the same taxa are being evaluated. The second column shows the percentage of shared taxa that make up the source profile in the corresponding position from the first column (i.e., 9% of the lake water source profile consists of shared taxa with effluent). The third column shows which OKS configuration (lake water as a source or lake water as a sink) the corresponding source profile in the first column is from.

Table 3.2. The shared relative abundances of selected source profiles within certain mesocosms

Source Profiles	Source Profile in Mesocosm	Shared Relative Abundance (%)
FTL Configuration: OKS with lake water as a sink		
Cow/Horse	Cow in Cow	46
Effluent/Unknown	Effluent in Effluent	97
Effluent/Unknown	Effluent in Lake Water	99
FTL Configuration: OKS with lake water as a source		
Cow/Horse	Cow in Cow	25
Cow/Horse	Cow in Mix	16
Lake Water/Effluent	Effluent in Effluent	37

This table depicts shared relative abundances of different source profiles used in a variety of mesocosms in the OKS configurations. The first column shows the two source profiles whose shared taxa are being compared to the taxa in the source profile in the second column. The second column indicates the source profile within a certain mesocosm of which whose taxa are being compared to the shared taxa from the first column. The third column shows how much of the source profile within a specific mesocosm in the second column is composed of taxa shared with the two source profiles in the first column.

DISCUSSION

SourceTracker was able to predict the presence of nearly all sources that were added to each of the mesocosms. Across the AAS and OKS FTL configurations with lake water as a sink, cows were consistently the largest source in the cow mesocosms. In the mix mesocosms of the OKS configuration, cows and horses were the largest sources while in the AAS configuration, cows,

horses and geese were the major sources. In the effluent and lake water mesocosms in the AAS and OKS configurations, effluent was predicted as the primary source of contamination. In the AAS and OKS FTL configurations that used lake water as a source, both cows and lake water were the major sources in the cow mesocosms, while in the mix mesocosms, cows, lake water, and horse were the largest sources. Given this data, less spurious predictions occurred when lake water was used as a source. Effluent and waterfowl were no longer predicted in the cow mesocosms and there were lower levels of geese predictions in the AAS lake water as a sink configuration results. The high overall RSD values indicate low SourceTracker model confidence in predictions and thus these observations are removed. Moreover, when lake water was used as a source, all of the mesocosms sum to or around 100% in both experiments indicating consistent source predictions.

Across the AAS and OKS configurations that used lake water as a source, the effluent source profile was not detected in the mixed mesocosms. This may be due to the fact that bacterial cell counts in effluent (10^6) (139,140) are expected be much lower than those corresponding to feces (>10^{10}) (141). Additionally, effluent was predicted to be a minor source in the effluent mesocosms. While this may be due to the effluent source used in this study's FTL being different from the effluent source spiked into the mesocosms, we hypothesized that it could also be because of the similar bacterial community structures between effluent and freshwater (130) leading to SourceTracker struggling to distinguish between these two sources when present together.

To explore whether SourceTracker conflated lake water and effluent with each other, we examined the source profiles more in depth. First, we looked at two distinct cases with both of the OKS configurations. The first case we examined was the OKS configuration when lake water was used as a source. We looked at the level of shared taxonomic composition between effluent and lake water. Nearly 40% of the effluent source profiles in the effluent mesocosm consisted of taxa almost entirely shared between effluent and lake water. For the second case in the OKS lake water as a sink configuration, we hypothesized that taxa associated with lake water would be classified to the unknown fraction. Therefore, if there were shared taxa between lake water and effluent, this could be partially reflected in the shared taxa between effluent and the unknown fraction. In the effluent source profile, a large percentage of taxa were shared with unknown confirming our hypothesis. These shared taxa made up nearly 100% of the effluent source profiles in both the effluent and lake water mesocosms. While this is striking, it should be noted that SourceTracker uses presence/absence and relative abundance in its calculations.

SourceTracker predicted higher levels of effluent in the effluent mesocosms when lake water wasn't available as a source. Additionally, when lake water was used as a mesocosm, roughly half of the contamination was attributed to effluent. Moreover, given the high number of shared taxa between lake water and effluent, SourceTracker did struggle to distinguish between these two sources. Therefore, care should be taken with interpreting SourceTracker predictions with sources that have similar bacterial community structures.

Throughout this study, the relative source contribution (~40%) of the cow source predicted by SourceTracker was mostly consistent across different FTL configurations. Interestingly, with lake water as a source, taxonomic assignments for the cow source profiles led to the same relative source contributions seen in configurations that used lake water as a sink. The cow source profiles in AAS and OKS configurations (including lake water as a source and as a sink) also had the greatest intra-source shared taxonomic composition, of approximately 50-60%. Given that horse was predicted to be present in the cow mesocosms, it was expected that the shared taxa between cow and horse would be large. While the shared taxonomic composition was low, the shared relative abundance ranged from 46% to 25% dependent on the FTL configuration used. Moreover, the intra-group variability within the cow source group was among the lowest of all groups. The tight clustering in the NMDS also highlights the lower variance within the source group. Given that the cow source profile had the most fecal samples associated with it, this suggest that more samples allow for a more consistent average profile of this source. This suggest that an ideal SourceTracker source profile could come from having a sizable library of fecal samples, where the necessary library size would be dependent on the intra-source variability of the source.

In contrast, the geese source in this study had the greatest intra-group variability. Quantitative-PCR results failed to corroborate the presence of geese in all of the mesocosms suggesting that the geese source profile was not present, or below the detection limit of our assay. Even though the geese source

profile could be present in the background lake water, it would be expected that the fresh deposits of cow and horse feces in the cow and mix mesocosms would heavily surpass the levels of the geese marker. Since SourceTracker uses an average profile of each source, high intra-group variability could make it more difficult for the program to identify geese or other sources that have similar bacterial community structures. To avoid such scenarios, we suggest that additional sources that are not expected to be present in the study area are not used in creation of the FTL. If a high-variability source must be used, it may be helpful to break the source into smaller groups of closely-related samples rather than use all samples as one source.

When present sources are missing, SourceTracker will yield erroneous predictions. We found that SourceTracker conflated sources that could easily be differentiated when all of the correct sources were present. This suggests that before undertaking MST studies, a knowledge of the fecal sources impacting a body of water are critical.

All of this data suggest that SourceTracker is a suitable tool for community-based MST. However, further work is needed to evaluate SourceTracker's quantitative predictive power. In so far as SourceTracker's predictive power is proven, quantitative risk assessments are needed to translate quantitative SourceTracker results into the potential for human health risks in waterways.

CHAPTER 4 : INVESTIGATION INTO THE IMPACTS OF AGE, DIET, AND GEOGRAPHY ON SOURCETRACKER PREDICTIONS

Brown C, Kleinheinz G, Briggs S, Peraud J, Harwood V, Raith M, Gilbert J, Edge T, Sadowsky M. Manuscript in preparation.

INTRODUCTION

Microbial source tracking (MST) arose, in large part, from a need to determine sources of fecal bacteria in waterways. Initial MST methods required use of genetic and phenotypic libraries to determine fecal sources (44,45). These library-dependent methods included profiling fecal contamination sources by the presence of certain antibiotic resistance genes, DNA fingerprinting (52), and ribotyping (45). In evaluating these methods, researchers examined if these libraries could be shared across various geographic regions. Hartle and colleagues reported that some ribotypes had decreased sensitivity when the library used was created from animals in a geographically different region from the sources being examined (142).

Due to many of the limitations of library-based methods, library-independent techniques such as qPCR were developed to determine sources of fecal bacteria. A variety of studies have examined the geographic stability of these markers and found variations in sensitivity and specificity (143). This suggests that animals of the same species may have microbiomes affected by diet and geographically distinct factors.

The recent use of high-throughput DNA sequencing has enabled a rapid proliferation of sequencing technologies that have decreasing costs. This subsequently led to the use of microbial community analysis on the fecal microbiota of different animals to evaluate different variables that may impact the composition of their microbiomes. Diet has been shown to influence the microbiomes of humans (144), gnotobiotic mice (145,146), horses (147) and

cows (148–150). In humans (151), cows (152), and horses (149), the microbiome varies in an age-dependent manner. Recently, community-based MST methods have been further enabled by use of the widely-accepted Bayesian statistical program SourceTracker (94). SourceTracker has now been used in several MST projects (84,95,130,153) and searches for the community structure of source samples in environmental sink samples (56).

It has been reported that SourceTracker is sensitive to relative differences in concentrations of fecal contaminants present in waterways (95,154). Additionally, it has been demonstrated that geography plays a role in the accuracy of SourceTracker predictions (154). To our knowledge, however, no studies to date have examined how diet and the age of source animals impacts SourceTracker predictions. To examine this in detail, we challenged SourceTracker to identify spiked cow feces from a farm in St. Paul, Minnesota (MN), USA using DNA-based fecal taxon libraries created from approximately 190 cow fecal samples from across the United States and Canada.

METHODS

Sample collection. Ten unique cow fecal samples were collected from each farm across the United States in Minnesota (n = 1), Wisconsin (n = 10), Michigan (n = 3), California (n = 3), Florida (n = 1), and in Canada (n = 1) in 2017 (Table 4.1 and Fig. 4.1). One farm from Wisconsin only had 7 samples. Samples from across the United States and Canada were shipped overnight on blue or dry ice in coolers. Received samples were frozen at -80°C until DNA extraction.

Spiking experiments. Five fresh cow samples collected the day of the experiment from the University of Minnesota Agricultural Extension were used for the spiking experiments. Approximately 1.6 g was taken from each cow fecal sample, mixed together and blended with 400ml of sterile PBS to make a fecal slurry. The fecal slurry was diluted with PBS to four different dilutions: 1:10, 1:100, 1:1000, and 1:10,000. Four 1:10, 1:100, and 1:1000 dilutions were made, allowing for three replicates and one seed dilution that was subsequently diluted to the next lowest dilution. All triplicate dilutions were then filtered, and immediately frozen at -80°C until cell removal with vortexing in PBS buffer as previously described (130). Only the 0.45μm and 0.22μm filters were used in this study.

DNA extraction and sequencing. DNA from the spiking experiment dilutions was extracted from all filters using the Qiagen DNeasy PowerSoil Kit (Qiagen, Hilden, Germany), per kit instructions. The same kit was used to extract DNA from all of the cow source samples, except those from Canada, which were previously isolated by Tom Edge with a Qiagen DNeasy PowerSoil Kit (Qiagen, Hilden, Germany). Sequencing for all samples was performed on the V4 region, using dual-indexing, paired-end high-throughput Illumina DNA sequencing (Illumina, San Diego, CA) at the University of Minnesota Genomics Center (103). Sequencing results from this study can be accessed under BioProject PRJNA473284.

Bioinformatic processing and statistical analysis. High-throughput sequencing data was analyzed in QIIME v. 1.9.1, mothur v. 1.37.6, and in python

3.6. In QIIME, sequencing results were analyzed as previously described (130). An OTU-level table, rarefied to 30,000 sequences per sample, was used as input into SourceTracker v. 1.0.1 which was performed via QIIME (56). Pairwise analysis of molecular variance (AMOVA) comparisons were performed in mothur. SourceTracker runs were set up to compare all of the cows from one farm as a source to all of the diluted sink samples. Principal coordinates and diversity analyses were performed on the rarefied OTU table by using the ecopy library in python.

Table 4.1. Overview of fecal sample collection

Locations	No. of Farms	Total Samples
Wisconsin, USA	10	97
Minnesota, USA	1	10
Michigan, USA	3	30
California, USA	3	30
Ontario, Canada	1	10
Florida, USA	1	10

One Wisconsin farm that had only seven samples.

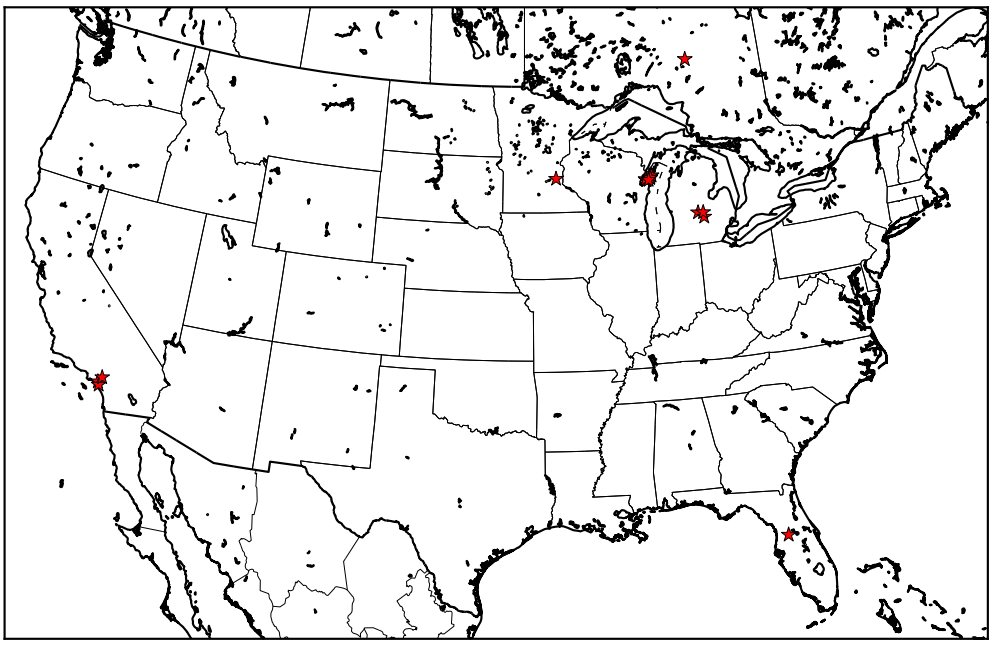

Figure 4.1. Map of sample locations

Map of locations where fecal samples were collected from. The locations are represented by red stars.

RESULTS

The bacterial community structure of cow fecal samples from across North America. Approximately 190 cattle fecal samples were collected from across North America and their DNA was extracted. These samples, which constituted the fecal taxon library (FTL), were compared against dilutions of cow fecal samples obtained in Minnesota. Cow source samples and fecal spiked samples had Shannon indices that ranged from 4.1 to 7.15, with an average of 6.2 There were on average 3,374 OTUs, ranging from 1,597 to 5,192.

Cows from across the United States and Canada generally possessed high levels of *Ruminococcaceae*, *Bacteroidaceae*, and *Lachnospiraceae* (Fig.

4.2). *Ruminococcaceae* levels ranged from 20-40%, with consistent average levels seen in nearly all cow fecal samples (Fig. 4.3). Members of the *Ruminococcaceae* accounted for less than 10% of the sequenced bacterial community in the dilution samples, with levels increasing slightly with dilution factor (Fig. 4.3). The levels of *Bacteroidaceae*, and *Lachnospiraceae* ranged from approximately 2% to 25% across the cow source samples (Fig. 4.2) and were in general similar to each other.

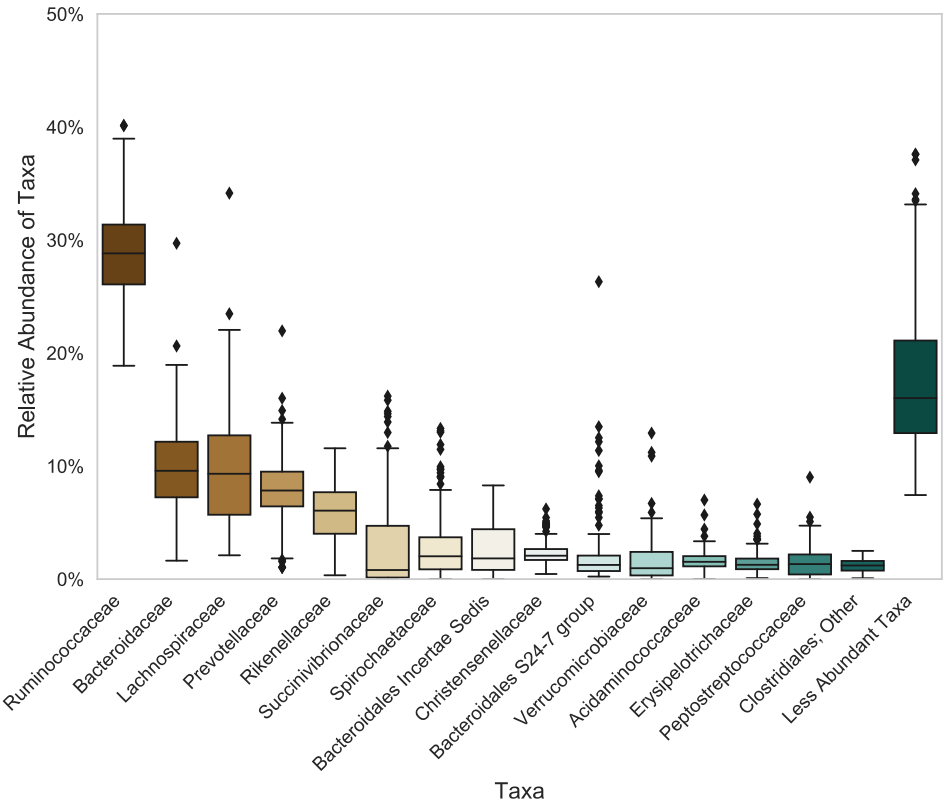

Figure 4.2. Distribution of relative abundances of the fifteen most abundant family-level taxa

The fifteen most abundant taxa were selected from the cow source samples. These taxa were selected from a rarefied OTU table containing 30,000 sequences per sample.

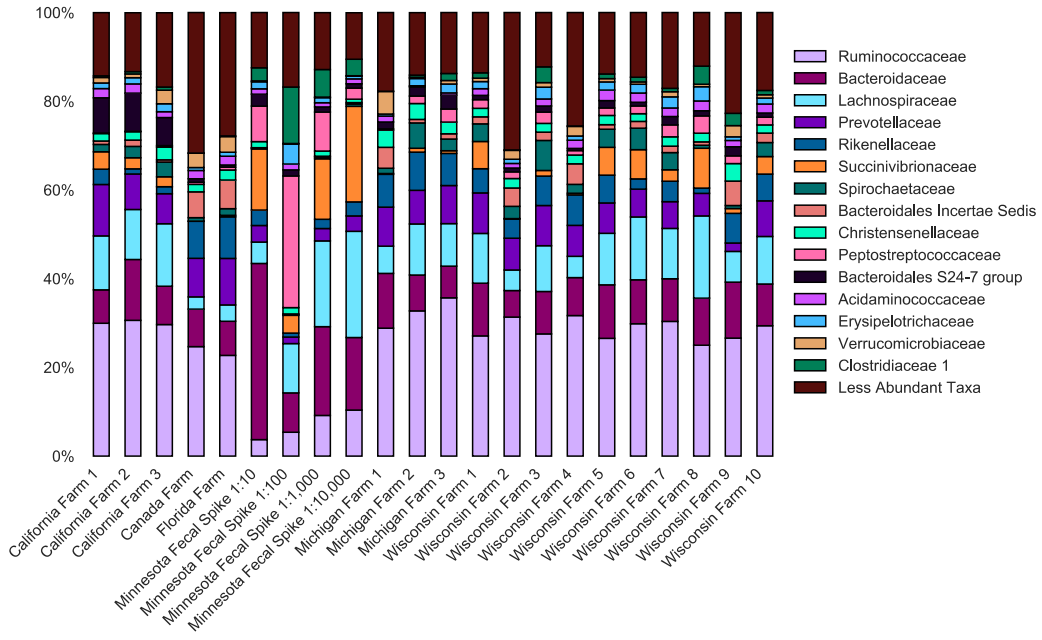

Figure 4.3. Averaged taxonomic bar charts of all cow samples

Distribution of relative abundances of the fifteen most abundant taxa at family-level across all cow source samples.

The effects of age, diet, and location on the bacterial community structure in cattle feces. To evaluate the influence of age, diet and location, principal coordinates analysis was used to visualize relationships amongst variables at the OTU level (Fig. 4.4). The observed clusters were mostly coherent when viewing samples grouped by farm location (Fig. 4.4a and b). While cow samples from close locations tended to cluster together, there were three main clusters, two of which were subdivided by state groups. The Wisconsin samples were located in two clusters, one of which was comprised by fecal spike samples from Minnesota and Michigan samples, and the other comprised of samples from Michigan, Canada and Florida (Fig. 4.4a).

The OTU composition in the cow samples from California were most distinct, clustering apart from other samples from Wisconsin, Florida, Michigan, Canada, and Minnesota (Fig. 4.4a). Cow fecal samples from Wisconsin and Michigan possessed more similar bacterial community structures, at the OTU level, than to those from the fecal spikes prepared from Minnesotan cows (Fig. 4.4a). Pairwise AMOVA comparisons revealed that samples from each state were significantly different (p-value < 0.001) from each other.

Clusters were even more coherent when cow samples were examined by farm. Samples from the same farm tended to cluster together (Fig. 4.4b). When the locations of farms were examined by pairwise AMOVA comparisons, most OTU differences were found to be significant (p-value < 0.001), except for a few farms in Wisconsin.

Cows within this dataset from the different farms had varying ages. When age was evaluated as a possible variable to explain OTU variance, there were no discernable trends observed in this study (Fig. 4.4c). Pairwise AMOVA analyses found no significant differences in OTU abundance or presence (p-value < 0.001) between cows of different ages, except between 1- and 2-year-old cows and 1- and 4-year-old cows. The clusters were also mostly coherent when visualizing samples by diet (Fig. 4.4d). The grain and haylage diet group formed two distinct clusters as did the haylage diet group (Fig. 4.4d). Diets dominated by haylage were generally not significantly different from those rich in alfalfa silage, corn silage, grass/hay, alfalfa, or high energy total mixed rations (Fig. 4.4d).

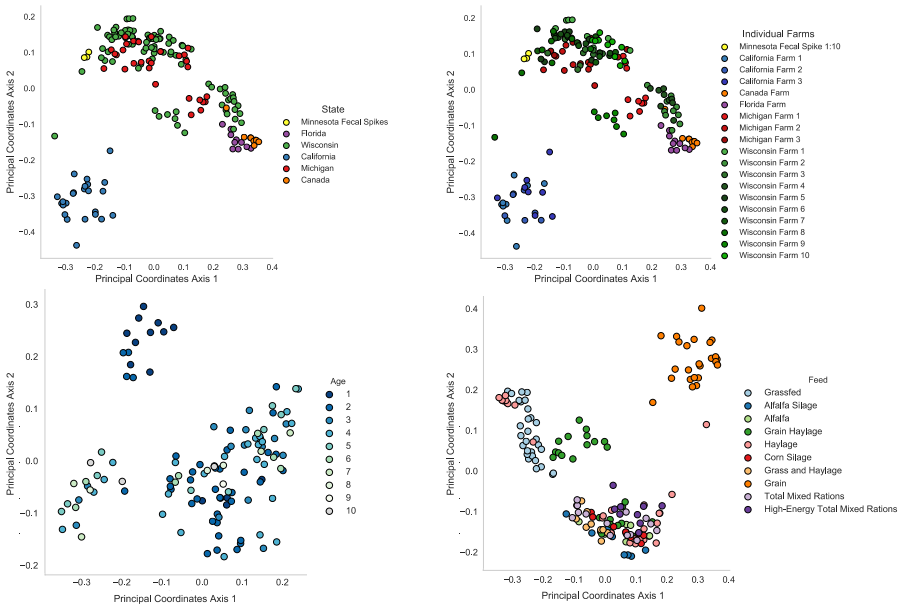

Figure 4.4. Influence of age, diet and location on sample relatedness

Principal coordinates plots of state (a), individual farms (b), age (c) and diet (d) variables. If metadata for samples was missing for certain variables, they were removed from the plotted dataset for that variable.

Influence of geography on SourceTracker predictions. SourceTracker was able to predict that the majority of the Minnesota dilution samples were from cows when source samples from farms across North America were used (Fig. 4.5). All samples had at least a 60% attribution to cow (Fig. 4.5).

Farms from Wisconsin and Michigan consistently had the highest SourceTracker cow predictions when compared to other farms (Fig. 4.5). Predictions of Minnesota cows in spiked samples ranged from 90% using a FTL created with cow feces from Wisconsin and Michigan to 60% for cows from farms in California, Canada, and Florida (Fig. 4.5).

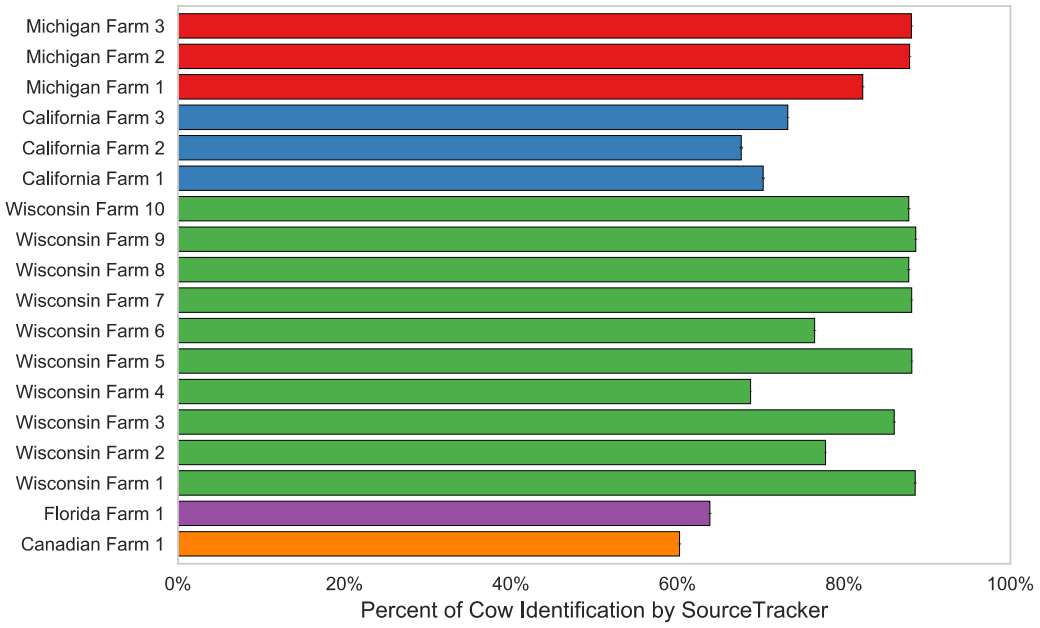

Figure 4.5. SourceTracker analysis of fecal cow spikes from Minnesota

Bar charts showing SourceTracker predictions for each farm using only the most concentrated samples (1:10 diluted samples). Error bars are standard deviation between biological replicates.

Influence of concentration on SourceTracker predictions. The distribution of SourceTracker predictions in each farm was tightly clustered (Fig. 4.6). Distributions tended to become more clustered when SourceTracker predictions were greatest (Fig. 4.6), and the most dilute fecal spike samples had lower SourceTracker predictions, on average (Fig. 4.6).

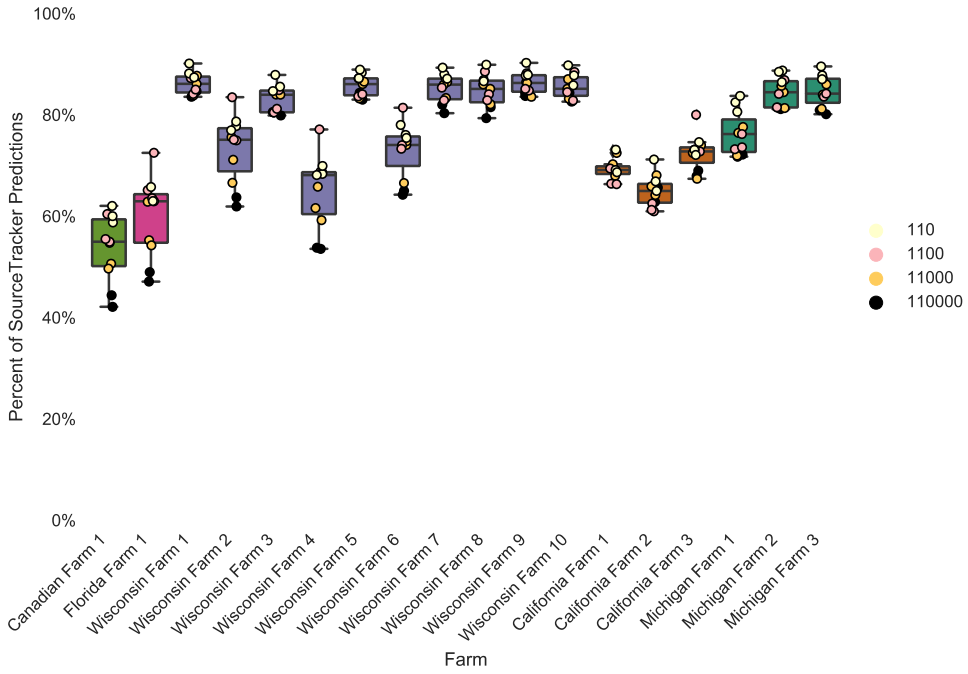

Figure 4.6. Boxplots of SourceTracker predictions

A jitter plot with multiple colored categories shows the influence of concentration of feces on SourceTracker predictions. Dilutions are shown in different colored dots.

DISCUSSION

The approximately 190 cow fecal samples from 19 different farms located across the United States and from Canada contained cow fecal microbiomes that were mostly consistent at the bacterial family level. Previous studies have shown that the cow fecal microbiome is dominated by members of the phyla *Firmicutes* and *Bacteroidetes* (78,148,150). At the family-level, however, there are inconsistent results. While one study reported that the most abundant bacterial families present in cow feces were *Ruminococcaceae*, *Bacteroidaceae*, and *Lachnospiraceae* (130), another study reported that the most dominant families

in cow feces were *Ruminococcaceae*, *Prevotella*, and *Lachnospiraceae* (150). Our results presented here were more consistent with those reported by Shanks and colleagues (150) and *Prevotella* had comparable levels to some of the more abundant family members. The levels of *Ruminococcaceae* in the samples spiked with Minnesota cattle feces were at a much lower abundance when compared to cow fecal samples obtained from across North America. Despite this, however, *Ruminococcaceae* levels were still consistent with previous findings from cow feces from across the United States (150).

The goal of this current study was to investigate how biogeography influences the cow fecal microbiome, and thus predictions of source attribution by the SourceTracker program. The most coherent clustering was observed when samples were grouped by the specific farms they came from, despite age differences and that some cows on the same farm were fed different diets (Michigan Farm 1). While taxonomic families were consistent across cow populations, there were statistically significant differences at the OTU level amongst cows from different states and farms. Lower intra-population variability in cows from the same farms, like what was found in this study, has been previously observed (150).

Moreover, when evaluating how age played a role in shaping the cow fecal microbiome, it was observed that there were minimal statistically significant relationships found between cows of different ages. Previous research on this subject focused on the fecal microbiome changes in the first year of life from pre-weaned calves to adult cows, but not as young adult cows subsequently mature

into older adult cows (152,155). When evaluating whether targeted qPCR-based genetic markers commonly used to identify fecal contamination from cows were detectable, it was found that *E.coli*, *Enterococcus* spp., and *Bacteroidales* levels were significantly different between adult cows and calves (152). Additionally, there were differing abundance trends for the three genetic markers at different times within the calves' lives (152). However, there have been no reported studies, to our knowledge, where the microbiome is explored between young and older adult cows. While our findings could have been influenced by the low numbers of cows in some age categories, it is more likely that the majority of variance in this dataset is explained by the farm to which the cows belong to.

When the effects of fecal concentration on SourceTracker performance were evaluated, we found that predictions for the more diluted samples were lower than those found for the more concentrated samples. Previous studies reported similar results, with SourceTracker being able to detect source contribution changes in a sample (95,154).

When examining cows from different farms that were fed the same diets (haylage, or grain and haylage), there was coherent clustering by farm and state, but not by diet. Interestingly, the cow fecal microbiome from Ontario, Canada mostly clustered with cows from Florida and some from Wisconsin. Shanks and colleagues previously observed that diet provided the most profound influences on the cow fecal microbiome, that were divided based on whether the cows were fed a diet rich in plant material (i.e., silage, corn, alfalfa) or possessed high levels of grains (150). Kim et al, also observed significant taxonomic differences among

cows fed forage versus different total mixed ration diets (148). Since the majority of cows used in this study were fed diets rich in plant materials, this could explain the lack of discernable microbiome differences between these different cattle populations.

Care must be taken, however, when using fecal source samples from locations different from the contaminated area. As was found with previous evaluations of library-dependent MST method sensitivity with concern to geographic variation (44,154), there were on average, higher levels of detection when cows closer to Minnesota were used as sources. Locations farther away from Minnesota, like Florida and California had lower prediction levels. Surprisingly, cow fecal samples from a farm in Ontario, Canada produced SourceTracker predictions comparable to those obtained using source samples from a farm in Florida. Additionally, the microbiomes in cattle from farms in Canada and Florida were observed to be closely related. This suggests that although they were significantly different at the OTU level, they were both more similar to each other when compared to the Minnesota fecal spike samples. This suggests that while geography plays a role in influencing SourceTracker predictions, other variables can be important as well.

CHAPTER 5 : CONCLUSIONS AND FUTURE DIRECTIONS

This thesis explores the use and limitations of community-based microbial source tracking (MST) using SourceTracker for identifying sources of fecal bacteria in waterways. Due to the relative novelty of SourceTracker, few studies had evaluated its limitations as a MST tool before the beginning of this thesis.

Chapter 2 describes a study done in the Lake Superior-St. Louis River estuary using SourceTracker, a program that calculates the source contribution to an environment. High-throughput DNA sequencing analysis of microbiota from a diverse collection of fecal and environmental samples revealed that the community compositions in water and fecal samples were significantly different, allowing for determination of the presence of fecal inputs and identification of specific sources. SourceTracker results indicated that fecal bacterial inputs into the Lake Superior estuary were primarily attributed to wastewater effluent, and to a lesser extent geese and gull wastes. As with previous community-based MST studies, SourceTracker results were logical and corresponded well with metadata (nearby wastewater treatment plant, gull sightings).

Chapter 3 investigated the ability of SourceTracker to correctly determine sources of known fecal contamination with *in situ* mesocosms when different combinations of library configurations were used. The structure and composition of the source profiles SourceTracker used to discern sources were also examined. SourceTracker was able to predict most sources in the *in situ* mesocosms. These results were most reliable when the fecal source library contained only the predicted sources. Sources missing from the library resulted in erroneous classifications not seen when all known sources were present.

Results of this chapter also indicated that the ideal SourceTracker source profile has low-intragroup variability and shares few taxa with other sources.

In Chapter 4, SourceTracker was challenged with identifying spiked cow feces from a farm in St. Paul, Minnesota (MN), USA using fecal source libraries created from cow fecal samples from across the United States and Canada. While cows from across North America were dominated by members of the families *Ruminococcaceae*, *Bacteroidaceae*, and *Lachnospiraceae*, there were statistically significant differences in taxa at the OTU level when evaluating sample relationships by farm, state, age groups, and diet. Most of the OTU variance was attributed to the individual farms the cows came from, and not to age and diet differences. All source samples yielded high SourceTracker predictions. On average, greater predictions in source attribution were associated with more concentrated samples and with cow feces from animals closer to the spiked animal source location. While SourceTracker was able to detect fecal contamination with source animals that were not from the original location, animals sourced closer to the contamination site provided the most predictions.

Not only was SourceTracker able to detect the occurrence of fecal contamination, but it was also able to determine sources that either corresponded well with metadata or were known sources. In Chapter 2, SourceTracker identified wastewater effluent as the predominant source nearby a wastewater treatment plant. In addition, previous research using other MST methods also found waterfowl and wastewater effluent to be major sources in the area. In Chapter 3, SourceTracker was mostly able to identify wastewater

effluent, cow and horse feces when they were used in different mesocosms when the known sources were present in the fecal source library. This provides more evidence for SourceTracker's ability to be used as a tool for community-based MST.

In Chapter 3, using only the known sources in the fecal source library yielded the most expected SourceTracker predictions. In Chapter 4, only cow was used as a source in the fecal source library. The higher source predictions would have most likely been negatively impacted by using other fecal sources (i.e., waterfowl, wastewater effluent).

There are still large knowledge gaps about several aspects of community-based MST with SourceTracker. In the past two years, studies have begun exploring the limitations of this program. Part of this work addressed how age, diet, the concentration of feces and geography impacted SourceTracker predictions for cattle populations across North America. However, other animals' feces also harbor pathogens that pose health risks and are worth investigation. Animal sources like waterfowl or chickens should have similar work performed to evaluate how these factors affect SourceTracker's accuracy. In previous MST methods, libraries for certain animals were found to be geographically-stable while geographically-specific for others. The sharing of fecal source libraries across certain regions, or animals could further facilitate the use of this method.

Future studies exploring community-based MST with SourceTracker should explore the stability of fecal source libraries over time. Studies addressing this should pay attention to how the gut microbiomes of the animal or wastewater

treatment sources change over time. These studies should include challenging SourceTracker to identify known sources with fecal source libraries that have different ages.

Currently, the inability to obtain absolute quantification of fecal pollution sources when performing community-based MST with SourceTracker is a significant drawback. Future work should involve finding complementary methods to allow absolute quantification of fecal pollution for proper health risk evaluations.

Additionally, new sequencing technology such as the Oxford Nanopore is allowing ultra-rapid sequencing. The MinION Nanopore sequencer could allow instantaneous community analysis of potentially contaminated water samples. This method could allow recently captured DNA sequencing results from water samples to be compared against a reference fecal source library with SourceTracker.

BIBLIOGRAPHY

1. WHO. Preventing diarrhoea through better water, sanitation and hygiene. WHO Libr Cat Data [Internet]. 2014;1–48. Available from: http://apps.who.int/iris/bitstream/10665/150112/1/9789241564823_eng.pdf

2. Pruss A, Kay D, Fewtrell L, Bartram J. Estimating the Burden of Disease from Water, Sanitation, Hygene at a Global Level. Environ Health Perspect [Internet]. 2002;110(5):537. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=6766240&site=ehost-live

3. German RR, Lee LM, Horan JM, Milstein RL, Pertowski CA, Waller MN. Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. MMWR Recomm Rep [Internet]. 2001;50(RR-13):1-35-7. Available from: http://www.columbia.edu/itc/hs/pubhealth/p8475/readings/cdc-updated-guidelines.pdf

4. Corso PS, Kramer MH, Blair KA, Addiss DG, Davis JP, Haddix AC. Cost of illness in the 1993 waterborne Cryptosporidium outbreak, Milwaukee, Wisconsin. Emerg Infect Dis. 2003;9(4):426–31.

5. Auld H, MacIver D, Klaassen J. Heavy rainfall and waterborne disease outbreaks: The Walkerton example. J Toxicol Environ Heal - Part A. 2004;67(20–22):1879–87.

6. Hall HI, An Q, Tang T, Song R, Chen M, Green T, et al. Prevalence of Diagnosed and Undiagnosed HIV Infection--United States, 2008-2012. MMWR Morb Mortal Wkly Rep [Internet]. 2015;64(24):657–80. Available

from: http://www.ncbi.nlm.nih.gov/pubmed/26110835

7. Eisenberg JNS, Bartram J, Wade TJ. Perspectives | Brief Communication The Water Quality in Rio Highlights the Global Public Health Concern. 2016;(10):180–1.

8. Delahoy MJ, Wodnik B, McAliley L, Penakalapati G, Swarthout J, Freeman MC, et al. Pathogens Transmitted in Animal Feces in Low- and Middle-Income Countries. Int J Hyg Environ Health [Internet]. 2018;(December 2017):1–16. Available from: https://www.sciencedirect.com/science/article/pii/S1438463917308635

9. Leclerc H, Schwartzbrod L, Leclerc H, Schwartzbrod L. Microbial Agents Associated with Waterborne Diseases Microbial Agents Associated with Waterborne. 2008;7828(4):371–409.

10. Whiley H, van den Akker B, Giglio S, Bentham R. The role of environmental reservoirs in human campylobacteriosis. Int J Environ Res Public Health. 2013;10(11):5886–907.

11. Field KG, Samadpour M. Fecal source tracking, the indicator paradigm, and managing water quality. Water Res. 2007;41(16):3517–38.

12. Ashbolt N, Grabow W, Snozzi M. Indicators of microbial water quality. Water Qual Guidel Stand Heal [Internet]. 2001;(Grabow 1996):289–316. Available from: www.who.int/water_sanitation_health/dwq/whoiwa/en/

13. Whitacre D. Reviews of Environmental Contamination and Toxicology. Vol. 192, Reviews of enviornmental contamination and toxicology. 2008. 1-212

p.

14. Schuettpelz DH. Fecal and Total Coliform Tests in Water Quality Evaluation. 1969;1–25.

15. Malcolm JF. The classification of coliform bacteria. J Hyg (Lond). 1938;38(4):395–423.

16. Leclerc H, Mossel DAA, Edberg SC, Struijk CB. Advances in the Bacteriology of the Coliform Group: Their Suitability as Markers of Microbial Water Safety. Annu Rev Microbiol [Internet]. 2001;55(1):201–34. Available from: http://www.annualreviews.org/doi/10.1146/annurev.micro.55.1.201

17. USEPA. Ambient Water Quality Criteria for Bacteria. U.S. Environmental Protection Agency Report EPA-440/5-84-002. 1986.

18. Doyle MP, Erickson MC. Closing the Door on the Fecal Coliform Assay. Microbe [Internet]. 2006;1(4):162–3. Available from: http://www.asmscience.org/content/view.action

19. EPA. Recreational Water Quality Criteria. U S Environ Prot Agency. 2012;1–69.

20. Le H, Fairchild ROY. Streptococci Characteristic of Sewage and Sewage-Polluted Waters Apparently not Hitherto Reported in America Author (s): C . -E . A . Winslow and M . P . Hunnewell Published by : American Association for the Advancement of Science Stable URL : http://ww. 2018;15(386):827–9.

21. Taylor P. Faecal streptococci as faecal pollution indicators : A review . Part I : Taxonomy and enumeration Faecal streptococci as faecal pollution indicators : a review . Part I : Taxonomy and enumeration. 2010;8330(May 2011).

22. Kenner B, Clark H, Kabler P. Fecal Streptococci. I. Cultivation and enumeration of Streptococci in surface waters. Appl Environ Microbiol. 1961;9(table 4):15–20.

23. Paredes-Sabja D, Torres JA, Setlow P, Sarker MR. Clostridium perfringens spore germination: Characterization of germinants and their receptors. J Bacteriol. 2008;190(4):1190–201.

24. Mueller-Spitz SR, Stewart LB, Val Klump J, McLellan SL. Freshwater suspended sediments and sewage are reservoirs for enterotoxin-positive clostridium perfringens. Appl Environ Microbiol. 2010;76(16):5556–62.

25. Payment P, Franco E. Clostridium perfringens and somatic coliphages as indicators of the efficiency of drinking water treatment for viruses and protozoan cysts. Appl Environ Microbiol. 1993;59(8):2418–24.

26. Anderson KL, Whitlock JE, Harwood VJ. Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments. Appl Environ Microbiol. 2005;71(6):3041–8.

27. Harwood VJ, Levine AD, Scott TM, Chivukula V, Lukasik J, Farrah SR, et al. Validity of the Indicator Organism Paradigm for Pathogen Reduction in Reclaimed Water and Public Health Protection Validity of the Indicator

Organism Paradigm for Pathogen Reduction in Reclaimed Water and Public Health Protection †. Appl Environ Microbiol. 2005;71(6):3163.

28. Bonadonna L, Briancesco R, Ottaviani M, Veschetti E. Occurrence of Cryptosporidium oocysts in sewage effluents and correlation with microbial, chemical and physical water variables. Environ Monit Assess. 2002;75(3):241–52.

29. Havelaar AH, Van Olphen M, Drost YC. F-specific RNA bacteriophages are adequate model organisms for enteric viruses in fresh water. Appl Environ Microbiol. 1993;59(9):2956–62.

30. USEPA. Method 1611.1: Enterococci in Water by TaqMan® Quantitative Polymerase Chain Reaction (qPCR). 2015;(April):41.

31. Chern EC, Siefring S, Paar J, Doolittle M, Haugland RA. Comparison of quantitative PCR assays for Escherichia coli targeting ribosomal RNA and single copy genes. Lett Appl Microbiol. 2011;52(3):298–306.

32. WHO. Guidelines for drinking-water quality. 2017. 631 p.

33. Tallon P, Magajna B, Lofranco C, Kam TL. Microbial indicators of faecal contamination in water: A current perspective. Water Air Soil Pollut. 2005;166(1–4):139–66.

34. Gundry S, Wright J, Conroy R, others. A systematic review of the health outcomes related to household water quality in developing countries. J Water Health [Internet]. 2004;2:1–14. Available from: http://www.bristol.ac.uk/aquapol%5Cnhttp://www.iwaponline.com/jwh/002/0

001/0020001.pdf

35. Gruber JS, Ercumen A, Colford JM. Coliform bacteria as indicators of diarrheal risk in household drinking water: Systematic review and meta-analysis. PLoS One. 2014;9(9).

36. Ishii S, Ksoll WB, Hicks RE, Sadowsky MJ. Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds. Appl Environ Microbiol [Internet]. 2006 Jan;72(1):612–21. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1352292&tool=pmcentrez&rendertype=abstract

37. Meays CL, Broersma K, Nordin R, Mazumder A. Source tracking fecal bacteria in water: a critical review of current methods. J Environ Manage [Internet]. 2004 Oct [cited 2014 Aug 29];73(1):71–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15327848

38. Hagedorn C, Blanch AR, Harwood VJ. Microbial Source Tracking: Methods, Applications, and Case Studies. Springer US; 2011. 645 p.

39. Santo Domingo JW, Bambic DG, Edge TA, Wuertz S. Quo vadis source tracking? Towards a strategic framework for environmental monitoring of fecal pollution. Water Res. 2007;41(16):3539–52.

40. Hagedorn C, Crozier JB, Mentz KA, Booth AM, Graves AK, Nelson NJ, et al. Carbon source utilization profiles as a method to identify sources of faecal pollution in water. J Appl Microbiol. 2003;94(5):792–9.

41. Griffith JF, Weisberg SB, McGee CD. Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. J Water Heal. 2003;1(4):141–51.

42. Parveen S, Portier KM, Robinson K, Edmiston L, Tamplin ML. Discriminant analysis of ribotype profiles of *Escherichia coli* for differentiating human and non-human sources of fecal pollution. Appl Envir Microbiol [Internet]. 1999;65(7):3142–7. Available from: http://aem.asm.org/content/65/7/3142.short

43. Parveen S, Hodge NC, Stall RE, Farrah SR, Tamplin ML. Phenotypic and genotypic characterization of human and nonhuman Escherichia coli. Water Res. 2001;35(2):379–86.

44. Scott TM, Rose JB, Jenkins TM, Samuel R, Lukasik J, Farrah SR. Microbial Source Tracking : Current Methodology and Future Directions Microbial Source Tracking : Current Methodology and Future Directions †. Appl Environ Microbiol. 2002;68(12):5796–803.

45. Simpson JM, Domingo JWS, Reasoner DJ. Critical Review Microbial Source Tracking : State of the Science. 2002;36(24):5279–88.

46. Reischer GH, Ebdon JE, Bauer JM, Schuster N, Ahmed W, Åström J, et al. Performance characteristics of qPCR assays targeting human- and ruminant-associated bacteroidetes for microbial source tracking across sixteen countries on six continents. Environ Sci Technol. 2013;47(15):8548–56.

47. Boehm AB, Van De Werfhorst LC, Griffith JF, Holden PA, Jay JA, Shanks OC, et al. Performance of forty-one microbial source tracking methods: a twenty-seven lab evaluation study. Water Res [Internet]. 2013 Nov 15 [cited 2014 Oct 16];47(18):6812–28. Available from: http://www.sciencedirect.com/science/article/pii/S0043135413005496

48. Wiggins BA, Andrews RW, Conway RA, Corr CL, Dobratz EJ, Dougherty DP, et al. Use of antibiotic resistance analysis to identify nonpoint sources of fecal pollution. Appl Environ Microbiol. 1999;65(8):3483–6.

49. Seurinck S, Deschepper E, Deboch B, Verstraete W, Siciliano S. Characterization of Escherichia coli isolates from different fecal sources by means of classification tree analysis of fatty acid methyl ester (FAME) profiles. Environ Monit Assess. 2006;114(1–3):433–45.

50. Stoeckel DM, Mathes M V., Hyer KE, Hagedorn C, Kator H, Lukasik J, et al. Comparison of seven protocols to identify fecal contamination sources using Escherichia coli. Environ Sci Technol. 2004;38(22):6109–17.

51. Wheeler AL, Hartel PG, Godfrey DG, Hill JL, Segars WI. Potential of Enterococcus faecalis as a human fecal indicator for microbial source tracking. J Environ Qual. 1991;31(4):1286–93.

52. Sadowsky MJ, Kinkel LL, Bowers JH, Schottel JL. Use of repetitive intergenic DNA sequences to classify pathogenic and disease-suppressive Streptomyces strains. Appl Environ Microbiol [Internet]. 1996 Sep [cited 2014 Aug 29];62(9):3489–93. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=168149&tool=p

mcentrez&rendertype=abstract

53. Bernhard AE, Field KG. A PCR Assay To Discriminate Human and Ruminant Feces on the Basis of Host Differences in Bacteroides-Prevotella Genes Encoding 16S rRNA A PCR Assay To Discriminate Human and Ruminant Feces on the Basis of Host Differences in Bacteroides-Prevotella Genes E. 2000;

54. Green HC, Dick LK, Gilpin B, Samadpour M, Field KG. Genetic markers for rapid PCR-based identification of gull, Canada goose, duck, and chicken fecal contamination in water. Appl Environ Microbiol. 2012;78(2):503–10.

55. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, et al. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 2013;4(12):1111–9.

56. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods [Internet]. 2011 Sep 17 [cited 2014 Jul 16];8(9):761–3. Available from: http://www.nature.com.ezp2.lib.umn.edu/nmeth/journal/v8/n9/full/nmeth.1650.html

57. Rappé MS, Giovannoni SJ. The Uncultured Microbial Majority. Annu Rev Microbiol [Internet]. 2003;57(1):369–94. Available from: http://www.annualreviews.org/doi/10.1146/annurev.micro.57.030502.090759

58. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007;449(7164):804–10.

59. Rastogi G, Coaker GL, Leveau JHJ. New insights into the structure and function of phyllosphere microbiota through high-throughput molecular approaches. FEMS Microbiol Lett. 2013;348(1):1–10.

60. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci [Internet]. 1977;74(11):5088–90. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.74.11.5088

61. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci [Internet]. 1977;74(12):5463–7. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.74.12.5463

62. Giovannoni SJ, DeLong EF, Schnmidt TM, Pace NR. Tangential flow filtration and preliminary phylogenetic analysis of marine picoplankton. Appl Environ Microbiol. 1990;56(8):2572–5.

63. Clarridge JE, Alerts C. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin Microbiol Rev. 2004;17(4):840–62.

64. McCabe KM, Zhang Y-H, Huang B-L, Wagar EA, McCabe ERB. Bacterial Species Identification after DNA Amplification with a Universal Primer Pair. Mol Genet Metab [Internet]. 1999;66(3):205–11. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1096719298927950

65. Baker GC, Smith JJ, Cowan DA. Review and re-analysis of domain-

specific 16S primers. J Microbiol Methods. 2003;55(3):541–55.

66. Van de Peer Y. A quantitative map of nucleotide substitution rates in bacterial rRNA. Nucleic Acids Res [Internet]. 1996;24(17):3381–91. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/24.17.3381

67. Chakravorty S, Helb D, Burday M, Connell N. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods. 2007;69(2):330–9.

68. Schloss PD, Handelsman J. Introducing DOTUR , a Computer Program for Defining Opera : onal Taxonomic Units and Es : ma : ng Species Richness. Appl Environ Microbiol. 2005;71(3):1501–6.

69. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, et al. The Ribosomal Database Project (RDP-II): Sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. 2005;33(DATABASE ISS.).

70. DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, et al. NAST: A multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res. 2006;34(WEB. SERV. ISS.):394–9.

71. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high- throughput community sequencing data. Nat Publ Gr [Internet]. 2010;7(5):335–6. Available from: http://dx.doi.org/10.1038/nmeth0510-335

72. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol [Internet]. 2009 Dec [cited 2014 Jul 9];75(23):7537–41. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2786419&tool=pmcentrez&rendertype=abstract

73. Caporaso JG, Bittinger K, Bushman FD, Desantis TZ, Andersen GL, Knight R. PyNAST: A flexible tool for aligning sequences to a template alignment. Bioinformatics. 2010;26(2):266–7.

74. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics [Internet]. 2011 Aug 15 [cited 2014 Jul 10];27(16):2194–200. Available from: http://bioinformatics.oxfordjournals.org/content/27/16/2194

75. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res [Internet]. 2014 Jan [cited 2014 Jul 10];42(Database issue):D633-42. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965039&tool=pmcentrez&rendertype=abstract

76. Dowd SE, Callaway TR, Wolcott RD, Sun Y, McKeehan T, Hagevoort RG, et al. Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP).

BMC Microbiol [Internet]. 2008;8(1):125. Available from: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/1471-2180-8-125

77. McLellan SL, Huse SM, Mueller-Spitz SR, Andreihcheva EN, Sogin ML. Diversity and Population Structure of Sewage Derived Microorganisms in Wastewater Treatment Plant Influent. Environ Microbiol. 2011;12(2):378–92.

78. Unno T, Jang J, Han D, Kim JH, Sadowsky MJ, Kim O-S, et al. Use of barcoded pyrosequencing and shared OTUs to determine sources of fecal bacteria in watersheds. Environ Sci Technol [Internet]. 2010 [cited 2014 Aug 29];44(20):7777–82. Available from: http://pubs.acs.org/doi/full/10.1021/es101500z

79. Flores GE, Bates ST, Knights D, Lauber CL, Stombaugh J, Knight R, et al. Microbial biogeography of public restroom surfaces. PLoS One. 2011;6(11).

80. Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet [Internet]. 2014;46(4):336–44. Available from: http://www.nature.com/doifinder/10.1038/ng.2906

81. Longo MS, O'Neill MJ, O'Neill RJ. Abundant human DNA contamination identified in non-primate genome databases. PLoS One. 2011;6(2):1–4.

82. Newton RJ, Bootsma MJ, Morrison HG, Sogin ML, McLellan SL. A Microbial Signature Approach to Identify Fecal Pollution in the Waters Off

an Urbanized Coast of Lake Michigan. Microb Ecol. 2013;65(4):1011–23.

83. Shanks OC, Newton RJ, Kelty CA, Huse SM, Sogin ML, McLellan SL. Comparison of the microbial community structures of untreated wastewaters from different geographic locales. Appl Environ Microbiol. 2013;79(9):2906–13.

84. Ahmed W, Staley C, Sadowsky MJ, Gyawali P, Sidhu J, Palmer A, et al. Toolbox approaches using molecular markers and 16S rRNA gene amplicon data sets for identification of fecal pollution in surface water. Appl Environ Microbiol. 2015;81(20):7067–77.

85. Neave M, Luter H, Padovan A, Townsend S, Schobben X, Gibb K. Multiple approaches to microbial source tracking in tropical northern Australia. Microbiol Open. 2014;3(6):860–74.

86. Staley C, Reckhow KH, Lukasik J, Harwood VJ. Assessment of sources of human pathogens and fecal contamination in a Florida freshwater lake. Water Res [Internet]. 2012 Nov 1 [cited 2014 Aug 18];46(17):5799–812. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22939220

87. Ebentier DL, Hanley KT, Cao Y, Badgley BD, Boehm AB, Ervin JS, et al. Evaluation of the repeatability and reproducibility of a suite of qPCR-based microbial source tracking methods. Water Res [Internet]. 2013 Nov 15 [cited 2014 Aug 13];47(18):6839–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23911226

88. Ryu H, Griffith JF, Khan IUH, Hill S, Edge TA, Toledo-Hernandez C, et al.

Comparison of gull feces-specific assays targeting the 16S rRNA Genes of Catellicoccus marimammalium and Streptococcus spp. Appl Environ Microbiol. 2012;78(6):1909–16.

89. Shanks OC, White K, Kelty CA, Hayes S, Sivaganesan M, Jenkins M, et al. Performance assessment PCR-based assays targeting Bacteroidales genetic markers of bovine fecal pollution. Appl Environ Microbiol. 2010;76(5):1359–66.

90. Odagiri M, Schriewer A, Hanley K, Wuertz S, Misra PR, Panigrahi P, et al. Validation of Bacteroidales quantitative PCR assays targeting human and animal fecal contamination in the public and domestic domains in India. Sci Total Environ [Internet]. 2015;502:462–70. Available from: http://dx.doi.org/10.1016/j.scitotenv.2014.09.040

91. Ahmed W, Staley C, Hamilton KA, Beale DJ, Sadowsky MJ, Toze S, et al. Amplicon-based taxonomic characterization of bacteria in urban and peri-urban roof-harvested rainwater stored in tanks. Sci Total Environ [Internet]. 2017;576:326–34. Available from: http://dx.doi.org/10.1016/j.scitotenv.2016.10.090

92. Handl S, Dowd SE, Garcia-Mazcorro JF, Steiner JM, Suchodolski JS. Massive parallel 16S rRNA gene pyrosequencing reveals highly diverse fecal bacterial and fungal communities in healthy dogs and cats. FEMS Microbiol Ecol. 2011;76(2):301–10.

93. Staley C, Unno T, Gould TJ, Jarvis B, Phillips J, Cotner JB, et al. Application of Illumina next-generation sequencing to characterize the

bacterial community of the Upper Mississippi River. J Appl Microbiol [Internet]. 2013 Nov [cited 2014 Jul 14];115(5):1147–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23924231

94. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods [Internet]. 2011 Sep 17 [cited 2014 Jul 16];8(9):761–3. Available from: http://www.nature.com.ezp1.lib.umn.edu/nmeth/journal/v8/n9/full/nmeth.1650.html

95. Henry R, Schang C, Coutts S, Kolotelo P, Prosser T, Crosbie N, et al. Into the deep: Evaluation of SourceTracker for assessment of faecal contamination of coastal waters. Water Res. 2016;93:242–53.

96. la Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, et al. Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. PLoS One. 2012;7(12):1–13.

97. Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. PLoS One. 2012;7(2).

98. Lapara TM, Burch TR, McNamara PJ, Tan DT, Yan M, Eichmiller JJ. Tertiary-Treated Municipal Wastewater is a Significant Point-Source of Antibiotic Resistance Genes into Duluth-Superior Harbor. Environ Sci Technol [Internet]. 2011;45(22):9543–9. Available from: http://dx.doi.org/10.1021/es202775r

99. Eichmiller JJ, Hicks RE, Sadowsky MJ. Distribution of genetic markers of fecal pollution on a freshwater sandy shoreline in proximity to wastewater effluent. Environ Sci Technol [Internet]. 2013 Apr 2 [cited 2014 Aug 22];47(7):3395–402. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3629727&tool=pmcentrez&rendertype=abstract

100. Staley C, Gould TJ, Wang P, Phillips J, Cotner JB, Sadowsky MJ. Evaluation of water sampling methodologies for amplicon-based characterization of bacterial community structure. TL - 114. J Microbiol Methods [Internet]. 2015;114 VN-:43–50. Available from: file:///Users/JCThrash/Documents/ReadCube Media/Staley - Evaluation of water sampling methodologies for amplicon-based characterization of bacterial community structure..pdf%5Cnhttp://dx.doi.org/10.1016/j.mimet.2015.05.003

101. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 2007;35(18).

102. Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, et al. Microbial population structures in the deep marine biosphere. Science [Internet]. 2007;318(5847):97–100. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17916733

103. Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, et al. Systematic improvement of amplicon marker gene methods for increased

accuracy in microbiome studies. Nat Biotech [Internet]. 2016 Sep;34(9):942–9. Available from: http://dx.doi.org/10.1038/nbt.3601

104. Bolger a. M, Lohse M, Usadel B. Trimmomatic: A flexible read trimming tool for Illumina NGS data. Bioinformatics [Internet]. 2014;30(15):2114–20. Available from: http://bioinformatics.oxfordjournals.org/content/30/15/2114

105. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics [Internet]. 2012;13(1):31. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-31

106. Aronesty E. Comparison of sequencing utility programs. Open Bioinforma J. 2013;31(7):1–8.

107. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1.

108. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res [Internet]. 2013 Jan [cited 2014 Jul 10];41(Database issue):D590-6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531112&tool=pmcentrez&rendertype=abstract

109. R Core team. R Core Team [Internet]. Vol. 55, R: A Language and Environment for Statistical Computing. R Foundation for Statistical

Computing , Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. 2015. p. 275–86. Available from: http://www.mendeley.com/research/r-language-environment-statistical-computing-96/%5Cnpapers2://publication/uuid/A1207DAB-22D3-4A04-82FB-D4DD5AD57C28

110. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol Monogr [Internet]. 1957;27(4):325–49. Available from: http://www.jstor.org/stable/1942268

111. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics. 1992;131(2):479–91.

112. Suzuki R, Shimodaira H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22(12):1540–2.

113. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol. 2008;6:776–88.

114. Biddle A, Stewart L, Blanchard J, Leschine S. Untangling the genetic basis of fibrolytic specialization by lachnospiraceae and ruminococcaceae in diverse gut communities. Diversity. 2013;5(3):627–40.

115. Lu J, Domingo JS. Turkey fecal microbial community structure and functional gene diversity revealed by 16S rRNA gene and metagenomic sequences. J Microbiol. 2008;46(5):469–77.

116. Wei S, Morrison M, Yu Z. Bacterial census of poultry intestinal microbiome. Poult Sci [Internet]. 2013;92(3):671–83. Available from: http://ps.fass.org/content/92/3/671.abstract

117. Videnska P, Faldynova M, Juricova H, Babak V, Sisak F, Havlickova H, et al. Chicken faecal microbiota and disturbances induced by single or repeated therapy with tetracycline and streptomycin. BMC Vet Res [Internet]. 2013;9(1):30. Available from: http://bmcvetres.biomedcentral.com/articles/10.1186/1746-6148-9-30

118. Li Q, Lauber CL, Czarnecki-Maulden G, Pan Y, Hannah SS. Effects of the dietary protein and carbohydrate ratio on gut microbiomes in dogs of different body conditions. MBio. 2017;8(1):1–14.

119. Gruninger RJ, McAllister TA, Forster RJ. Bacterial and archaeal diversity in the gastrointestinal tract of the North American beaver (Castor canadensis). PLoS One. 2016;11(5):1–17.

120. Eshar D, Weese JS. Molecular analysis of the microbiota in hard feces from healthy rabbits ({Oryctolagus} cuniculus) medicated with long term oral meloxicam. BMC Vet Res [Internet]. 2014;10:62. Available from: http://apps.webofknowledge.com/full_record.do?page=4&excludeEventConfig=ExcludeIfFromFullRecPage&qid=157&log_event=yes&viewType=fullRecord&SID=Z2Y2YGVgZd3Cb4dUqge&product=UA&doc=197&search_mode=Refine%5Cnhttp://apps.webofknowledge.com/full_record.do?produ

121. Niu Q, Li P, Hao S, Zhang Y, Kim SW, Li H, et al. Dynamic Distribution of the Gut Microbiota and the Relationship with Apparent Crude Fiber

Digestibility and Growth Stages in Pigs. Sci Rep [Internet]. 2015;5(1):9938. Available from: http://www.nature.com/articles/srep09938

122. Eichmiller JJ, Hicks RE, Sadowsky MJ. Distribution of genetic markers of fecal pollution on a freshwater sandy shoreline in proximity to wastewater effluent. Environ Sci Technol. 2013;47(7):3395–402.

123. Ishii S, Hansen DL, Hicks RE, Sadowsky MJ. Beach sand and sediments are temporal sinks and sources of Escherichia coli in Lake Superior. Environ Sci Technol [Internet]. 2007 Apr 1;41(7):2203–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17438764

124. Toze S. Reuse of effluent water - Benefits and risks. Agric Water Manag. 2006;80(1–3 SPEC. ISS.):147–59.

125. Stoeckel DM, Harwood VJ. Performance, design, and analysis in microbial source tracking studies. Appl Environ Microbiol. 2007;73(8):2405–15.

126. Stewart JR, Ellender RD, Gooch JA, Jiang S, Myoda SP, Weisberg SB. Recommendations for microbial source tracking: lessons from a methods comparison study. J Water Heal [Internet]. 2003;1(4):225–31. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15382726

127. Eichmiller JJ, Borchert AJ, Sadowsky MJ, Hicks RE. Decay of genetic markers for fecal bacterial indicators and pathogens in sand from Lake Superior. Water Res [Internet]. 2014 Aug 1 [cited 2014 Jul 31];59:99–111.

Available from:
http://www.sciencedirect.com/science/article/pii/S0043135414002887

128. Newton RJ, VandeWalle JL, Borchardt MA, Gorelick MH, McLellan SL. Lachnospiraceae and bacteroidales alternative fecal indicators reveal chronic human sewage contamination in an Urban harbor. Appl Environ Microbiol. 2011;77(19):6972–81.

129. Green HC, Haugland RA, Varma M, Millen HT, Borchardt MA, Field KG, et al. Improved HF183 quantitative real-time PCR assay for characterization of human fecal pollution in ambient surface water samples. Appl Environ Microbiol. 2014;80(10):3086–94.

130. Brown C, Staley C, Wang P, Dalzell B, Chun CL, Sadowsky M. A High-Throughput DNA Sequencing-Based Approach for Determining Sources of Fecal Bacteria in the Lake Superior Watershed. Environ Sci Technol. 2017;51:8263–71.

131. Newton RJ, McLellan SL. A unique assemblage of cosmopolitan freshwater bacteria and higher community diversity differentiate an urbanized estuary from oligotrophic Lake Michigan. Front Microbiol. 2015;6(SEP):1–13.

132. Soller JA, Schoen ME, Bartrand T, Ravenscroft JE, Ashbolt NJ. Estimated human health risks from exposure to recreational waters impacted by human and non-human sources of faecal contamination. Water Res [Internet]. 2010;44(16):4674–91. Available from: http://dx.doi.org/10.1016/j.watres.2010.06.049

133. Korajkic A, McMinn BR, Shanks OC, Sivaganesan M, Fout GS, Ashbolt NJ. Biotic interactions and sunlight affect persistence of fecal indicator bacteria and microbial source tracking genetic markers in the upper mississippi river. Appl Environ Microbiol. 2014;80(13):3952–61.

134. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J [Internet]. 2012 Aug [cited 2014 Jul 9];6(8):1621–4. Available from: http://dx.doi.org/10.1038/ismej.2012.8

135. R Core Team. R: A language and environment for statistical computing. R Found Stat Comput [Internet]. 2013; Available from: http://www.r-project.org/

136. Anderson MJ. Distance-based tests for homogeneity of multivariate dispersions. Biometrics. 2006;62(1):245–53.

137. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci [Internet]. 2003 Dec 9 [cited 2014 Dec 29];14(6):927–30. Available from: http://doi.wiley.com/10.1111/j.1654-1103.2003.tb02228.x

138. Levene H. Robust tests for equality of variances. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford Univ Press. 1960;278–92.

139. Ma L, Mao G, Liu J, Yu H, Gao G, Wang Y. Rapid quantification of bacteria and viruses in influent, settled water, activated sludge and effluent from a wastewater treatment plant using flow cytometry. Water Sci Technol.

2013;68(8):1763–9.

140. Dutton RJ, Bitton G, Koopman B. Malachite green-INT (MINT) method for determining active bacteria in sewage. Appl Environ Microbiol. 1983;46(6):1263–7.

141. Franks AH, Harmsen HJM, Raangs GC, Jansen GJ, Schut F, Welling GW. Variations of bacterial populations in human feces measured by fluorescent in situ hybridization with group-specific 16S rRNA-targeted oligonucleotide probes. Appl Environ Microbiol. 1998;64(9):3336–45.

142. Hartel PG, Summer JD, Hill JL, Collins J V, Entry JA, Segars WI. Geographic variability of Escherichia coli ribotypes from animals in Idaho and Georgia. JEnvironQual. 2002;31(4):1273–8.

143. Hamilton MJ, Yan T, Sadowsky MJ. Development of goose- and duck-specific DNA markers to determine sources of Escherichia coli in waterways. Appl Environ Microbiol. 2006;72(6):4012–9.

144. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature [Internet]. 2014;505(7484):559–63. Available from: http://dx.doi.org/10.1038/nature12820

145. Turnbaugh PJ. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med. 2009;1(6).

146. Faith JJ, Mcnulty NP, Rey FE, Gordon JI. Response to Diet in Gnotobiotic

Mice. Science (80-). 2011;333(July):101–5.

147. Metcalf JL, Song SJ, Morton JT, Weiss S, Seguin-Orlando A, Joly F, et al. Evaluating the impact of domestication and captivity on the horse gut microbiome. Sci Rep. 2017;7(1):1–9.

148. Kim M, Kim J, Kuehn LA, Bono JL, Berry ED, Kalchayanand N, et al. Investigation of bacterial diversity in the feces of cattle fed different diets Investigation of bacterial diversity in the feces of cattle fed different diets. J Anim Sci [Internet]. 2014;92(April):683–94. Available from: http://www.journalofanimalscience.org/content/92/2/683%5Cnwww.asas.org%5Cnwww.journalofanimalscience.org

149. Dougal K, De La Fuente G, Harris PA, Girdwood SE, Pinloche E, Geor RJ, et al. Characterisation of the faecal bacterial community in adult and elderly horses fed a high fibre, high oil or high starch diet using 454 pyrosequencing. PLoS One. 2014;9(2).

150. Shanks OC, Kelty CA, Archibeque S, Jenkins M, Newton RJ, McLellan SL, et al. Community structures of fecal bacteria in cattle from different animal feeding operations. Appl Environ Microbiol. 2011;77(9):2992–3001.

151. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012;486(7402):222–7.

152. Shanks OC, Kelty CA, Peed L, Sivaganesan M, Mooney T, Jenkins M. Age-related shifts in the density and distribution of genetic marker water

quality indicators in cow and calf feces. Appl Environ Microbiol. 2014;80(5):1588–94.

153. Staley C, Kaiser T, Gidley ML, Enochs IC, Jones PR, Goodwin KD, et al. Differential impacts of land-based sources of pollution on the microbiota of southeast Florida coral reefs. Appl Environ Microbiol. 2017;83(10).

154. Staley C, Kaiser T, Lobos A, Ahmed W, Harwood VJ, Brown CM, et al. Application of SourceTracker for Accurate Identification of Fecal Pollution in Recreational Freshwater: A Double-Blinded Study. Environ Sci Technol. 2018;52(7):4207–17.

155. Oikonomou G, Teixeira AGV, Foditsch C, Bicalho ML, Machado VS, Bicalho RC. Fecal Microbial Diversity in Pre-Weaned Dairy Calves as Described by Pyrosequencing of Metagenomic 16S rDNA. Associations of Faecalibacterium Species with Health and Growth. PLoS One. 2013;8(4).