The Pennsylvania State University

The Graduate School

MICROBIAL ECOLOGY OF SURFACE WATER FROM THE NORTHEAST U.S. AND ITS ASSOCIATION WITH ENVIRONMENTAL FACTORS AND FOODBORNE PATHOGENS

A Thesis in

Food Science

by

Taejung Chung

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2019

The thesis of Taejung Chung was reviewed and approved* by the following:

Jasna Kovac
Assistant Professor of Food Science
Thesis Advisor

Darrell W. Cockburn
Assistant Professor of Food Science

Edward G. Dudley
Professor of Food Science

Luke LaBorde
Professor of Food Science

Robert F. Roberts
Professor of Food Science
Head of the Department of Food Science

*Signatures are on file in the Graduate School

ABSTRACT

Fresh produce is a common food vehicle associated with foodborne outbreaks in the U.S. Moreover, recent foodborne outbreaks demonstrated that surface water used for irrigation can act as an important source of pre-harvest produce contamination with foodborne pathogens. The US Food and Drug Administration (FDA) has outlined standard protocols for direct pathogen detection methods in variety of food and environmental sources. However, direct detection of pathogens in routine analyses of irrigation water is recognized as inefficient because it requires multiple complex test for detection of each individual target pathogen. Thus, indirect methods based on indicator microorganisms, such as generic *Escherichia coli* are used for monitoring of the microbiological quality of agricultural water. The FDA Produce Safety Rule (PSR) of the Food Safety Modernization Act (FSMA) proposed quantitative detection of generic *E. coli* for indirect detection of potential microbiological food safety hazards. However, several studies suggested that indicator microorganism that are currently used for evaluation of microbiological quality and safety of irrigation waters do not correlate well with occurrence of relevant pathogens under all relevant environmental conditions. In this thesis we sought to better understand the microbial ecology of surface waters collected in streams located in the upstate New York. The goal of this thesis was to characterize microbial communities of surface water and investigate potential associations between microbial communities' profiles, presence of foodborne pathogens and environmental factors. We characterized the composition of the bacterial and fungal communities in 68 water samples collected between May and August 2017 from six streams located in the upstate New York. Microbial communities were

determined by Illumina sequencing of PCR-amplified 16S rRNA gene V4 region and ITS2 sequences. Alpha and beta diversity indices were used to analyze and compare the microbial communities' characteristics of the samples. Moreover, Random forest (RF) machine learning was implemented to predict the presence of pathogens based on the microbial community composition.

According to the principal coordinate (PCoA) analysis and permutational multivariate analysis of variance (PERMANOVA), microbial communities differed significantly ($p<0.01$) between suspended sediment and water fractions. Moreover, specific physicochemical properties of water (i.e., average flow rate, pH, turbidity, and conductivity) were significantly associated with microbial composition. Additionally, the microbial communities of sediment fractions differed significantly among the sampling sites based on both alpha and beta diversities measurement. The observed differences may be due to the upstream land use of the sites; however these are solely descriptive observations. Furthermore, certain microbial families (e.g., *Kallotenuaceae*, *Flavobacteriaceae,* and *Sericytochromatia*) were found to be predictive of *Salmonella* presence using RF; however, the predictions had a very low accuracy (AUC of 0.55) and therefore cannot be used.

Overall, this study provided a new insight into microbial ecology of surface waters in a limited geographic area in the northeast U.S., and its relationships with pathogen occurrence and environmental factors including physicochemical properties.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

16S	16 Small Subunit
AIEC	Adherent Invasive *E. coli*
ANOVA	Analysis of Variance
BAM	Bacteriological Analytical Manual
CCA	Canonical Correspondence Analysis
CDC	Center for Disease Control
CFU	Colony Forming Unit
DAEC	Diffusely Adherent *E. coli*
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleoside Triphosphate
EAEC	Enteroaggregative *E. coli*
EHEC	Enterohemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
ELISA	Enzyme-linked Immunosorbent assay
EPA	Environmental Protection Agency
EPEC	Enteropathogenic *E. coli*
ETEC	Enterotoxigenic *E. coli*
FDA	U.S. Food and Drug Administration
FSMA	Food Safety Modernization Acts
GM	Geometric Mean
HC	Hemorrhagic Colitis
HUS	Hemolytic Uremic Syndrome
ITS	Internal Transcribed Spacer
kb	Kilobase
LFA	Lateral Flow Immunochromatographic assay
MB	Megabase
MPN	Most Probable Number
mTEC	Membrane Thermotoloerant *E. coli*
NGS	Next Generation Sequencing
NMDS	Non-metric Multidimensional Scailing
NORS	National Outbreak Reporting System
NTS	Nontyphoidal *Salmonella*
OTU	Operational Taxonomic Unit
PCoA	Principal Coordination Analysis
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
PFGE	Pulsed-field Gel Electrophoresis
PSR	Produce Safety Rule
QIIME2	Quantitative Insights Into Microbial Ecology 2
qPCR	Quatitative Polymerase Chain Reaction
RF	Random Forest
RNA	Ridonucleic Acid
RPLA	Reverse Passive Latex Agglutination

rRNA	Ribosomal Ridonucleic Acid
SBS	Sequencing by Synthesis
SMRT	Single Molecule Realtime
STEC	Shiga-toxin producing *E. coli*
STV	Statistical Threshold Value
WGS	Whole Genome Sequencing

ACKNOWLEDGEMENTS

First of all, I would like to thank my parents in Korea. Their unconditional support and guidance lead me to study today. I would also like to thank my sister and brother-in-law in Japan, being my friends, guardians, supporters.

I sincerely express my thanks to my advisor Dr. Jasna Kovac for allowing me to be here and teach me how to be a researcher. Her insights and attitude inspired me a lot toward not only the academic prospects but also personal development.

I would like to express my gratitude to my committee members, Dr. Edward Dudley, Luke LaBorde, and Dr. Darrell Cockburn for their intellectual insight, feedback on the thesis, and support for this work. I also thank the Department of Food Science here at the Penn State, especially to the administrative staff of the department.

Big up for my lab members over the past years; Manjari, Ellie, Runan, Laura, Naomi for their continuous friendship, support, and encouragement throughout this experience. I also thank to all of the undergraduate researchers in our lab, especially someone who helped me or worked with me; Meg as my first and the best mentee, Sarah as my first experimental failure mate, and Hepzibah. I know no one of you will read this, but anyway I am appreciated with the memories we made together.

I also would like to thank Dr. Daniel Weller from Cornell University for not only being an exceptional collaborator but for always providing encouraging feedback and advice.

Lastly, I am deeply thankful to all my friend in here and in Korea being my life jacket, parachute, and safeguard every day. I am especially grateful to Dr. Ben D. Tall from the US FDA for initiating my passion for pursuing this opportunity from my internship.

Chapter 1

Statement of the Problem

Fresh produce, including fruits and vegetables, is the leading cause of foodborne outbreaks in the U.S. Consumption of fresh produce without thermal treatment increases the risk of infections with foodborne pathogens such as pathogenic *Escherichia coli*, *Salmonella* spp., and *Listeria monocytogenes*. Contaminated irrigation water is one of the common sources for introduction of foodborne pathogens onto pre-harvest fresh produce, and surface water is reported as the most commonly used source of irrigation water in the U.S. Hence, the subject of investigation in this thesis was water collected from surface streams.

The US Food and Drug Administration (FDA) has established standard protocols for pathogen detection methods in water and food based on conventional microbiological and biochemical methods, with optional immunological and molecular confirmation methods outlined in the Bacteriological Analytical Manual (BAM). However, direct detection of pathogens is recognized as inefficient because it requires multiple tests for detection of individual pathogens, resulting in laborious and lengthy workflows. Thus, a more practical way of monitoring the microbiological quality of irrigation water is by using indirect methods based on indicator microorganisms, such as generic *E. coli*.

The FDA Produce Safety Rule (PSR) of the Food Safety Modernization Act (FSMA) recommends quantitative detection of generic *E. coli* for indirect detection of potential microbiological food safety hazards. However, several studies pointed out that the

correlation between *E. coli* and foodborne pathogens is low. In this thesis we sought to better understand the microbial ecology of surface waters collected in streams located in the upstate New York and investigate potential associations between microbial community profiles, presence of foodborne pathogens, and environmental factors. The first objective of this study was to utilize the amplicon sequencing for microbial community profiling to investigate potential associations between microbial communities and multiple factors, including the sample fraction, geospatial factors, environmental factors, and foodborne pathogen and indicator presence in surface waters from the upstate New York. The second objective was to apply statistical and machine learning approaches to investigate whether it is possible to identify microbial groups predictive of presence of foodborne pathogens. The data generated in this study provide a baseline for describing the northeast U.S. surface water microbial ecology and the relationships between microbial community composition, detected pathogen occurrence, and environmental factors.

Chapter 2

Literature Review

2.1. Foodborne pathogens related to fresh produce

According to the National Outbreak Reporting System (NORS) from the Centers for Disease Control and Prevention (CDC), there were 9,594 outbreaks, 174,620 illnesses, 9,681 hospitalizations, and 238 deaths attributed to foodborne pathogens in the United States (US) between 2007 and 2017 (1). Fresh produce is a leading cause of foodborne outbreaks in the US. According to the Food and Drug Administration (FDA) and CDC, there were 19 multistate outbreaks associated with fresh produce in the last five years (Table 2.1). A majority of cases (15 out of 19) were associated with pathogenic *Escherichia coli*, *Salmonella spp*., and *Listeria monocytogenes* (2). Investigation of those outbreaks could not identify a source of the contamination in most cases, whereas a few cases identified that the source of contamination was environmental, from either irrigation water or the sediment from the source of irrigation water (3, 4). Thus, it is important to understand the characteristics and prevalence of foodborne pathogens in irrigation water in order to reduce potential risks of fresh produce contamination.

Table 2.1 Summary of Outbreaks associated with fresh fruits and vegetables in United Stated from 2015 to 2019

Foodborne pathogen	Food associated with and outbreak	Year
Escherichia coli	Romaine Lettuce	2018
Salmonella spp.	Romaine Lettuce	2018
	Leafy Greens	2017
	Alfalfa sprouts	2016
	Pre-cut melons	2019
	Salad mix	2018
	Pre-cut melons	2018
	Sprouts	2018
	Frozen shredded coconut	2017
	Papayas	2017
	Alfalfa sprouts	2016
	Alfalfa sprouts	2016
	Cucumbers	2016
Listeria monocytogenes	Frozen vegetables	2016
	Leafy Greens	2016

2.1.1. Pathogenic *Escherichia coli*

E. coli is a Gram-negative, rod-shaped, facultatively anaerobic bacterium belonging to a family of Enterobacteriaceae, a phylum of Proteobacteria, class of Gammaproteobacteria, and an order of Enterobacteriales. *E. coli* was first identified by Theodor Escherich in 1885 and has been studied in detail ever since its identification. The natural reservoir of *E. coli* is the gastrointestinal tract of warm-blooded organisms, such as mammals (5). Most *E. coli* strains are not harmful and do not cause illness in humans; however, several groups of strains can cause foodborne disease (6). Pathogenic *E. coli* are classified into several pathotypes based on their serological and virulence properties (7). Pathotypes of pathogenic *E. coli* include enteropathogenic *E. coli* (EPEC), enterohemorrhagic *E. coli* (EHEC), enteroinvasive *E. coli* (EIEC), enteroaggregative *E. coli* (EAEC), diffusely adherent *E. coli* (DAEC), and enterotoxigenic *E. coli* (ETEC), in addition to adherent invasive *E. coli*

(AIEC) are known to be associated with gastrointestinal diseases (7). One of the pathogenic *E. coli* groups of highest food safety concern in the US are the Shiga toxin-producing *E. coli* (STEC). STEC may cause foodborne infections leading to life-threatening diseases in humans. Specifically, some STEC strains are subclassified as enterohemorrhagic *E. coli* (EHEC), which includes *E. coli* O157:H7 that is frequently associated with outbreaks and clinical illness. (8).

STEC are strains of *E. coli* that produce potent cytotoxins called Shiga toxins (9). The name Shiga toxin was derived from a cytotoxin produced by *Shigella dysenteriae* Serotype1 due to their similarity (10). Shiga toxins are represented by two toxins, Stx1 and Stx 2, each with different characteristics (11). Another virulence factor that contributes to STEC or EHEC pathogenesis is intimin encoded by *eaeA* (14). Intimin is a 97-kDa attachment-and-effacement protein (A/E protein) that is required for A/E lesion formation that helps EHEC and also EPEC to attach to the intestinal cells to promote colonization in mammals (14,15). STEC is known for causing hemorrhagic colitis and hemolytic uremic syndrome (HC and HUS, respectively) (17, 18). Epidemiological data support that the expression of Stx2 has a stronger association with development of HUS than Stx1 alone (12, 13). EHEC strains express Stx1, Stx2, or both (12). The first hemolysis-associated response is bloody diarrhea (19). STEC strains have been detected in the large intestine in a wide diversity of animals, such as cattle (20). Hence, STEC can be introduced to food crops through the cattle feces from the dairy farms near the watersheds. Thus, it is important to understand the prevalence of STEC in surface water in order to prevent contamination of fresh produce during irrigation.

A multistate outbreak associated with romaine lettuce that had occurred in April 2018 and caused 210 cases of foodborne illness and five deaths (3). The FDA environmental assessment team collected samples from agricultural water, soil amendments, processing and manufacturing facilities, and cattle feeding operations near the farm and found that three samples contained *E. coli* O157:H7 with the same rare genetic fingerprint associated with the outbreak, as determined by whole genome sequencing (WGS). Those three samples were collected at the irrigation canal near Yuma County in which the lettuce farm was located. Thus, FDA concluded that the source of contamination was most likely the water from the irrigation canal (3). Multiple *E. coli* outbreaks linked with fresh produce have been reported in other countries as well. In 2005, lettuce contaminated with *E. coli* O157: H7 was identified as the cause of an outbreak that affected 135 individuals in Sweden (21). The investigation revealed that the same pulse field gel electrophoresis (PFGE) patterns corresponded to patients and cattle feces from the farm located upstream of the farm from which the lettuce was grown (21), indicating that land use and anthropogenic activities upstream of the produce fields affected the microbiological quality of the irrigation water used downstream. More recently, an *E. coli* O157:H7 outbreak linked to romaine lettuce grown in California was reported in fall 2018. During the investigation using WGS, the FDA identified a sample that contained the *E. coli* O157:H7 strain with the same genetic fingerprint as the outbreak strains (22). The sample was collected from the sediment of the water reservoir that was used as the source of the irrigation water for the adjacent lettuce field. Thus, the FDA concluded that irrigation water was the most probable source of romaine lettuce contamination that was implicated in the multistate foodborne outbreak (22).

2.1.2 *Salmonella spp.*

The CDC estimates that *Salmonella* causes approximately 1.2 million illness, 23,000 hospitalizations, and 450 deaths in the US every year (23). *Salmonella* is a genus of rod-shaped, Gram-negative bacteria belonging to the family Enterobacteriaceae, a phylum of Proteobacteria, class of Gammaproteobacteria, and order of Enterobacteriales. *Salmonellae* are motile facultative aerobes belonging to two species: (1) *S. enterica* and (2) *S. bongori* (24). *S. enterica* has six subspecies: (1) *S. enterica* subsp. *enterica*; (2) *salamae;* (3) *arizonae*; (4) *diarzonae*; (5) *indica*; and (6) *houtenae* (25). In those six subspecies, *S. enterica* subsp. *enterica* is associated with disease in warm-blooded animals (25). There are over 2,500 different serovars identified within *S. enterica* subsp. enterica (26). Furthermore, *Salmonella enterica* subsp. *enterica* infections manifest in a variety of clinical symptoms, whereas *Salmonella bongori* is known as an animal pathogen in dogs and birds (27, 28). The nontyphoidal *Salmonella* (NTS) strains cause gastroenteritis or bacteremia, which can be harmful not only for the humans but also for the wide variety of animals. For example, *S.* Dublin and *S*. Typhimurium serovars have been associated with rabbit, mice, and dairy cattle infections, and *S*. Cerro has been identified as a cause of disease in cattle (29, 30). NTS is a leading cause of foodborne illness in the US.

There have been several produce-related outbreaks caused by *Salmonella* in the past five years. In August 2014, a multistate outbreak of *S. enterica* serotype Newport was detected and confirmed using PFGE (31). A total of 275 illness from 30 different states were identified (31). In 2016, a total of 907 people from 40 states got infected with *S.* Poona. Epidemiological investigations by state and local agencies confirmed that 75% of infected

people reported eating cucumbers in the week before their illness symptoms started. Several state health and agricultural departments collected retail cucumbers and successfully isolated the outbreak strains of *S. Poona*. Additionally, raw tomatoes contaminated with different *Salmonella* serovars have caused multistate outbreaks in the past. From 1990 to 2010, 15 multistate outbreaks associated with *Salmonella enterica*-contaminated fresh tomatoes were reported, causing 1952 illness, 384 hospitalizations, and three deaths (32). According to the traceback investigation, in 14 of these outbreaks, most of the results indicate that the contamination occurred either on the farm or in the packing house (32).

2.1.3 *Listeria monocytogenes*

L. monocytogenes is a gram-positive, facultative anaerobic bacterial species that can grow at refrigeration temperatures; hence, it represents a serious food safety concern in the fresh and frozen produce industries (33–35). In Canada, Denis et al. reported that 0.3% (n=4435) of leafy green samples contained *L. monocytogenes* (34), and Mritunjay and Kumar found 3.5% (n= 480) in eight different type of fresh produce, such as tomatoes, cucumbers, carrots, radishes, coriander, beet-root, cabbage, and spinach. *L. monocytogenes* causes an infectious disease, listeriosis, with its primary symptoms similar to other foodborne illnesses, including fever and diarrhea (33). However, invasive listeriosis can also lead also to miscarriages, stillbirths, premature deliveries, or life-threatening infections of the newborn if pregnant women contract listeriosis (33). According to the CDC's NORS, 80 outbreaks of listeriosis, including 945 cases, 691 hospitalizations, and 140 deaths, were reported between 1998 and 2017. In most of these cases (78/80), food was identified as the primary source of the outbreak (1, 36–39). The relatively high rate of hospitalization and mortality in addition to its capability of surviving and growing in foods and food processing

environments at cool temperatures necessitate improved detection methods in order to reduce the occurrence of *L. monocytogenes* (23, 40, 41).

2016 *L. monocytogenes* outbreak caused by contaminated packaged salad produced by Dole in Ohio (42) resulted in nineteen sickened individuals in nine. Whole genome sequencing (WGS) confirmed that all 19 patient isolates were genetically closely related (42). Other than those large-scale multistate outbreaks, several *Listeria*-associated outbreaks in fresh produce, such as cabbage, corn, carrots, lettuce, cucumbers, parsley, and salad vegetables were reported not only in the U.S., but also in other countries (43). Even though there was fewer disease cases associated with *L. monocytogenes* outbreaks in fresh produce compared to other pathogens, listeriosis is more severe compared to illnesses caused by other foodborne pathogens. This severity is reflected in approximately one in five cases of listeriosis resulting in death (44).

2.1.4 Potential environmental sources of fresh produce contamination with foodborne pathogens

Foodborne pathogens can be introduced to produce grown outdoors primarily through soil (45–47), water (45–48), organic amendments (49, 50), and wildlife feces (47, 51). The risk of introducing foodborne pathogens to produce increases when inefficient manure decontamination processes are used prior to its use as a natural fertilizer (52–54). Once foodborne pathogens are introduced into the soil environment, they may survive in the soil for extended periods of time. For example, Sharma et al. found that *E. coli* can survive more than 90 days in manure-amended soils (55). While pathogens are living in the soil, they can be transferred to the crops through the water splashing during irrigation activities (56, 57).

In addition to soil, irrigation water is known as one of the common sources of produce contamination by foodborne pathogens. Surface water is often not treated before being used for irrigation of agricultural crops, and it may harbor foodborne pathogens that can be transmitted to the crops (58–60). Several studies reported the presence of Shiga toxin-producing *E. coli* (STEC) in surface waters in different geographic regions in the US. The studies conducted in New York state found 3% prevalence of STEC. Other studies from California (11%, n=1386) and Georgia (0.2%, n=496) were also published in which untreated surface water was reported to be contaminated with pathogenic microorganisms (47, 61, 62). Thus, ensuring that the water is not contaminated before being used for irrigation can prevent pre-harvest contamination of fresh produce. Factors that contribute to an increased risk of introducing pathogens onto produce through the irrigation water include weather, irrigation system, upstream land use, and anthropogenic activities (63, 64). Upstream activities, such as livestock operations and pastures, may cause water used for crop irrigation to become contaminated through fecal deposition and run-off into the water streams after rain storms (65–68). The fecal matter can be directly introduced to crops by using natural fertilizers with improper decontamination process or indirectly to the irrigation water from which it is then transmitted to the crops during irrigation.

2.2 Foodborne pathogen detection methods

2.2.1 Direct pathogen detection methods

Detection of pathogenic bacteria is key to the prevention and identification of the foodborne pathogen outbreaks and infections. The U.S. FDA Bacteriological Analytical Manual (BAM) provides standard protocols for conventional microbiological and

biochemical methods in combination with relatively recently established immunological and molecular methods for isolation, characterization, and identification of known pathogenic bacteria such as pathogenic *E. coli, Salmonella spp*., and *L. monocytogenes*.

2.2.1.1 Microbiological foodborne pathogen detection methods

Microbiological foodborne pathogen detection methods outlined in the FDA BAM are based on enrichment of targeted foodborne pathogens, their isolation via use of selective and often differential media, and identification using a combination of biochemical methods with an option of using immunological and/or molecular identification methods that are not specifically prescribed in the BAM protocols (69–71). Different primary and often secondary enrichment media are used for different target pathogens (Table 2.2). Enrichments typically take 24 to 48 h, depending on the organisms (69) (Table 2.2). Additionally, isolation also takes another 24 h. After isolation, typically 2–5 biochemical tests are recommended in order to provide sufficient information for pathogen identification. Biochemical tests usually require multiple media and at least a 24-h incubation, which also prolongs pathogen identification. Overall, pathogen enrichment, isolation, characterization, and identification require a minimum of four days when standard microbiological protocols are followed. These protocols also require multiple complex media and are laborious, but extremely sensitive, as they can detect a single viable cell per tested unit of a food sample. Thus, pathogen detection and identification methods with comparable sensitivity and specificity are desired to improve time and cost efficiency (72).

Table 2.2 FDA BAM methods for detection of foodborne pathogens

	Pathogenic *E. coli* except STEC and *E. coli* O157:H7	STEC and *E. coli* O157:H7	*Salmonella* spp.	*Listeria monocytogenes*
Enrichment	35°C for 3 hours 44± 0.2°C for 20 hours	37°C ± 1°C static for 5 hours 42°C ± 1 °C static overnight	35ºC ± 2.0ºC C for 24 ± 2.0 hours	30° C for 4 hours 30° C for 24 to 48 hours
Isolation	35°C for 20 hours	37°C ± 1°C for 18 - 24 h 18-24 h at 37°C ± 1°C	42 or 35 ± 0.2°C (depends on media) for 24 ± 2 hours 35°C for 24 ± 2 hours	Range of 30 to 37 °C for 24 to 48 hours (depends on the media)
Identification	35°C for 26 hours (primary screening) 35°C for 48 hours (secondary screening)	qPCR assay	35°C for 24 ± 2 hours	35° C for 24 h
Additional/ Alternative test for identification	Additional biochemical tests are required for different class of *E. coli*	Additional presumptive isolate screening is required (37°C ± 1°C for 18 - 24 h)	Alternative rapid biochemical assays or qPCR method is described	Alternative rapid biochemical assays or qPCR method is described

2.2.1.2 Nucleic acid- and immunologically-based foodborne pathogen detection methods

Biochemical assays may be replaced by rapid nucleic acid-based on immunological methods. These include reverse passive latex agglutination test (RPLA), enzyme-linked immunosorbent assay (ELISA), lateral flow immunochromatographic assays (LFA), the molecular level polymerase chain reaction (PCR), quantitative PCR tests, among others. These assays usually generate results in a matter of hours (Table 2.2) (69–71).

With respect to immunological methods, RPLA is a type of agglutination test in which known antibody is bound to the antigen that belongs to the pathogen of interest. The latex particle coated with a selected antibody reacts with a targeted antigen (such as the EspB protein in EPEC/EHEC diagnosis) (73). ELISA is one of the most commonly used immunological methods for foodborne pathogen detection. The detection limit of ELISA is 10^5 to 10^7 CFU/ml; however, the recently used sandwich ELISA assays are more sensitive and have a better limit of detection, which is as low as 10^3 CFU/ml in food samples. The sandwich ELISA is based on two different antibodies. The primary antibody is used to immobilize bacterial cells from the sample binding to the surface. After that step, a secondary antibody, which is conjugated with an enzyme, is attached to the cell that was already bound to the primary antibody. This multi-antibody structure can be detected by adding a colorless substrate that will be converted into a colored form by the conjugated enzyme on the secondary antibody. Fluorophores can also be used instead of enzymes. The advantage of ELISA is its high throughput that allows for concurrent testing of a larger number of samples.

Another immunological assay is the LFA, which requires relatively less equipment and supplies than the ELISA. The sample fluid goes through the assay via capillary action and it encounters the conjugate, a colored particle (e.g., colloidal gold) labeled antibody or antigen, which will be used for visualization of specific antibody-antigen binding. Jung et al. reported that the limit of detection for *E. coli* O157 in enriched samples using LFA was 1.8 x 10^5 CFU/ml. LFA also can be used for detection of other pathogens, such as *Salmonella* spp. and *L. monocytogenes.* Nevertheless, immunological assays work best when there are

no interfering molecules in the samples such as off-target cells, DNA, or proteins, which may inhibit antibody-antigen binding.

PCR and qPCR are molecular techniques that are widely used for amplification of the target DNA or RNA sequences using a heat-stable DNA polymerase. Typically, PCR consists of 20 to 40 cycles of repeated temperature changes for denaturation, annealing, and extension of the DNA to perform the amplification. qPCR is an advanced PCR-based technique that enables real-time fluorescence monitoring of amplified DNA fragments using a fluorophore hybridization with the double-stranded DNA product to measure the fluorescence in real time (74). In pathogen detection, specific genes (such as *stx1*, *stx2* for STEC, *invA* for *Salmonella* spp.) are used to identify the presumptive pathogen isolates (74). Several different fluorescent systems have been used for qPCR: the cyanine dye, SYBR green, TaqMan probes, and molecular beacons. SYBR green is a double-stranded DNA-binding fluorescent dye, which is a non-sequence-specific dye. The fluorescence signal is amplified when bound to the minor groove of the DNA double helix, regardless of its sequence. TaqMan probe is complementary to a specific nucleotide sequence in one of the amplicon strands internal to both primers, and the system depends on the 5′-3′ exonuclease activity of Taq DNA polymerase that cleaves the probe and separates both dyes in order to generate the fluorescent signal, which improves the specificity of amplification and signal detection. The reported detection limit of TaqMan qPCR was approximately 100 CFU/ml when detected directly from the food sample and ~4 CFU/25 g when detected from enrichment. Comparing the two systems, SYBR green is simple and less expensive. On the other hand, the specificity of SYBR green is lower due to its ability to bind to any PCR product. Nevertheless, the specificity of the PCR-based method is mainly affected by primer

specificity rather than the type of system. Overall, qPCR is more specific, and also requires less time.

A number of rapid immunological and nucleic-acid based pathogen detection protocols have been developed and validated against standard protocols that allow for direct detection and identification of pathogens from primary enrichment. The goal of developing rapid detection assays is to reduce the time required to obtain a result. However, if the method does not ensure reliability toward its results, it is not useful.

The validated rapid immunological- and molecular-based pathogen detections are presently reducing the time required to get a result, which is very important for foodborne pathogen detection (77). This is especially true for foods that have short shelf life, such as fresh produce. Nevertheless, immunological-based rapid detection methods often have lower sensitivity compared to conventional and DNA-based methods. Molecular, DNA-based methods require specialized instruments (e.g., qPCR instrument) and relatively expensive reagents (e.g., TaqMan reagents) (78). Due to the heavy work load and high cost of direct methods, indirect detection methods using indicator and/or index organisms are widely implemented, especially for testing of water microbiological quality.

2.2.2 Indirect foodborne pathogen detection methods

2.2.2.1 Indicators, indices, and surrogates

Direct detection of each pathogen is costly and is not time efficient; thus, it is not practical to implement in routine monitoring of microbiological quality of irrigation water. For this reason, water microbial quality tests rely on indirect pathogen detection methods

based on the indicator organisms. There are a number of indirect pathogen detection methods that have been introduced and implemented. Indicator organisms can generally be divided into two categories: (1) process indicators and (2) fecal indicators. Process indicators are a group of organisms that are used to evaluate the efficiency of a process (such as total heterotrophic bacteria or total coliforms for chlorine disinfection). The fecal indicators are a group of microorganisms that indicate fecal contamination of the sample, such as thermotolerant coliforms or *E. coli* (79). For water testing, generic *E. coli*, coliforms, and Enterobacteriaceae are most widely used as indicators of fecal contamination (80). Presence of fecal indicators suggests that pathogens may be present in water (79).

2.2.2.1.1 Indicators

Indicator organisms are microorganisms that can be used to indicate the hygienic condition of a particular environment or a sample (81–84). Indicator organisms are considered non-pathogenic, are assumed to occur consistently occur in water contaminated with foodborne pathogens, and are easily detectable at low concentrations (81, 84). Currently, several indicator organisms are used for the evaluation of the microbiological quality of water (84). Two of the most widely used indicators in environmental water testing are coliforms and *E. coli* due to their association with fecal matter (85, 86). In addition to those, fecal coliform bacteria, *Enterococci,* and Enterobacteriaceae are also used as indicator organisms worldwide (Figure 2.1) (79, 87).

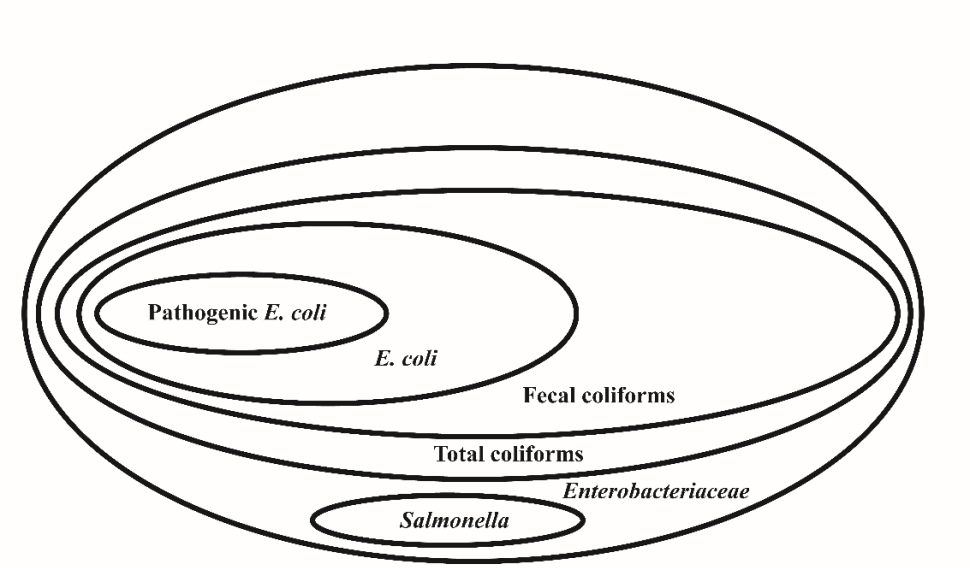

Figure 2.1 Diagram showing the relationship between indicator organisms and select foodborne pathogens

Coliforms are defined as aerobic or facultatively anaerobic, Gram-negative, non-spore-forming rod-shaped bacterial group that can ferment lactose and produce acid and gas at temperatures from 32 to 35°C (85). This classification is not based on taxonomy, but rather based on the above listed phenotypic trait. There are currently 19 different genera classified into the coliform group. These also include *Escherichia*, *Citrobacter*, *Enterobacter*, and *Klebsiella*. Most of the genera in the coliform group are representatives of a single family of Enterobacteriaceae with the exception of *Aeromonas* that belongs to the family Aeromonadaceae (85). Coliforms can be further divided in three groups: thermotrophic, thermotrophic and ubiquitous, and psychrotrophic coliforms, as described by the Leclerc et al. (88). The thermophilic coliform group is also known as fecal coliforms, which is represented by *E. coli* (88). The unique characteristic of fecal coliforms is their ability to grow and ferment lactose at 44.5°C (88). This characteristic is also used to

discriminate fecal coliforms from other types of coliforms. Two distinct characteristics of *E. coli* as an indicator of fecal contamination are its persistence in the environment and its main reservoir, which is the large intestine of warm-blooded animals (89–92). *E. coli* is therefore used as an indicator of fecal contamination (85).

Enterobacteriaceae are another indicator of microbiological water quality and safety. The family of Enterobacteriaceae contains approximately 20 genera, including *Escherichia* while coliforms belonging to the previously mentioned genera. Enterobacteriaceae were first proposed as alternative indicators to the coliforms because they are considered to be more inclusive indicators of a broader range of different pathogens. For example, *Salmonella*, *Shigella*, or *Yersinia* cannot ferment lactose. Thus, the rationale for using the Enterobacteriaceae as an indicator was based on the inability of lactose fermentation by specific pathogens, which led the low or negative coliform test (93). Enterobacteriaceae are widely used as an indicator in the European Union, whereas coliforms remain as recommended indicators in the U.S. (93, 94).

2.2.2.1.2 Index organisms

Index organisms can predict the presence of pathogenic microorganisms because they originate from the same source and are capable of surviving and persisting in similar environments as the pathogenic microorganisms of interest (86, 95). The difference between indicator and index organisms is that indicator organisms are suggesting the inadequate hygienic conditions (such as fecal contamination) or failure in a process (such as inadequate sanitation or thermal treatment) by indicators and index organisms are suggesting conditions suitable for survival and/or growth of pathogens or possible presence of specific pathogens

by indices. For example, *Listeria* spp. are used as an index for the presence of *Listeria monocytogenes* (95). Furthermore, generic *E. coli* has been proposed as the index organism for EHEC, *Salmonella* spp., and *Shigella* spp. However the reliability of *E. coli* as an indicator/index organism is still questionable (88, 96–99).

2.2.3 Federal agricultural water standard

Currently, most of the water regulation from all around the world is based on the indirect methods for testing of microbiological quality (100). The FDA's Food Safety Modernization Act Produce Safety Rule (FSMA-PSR) established agricultural water standards and recommended water microbiological quality testing procedures (60, 101). The PSR set the first science-based standards for the safe growing, harvesting, packing, and storing of fresh produce, such as fruits and vegetables (101). Agricultural water is addressed in a subsection of the PSR that defines the minimum requirements and compliance for the agricultural water that is intended or likely to contact edible parts of the fresh produce except sprouts due to their unique vulnerability for contamination. Irrigation water testing for harvesting sprouts is based on the direct detection of select pathogens (i.e., pathogenic *E. coli*, *Salmonella* spp.). (101).

The rule established two different sets of criteria based on intended water use (60, 102, 103). No detectable *E. coli* is allowed for post-harvest water applied directly to food contact surfaces, hand washing, and water that will come into direct contact with produce during or after the harvest (101). The second criterion for the agricultural water quality applies directly to produce growing, pre-harvest. It is based on two statistical measurements of generic *E. coli* per 100 ml of a water sample: (1) the geometric mean (GM) and (2) the

statistical threshold (STV) (101). According to the FDA's FSMA-PSR, the untreated surface water that is going to be directly used for produce growing needs to be tested initially using a minimum of 20 samples over the course of two to four years (101). GM and STV metrics are then calculated based on multiple samples collected on a rolling basis (104). GM reflects the central tendency of the quantities of *E. coli* in samples collected over time, and the STV is a measure of the amount of variability in the *E. coli* quantities in the water. The latter minimizes the effects of substantially increased counts of *E. coli* in certain sampling times due to unusual conditions, such as heavy rain. A GM of <126 CFU/100 ml of generic *E. coli* and STV of ≤410 CFU/100 ml is allowed in water that is intended to be used for irrigation of food crops (101). The GM and STV are calculated to check if the quality of the water source is acceptable for irrigation of food crops. After the initial survey, at least five new samples must be collected and tested on a yearly basis and analyzed together with 15 other most recently collected samples (101). In the FSMA Final Rule on Produce Safety such as Method 1603: *Escherichia coli* (*E. coli*) in Water by Membrane Filtration Using Modified membrane-Thermotolerant *Escherichia coli* Agar (Modified mTEC). As the title implies, this method is based on membrane filtration of water samples. A sample is filtered through a membrane with a 0.45 µm pore size that captures the bacteria. After filtration, the membrane needs to be placed on the selective agar (mTEC agar) and incubated for a short period of 2 ± 0.5 hours at 35°C for retrieving injured or stressed cells and then incubated again at 44.5°C for 22 ± 2 hours (105). In addition to that method, the FDA also defined four other membrane filter based methods and three concurrent methods using m-ColiBlue24® Broth PourRite Ampules or the IDEXX Colilert® Test Kit (106). Those tests are widely used, whereas limitations with respect to the current method have been reported. In the following

section, we will outline and discuss several studies that had suggested that current indicators are poorly associated with the presence of foodborne pathogens under their test conditions.

2.2.4 Limitation of current indirect methods for detection of microbiological food safety hazards

After the concept of indirect methods was introduced, a number of studies investigated the reliability of the indicator microorganisms and/or their relationship with pathogens. A study by Goyal et al. revealed that there is a strong correlation between *Salmonella* from the sediment and fecal coliforms in water samples (107). That study was the first published report that demonstrated the relationship between indicators and the presence of pathogens (107). Similar to Goyal et al., several studies also indicated that there is a strong correlation or association between the level of indicator organisms and pathogens (108–110). Mansilha et al. classified water samples from north Portugal according to the concentrations of *E. coli* and fecal *Enterococci* by using the EU's directive standard (111). Significant correlations ($p < 0.05$) were obtained between *Salmonella* and the indicators when the samples were under to the set microbiological criteria. However, *Salmonella* was also detected when an indicator level was below the defined threshold value. Thus, they concluded that a significant correlation was found only in a particular state, such as an abnormally high concentration of indicator microorganisms (108). Poma et al. showed that the concentration of *E. coli* was associated with the presence of enteric pathogens (such as pathogenic *E. coli* and *Salmonella*) from the La Paz River basin in Bolivia. Based on their results, about 85% of the samples that showed higher level of *E. coli* (10^5 to 10^6 MPN/100 ml) also contained *Salmonella* (109). A study by Xiao et al. concerning the infection risk of pathogens in

Wanzhou Watershed in China found a weak but significant correlation between the levels of generic *E. coli* and *Salmonella* ($R = 0.39$ to 0.43; $p < 0.05$) (110).

Even though there are multiple studies in which the reliability of the current indicators is mentioned, a number of other publications indicate that there are weak or no correlations between indicator organisms and pathogenic bacteria. *E. coli* as an indicator organism was used by numerous studies that had found either weak or no correlation between *E. coli* and pathogens using both culture-dependent microbiological enumeration (such as most probable number [MPN]) and culture-independent molecular methods (such as qPCR). Tingting et al. used qPCR for *E. coli* (*uidA*), and *Salmonella* (*invA*) detection for samples collected from the three lakes in Beijing, China in 2014 (112). The study revealed that there was no correlation between the level of *uidA* and *invA*, which also implied no correlation between *E. coli* and *Salmonella* presence in the sample (112). This finding was consistent different studies conducted in Florida, Georgia, California, and many other region's natural watersheds (such as lakes, rivers, canals) (82, 96, 113–117). For example, Haley et al. collected 72 water samples from six different sampling sites in Little River, Georgia. Using logistic regression, they found that *E. coli* does not provide significant predictive value for the presence of *Salmonella*. More notably, a high frequency of *Salmonella*-positive samples (76%, 42/55) was still below the threshold standard for the indicator *E. coli* from the US Environmental Protection Agency ([EPA] 576 CFU/100ml) (114). Similarly, McEgan et al. revealed factors related to the concentration of *Salmonella* in Florida surface water from 18 different sites, including lakes, ponds, creeks, streams, and canals. They collected 202 samples over a 12-month sampling period and found that the sampling month did not have a

strong correlation with *Salmonella* concentration ($R^2 = 0.2$). *E. coli* concentrations were weakly correlated with *Salmonella* concentration ($R^2 = 0.1$) (96). Even though both studies reported weak correlations between *E. coli* and *Salmonella* concentrations, Haley et al. revealed that the prevalence of *Salmonella* is much higher in April, whereas McEgan et al. could not find any correlation with bacterial concentration and sampling month. From their review paper based on 81 previously published datasets, Pachepsky et al. also indicate that those environmental, geospatial, and seasonal factors might blur the sole correlation between pathogens and indicator organisms (113). They concluded that the pathogen prevalence and concentration predictions need to be done with combinations of indicator organisms and hydrological, environmental, spatial parameters (113). Additionally, Truchado et al. reported that a strong correlation was only found in heavily polluted water, such as untreated waste water (115). In addition to that finding, other studies indicated that a low concentration of pathogen in the sample showed insignificant/weak correlation with the concentration of indicator organism (116, 117). These results from different studies indicate that *E. coli* by itself is not a reliable indicator for predicting foodborne pathogen presence in many relevant conditions and environments, and thus, alternative or supplementary indicators are needed.

2.3 Microbial community profiling

Microbial communities or microbiomes are one of the novel approaches in the field of microbiology for enabling the study of microbial communities (118, 119). The word ‘microbiome’ is combined with ‘micro-’ and ‘-biome,’ which means a community of microbiological organisms (micro-) that have common characteristics of the environment in

which they exist (-biome) (120). The human microbiome project ignited researchers' interest in the microbial communities residing in a particular environment, not only in the human body, but also in a variety of environmental samples (119). For example, the Earth microbiome project was founded in 2010 with the aim of collaborative collection of samples in order to advance the understanding of the microbial community structure in different environments such as earth, including water, soil, sediment, and plants (121).

2.3.1 Water microbiomes and factors influencing their composition

To date, several studies have been conducted with respect to the water microbiome. These studies include characterization of drinking water microbial communities (122–124) in addition to natural, untreated water microbial communities (125–130). Drinking water microbial community studies have identified several factors, such as chlorination treatments (131), filtration systems (123), and distribution systems, which influence a microbial community's composition (132). Studies that characterized microbial communities of natural untreated waters have mostly focused on river waters. Staley et al. found that upper Mississippi river microbial communities changed among 10 sampling sites. They found that environmental factors, such as pH and water sample temperatures had a significant effect on microbial community composition and diversity (125). In addition to Staley et al., a study conducted in another geographical region along the Mississippi river in a different year reported that different sampling sites along the same river had a significant impact on the discovered microbial community composition (126). Based on the results from two different studies, the major phyla detected in both studies were consistent and included alpha- and beta-Proteobacteria, Actinobacteria, Bacteroidetes, Cyanobacteria, and Verrucomicrobia (125, 126). Those phyla have previously been described as common in freshwater bacterial

assemblages (127,128,133). In addition to those studies, year-long metagenomic studies of surface water microbial communities revealed that not only the compositional changes but also the metagenome gene functional groups were different and associated with land use over time in microbiomes (129). The results suggested that samples collected near the agricultural site showed more distinct microbial community composition than both urban and protected sites (129). In this study, the author also suggest that one potential factor associated with the microbial community composition could be the sampling month. Wang at al. also suggested that urban landscapes can influence microbial community composition because of urban wastewater treatment. They concluded that untreated water samples contain unique microorganisms different from the conventional organisms found in water (134). Yong et al. indicated that long-term and high-concentrations of heavy metal contamination strongly influences the microbial community composition of natural water (135). Wang et al., using a multivariate ordination analysis, revealed that geographical factors (such as latitude, longitude, and altitude) and environmental factors (such as conductivity, total phosphorus concentration) affect bacterial community composition (136). According to these findings from several studies, it was shown that multiple factors can influence the microbial communities of environmental water.

2.4 Methods for microbial community characterization

2.4.1 Next generation sequencing

Microbiome characterization can be done using the next generation sequencing (NGS). NGS fundamentally differs from Sanger sequencing (first generation sequencing) because of its capability of simultaneously sequencing millions of DNA fragments from

different samples. NGS (second-generation sequencing) includes short-read sequencing technologies that sequence by synthesis (SBS) (137). Illumina sequencing, which currently dominates the market, is based on SBS combined with bridge amplification (138). Library preparation is essential for NGS and allows for barcoding of multiple fragments from different samples that can be sequenced in the same sequencing run. Barcoding is carried out by attaching adapter sequences onto the fragmented DNA, which allows for binding of fragments onto the flow cell prior to sequencing in addition to linking individual DNA fragment sequences with individual samples after sequencing (138). Single-stranded DNA attached to the flow cell is synthesized by isothermal amplification (139). During this amplification, fluorescently labeled dNTPs are incorporated to form double-stranded DNA (139). Each of the four dNTPs has a different fluorescent label, which serves to identify which base pair is incorporated during the DNA synthesis. The order of different fluorescent labels is captured and analyzed in order to detect the sequence of the attached DNA fragment (140, 141). NGS technologies enable generation of millions of fragment sequences in a single run (142). This technological feature also leads DNA sequencing technology implementation from short DNA to the entire genome. Illumina technology was further developed for microbiome characterization using targeted amplicon sequencing and whole genome sequencing of the multiple species present in a single sample (143, 144).

Following the second generation of NGS technologies, third generation technologies have been developed that allow for long-read sequencing (142, 145–150). Two leading technologies considered as third generation sequencing are PacBio SMRT and Oxford Nanopore. PacBio sequencing uses a nanophotonic surface containing a single DNA

polymerase with a single molecule of single-stranded DNA template in order to capture the fluorescence specified for each nucleobase. Nanopore is based on detection of different ionic flow generated from different dNTPs moving through the nanoscale pores that are embedded in their synthetic membrane (151, 152). These technologies allow for long reads of the DNA to be sequenced, for example, the PacBio machine allows sequencing of up to 20 kb reads in a single run, and Oxford Nanopore can sequence up to 2 MB (151, 152). Despite the relatively low-quality profiles compared to second generation NGS, it is expected that the current technical challenges will be overcome and these methods will soon become broadly applied in DNA sequencing field (137).

2.4.2 Microbiome sequencing

Metagenome sequencing is the universally used approach for identifying microbial communities in different systems. Since over 90% of the microbes cannot be cultured with biological culturing, metagenomics offers a variety of possibilities for identifying these communities (153). As mentioned above, metagenome sequencing was developed in parallel with NGS. Currently, the metagenomic methods are broadly divided into two approaches: (1) amplicon sequencing and (2) shotgun metagenomics based on its mechanism of identifying microbial communities (153). Amplicon sequencing allows characterization of the microorganisms present in a given sample based on the specific target sequences present in a specific group of microorganisms (e.g., 16S rRNA for bacteria or ITS for fungi). Shotgun metagenomics, on the other hand, is non-targeted sequencing of all DNA present in a samples and can be used to identify not only the collections of all microorganisms in a sample, but also their genetic materials by sequencing the whole genome of each

microorganism in parallel (154–156). In this study, we used the amplicon sequencing approach due to its cost-effectiveness.

2.4.2.1 Amplicon sequencing

In amplicon sequencing, specific genes or spacers are used as DNA fingerprints in order to identify the taxonomic profile of the selected microorganisms (143). The bacterial 16S ribosomal RNA (rRNA) gene contains nine hypervariable regions (V1 through V9) containing approximately 30 to 100 base pairs in each region (143, 157). Some of the regions are reliable enough to allow for identification of the taxonomic profile of the bacteria even though many of the regions by themselves can be used to accurately classify the bacteria from domain to species (158). Although ideally one would sequence the entire 16S rRNA gene in order to obtain the highest resolution of classification based on this gene, conventional NGS sequencers allow for sequencing of only a shorter fragment of DNA with sufficient quality. Thus, most studies choose either one region or a combination of two regions in order to ensure sufficient sequencing quality (157). While 16S rRNA gene sequencing is a powerful tool for identifying a microbial community, it does not correctly classify all microorganism, especially when microorganisms are closely and taxonomically related (143). Juan et al. showed that in the Enterobacteriaceae, Clostridiaceae and Peptostreptococcaceae species share up to 99% sequence similarity even across the full 16S rRNA gene (159). Thus, this technique cannot provide reliable classification at lower taxonomic levels, such as at the species level (159). While it is not the ideal approach for classifying bacterial communities, it is evident that analysis of the 16S rRNA hypervariable regions is considered as a broadly accepted and used practical approach for profiling of bacterial communities (160).

Internal transcribed spacers (ITS) are most commonly used sequencing method for characterization of fungal communities (161–163). ITS is an internal spacer DNA between the small and large subunit rRNA genes in the chromosome of the microorganism (161). Multiple studies have identified the variability of the ITS region sequence based on the fungal taxonomic classification. Otherwise, ITS, in the same manner as 16S rRNA gene, cannot reliably differentiate specific species due to a higher similarity of the sequences (164). Still, ITS is believed to be one of the most powerful tools for molecular classification of fungal communities up to the genus level. In that end, sequences need to be classified into unique groups or units in order to be considered as the same taxonomic profile. The most common classification of the sequence is based on defining operational taxonomic unit (OTU) (165). Most of the time, when two sequences have >97% similarity, they are considered as the same OTU; however, for more discriminatory classification, some researchers are using 99% similarity threshold for an OTU (165). Higher similarity thresholds do not always ensure higher resolution because those small changes in sequences might be due to sequencing error, and moreover, a particular region of hypervariable sequences cannot differentiate two taxonomic profiles as mentioned above.

One advantage of amplicon sequencing over shotgun metagenomics is its practical efficiency (153, 159). Since amplicon sequencing only produces a short fragment (up to 500bp) of DNA, it is cost-efficient, and also generates more data in a single sequencing run (166). On the other hand, amplicon sequencing also has several pitfalls and limitations stemming from its nature of the design. Since it is based on amplification via PCR, a bias may be created during amplification. In addition to PCR-induced bias, amplicon sequencing

also faces sequencing reproducibility issues. Since only a specific concentration of template can be sequenced during the library preparation of NGS, concentration normalization of the template, which generates sequencing bias toward the most abundant species occurs (167). Nevertheless, amplicon sequencing is widely used because of its simplicity and low cost, although shotgun metagenomics sequencing is emerging in environmental microbial community profiling (155).

2.4.2.2 Shotgun metagenomics

Shotgun metagenomic sequencing is based on whole-genome sequencing that does not focus on one strain or isolate but on all organisms in the analyzed sample (155). It is accomplished by unrestricted genomic sequencing of all microorganisms in a sample. Shotgun metagenomic sequencing can be used to study the species and strain-level compositions of microbial communities (154, 155, 159, 166). It can also provide insight in not only microbial community characteristics but also its functional capacity since the whole genome is sequenced, and it is possible to extract functional properties based on the gene content (156, 168, 169). Additionally, this approach does not rely on PCR amplification; thus, it can minimize PCR-associated bias (166). However, shotgun metagenomics is relatively expensive compared to amplicon sequencing and generates substantially larger amounts of data, which leads to a more complicated analysis. Additionally, most of the environmental communities are too diverse to be comprehensively characterized due to insufficient sequencing depth (170).

2.5 Amplicon sequencing data analyses

2.5.1 Overview

Analyses of microbial community profiles are traditionally based on biodiversity analyses using alpha and beta diversity (171–174). Biodiversity analyses are derived from ecological studies and have been adapted to the microbial community data (171, 174–176). In addition to that use, multivariate statistical analysis can also be utilized to analyze the high dimensional microbial community data (177). However, multivariate statistical analyses have many assumptions and specific conditions required to run, including multivariate normality, equal scaling of each variable, and multicollinearity. Hence, a thoroughly suitable method for microbial community data analysis does not exist at this time (177). The recently established machine learning approach is an emerging technique for analyzing microbial community compositions by overcoming the limitations of multivariate methods (178, 179).

2.5.2 OTU and taxonomic classification

In order to analyze data from massively parallel sequencing of amplicons, bioinformatic processing is required to classify raw sequences into OTUs. Various bioinformatic tools have been developed, each with unique pipelines (180, 181). Currently, two platforms, Mothur and Quantitative Insights into Microbial Ecology (QIIME2), are widely used (172, 182). Even though these two platforms have their own specified workflows, several core conceptually similar steps are shared across the platforms, and OTUs will are assigned at the end of both pipelines (181). Both QIIME and Mothur start with quality control of raw sequences. Unlike QIIME, Mothur has its own steps between quality control and classification, including sequence alignment and pre-clustering (172).

Sequence alignment allows for matching of sequences to the designated database (such as SILVA, UNITE). If a sequence is not aligned to the database, it will be discarded. During pre-clustering, rare sequences are merged into more abundant sequence within a given threshold. Those steps are computationally intensive; therefore, the overall sequence processing using Mothur is longer than that using QIIME (181). After the pre-processing, both platforms classify sequences based on the similarity and clustering into OTUs. Finally, each OTU is assigned a taxonomy and presented as an OTU table and corresponding taxonomic profile (181). Based on the previous review from Plummer et al., there are no significant differences between the results from Mothur and QIIME (181). Thus, the selection of the tools depends on users' preference and individual computational environments.

2.5.3 Normalization of OTUs

Two common pitfalls of the microbial community profile data include various sequencing depth among samples and data sparsity, meaning that the data has too many zeros (183). Different number of sequences per sample can cause an incorrect calculation of biodiversity because it does not reflect true biological variation but rather reflects differences in sequencing depth where samples sequenced at a greater depth will more likely have greater biodiversity discovered in them (183). The sparsity of the data implies that the counts of unique taxa are uncertain since there is a higher chance for detecting the unique taxa with a large library size and with greater sequencing depth (183). Those characteristics lead to erroneous diversity measurements and result in uncertainty regarding unique taxa counts. Thus, in most cases, the data need to be normalized prior to alpha diversity estimation. Because normalization is intended to enable accurate downstream analysis by

comparing each sample, it is important to use valid methods. Rarefying is the most common normalization method. Rarefying allows random resampling without replacement so that all samples have the same standardized library size. This approach is relatively simple. However, if the library size variation between samples is very large, it also ensures a higher chance of discarding important information by reducing library size into the minimum number of reads among samples. Still, rarefying is considered a default normalization method for microbial ecologists in the most commonly used platforms for analyzing microbial community data, such as Mothur, QIIME, Vegan, and Phyloseq (172, 182, 184, 185). Inadmissibility of rarefying data was first proposed by McMurdie et al. (186). By comparing the rarefied data to other normalization based on the statistical models such as variance stabilization implemented for the RNA sequencing analysis, upper-quartile log-fold change, and proportional normalization (186, 187), the researchers concluded that rarefying was associated with both higher false negative and positive rates than other normalizations (186). Thus, they suggested not to rarefy microbial community data, while other studies have argued that rarefying is the only normalization that ensures a fully normalized library size and also provides an accurate comparison among communities with ecological approaches (such as biodiversity) (188). Both studies did not draw a concrete conclusion one way or the other, but they encouraged researchers to consider their hypotheses or at least compare the results from different normalization and select the most suitable for addressing the questions that they are asking (183, 186, 188)

2.5.4 Analysis of microbial community diversity

2.5.4.1 Alpha-diversity within sample microbial diversity

Alpha diversity measures the diversity within a sample (171). It is an measurement independent of other samples in the same system or environment (189). The term alpha diversity was first introduced by R. H. Whittaker in the mid-1990s for the purpose of measuring species diversity in a particular ecosystem (171). Most of the time, alpha diversity measures species richness and the evenness of the sample. The richness of the sample represents the number of unique species present in the sample. For example, if one sample contains a large number of one species, the richness of the sample is very low, whereas if one sample has only two detected species and the two are different as opposed to two species that are the same, the richness is higher than in the first case. Richness is meant to identify the diversity of the sample, but the abundance of each species is not taken into account. Evenness of the sample measures the proportions of species present at a site or in a sample. Higher evenness indicates that the proportion of each species abundance is similar. In other words, low evenness means there are a few species that dominate the sample (176).

In order to overcome the limitation of using either richness or evenness for analyses, several alpha diversity indices have been developed that combine both components in a different ratio. The most common three alpha diversity indices are Chao1, Shannon, and Simpson (190–192). Those three measurements have their own characteristics. Chao1 describes species richness in a sample (193). Shannon and Simpson's indices describe both richness and evenness when they are calculated in a different ratio. For example, the Simpson index places more weight on the more dominant species in the sample (173, 191).

Still, there is no absolute standard of measuring alpha diversity using only one index. Therefore, most of the research relies on multiple indices.

The alpha diversity index allows for interpretation of microbial diversity in a unidimensional space since each measurement will come out with a single value. As a result, statistical analysis suitable for non-parametric univariate tests, such as the Kruskal-Wallis test, can be utilized in order to find the statistical difference of alpha diversity between two or more groups of samples (194, 195). If more than two groups are compared, a post-hoc Dunn's test needs to be applied in order to identify specific pairs of groups that are significantly different (196, 197).

2.5.4.2 Estimating total diversity using rarefaction curves

In order to identify the overall diversity of the samples, several studies have attempted to estimate the total diversity that sequencing results could not cover. Based on this aspect, rarefaction curves and their asymptotes are utilized to estimate the total diversity of the sample, whereas several studies have argued that the estimation does not have any statistical power (198, 199). Rarefaction curves are constructed based on random subsampling without replacement from the produced sequencing data (200). From the OTU table, a random number of OTUs can be subsampled several times. An average number of unique OTUs resulted are then plotted (198). Thus, the rarefaction curve generates the expected number of unique OTUs in a certain number of OTUs in a large pool of matrices (198). When the curve reaches the plateau that is mostly parallel to the x-axis, mean unique OTUs are detected by increasing the number of OTUs selected for subsampling. At this

point, sequencing depth can be considered saturated, which reflects the microbial community more accurately (198, 201).

2.5.4.3 Beta-diversity describes among-sample diversity

Unlike alpha-diversity, beta-diversity refers to diversity between samples (171, 176, 202). The concept of the beta-diversity was first introduced by R. H. Whittaker in conjunction with the concept of alpha-diversity (171). Most of the microbial studies rely on a beta-diversity comparison between samples to compare the compositional changes/differences in the communities (203). There are several beta-diversity indices that account for both the presences and abundance of each species (174). Several different beta-diversity indices were established and are widely applied not only to ecological studies but also to microbial ecologies. The Bray-Curtis dissimilarity measurement is the most common index for quantifying the compositional dissimilarity between two samples (204). The Bray-Curtis dissimilarity is calculated using the sum of the lesser counts for only those species in common between two samples and the average number of species in the two samples. Thus, the dissimilarity can consider both the abundance of each species and the total number of species in each sample. This index is bound to the range of 0 to 1 in which 0 means there are no overlapping species that exist in the two different ecosystems, whereas 1 means the two samples' composition is identical (204). Even though the Bray-Curtis dissimilarity index can widely be used for calculating the compositional difference between the two samples, phylogenetic distance measures can provide more insight toward the compositional differences between the two samples. Thus, UniFrac, which measures the distance between two communities based on the phylogenetic distance of each taxon along with its abundance, was introduced (205). Multivariate analyses with UniFrac frequently retrieve

meaningful biological patterns or biological distances between two communities that other beta-diversity indices cannot capture (205). One possible pitfall of using UniFrac is the possibility of boosting the importance of unique taxa. If one taxon, which is not commonly present in all of the samples, has a close relationship with the most abundant taxa, the distance does not accurately reflect the real dissimilarity between two communities. To that end, a weighted UniFrac can be adopted (205). A weighted UniFrac is fundamentally the same as an un-weighted UniFrac, but the weighted version gives more weight to the most abundant taxa. Thus, it can balance the weight of abundance of taxa to the phylogenetic distance between each taxon (206).

2.5.4.4 Ordination analysis and visualization of microbiome clusters

Beta diversity does not have any capability for reducing the complexity of the data because the results of calculating beta diversity are also constructed as a matrix. However, either ordination of the dissimilarity between samples or statistical analysis of the beta-diversity measurement is widely used (189). Ordination of the sample is based on projecting each sample in 2D dimension with respect of their distance between each other. Principal coordination analysis (PCoA), non-metric multidimensional scaling (NMDS), and canonical correspondence analysis (CCA) are commonly used ordination methods (207–209). PCoA and NMDS compare the sole distance between samples, whereas CCA was developed to analyze two multivariate sets of variables (such as OTU tables and metadata for samples reflecting the sample characteristics) and their relationships with each other. The main difference between PCoA and NMDS is the construction and projection into the two-dimensional (2D) space (184). PCoA uses the eigen decomposition of the transformed dissimilarity matrix (207), whereas NMDS is utilizes an iterative approximation algorithm

(208). Eigenvalue decomposition is a widely used multivariate approach for reducing the dimension by selecting two or three of the most informative coordinates and projecting them onto a 2D or 3D dimension (207). On the other hand, NMDS does select a specific number from the axis to be used for projection, but the dissimilarities between samples are distorted during calculation and projection in the 2D aspect, which means the relative distance on the plot does not indicate the actual distance between two samples. Furthermore, depending on the first sample, which is randomly selected and projected onto the plot, the results of NMDS can be different every time even though the clustering and ordination would not be significantly affected (208). NMDS is comparable to PCoA when the calculated eigenvalue for the eigenvalue deposition method in the first few dimensions are similar and there is only a low proportion of the variance among samples. The ordination itself does not have any statistical meaning; however, it does help in visualization of potential clusters of samples. This form of visualization can guide further analyses aimed at testing the effects of factors of interest on the microbial communities' composition (189, 210).

2.5.4.5 Multivariate statistical analyses

One of the most widely used statistical analysis of beta-diversity is a permutational multivariate analysis of variance (PERMANOVA), which is designed for the space of a chosen dissimilarity measure (211). PERMANOVA originates from the conventional analysis of variance (ANOVA) and its multivariate analysis of variance (MANOVA), but PERMANOVA is a non-parametric analysis that does not assume a normal variable distribution. This assumption is possible due to its permutational testing mechanism. Geometric partitioning of the samples is calculated according to the ANOVA design in which all of the labels of the samples are shuffled randomly. The F-statistic is calculated

repeatedly, usually either 999 or 9999 times depending on the sample size (211). All new F-values are compared and checked in order to determine if it is significantly changed each time. Thus, if the statistical value change significantly, it means that random shuffling affects the pattern or ordination of the samples (211). If there was no significant pattern from the beginning, it does not matter if the labels were shuffled, as this does not result in changes in statistical F-value (211).

2.5.5 Machine learning microbiome data analyses

One of the reasons why microbiome analyses rely on biodiversity and statistical analyses approach are their high-dimensional structures. In order to overcome this limitation, machine learning approaches have recently been adapted to microbiome data analyses (178, 179, 212–215). Machine learning is mathematical modeling used for predictions or decisions without programmed testing (216). The training set is used to choose the overall data patterns, and the validation set of data confirms the accuracy of the prediction of the choice (216). For microbiome data, random forest classification and regression are suitable for classifying samples or for finding predictive regression between independent and dependent variables (217, 218).

Random forest (RF) classification and regression are most commonly used machine learning algorithms in microbial community data analyses. RF is an ensemble learning method for classification, regression, and decision making based on the data set (218). Random forest involves construction of a random number of decision trees in order to predict potential classifications or regressions (218). A decision tree is a simple classification model obtained by splitting the data. After each splitting of the nodes in the

tree, the samples are split by increasing the purity of each result. Thus, multiple decision trees can validate the factors that allow classifying the samples into the groups (217). With those factors, the RF classification model is constructed, and when the new set of samples is introduced, RF predicts the group to which the new sample belongs. Based on the accuracy of the prediction, overall model validity can be calculated (219). In microbiomes, each of the taxa counts is a potential classifier. Delphine et al. successfully classified patients with irritable bowel syndrome into the specific subtypes using an RF classification with >95% accuracy based on the patient's microbiome (213). Based on these results, they can also extract the proportion of specific genus that may have a role as a classifier (213). RF classifications were used to predict the disease states of the patient using metagenomics data. They predicted liver cirrhosis with 95% accuracy, colorectal cancer with 87% accuracy, irritable bowel syndrome with 89% accuracy, type 2 diabetes with 74% accuracy, and obesity with 65% accuracy (179, 212, 213, 215). Those studies focused on the human gut microbiome, which is considered to have less variety than environmental microbiome. However, there are no significant results involving the environmental microbiome and prediction using the RF or any other machine learning approaches.

2.6 References

1. CDC. National Outbreak Reporting System (NORS). 2017. https://wwwn.cdc.gov/norsdashboard/

2. CDC. Foodborne Illness and Outbreaks. 2019. https://www.cdc.gov/foodsafety/outbreaks/multistate-outbreaks/outbreaks-list.html

3. FDA. Environmental Assessment of Factors Potentially Contributing to the Contamination of Romaine Lettuce Implicated in a Multi-State Outbreak of *E. coli* O157:H7. 2018. https://www.fda.gov/Food/RecallsOutbreaksEmergencies/Outbreaks/ucm624546.htm

4. FDA. FDA Continues Investigation into Source of *E. coli* O157:H7 Outbreak Linked to Romaine Lettuce Grown in CA; CDC Reports End to Associated Illnesses. 2019. https://www.fda.gov/Food/RecallsOutbreaksEmergencies/Outbreaks/ucm626330.htm

5. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. Nature Reviews Microbiology. 2010 Mar 1;8:207. https://doi.org/10.1038/nrmicro2298

6. Lim JY, Yoon J, Hovde CJ. A brief overview of *Escherichia coli* O157:H7 and its plasmid O157. J Microbiol Biotechnol. 2010 Jan;20(1):5–14. https://www.ncbi.nlm.nih.gov/pubmed/20134227

7. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in understanding enteric pathogenic *Escherichia coli.* Clin Microbiol Rev. 2013 Oct;26(4):822–80. https://www.ncbi.nlm.nih.gov/pubmed/24092857

8. Delannoy S, Beutin L, Fach P. Discrimination of enterohemorrhagic *Escherichia coli* (EHEC) from non-EHEC strains based on detection of various combinations of type III effector genes. J Clin Microbiol. 2013 Oct;51(10):3257–62. https://www.ncbi.nlm.nih.gov/pubmed/23884997

9. Gyles CL. Shiga toxin-producing *Escherichia coli*: An overview1. J Anim Sci. 2007 Mar 1;85:E45–62. https://dx.doi.org/10.2527/jas.2006-508

10. O'Brien AD, LaVeck GD, Thompson MR, Formal SB. Production of *Shigella dysenteriae* Type 1-Like Cytotoxin by *Escherichia coli*. J Infect Dis. 1982 Dec 1;146(6):763–9. https://dx.doi.org/10.1093/infdis/146.6.763

11. Melton-Celsa AR. Shiga Toxin (Stx) Classification, Structure, and Function. Microbiol Spectr. 2014/07/31. 2014;2(4):10.1128/microbiolspec.EHEC-0024-2013–2013.

12. Nataro JP, Kaper JB. Diarrheagenic *Escherichia coli*. Clin Microbiol Rev. 1998 Jan;11(1):142–201.

13. Pickering LK, Obrig TG, STAPLETON BF. Hemolytic-uremic syndrome and enterohemorrhagic *Escherichia coli*. Pediatr Infect Dis J. 1994;13(6):459–75.

14. Louie M, de Azavedo JC, Handelsman MY, Clark CG, Ally B, Dytoc M, et al. Expression and characterization of the eaeA gene product of *Escherichia coli* serotype O157:H7. Infect Immun . 1993 Oct;61(10):4085–92.

15. Nguyen Y, Sperandio V. Enterohemorrhagic *E. coli* (EHEC) pathogenesis. Front Cell Infect Microbiol. 2012 Jul 12;2:90.

16. Wang G, Clark CG, Rodgers FG. Detection in *Escherichia coli* of the genes encoding the major virulence factors, the genes defining the O157:H7 serotype, and components of the type 2 Shiga toxin family by multiplex PCR. J Clin Microbiol.

2002 Oct;40(10):3613–9.

17. Riley LW, Remis RS, Helgerson SD, McGee HB, Wells JG, Davis BR, et al. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. N Engl J Med. 1983;308(12):681–5.

18. Karmali M, Petric M, Steele B, Lim C. Sporadic cases of haemolytic-uraemic syndrome associated with faecal cytotoxin and cytotoxin-producing *Escherichia coli* in stools. Lancet. 1983;321(8325):619–20.

19. Villysson A, Tontanahal A, Karpman D. Microvesicle involvement in Shiga toxin-associated infection. Toxins (Basel). 2017;9(11):376.

20. Persad AK, Lejeune JT. Animal reservoirs of Shiga toxin-producing *Escherichia coli*. In: Enterohemorrhagic *Escherichia coli* and Other Shiga Toxin-Producing *E coli*. American Society of Microbiology; 2015. p. 231–44.

21. Soderstrom A, Osterberg P, Lindqvist A, Jonsson B, Lindberg A, Blide Ulander S, et al. A large *Escherichia coli* O157 outbreak in Sweden associated with locally produced lettuce. Foodborne Pathog Dis. 2008 Jun;5(3):339–49.

22. FDA. Investigation Summary: Factors Potentially Contributing to the Contamination of Romaine Lettuce Implicated in the Fall 2018 Multi-State Outbreak of *E. coli* O157:H7. 2019. https://www.fda.gov/Food/RecallsOutbreaksEmergencies/Outbreaks/ucm631243.html

23. Scallan E, Hoekstra RM, Angulo FJ, Tauxe R V, Widdowson M-A, Roy SL, et al.

Foodborne illness acquired in the United States--major pathogens. Emerg Infect Dis. 2011 Jan;17(1):7–15.

24. Brenner FW, Villar RG, Angulo FJ, Tauxe R, Swaminathan B. *Salmonella* Nomenclature. J Clin Microbiol. 2000 Jul 1;38(7):2465 LP – 2467.

25. Porwollik S, Boyd EF, Choy C, Cheng P, Florea L, Proctor E, et al. Characterization of *Salmonella enterica* Subspecies I Genovars by Use of Microarrays. J Bacteriol. 2004 Sep 1;186(17):5883 LP – 5898.

26. Andino A, Hanning I. *Salmonella enterica*: survival, colonization, and virulence differences among serovars. Scientific World Journal. 2015/01/13. 2015;2015:520179. https://www.ncbi.nlm.nih.gov/pubmed/25664339

27. Giammanco GM, Pignato S, Mammina C, Grimont F, Grimont PAD, Nastasi A, et al. Persistent Endemicity of *Salmonella bongori* 48:z₃₅:− in Southern Italy: Molecular Characterization of Human, Animal, and Environmental Isolates. J Clin Microbiol. 2002 Sep 1;40(9):3502 LP – 3505. http://jcm.asm.org/content/40/9/3502.abstract

28. Giannella RA. *Salmonella*. In: S B, editor. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston; 1996. https://www.ncbi.nlm.nih.gov/books/NBK8435/

29. Wiedemann A, Virlogeux-Payant I, Chaussé A-M, Schikora A, Velge P. Interactions of *Salmonella* with animals and plants. Front Microbiol. 2015 Jan 21;5:791.

https://www.ncbi.nlm.nih.gov/pubmed/25653644

30. Cummings KJ, Warnick LD, Elton M, Rodriguez-Rivera LD, Siler JD, Wright EM, et al. *Salmonella enterica* serotype Cerro among dairy cattle in New York: an emerging pathogen? Foodborne Pathog Dis. 2010 Jun;7(6):659–65. https://www.ncbi.nlm.nih.gov/pubmed/20187753

31. Angelo KM, Chu A, Anand M, Nguyen T-A, Bottichio L, Wise M, et al. Outbreak of *Salmonella* Newport infections linked to cucumbers--United States, 2014. MMWR Morb Mortal Wkly Rep. 2015 Feb 20;64(6):144–7. https://www.ncbi.nlm.nih.gov/pubmed/25695319

32. Bennet SD, Littrell KW, Hill TA, Mahovic M, Behravesh CB. Multistate foodborne disease outbreaks associated with raw tomatoes, United States, 1990–2010: a recurring public health problem. Epidemiol Infect. 2014/08/28. 2015;143(7):1352–9. https://www.cambridge.org/core/article/multistate-foodborne-disease-outbreaks-associated-with-raw-tomatoes-united-states-19902010-a-recurring-public-health-problem/BDE96249108657A04B53A36F3501D050

33. Farber JM, Peterkin PI. *Listeria monocytogenes*, a food-borne pathogen. Microbiol Rev. 1991 Sep;55(3):476–511. https://www.ncbi.nlm.nih.gov/pubmed/1943998

34. Denis N, Zhang H, Leroux A, Trudel R, Bietlot H. Prevalence and trends of bacterial contamination in fresh fruits and vegetables sold at retail in Canada. Food Control. 2016;67:225–34.

35. Ponniah J, Robin T, Paie MS, Radu S, Ghazali FM, Kqueen CY, et al. *Listeria monocytogenes* in raw salad vegetables sold at retail level in Malaysia. Food Control. 2010;21(5):774–8.

36. CDC. Jansen Farms Cantaloupes (Outbreak summary). 2012. https://www.cdc.gov/*Listeria*/outbreaks/cantaloupes-jensen-farms/index.html

37. Dawson SJ, Evans MR, Willby D, Bardwell J, Chamberlain N, Lewis DA. *Listeria* outbreak associated with sandwich consumption from a hospital retail shop, United Kingdom. Euro Surveill. 2006;11(6):89–91.

38. Dalton CB, Austin CC, Sobel J, Hayes PS, Bibb WF, Graves LM, et al. An outbreak of gastroenteritis and fever due to *Listeria monocytogenes* in milk. N Engl J Med. 1997;336(2):100–6.

39. Salamina G, Dalle Donne E, Niccolini A, Poda G, Cesaroni D, Bucci M, et al. A foodborne outbreak of gastroenteritis involving *Listeria monocytogenes*. Epidemiol Infect. 1996;117(3):429–36.

40. Miettinen MK, Björkroth KJ, Korkeala HJ. Characterization of *Listeria monocytogenes* from an ice cream plant by serotyping and pulsed-field gel electrophoresis. Int J Food Microbiol. 1999;46(3):187–92.

41. Farber JM, Sanders GW, Johnston MA. A survey of various foods for the presence of *Listeria* species. J Food Prot. 1989;52(7):456–8.

42. Prevention C for DC and. Multistate outbreak of listeriosis linked to packaged salads

produced at Springfield, Ohio Dole processing facility (final update). Centers Dis Control Prev Atlanta, GA. http//www cdc gov/*Listeria*/outbreaks/bagged-salads-01-16/index html Accessed August. 2016;

43. Zhu Q, Gooneratne R, Hussain AM. *Listeria monocytogenes* in Fresh Produce: Outbreaks, Prevalence and Contamination Levels. Vol. 6, Foods . 2017.

44. de Noordhout CM, Devleesschauwer B, Angulo FJ, Verbeke G, Haagsma J, Kirk M, et al. The global burden of listeriosis: a systematic review and meta-analysis. Lancet Infect Dis. 2014/09/15. 2014 Nov;14(11):1073–82. https://www.ncbi.nlm.nih.gov/pubmed/25241232

45. Ceuppens S, Johannessen GS, Allende A, Tondo EC, El-Tahan F, Sampers I, et al. Risk Factors for *Salmonella*, Shiga Toxin-Producing *Escherichia coli* and Campylobacter Occurrence in Primary Production of Leafy Greens and Strawberries. Int J Environ Res Public Health. 2015 Aug 18;12(8):9809–31. https://www.ncbi.nlm.nih.gov/pubmed/26295251

46. Holvoet K, Sampers I, Callens B, Dewulf J, Uyttendaele M. Moderate prevalence of antimicrobial resistance in *Escherichia coli* isolates from lettuce, irrigation water, and soil. Appl Environ Microbiol. 2013 Nov;79(21):6677–83. https://www.ncbi.nlm.nih.gov/pubmed/23974140

47. Weller D, Wiedmann M, Strawn LK. Irrigation Is Significantly Associated with an Increased Prevalence of *Listeria monocytogenes* in Produce Production Environments in New York State. J Food Prot. 2015 Jun;78(6):1132–41.

http://jfoodprotection.org/doi/10.4315/0362-028X.JFP-14-584

48. Castillo A, Mercado I, Lucia LM, Martinez-Ruiz Y, De Leon JP, Murano EA, et al. *Salmonella* contamination during production of cantaloupe: a binational study. J Food Prot. 2004;67(4):713–20.

49. Ceuppens S, Hessel CT, de Quadros Rodrigues R, Bartz S, Tondo EC, Uyttendaele M. Microbiological quality and safety assessment of lettuce production in Brazil. Int J Food Microbiol. 2014 Jul 2;181:67–76. https://www.sciencedirect.com/science/article/pii/S0168160514002013

50. Chen Z, Jiang X. Microbiological safety of chicken litter or chicken litter-based organic fertilizers: a review. Agriculture. 2014;4(1):1–29.

51. Jay-Russell MT, Hake AF, Bengson Y, Thiptara A, Nguyen T. Prevalence and characterization of *Escherichia coli* and *Salmonella* strains isolated from stray dog and coyote feces in a major leafy greens production region at the United States-Mexico border. PLoS One. 2014 Nov 20;9(11):e113433–e113433. https://www.ncbi.nlm.nih.gov/pubmed/25412333

52. Erickson MC, Webb CC, Diaz-perez JC, Phatak SC, Silvoy JJ, Davey L, Infrequent Internalization of *Escherichia coli* O157:H7 into Field-Grown Leafy Greens. J Food Prot. 2010 Mar 1;73(3):500–6. https://doi.org/10.4315/0362-028X-73.3.500

53. Zhang G, Ma LI, Beuchat LR, Erickson MC, Phelan VH, Doyle MP. Lack of Internalization of *Escherichia coli* O157:H7 in Lettuce (Lactuca sativa L.) after Leaf

Surface and Soil Inoculation. J Food Prot. 2009 Oct 1;72(10):2028–37. https://doi.org/10.4315/0362-028X-72.10.2028

54. Holley RA, Arrus KM, Ominski KH, Tenuta M, Blank G. *Salmonella* Survival in Manure-Treated Soils during Simulated Seasonal Temperature Exposure. J Environ Qual. 2006;35:1170–80. http://dx.doi.org/10.2134/jeq2005.0449

55. Sharma M, Millner PD, Hashem F, Vinyard BT, East CL, Handy ET, et al. Survival of *Escherichia coli* in Manure-Amended Soils Is Affected by Spatiotemporal, Agricultural, and Weather Factors in the Mid-Atlantic United States. Schaffner DW, editor. Appl Environ Microbiol. 2019 Mar 1;85(5):e02392-18. http://aem.asm.org/content/85/5/e02392-18.abstract

56. Weller DL, Kovac J, Kent DJ, Roof S, Tokman JI, Mudrak E, et al. *Escherichia coli* transfer from simulated wildlife feces to lettuce during foliar irrigation: A field study in the Northeastern United States. Food Microbiol. 2017;68:24–33. http://www.sciencedirect.com/science/article/pii/S0740002016310310

57. Atwill ER, Chase JA, Oryang D, Bond RF, Koike ST, Cahn MD, Transfer of *Escherichia coli* O157: H7 from simulated wildlife scat onto romaine lettuce during foliar irrigation. J Food Prot. 2015;78(2):240–7.

58. Jones LA, Worobo RW, Smart CD. Plant-Pathogenic Oomycetes, *Escherichia coli* Strains, and *Salmonella* spp. Frequently Found in Surface Water Used for Irrigation of Fruit and Vegetable Crops in New York State. Elkins CA, editor. Appl Environ Microbiol. 2014 Aug 15;80(16):4814 LP – 4820.

http://aem.asm.org/content/80/16/4814.abstract

59. Delaquis P, Bach S, Dinu L-D. Behavior of *Escherichia coli* O157:H7 in Leafy Vegetables. J Food Prot. 2007 Aug 1;70(8):1966–74. https://doi.org/10.4315/0362-028X-70.8.1966

60. Markland SM, Ingram D, Kniel KE, Sharma M. Water for Agriculture: the Convergence of Sustainability and Safety. Microbiol Spectr. 2017 May;5(3).

61. Cooley MB, QuiÃ±ones B, Oryang D, Mandrell RE, Gorski L. Prevalence of shiga toxin producing *Escherichia coli*, *Salmonella* enterica, and *Listeria monocytogenes* at public access watershed sites in a California Central Coast agricultural region. Front Cell Infect Microbiol. 2014 Mar 4;4:30. http://journal.frontiersin.org/article/10.3389/fcimb.2014.00030/abstract

62. Cho S, Hiott LM, Barrett JB, McMillan EA, House SL, Humayoun SB, et al. Prevalence and characterization of *Escherichia coli* isolated from the Upper Oconee Watershed in Northeast Georgia. PLoS One. 2018;13(5):e0197005.

63. Sidabutar N V, Namara I, Hartono DM, Soesilo TEB. The effect of anthropogenic activities to the decrease of water quality. In: IOP Conference Series: Earth and Environmental Science. IOP Publishing; 2017. p. 12034.

64. Khatri N, Tyagi S. Influences of natural and anthropogenic factors on surface and groundwater quality in rural and urban areas. Front Life Sci. 2015 Jan 2;8(1):23–39. https://doi.org/10.1080/21553769.2014.933716

65. Van Beneden CA, Keene WE, Strang RA, Werker DH, King AS, Mahon B, et al. Multinational Outbreak of *Salmonella* enterica Serotype Newport Infections Due to Contaminated Alfalfa Sprouts. JAMA. 1999 Jan 13;281(2):158–62. https://dx.doi.org/10.1001/jama.281.2.158

66. Hooda PS, Edwards AC, Anderson HA, Miller A. A review of water quality concerns in livestock farming areas. Sci Total Environ. 2000;250(1):143–67. http://www.sciencedirect.com/science/article/pii/S0048969700003739

67. Islam M, Doyle MP, Phatak SC, Millner P, Jiang X. Persistence of enterohemorrhagic *Escherichia coli* O157: H7 in soil and on leaf lettuce and parsley grown in fields treated with contaminated manure composts or irrigation water. J Food Prot. 2004;67(7):1365–70.

68. Thurston-Enriquez JA, Gilley JE, Eghball B. Microbial quality of runoff following land application of cattle manure and swine slurry. J Water Health. 2005;3(2):157–71.

69. Feng P, Weagant SD, Grant MA, Burkhardt W, Shellfish M, Water B. BAM: Enumeration of *Escherichia coli* and the Coliform Bacteria. Bacteriol Anal Man. 2002;13.

70. Andrews WH, Jacobson A, Hammack T. Bacteriological Analytical Manual (BAM). Chapter 5 *Salmonella*. 2011.

71. Hitchins AD, Jinneman K, Chen Y. BAM: detection and enumeration of *Listeria monocytogenes*. Bacteriol Anal Man. 2016;

72. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G. Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. Arch Pathol Lab Med. 2017 Feb 7;141(6):776–86. https://doi.org/10.5858/arpa.2016-0539-RA

73. Rocha LB, Santos ARR, Munhoz DD, Cardoso LTA, Luz DE, Andrade FB, et al. Development of a rapid agglutination latex test for diagnosis of enteropathogenic and enterohemorrhagic *Escherichia coli* infection in developing world: defining the biomarker, antibody and method. PLoS Negl Trop Dis. 2014 Sep 25;8(9):e3150–e3150. https://www.ncbi.nlm.nih.gov/pubmed/25254981

74. Postollec F, Falentin H, Pavan S, Combrisson J, Sohier D. Recent advances in quantitative PCR (qPCR) applications in food microbiology. Food Microbiol. 2011;28(5):848–61. http://www.sciencedirect.com/science/article/pii/S0740002011000505

75. Feldsine P, Abeyta C, Andrews WH. AOAC International methods committee guidelines for validation of qualitative and quantitative food microbiological official methods of analysis. J AOAC Int. 2002;85(5):1187–200.

76. Jasson V, Jacxsens L, Luning P, Rajkovic A, Uyttendaele M. Alternative microbial methods: An overview and selection criteria. Food Microbiol. 2010;27(6):710–30.

77. Rajapaksha P, Elbourne A, Gangadoo S, Brown R, Cozzolino D, Chapman J. A review of methods for the detection of pathogenic microorganisms. Analyst. 2019;144(2):396–411. http://dx.doi.org/10.1039/C8AN01488D

78. Law JW-F, Ab Mutalib N-S, Chan K-G, Lee L-H. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. Front Microbiol. 2015 Jan 12;5:770. https://www.ncbi.nlm.nih.gov/pubmed/25628612

79. Ashbolt NJ, Grabow WOK, Snozzi M. Indicators of microbial water quality. 2001;

80. Council NR. Introduction and Historical Background. In: Indicators for Waterborne Pathogens. National Academies Press (US); 2004.

81. Birks R, Hills S. Characterisation of indicator organisms and pathogens in domestic greywater for recycling. Environ Monit Assess. 2007;129(1–3):61–9.

82. Topalcengiz Z, Strawn LK, Danyluk MD. Microbial quality of agricultural water in Central Florida. PLoS One. 2017;12(4):e0174889. 10.1371/journal.pone.0174889

83. Castro-Ibáñez I, Gil MI, Tudela JA, Ivanek R, Allende A. Assessment of microbial risk factors and impact of meteorological conditions during production of baby spinach in the Southeast of Spain. Food Microbiol. 2015 Aug 1 [cited 2018 Aug 17];49:173–81. https://www.sciencedirect.com/science/article/pii/S0740002015000283

84. Thomas JC, Lutz MA, Bruce JL, Graczyk DJ, Richards KD, Krabbenhoft DP, et al. Water-quality characteristics for selected sites within the Milwaukee Metropolitan Sewerage District planning area, Wisconsin, February 2004–September 2005. Magnesium. 34(6).

85. Martin NH, Trmčić A, Hsieh T-H, Boor KJ, Wiedmann M. The Evolving Role of Coliforms As Indicators of Unhygienic Processing Conditions in Dairy Foods. Front Microbiol. 2016 Sep 30;7:1549. https://www.ncbi.nlm.nih.gov/pubmed/27746769

86. Odonkor ST, Ampofo JK. *Escherichia coli* as an indicator of bacteriological quality of water: an overview. Microbiol Res (Pavia). 2013;4(1):e2–e2.

87. Davis K, Anderson MA, Yates M V. Distribution of indicator bacteria in Canyon Lake, California. Water Res. 2005;39(7):1277–88. http://www.sciencedirect.com/science/article/pii/S0043135405000321

88. 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. 2001;55(1):201–34.

89. Ishii S, Sadowsky MJ. *Escherichia coli* in the environment: implications for water quality and human health. Microbes Environ. 2008;23(2):101–8.

90. Byappanahalli M, Fowler M, Shively D, Whitman R. Ubiquity and persistence of *Escherichia coli* in a Midwestern coastal stream. Appl Environ Microbiol. 2003;69(8):4549–55.

91. NandaKafle G, Christie AA, Vilain S, Brözel VS. Growth and Extended Survival of *Escherichia coli* O157:H7 in Soil Organic Matter . Vol. 9, Frontiers in Microbiology. 2018. p. 762. https://www.frontiersin.org/article/10.3389/fmicb.2018.00762

92. Van Elsas JD, Semenov A V, Costa R, Trevors JT. Survival of *Escherichia coli* in the environment: fundamental and public health aspects. ISME J. 2010/06/24. 2011 Feb;5(2):173–83. https://www.ncbi.nlm.nih.gov/pubmed/20574458

93. Tortorello ML. Indicator organisms for safety and quality—uses and methods for detection: minireview. J AOAC Int. 2003;86(6):1208–17.

94. Baylis C, Uyttendaele M, Joosten H, Davies A. The Enterobacteriaceae and their significance to the food industry. Enterobact their significance to food Ind. 2011;

95. Chapin TK, Nightingale KK, Worobo RW, Wiedmann M, Strawn LK. Geographical and Meteorological Factors Associated with Isolation of *Listeria* Species in New York State Produce Production and Natural Environments. J Food Prot. 2014 Nov 1;77(11):1919–28. https://doi.org/10.4315/0362-028X.JFP-14-132

96. McEgan R, Mootian G, Goodridge LD, Schaffner DW, Danyluk MD. Predicting *Salmonella* Populations from Biological, Chemical, and Physical Indicators in Florida Surface Waters. Appl Environ Microbiol. 2013 Jul 1;79(13):4094–105. http://aem.asm.org/lookup/doi/10.1128/AEM.00777-13

97. Burton GA, Gunnison D, Lanza GR. Survival of pathogenic bacteria in various freshwater sediments. Appl Environ Microbiol. 1987;53(4):633–8.

98. Rhodes MW, Kator H. Survival of *Escherichia coli* and *Salmonella* spp. in estuarine environments. Appl Environ Microbiol. 1988;54(12):2902–7.

99. Chandran A, Varghese S, Kandeler E, Thomas A, Hatha M, Mazumder A. An

assessment of potential public health risk associated with the extended survival of indicator and pathogenic bacteria in freshwater lake sediments. Int J Hyg Environ Health. 2011;214(3):258–64.

100. Korajkic A, McMinn B, Harwood V. Relationships between Microbial Indicators and Pathogens in Recreational Water Settings. Int J Environ Res Public Health. 2018;15(12):2842.

101. FDA. FSMA Final Rule on Produce Safety. 2017. https://www.fda.gov/Food/GuidanceRegulation/FSMA/ucm334114.htm

102. FDA. United States Food and Drug Administration. Standards for the growing, harvesting, packing, and holding of produce for human consumption. 2013. http://www.fda.gov/downloads/Food/GuidanceRegulation/FSMA/UCM360734.pdf.

103. TRUITT LN, VAZQUEZ KM, PFUNTNER RC, RIDEOUT SL, HAVELAAR AH, STRAWN LK. Microbial Quality of Agricultural Water Used in Produce Preharvest Production on the Eastern Shore of Virginia. J Food Prot. 2018 Sep 13;81(10):1661–72. https://doi.org/10.4315/0362-028X.JFP-18-185

104. Lothrop N, Bright KR, Sexton J, Pearce-Walker J, Reynolds KA, Verhougstraete MP. Optimal strategies for monitoring irrigation water quality. Agric Water Manag. 2018;199:86–92.

105. US EPA. Method 1603: *Escherichia coli* (*E. coli*) in water by membrane filtration using modified membrane-thermotolerant *Escherichia coli* agar (modified mTEC).

US Environmental Protection Agency, Office of Water; 2002.

106. Rock CM, Brassill N, Dery JL, Carr D, McLain JE, Bright KR, et al. Review of water quality criteria for water reuse and risk-based implications for irrigated produce under the FDA Food Safety Modernization Act, produce safety rule. Environ Res. 2019;172:616–29. http://www.sciencedirect.com/science/article/pii/S0013935118306856

107. Goyal SM, Gerba CP, Melnick JL. Occurrence and distribution of bacterial indicators and pathogens in canal communities along the Texas coast. Appl Environ Microbiol. 1977 Aug 1;34(2):139 LP – 149. http://aem.asm.org/content/34/2/139.abstract

108. Mansilha CR, Coelho CA, Reinas A, Moutinho A, Ferreira S, Pizarro C, et al. *Salmonella*: The forgotten pathogen: Health hazards of compliance with European Bathing Water Legislation. Mar Pollut Bull. 2010;60(6):819–26. http://www.sciencedirect.com/science/article/pii/S0025326X10000160

109. Poma V, Mamani N, Iñiguez V. Impact of urban contamination of the La Paz River basin on thermotolerant coliform density and occurrence of multiple antibiotic resistant enteric pathogens in river water, irrigated soil and fresh vegetables. Springerplus. 2016 Apr 22;5:499. https://www.ncbi.nlm.nih.gov/pubmed/27186463

110. Xiao G, Wang Z, Chen J, Qiu Z, Li Y, Qi J, et al. Occurrence and infection risk of waterborne pathogens in Wanzhou watershed of the Three Gorges Reservoir, China. J Environ Sci. 2013;25(9):1913–24. http://www.sciencedirect.com/science/article/pii/S1001074212602411

111. Directive C. 75/440/EEC of 16 June 1979 concerning the quality required of surface water intended for the abstraction of drinking water in the Member States. Off J L. 1975;194:25.

112. Fang T, Cui Q, Huang Y, Dong P, Wang H, Liu W-T, et al. Distribution comparison and risk assessment of free-floating and particle-attached bacterial pathogens in urban recreational water: Implications for water quality management. Sci Total Environ. 2018;613–614:428–38. http://www.sciencedirect.com/science/article/pii/S0048969717323501

113. Pachepsky Y, Shelton D, Dorner S, Whelan G. Can *E. coli* or thermotolerant coliform concentrations predict pathogen presence or prevalence in irrigation waters? Crit Rev Microbiol. 2016 May 3;42(3):384–93. https://doi.org/10.3109/1040841X.2014.954524

114. Haley BJ, Cole DJ, Lipp EK. Distribution, diversity, and seasonality of waterborne *Salmonella*e in a rural watershed. Appl Environ Microbiol. 2009 Mar;75(5):1248–55.

115. Truchado P, Hernandez N, Gil MI, Ivanek R, Allende A. Correlation between *E. coli* levels and the presence of foodborne pathogens in surface irrigation water: Establishment of a sampling program. Water Res. 2018 Jan 1 [cited 2018 Aug 17];128:226–33. https://www.sciencedirect.com/science/article/pii/S0043135417308643

116. Draper AD, Doores S, Gourama H, LaBorde LF. Microbial Survey of Pennsylvania Surface Water Used for Irrigating Produce Crops. J Food Prot. 2016;79(6):902–12.

http://www.ncbi.nlm.nih.gov/pubmed/27296593%0Ahttp://jfoodprotection.org/doi/10.4315/0362-028X.JFP-15-479

117. Henry R, Schang C, Kolotelo P, Coleman R, Rooney G, Schmidt J, et al. Effect of environmental parameters on pathogen and faecal indicator organism concentrations within an urban estuary. Estuar Coast Shelf Sci. 2016;174:18–26.

118. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007 Oct 17;449:804. https://doi.org/10.1038/nature06244

119. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biol. 2014;12(1):69.

120. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. BioMed Central; 2015.

121. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017 Nov 1;551:457. https://doi.org/10.1038/nature24621

122. Gomez-Smith CK, Tan DT, Shuai D. Research highlights: functions of the drinking water microbiome – from treatment to tap. Environ Sci Water Res Technol. 2016;2(2):245–9. http://dx.doi.org/10.1039/C6EW90007K

123. Pinto AJ, Xi C, Raskin L. Bacterial Community Structure in the Drinking Water Microbiome Is Governed by Filtration Processes. Environ Sci Technol. 2012 Aug

21;46(16):8851–9. https://doi.org/10.1021/es302042t

124. Ling F, Whitaker R, LeChevallier MW, Liu W-T. Drinking water microbiome assembly induced by water stagnation. ISME J. 2018;12(6):1520–31. https://doi.org/10.1038/s41396-018-0101-5

125. 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. 2013 Aug 8;115(5):1147–58. https://doi.org/10.1111/jam.12323

126. Jackson CR, Millar JJ, Payne JT, Ochs CA. Free-Living and Particle-Associated Bacterioplankton in Large Rivers of the Mississippi River Basin Demonstrate Biogeographic Patterns. Wommack KE, editor. Appl Environ Microbiol. 2014 Dec 1;80(23):7186 LP – 7195. http://aem.asm.org/content/80/23/7186.abstract

127. Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R, et al. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. Appl Environ Microbiol. 2011;77(17):6000–11.

128. Ghai R, Rodŕíguez-Valera F, McMahon KD, Toyama D, Rinke R, de Oliveira TCS, et al. Metagenomics of the water column in the pristine upper course of the Amazon river. PLoS One. 2011;6(8):e23785.

129. Van Rossum T, Peabody MA, Uyaguari-Diaz MI, Cronin KI, Chan M, Slobodan JR,

et al. Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality . Vol. 6, Frontiers in Microbiology. 2015. p. 1405. https://www.frontiersin.org/article/10.3389/fmicb.2015.01405

130. Wang P, Zhao J, Xiao H, Yang W, Yu X. Bacterial community composition shaped by water chemistry and geographic distance in an anthropogenically disturbed river. Sci Total Environ. 2019;655:61–9. http://www.sciencedirect.com/science/article/pii/S004896971834590X

131. Hwang C, Ling F, Andersen GL, LeChevallier MW, Liu W-T. Microbial Community Dynamics of an Urban Drinking Water Distribution System Subjected to Phases of Chloramination and Chlorination Treatments. Appl Environ Microbiol. 2012 Nov 15;78(22):7856 LP – 7865. http://aem.asm.org/content/78/22/7856.abstract

132. El-Chakhtoura J, Prest E, Saikaly P, van Loosdrecht M, Hammes F, Vrouwenvelder H. Dynamics of bacterial communities before and after distribution in a full-scale drinking water network. Water Res. 2015;74:180–90. http://www.sciencedirect.com/science/article/pii/S0043135415000883

133. Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl Environ Microbiol. 2011;77(11):3846–52.

134. Wang P, Chen B, Yuan R, Li C, Li Y. Characteristics of aquatic bacterial community and the influencing factors in an urban river. Sci Total Environ. 2016;569–570:382–9.

http://www.sciencedirect.com/science/article/pii/S0048969716313043

135. Chen Y, Jiang Y, Huang H, Mou L, Ru J, Zhao J, et al. Long-term and high-concentration heavy-metal contamination strongly influences the microbiome and functional genes in Yellow River sediments. Sci Total Environ. 2018;637–638:1400–12. http://www.sciencedirect.com/science/article/pii/S0048969718317509

136. Wang Y, Yang J, Liu L, Yu Z. Quantifying the effects of geographical and environmental factors on distribution of stream bacterioplankton within nature reserves of Fujian, China. Environ Sci Pollut Res. 2015;22(14):11010–21.

137. Greenleaf WJ, Sidow A. The future of sequencing: convergence of intelligent design and market Darwinism. BioMed Central; 2014.

138. Harismendy O, Frazer KA. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. Biotechniques. 2009;46(3):229–31.

139. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta - Mol Basis Dis. 2014;1842(10):1932–41. http://www.sciencedirect.com/science/article/pii/S092544391400180X

140. Nyrén P, Lundin A. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. Anal Biochem. 1985;151(2):504–9.

141. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science (80-). 1998;281(5375):363–5.

142. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1–8. https://www.ncbi.nlm.nih.gov/pubmed/26554401

143. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol. 2007/07/11. 2007 Sep;45(9):2761–4. https://www.ncbi.nlm.nih.gov/pubmed/17626177

144. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428(6978):37–43. https://doi.org/10.1038/nature02340

145. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19(R2):R227–40.

146. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. Anal Chem. 2011;83(12):4327–41.

147. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52(4):413–35.

148. Gut IG. New sequencing technologies. Clin Transl Oncol. 2013;15(11):879–81.

149. Bell RL, Zheng J, Burrows E, Allard S, Wang CY, Keys CE, et al. Ecological prevalence, genetic diversity, and epidemiological aspects of *Salmonella* isolated from tomato agricultural regions of the Virginia Eastern Shore. Front Microbiol. 2015;6:415.

150. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science (80-). 2008;320(5872):106–9.

151. Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418–26.

152. Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. Nature Publishing Group; 2012.

153. Shah N, Tang H, Doak TG, Ye Y. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. In: Biocomputing 2011. World Scientific; 2011. p. 165–76.

154. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017 Sep 12;35:833. https://doi.org/10.1038/nbt.3935

155. Scholz M, Ward D V, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13(5):435.

156. Bengtsson-Palme J, Boulund F, Fick J, Kristiansson E, Larsson DG. Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. Front Microbiol. 2014;5:648.

157. Větrovský T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One. 2013 Feb

27;8(2):e57923–e57923. https://www.ncbi.nlm.nih.gov/pubmed/23460914

158. Zhang J, Ding X, Guan R, Zhu C, Xu C, Zhu B, et al. Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. Sci Total Environ. 2018;618:1254–67. http://www.sciencedirect.com/science/article/pii/S0048969717325792

159. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front Microbiol. 2016 Apr 20;7:459. https://www.ncbi.nlm.nih.gov/pubmed/27148170

160. Walters W, Hyde ER, Berg-Lyons D, Ackermann G, Humphrey G, Parada A, et al. Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. Bik H, editor. mSystems. 2016;1(1).

161. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi Proc Natl Acad Sci. 2012 Apr 17;109(16):6241 LP – 6246. http://www.pnas.org/content/109/16/6241.abstract

162. Martin KJ, Rygiewicz PT. Fungal-specific PCR primers developed for analysis of the ITS region of environmental DNA extracts. BMC Microbiol. 2005;5(1):28.

163. Gardes M, Bruns TD. ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. Mol Ecol. 1993;2(2):113–8.

164. Seifert KA, Samson RA, Houbraken J, Lévesque CA, Moncalvo J-M, Louis-Seize G, et al. Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case. Proc Natl Acad Sci. 2007;104(10):3901–6.

165. Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, et al. Defining operational taxonomic units using DNA barcode data. Philos Trans R Soc B Biol Sci. 2005;360(1462):1935–43.

166. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 2016;469(4):967–77.

167. Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, Konstantinidis KT. Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. PLoS One. 2014 Apr 8;9(4):e93827. https://doi.org/10.1371/journal.pone.0093827

168. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci. 2012;109(52):21390–5.

169. Campanaro S, Treu L, Kougias PG, De Francisci D, Valle G, Angelidaki I. Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy. Biotechnol Biofuels. 2016;9(1):26.

170. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data . Vol. 5, Frontiers in Plant Science. 2014. p. 209. https://www.frontiersin.org/article/10.3389/fpls.2014.00209

171. Whittaker RH. Evolution and Measurement of Species Diversity. Taxon. 1972;21(2/3):213–51. http://www.jstor.org/stable/1218190

172. 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. 2009 Dec;75(23):7537–41.

173. Morris EK, Caruso T, Buscot F, Fischer M, Hancock C, Maier TS, et al. Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. Ecol Evol. 2014/08/28. 2014 Sep;4(18):3514–24. https://www.ncbi.nlm.nih.gov/pubmed/25478144

174. Wang J, Soininen J, Zhang Y, Wang B, Yang X, Shen J. Patterns of elevational beta diversity in micro- and macroorganisms. Glob Ecol Biogeogr. 2012 Jul 1;21(7):743–50. https://doi.org/10.1111/j.1466-8238.2011.00718.x

175. Prosser JI, Bohannan BJM, Curtis TP, Ellis RJ, Firestone MK, Freckleton RP, et al. The role of ecological theory in microbial ecology. Nat Rev Microbiol. 2007 May 1;5:384. https://doi.org/10.1038/nrmicro1643

176. Sepkoski JJ. Alpha, beta, or gamma: where does all the diversity go? Paleobiology.

2016/02/08. 1988;14(3):221–34. https://www.cambridge.org/core/article/alpha-beta-or-gamma-where-does-all-the-diversity-go/7E82A0ADFA5C6A942C1410A26368A89C

177. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Genes Dis. 2017;4(3):138–48. http://www.sciencedirect.com/science/article/pii/S2352304217300351

178. Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale J V, DeBruyn JM, et al. A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. PLoS One. 2016 Dec 22;11(12):e0167370. https://doi.org/10.1371/journal.pone.0167370

179. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome. 2013;1(1):11. https://doi.org/10.1186/2049-2618-1-11

180. López-García A, Pineda-Quiroga C, Atxaerandio R, Pérez A, Hernández I, García-Rodríguez A, et al. Comparison of Mothur and QIIME for the Analysis of Rumen Microbiota Composition Based on 16S rRNA Amplicon Sequences. Front Microbiol. 2018 Dec 13;9:3010. https://www.ncbi.nlm.nih.gov/pubmed/30619117

181. Plummer E, Twin J, Bulach DM, Garland SM, Tabrizi SN. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. J Proteomics Bioinform. 2015;8(12):283.

182. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010/04/11. 2010 May;7(5):335–6. https://www.ncbi.nlm.nih.gov/pubmed/20383131

183. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017 Mar;5(1):27.

184. Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre D, McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos MHH, Stevens ES and HW. vegan: Community Ecology Package. R package. 2018. https://cran.r-project.org/package=vegan

185. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One. 2013 Apr 22;8(4):e61217. https://doi.org/10.1371/journal.pone.0061217

186. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10. https://doi.org/10.1371/journal.pcbi.1003531

187. Maza E. In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. Front Genet. 2016 Sep 16;7:164. https://www.ncbi.nlm.nih.gov/pubmed/27695478

188. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. Methods for normalizing microbiome data: An ecological perspective. Methods Ecol Evol. 2019 Mar 1;10(3):389–400. https://doi.org/10.1111/2041-210X.13115

189. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, et al. Conducting a Microbiome Study. Cell. 2014;158(2):250–62. http://www.sciencedirect.com/science/article/pii/S0092867414008642

190. Coddington JA, Young LH, Coyle FA. Estimating spider species richness in a southern Appalachian cove hardwood forest. J Arachnol. 1996;111–28.

191. Keylock CJ. Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. Oikos. 2005;109(1):203–7.

192. Chao A, Shen T-J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environ Ecol Stat. 2003;10(4):429–43.

193. Chao A. Nonparametric Estimation of the Number of Classes in a Population. Scand J Stat. 1984;11(4):265–70. http://www.jstor.org/stable/4615964

194. Zamora J, Verdú JR, Galante E. Species richness in Mediterranean agroecosystems: spatial and temporal analysis for biodiversity conservation. Biol Conserv. 2007;134(1):113–21.

195. Older CE, Diesel A, Patterson AP, Meason-Smith C, Johnson TJ, Mansell J, et al. The feline skin microbiota: The bacteria inhabiting the skin of healthy and allergic cats. PLoS One. 2017;12(6):e0178555.

196. Dinno A. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. Stata J. 2015;15(1):292–300.

197. McKight PE, Najab J. Kruskal-wallis test. corsini Encycl Psychol. 2010;1.

198. Hughes JB, Hellmann JJ. The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity. Methods Enzymol. 2005 Jan 1 [cited 2018 Jul 27];397:292–308. https://www.sciencedirect.com/science/article/pii/S0076687905970171

199. Bunge J, Willis A, Walsh F. Estimating the number of species in microbial diversity studies. Annu Rev Stat Its Appl. 2014;1:427–45.

200. Hughes JB, Hellmann JJ. The application of rarefaction techniques to molecular inventories of microbial diversity. Methods Enzymol. 2005;397:292–308.

201. Raup DM. Taxonomic diversity estimation using rarefaction. Paleobiology. 1975;1(4):333–42.

202. Lozupone CA, Knight R. Species divergence and the measurement of microbial diversity. FEMS Microbiol Rev. 2008 Jul 1;32(4):557–78. https://doi.org/10.1111/j.1574-6976.2008.00111.x

203. Tuomisto H. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography (Cop). 2010 Feb 1;33(1):2–22. https://doi.org/10.1111/j.1600-0587.2009.05880.x

204. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr. 1957;27(4):325–49.

205. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2011 Feb;5(2):169–72.

206. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol. 2007;73(5):1576–85.

207. Dray S, Legendre P, Peres-Neto PR. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecol Modell. 2006;196(3–4):483–93.

208. Taguchi Y-H, Oono Y. Relational patterns of gene expression via non-metric multidimensional scaling analysis. Bioinformatics. 2004;21(6):730–40.

209. Thompson B. Canonical correlation analysis: Uses and interpretation. Sage; 1984.

210. Xia Y, Sun J. Hypothesis testing and statistical analysis of microbiome. Genes Dis. 2017 Sep 1 [cited 2018 Jul 27];4(3):138–48. https://www.sciencedirect.com/science/article/pii/S2352304217300351

211. Anderson MJ. Permutational Multivariate Analysis of Variance (PERMANOVA). Wiley StatsRef: Statistics Reference Online. 2017. (Major Reference Works). https://doi.org/10.1002/9781118445112.stat07841

212. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLOS Comput Biol. 2016 Jul 11;12(7):e1004977. https://doi.org/10.1371/journal.pcbi.1004977

213. Saulnier DM, Riehle K, Mistretta T, Diaz M, Mandal D, Raza S, et al. Gastrointestinal Microbiome Signatures of Pediatric Patients With Irritable Bowel Syndrome. Gastroenterology. 2011;141(5):1782–91. http://www.sciencedirect.com/science/article/pii/S001650851100922X

214. Mason MR, Nagaraja HN, Camerlengo T, Joshi V, Kumar PS. Deep Sequencing Identifies Ethnicity-Specific Bacterial Signatures in the Oral Microbiome. PLoS One. 2013 Oct 23;8(10):e77287. https://doi.org/10.1371/journal.pone.0077287

215. Yazdani M, Taylor BC, Debelius JW, Li W, Knight R, Smarr L. Using machine learning to identify major shifts in human gut microbiome protein family abundance in disease. In: 2016 IEEE International Conference on Big Data (Big Data). 2016. p. 1272–80.

216. Michie D, Spiegelhalter DJ, Taylor CC, Campbell J. 1995. Machine learning, neural and statistical classification, Ellis Horwood, Upper Saddle River, NJ, USA.

217. Breiman L. Random Forests. Mach Learn. 2001 Oct;45(1):5–32. https://doi.org/10.1023/A:1010933404324

218. Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on

Document Analysis and Recognition. 1995. p. 278–82 vol.1.

219. Calle ML, Urrea V, Boulesteix A-L, Malats N. AUC-RF: a new strategy for genomic profiling with random forest. Hum Hered. 2011;72(2):121–32.

Chapter 3

Associations between microbial communities, environmental factors and microbiological quality of surface waters from the northeast U.S.

3.1 Abstract

Recent foodborne outbreaks demonstrated the importance of surface irrigation water as a source of produce contamination with foodborne pathogens. Several studies suggested that indicator microorganism that are currently used for evaluation of microbiological quality and safety of irrigation waters do not correlate well with occurrence of foodborne pathogens under all relevant environmental conditions. Hence, better understanding of microbial communities co-occurring with foodborne pathogens is needed to facilitate identification of alternative or supplementary microbial indicators. We characterized the composition of the bacterial and fungal communities in 68 water samples collected between May and August 2017 from six streams located in the upstate New York. According to the principal coordinate analysis and permutational multivariate analysis of variance (PERMANOVA), microbial communities differed significantly ($p<0.01$) between suspended sediment and water fractions obtained after sample centrifugation. The pairwise PERMANOVA indicated that microbial communities of sediment fractions differed significantly among sampling sites ($p<0.01$). Furthermore, average flow rate, pH, turbidity, and conductivity of water were significantly associated with microbial community composition. However, random forest binomial classification failed to predict the presence of cultured *Salmonella* based on the composition of microbial communities, as indicated by area under the curve of 0.55. The

low accuracy of prediction may be due to the small sample size (N=68) or lack of biological associations.

3.2 Introduction

Over 126 billion gallons of water were used for agricultural purposes, including irrigation and aquaculture, in the United States in 2010 (1) and 54% of all irrigation withdrawals in 2015 were from surface water sources (2). Irrigation water is commonly applied on fresh produce, which is one of the leading foods categories associated with foodborne outbreaks and represents a growing food safety concern (3–7). Ensuring microbiological quality of surface waters used for irrigation of produce consumed raw is therefore of critical food safety and public health importance.

Recent high-profile multistate foodborne outbreaks associated with fresh produce have increased concerns regarding microbiological safety of surface water used for irrigation of fruits and vegetables (3,8,9). In 2018, romaine lettuce contaminated with *E. coli* O157:H7 led to a multi-state outbreak that involved 62 individuals, of which 25 were hospitalized (10). The outbreak investigation identified the outbreak strain in a sediment collected from an on-farm water reservoir that was identified as a likely was a source of lettuce contamination (10). This and other multi-state outbreaks associated with fresh produce contaminated by *E. coli* O157:H7, *Salmonella* spp., and *Listeria monocytogenes* have increased concerns related to the highly variable microbiological quality of agricultural waters (11,12).

The temporal and spatial variability in factors affecting microbiological quality of surface waters, as well as a low-level contamination that is challenging to detect necessitate

longitudinal monitoring of water quality using indicators of fecal contamination. The Food and Drug Administration (FDA) implemented minimum safety standards for agricultural waters in the Produce Safety Rule (PSR) of the Food Safety Modernization Act (FSMA) (13,14). FSMA proposed longitudinal microbiological water quality assessment on a rolling basis using a generic *E. coli* as an indicator of inadequate microbiological quality of irrigation water. The regulation currently allows for the presence of a low level of generic *E. coli* in agricultural waters used during growing activities, although stringency of these criteria may not reduce the risk of foodborne illness sufficiently (14,15).

Generic *E. coli* has traditionally been used as an indicator of microbiological quality and safety, however, its sensitivity and specificity for prediction of contamination with foodborne pathogens depend on the established quantity thresholds as well as a number of environmental and physicochemical factors that affect microbial ecology of surface waters (11,12,16,17). A study of Florida surface waters investigated microbiological detection of samples from 18 sites away from animal agricultural activities over a period of 12 months reported only weak linear relationship between MPN quantities of generic *E. coli* and *Salmonella* in samples collected at 16 out of 18 sampled sites (12). In another study of surface waters collected in Georgia, *E. coli* was detected in all 72 tested samples, but only 79% of these samples were also contaminated with *Salmonella* (18). Furthermore, *E. coli* as an indicator was not found to be predictive of *Salmonella* presence when binary logistic regression was applied to the collected data (18). In contrast to the Georgia study, a microbial survey of Pennsylvania surface waters found that none of the three *Salmonella*-positive samples out of 73 collected samples had detectable levels of *E. coli* (11). These studies suggest that *E. coli* may not be reliable predictor of microbiological food safety

hazards in agricultural waters under all relevant conditions. A similar conclusion was drawn by Pachepsky et al. (19) who found that a significant relationship between *E. coli* or coliforms and foodborne pathogens existed only in 35% of the studies they had reviewed, while the relationship was insignificant in 65% of the reviewed studies. Findings from these reports warrant further investigation of factors predictive of pathogen presence in agricultural waters.

Characterization of the microbial ecology of surface waters that are potential sources of water for irrigation of food crops represents one of the approaches to identification of novel indicator microorganisms that could complement and enhance the utility of generic *E. coli*. Vierheilig et al. indicated that microbial communities data could potentially be applied in water microbiological quality monitoring (20). Even though they did not compare the associations between microbiomes and pathogens detected using culture-dependent approaches, they provided an insight into variability in the composition microbial communities across different types of samples, including water, soil, sediment, and feces.

The goal of this study was to characterize bacterial and fungal microbial communities in surface waters collected from six streams in the Northeast US in summer 2017 and investigate potential associations between microbial community composition, environmental factors and presence of foodborne pathogens.

3.3 Materials and Methods

3.3.1 Sample collection and processing

Moore swabs (n=68) were collected between June and August 2017 from 6 streams (i.e., A, B, C, D, E, and F) in upstate New York. Each stream was sampled in 3 weeks for 3 - 5 consecutive days, resulting in a total of 11 – 13 collected samples per stream (see Supplemental Materials Table A.1). At the start of each week, a Moore swab was anchored in the waterway for 24 h. After 24 h the Moore swab was removed, and a second Moore swab was placed in the waterway. After the sample collection, each Moore swab was transferred to a separate, sterile Whirl-Pak bag and transported on ice to Cornell University, where they were stored at 4°C and processed < 24 h after sample collection.

Moore swabs were processed by first hand-massaging each swab for 1 min. After one minute, each Moore swab was pressed while 100 mL of the water that was contained in the swab was transferred to two 50 ml sterile conical tubes using a sterile serological pipette. Each 50 mL tube was then centrifuged at 4,000 g at 4 °C for 5 min to separate microorganisms bound to suspended sediment (further referred to as sediment fraction) from those that remained suspended in the water (further referred to as water fraction). The water fraction was then filtered through a 0.45 µm 250 ml analytical filter system (VWR). After filtration, each filter was transferred to the collection tube included in the DNeasy PowerWater DNA extraction kit (Qiagen). Sediment pellets and filters with biomass were stored at -80°C until DNA extraction.

3.3.2 Metadata acquisition, waterway enrollment, and foodborne pathogen data.

Remotely sensed data were obtained from publicly-available datasets to facilitate waterway enrollment (Weller et al., 2019; submitted for publication). Briefly, hydrological and land use data were obtained from government data portals (21–24). These data were used to identify watersheds that were >15 km^2 in area, and <400 m from a field where covered produce was grown in ≥4 of the last 8 years (2009-2016). Data on upstream land use was collected from the National Land Cover dataset for 2011 (https://www.mrlc.gov); data on the amount of impervious land upstream of each site was generated using the Stream Stat Portal (https://streamstats.usgs.gov/ss). Publicly-accessible sites along these watersheds (e.g., road intersections, public rights-of-way) were identified, and 6 sites were randomly selected for enrollment so that all enrolled sites were in non-overlapping watersheds (Weller et al., 2019; submitted for publication).

Physicochemical water quality and flow data were collected 6 times (~ 2 h apart) while the Moore swab was deployed in the waterway (see Supplemental Materials Table A.2). These 6 values were then averaged to obtain the values used in the analyses reported here. Data on physicochemical water quality (i.e., dissolved oxygen, conductivity, pH, and water temperature) and flow rate in the stream at the time of sample collection were measured in the field using a Hach HQ40d meter (Loveland, CO) and a flow meter (Global Water Instrumentation Inc, Cordova, CA), respectively. Turbidity was measured in the lab using a Hach turbidity meter (Loveland, CO).

Pathogen data used in the analyses were previously published and were generated as described in Weller et al. (Weller et al., 2019; submitted for publication).

3.3.3 DNA extraction

Extraction of DNA from biomass collected on filters and from sediment pellets was performed at The Pennsylvania State University using the DNeasy PowerWater DNA isolation kits or DNeasy PowerSoil DNA isolation kits (Qiagen), respectively. DNA was extracted following manufacturers' instructions. The extracted DNA samples were eluted with 100 µl 10 mM Tris elution buffer provided by the manufacturer. Concentration and purity of DNA from each sample were assessed using NanoDrop One UV-Vis spectrophotometer (Thermo Fisher Scientific). All DNA extracts were stored at -80°C until PCR amplification.

3.3.4 PCR amplification and Illumina sequencing of the 16S rDNA V4 region and ITS2 sequences

Extracted DNA samples were adjusted to 10 µg/ml double-stranded DNA using the Qubit dsDNA High Sensitivity Assay Kit and Qubit 3. Samples with standardized concentration were used as templates in PCR amplification of bacterial 16S rDNA V4 and fungal Internal Transcribed Spacer 2 (ITS2) sequences. Primers 515FB (GTGYCAGCMGCCGCGGTAA) and 806RB (GGACTACNVGGGTWTCTAAT) were used for amplification of V4 sequence of the 16S rRNA gene (25). Primers ITS9F (GAACGCAGCRAAIIGYGARV) and ITS4R (TCCTCCGCTTATTGATATGC) were used for amplification of the ITS2 region (26). PCR was performed in 20 µl reactions containing 10 µl of KAPA HiFi HotStart ReadyMix PCR Kit (Kapa Biosystems), 1 µl of each, the forward and reverse primers in concentration of 10 µM, 1 µl of the template DNA in concentration of 10 µg/µl and 7 µl of PCR-grade ultrapure water. PCR thermal cycling

conditions for 16S rRNA V4 and ITS2 amplification were optimized based on the Illumina 16S rRNA gene metagenomic sequencing library preparation protocol (27), and Joint Genome Institute iTag Sample Amplification QC SOP (28). For the 16S rRNA V4 sequence, an initial denaturation of 2 min at 98°C, followed by 25 cycles of 20 s at 98°C, 20 s at 56.5°C, and 25 s at 72°C, and a final step at 72°C for 5 min was used. For ITS2 amplification, an initial denaturation of 5 min at 98°C, followed by 30 cycles of 45 s at 95°C, 60 s at 50°C, and 60 s at 72°C, and a final step at 72°C for 5 min was performed. PCR amplicons were visualized and confirmed by running a 1% agarose gel electrophoresis. Successful amplification of target sequences was confirmed using 2100 Bioanalyzer (Agilent) on sets of 12 pooled samples to confirm the expected fragment lengths. After this quality control step carried out by the Penn State Huck Institutes of Life Sciences' Genomics Core Facility, PCR amplicons were submitted for library preparation and Illumina sequencing to the core facility. Amplicons were barcoded in a second PCR and cleaned-up following the Illumina 16S rRNA amplicon library preparation guide (27). Libraries were quantified using quantitative PCR and normalized. Normalized libraries were pooled, denatured, spiked with PhiX and sequenced on an Illumina MiSeq platform at the Penn State Huck Institutes of the Life Sciences Genomics Core Facility using 500 cycle V2 kit (Illumina) to obtain 250 bp paired-end reads. The targeted number of reads per sample was 40,000 bp per library.

3.3.5 Sequence read quality control, assembly and taxonomic classification

Raw sequence reads were demultiplexed and further processed using Mothur v.1.40 with default settings and parameters provided by Mothur miseq SOP unless noted otherwise

(29,30). The sequence analyses workflow is described in the Supplementary Materials. Low-quality bases were removed, and sequence read pairs assembled in contigs using make.contigs. Contigs that were longer or shorter than the expected length of 282 bp for 16S rRNA V4, and were out of the 265 to 420 bp range for ITS2 were removed with screen.seqs. Contigs were aligned to the SILVA 16S rRNA (SILVA 132 release, July 2017) reference database for bacteria (31) and ITS UNITE (UNITE mothur release, v 7.2, December 2017) database for fungi (32) using align.seqs. Contigs that remained unmapped, as well as chimeric contigs, were discarded using filter.seqs and chimera.vsearch commands, respectively, to minimize the effects of sequencing errors on the results of data analyses (33). Subsequently, operational taxonomic units (OTUs) were generated using classify.seqs and cluster.split with a 97% sequence identity threshold corresponding to a taxonomic species (34).

3.3.6 Prediction of sample richness

To visualize the relationship between the sequencing depth and discovered species richness (i.e., the cumulative number of unique OTUs per every additional new OTU), rarefaction curves were constructed by random sub-sampling of reads from each library using ggplot2 packages (v 3.1.0) in R (v 3.5.2) (35). The maximum richness in each sample was estimated using Chao1 richness estimator with spadeR package (v0.1.1) (36). Sample F25 had 474 reads that resulted in estimated 47.69% discovered richness and was therefore excluded from the analyses. Remaining samples all had over 60% discovered richness.

3.3.7 Data Normalization

To test the effect of different normalization methods, three normalization methods were used and compared. These included rarefying, proportional transformation, and relative log expression normalization (RLE). Biodiversity analyses were conducted on datasets normalized using each of the three normalization methods to assess and compare the effect of normalization method on determined biodiversity. By rarefying, OTUs from each library were randomly subsampled to minimize the sequencing depth bias and obtain even number of OTUs across all analyzed samples. Libraries with less than 9,000 reads were excluded (n = 4). The rest of the samples were rarefied to the number of OTUs present in the smallest library, using the vegan package (v 2.5-3) in R (v 3.5.2) (37,38). In proportional transformation, the counts of individual OTUs were divided by the sum of all counts, without any additional normalization (37). Lastly, the RLE normalization, a scaling method based on geometric means of counts among samples was used (39). After reviewing the recommendations from published literature (40), and comparing the effects of different normalization methods, the RLE-normalized data were used in all further analyses.

3.3.8 Alpha diversity analyses

Within sample diversity (i.e., alpha diversity) was measured by Chao1(41), Shannon-Wiener (i.e., Shannon index) (42), and inverse Simpson indices (43). The Chao1 index describes species richness and is highly influenced by rare OTUs that can falsely inflate the diversity index (162). Shannon and inverse Simpson indices describe both species richness and evenness and result in a more realistic estimation of microbial diversity (42, 43). Hence, Shannon and inverse Simpson indices were calculated in addition to Chao 1. Inverse

Simpson index gives higher weight to species with a higher relative abundance and was therefore used as a primary index for comparative analyses of alpha diversity. The distribution of the inverse Simpson indices describing individual samples' bacterial and fungal diversities was plotted by each sampling stream, using R package phyloseq (v 1.16.1) ('plot_richness') and ggplot2 (v 3.1.0) (211,215).

3.3.9 Beta-diversity analyses

Beta-diversity analyses were carried out to investigate the among-sample diversity of microbiomes and mycobiomes in suspended sediment and water fractions based on 16S rRNA V4 and ITS2 sequences, respectively (46). OTU counts were binned into taxonomic families according to the taxonomic profiles in SILVA and UNITE databases (32,47). Phylogenetic distances were calculated using weighted UniFrac distance that accounts for the distance between each pair of samples with respect to the relative abundance of each taxonomic family and their phylogenetic relationship (48). Principle Coordinate Analysis (PCoA) was performed using weighted UniFrac distances to visualize the ordination and clustering of samples. Two principal coordinates were plotted to visualize the similarity among microbiomes and mycobiomes of analyzed samples using R packages phyloseq and ggplot2 (35,45). Samples were color-coded to visually assess putative associations between the i) sample type (i.e., suspended sediment or water fraction), ii) sampling site (A through F), and iii) presence of pathogens (i.e. *Salmonella* spp., STEC, *L. monocytogenes*) and the microbial community composition in PCoA plots. Sample clustering in PCoA plots also guided the selection of potential explanatory variables that were further tested in statistical

analyses using PERMANOVA to assess their association with the microbial and fungal communities' composition of analyzed samples (49).

3.3.10 Identification of factors associated with the microbiome and mycobiome composition

Permutational Multivariate Analysis of Variance (PERMANOVA) was used to test the significance of the associations between the microbiome distance measure and the factors of interest identified through PCoA analyses (181). PERMANOVA tests were performed using 'adonis' function in R package 'vegan' and pairwise PERMANOVA tests using 'pairwise.adonis' function derived from 'vegan' package and implemented in R package 'pairwiseAdonis' (v 0.0.1) (38,50). Pairwise PERMANOVA was used for multifactorial analyses, using 999 permutations. For pairwise PERMANOVA, Bonferroni correction for multiple comparisons was applied to adjust the p-value and correct for the type 1 error (51). Binomial differential abundance test implemented in the R package 'DESeq2' (v 1.22.2) was applied to identify phyla that are differentially abundant among sample types (i.e., suspended sediment, water fraction) based on Wald test (52). Specifically, the differential abundance test was performed to identify phyla that are significantly ($P < 0.5$) differentially abundant among suspended sediment samples compared to water fractions. P values were corrected using Bonferroni correction (51).

Canonical correspondence analysis (CCA) was used to assess the relationship between the composition of microbial communities, and environmental factors. CCA was performed using 'cca' function in R package 'vegan' (38). Six physicochemical measurements were used for the analysis (i.e., pH, dissolved oxygen, flow rate, air temperature, water temperature, conductivity, and turbidity). Permutation analysis was used to identify the

environmental factors, including physicochemical properties that were significantly associated with the composition of microbial communities. A matrix describing the significance of correlations between the abundance of each phylum and the environmental factors was calculated using 'taxa.env.correlation' function in R package 'microbiomeSeq' (v 0.1) (53).

3.3.11 Prediction of foodborne pathogen presence based on the microbiome composition

Random forest classification analysis was performed to identify bacterial and fungal families predictive of cultured foodborne pathogen presence (54–56). The data describing culture-based determined presence of *Salmonella*, *Listeria* spp. and *L. monocytogenes* have previously been published and were used in this study (Weller et al., 2019; submitted for publication). The presence of Shiga toxin producing *E. coli* (STEC) determined by a real-time PCR method was also predicted (Weller et al., 2019; submitted for publication). The OTU tables with 1111 unique bacterial and 603 unique fungal families were included in two separate random forest analyses. For each classification based on these OTU tables, two major parameters were tuned to improve the prediction accuracy. These parameters included the number of trees (ntree) and the number of variables (i.e., the number families used for each split) which were initially selected for each node split (mtry) using caret package in R (57). The tuning process was based on empirical trials carried out with the purpose of identifying the optimal mtry and ntree parameters that resulted in most accurate classification, as judged by the accuracy obtained after 10-fold cross-validation and kappa statistics (58,59). Conditional variable importance measures were used to calculate the variable importance of each classifier (60). Variable importance was measured by an area

under the curve (AUC) resulting from the magnitude of the variability in individual binary outcomes (i.e., positive or negative pathogen detection results). AUC-based variable importance measure was computed based on the AUC determined before and after using a variable for classification (61). Thus, it is possible to identify variables that have the greatest impact on the classification (i.e., those with a larger increase in AUC) by comparing the change in the AUC value for each variable (61). Variable importance measures were normalized to facilitate interpretation and visualization of the results. Normalized variable importance was calculated by min-max normalization for linear transformation of the original data (62).

3.4 Results

We hypothesized that (i) a choice of operational taxonomic unit (OTU) normalization method will have an effect on the alpha diversity; (ii) microbial communities associated with sedimented particles extracted from Moore swabs will be significantly different compared to communities remaining in a water fraction; (iii) microbial communities' composition will differ among streams; (iv) certain environmental factors will be associated with microbial composition; and that (v) the composition of microbial communities will be predictive of foodborne pathogen presence.

3.4.1 The sequencing effort captured a median of 74.6 percent estimated species richness

A total of 114 samples, including 68 samples of suspended sediments and 46 samples of water fractions were sequenced. A median of 30,844 reads per sample were obtained for 16S rRNA V4 PCR amplicons (min = 5,272, max = 47,078, standard deviation [SD] =

8,371.36) and a median of 33,297 reads for ITS2 PCR amplicons (min = 5,700, max = 79,301, SD = 13,814). Reads that passed the quality control were assigned operational taxonomic units (OTUs). A median of 19,437 OTUs per sample was obtained for 16S rRNA V4 PCR amplicons (min = 2,375, max = 21,889, SD = 5,418) and a median of 18,737 for ITS2 PCR amplicons (min = 4,207, max = 57,051, SD = 10,643). Rarefaction curves for individual samples were generated by random resampling of the subsets of OTUs and plotting of the number of OTUs against the number of unique OTUs (Figure 3.1).

To estimate the species richness in each of the sequenced samples, Chao1 index was used to estimate the total richness of each sample and compare it with an observed number of unique OTUs from each sample to calculate the percent discovered richness (161). A median of 72.2 % (min = 47.69%, max = 77.9%) and 77.9 % (min = 68.62%, max = 86.27%) bacterial and fungal species were estimated to be discovered at a sequencing depth used in the present study, respectively (36). We considered this sufficient for identifying highly abundant microbial taxa that were hypothesized to be most informative for potential new foodborne pathogen indicator discovery.

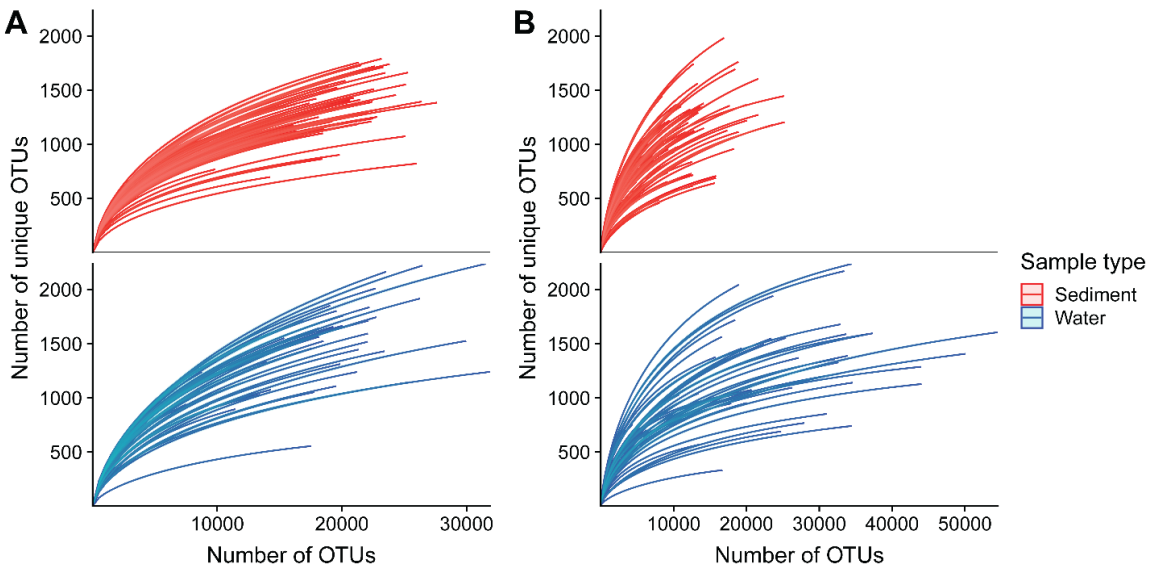

Figure 3.1 Rarefaction curves and associated standard errors (indicated in lighter color around each line) for (A) bacterial and (B) fungal species richness observed in suspended sediments (red lines) and water fractions (blue lines) of surface water samples. Each line indicates an individual sample.

3.4.2 Different normalization methods had an effect on alpha diversity, but not on the beta diversity

Observational and statistical analyses of microbial diversity were used to assess and compare the effect of different normalization methods on calculated alpha and beta diversities. Based on Kruskal-Wallis test using three different alpha diversity indices (i.e., Chao1, Shannon, inverse Simpson), significant differences in calculated alpha diversities using all three indices were observed when different normalization methods were applied ($p < 0.001$). Overall, alpha diversity measurement with relative log expression normalization (RLE) showed higher values and larger variance than other two methods for all three indices. Nevertheless, the results of PERMANOVA analyses of the effect of the sample type (i.e., sediment, water) on the microbial community composition did not show significant

differences when data were normalized using the three different methods, suggesting that beta diversity was not significantly affected by the normalization method (Supplemental Material Table A.1). This finding was confirmed also by similar clustering of sediment and water samples from different stream in PCoA plots (Supplemental Material Figure A.1). RLE with inverse Simpson for alpha diversity and weighted UniFrac for beta diversity was used in all further analyses, as it was deemed the most appropriate based on the previous studies that had compared different normalization methods (37,40).

3.4.3 Microbial communities in suspended sediments are significantly different compared to those in water fractions

Increased level of suspended sediment in water streams frequently occurs after the rainfall as a result of (i) surface water run-off that introduces soil from the adjacent watershed and (ii) suspension of bedded sediment due to increased stream flow rate. The Moore swabs collected in our study contained large amounts of sediment particles which necessitated centrifugation of samples. Pelleted sediment was therefore investigated separately from the water fraction that was filtered to retain microorganisms persisting in the supernatant. We assumed that the persisting microorganisms would also persist in a water column for a longer time after rain events that suspend particles from the stream sediment. No phyla were unique to either sediments or water fractions separated *in silico*, however, we identified 33 bacterial phyla with significantly different ($p<0.05$) relative abundance in sediments compared to water fractions using a binomial test for differential abundance (Figure 3.2A). The test identified 14 bacterial phyla with significantly higher relative abundance in sediment fractions and 16 bacterial phyla with higher relative abundance in water fractions, based on Wald test. Latescibacteria, Entotheonellaeota, Nitrospinae,

Thermosulfidibacteraeota, and Acidobateria were more abundant in sediment compared to water fractions. Conversely, Epsilonbacteraeota, Omnitrophicaeota, Lentisphaerae, Bacteroidetes, and Thermotogae were significantly more abundant in water fractions compared to sediments. In terms of fungal communities, five phyla were differentially abundant between sediment and water fractions (Figure 3.2B). One phylum, Mortierllomycota, was present in higher abundance in sediments compared to water fractions, and four phyla (i.e., Aphelidiomycota, Rozellomycota) were more highly abundant in water fractions compared to sediments.

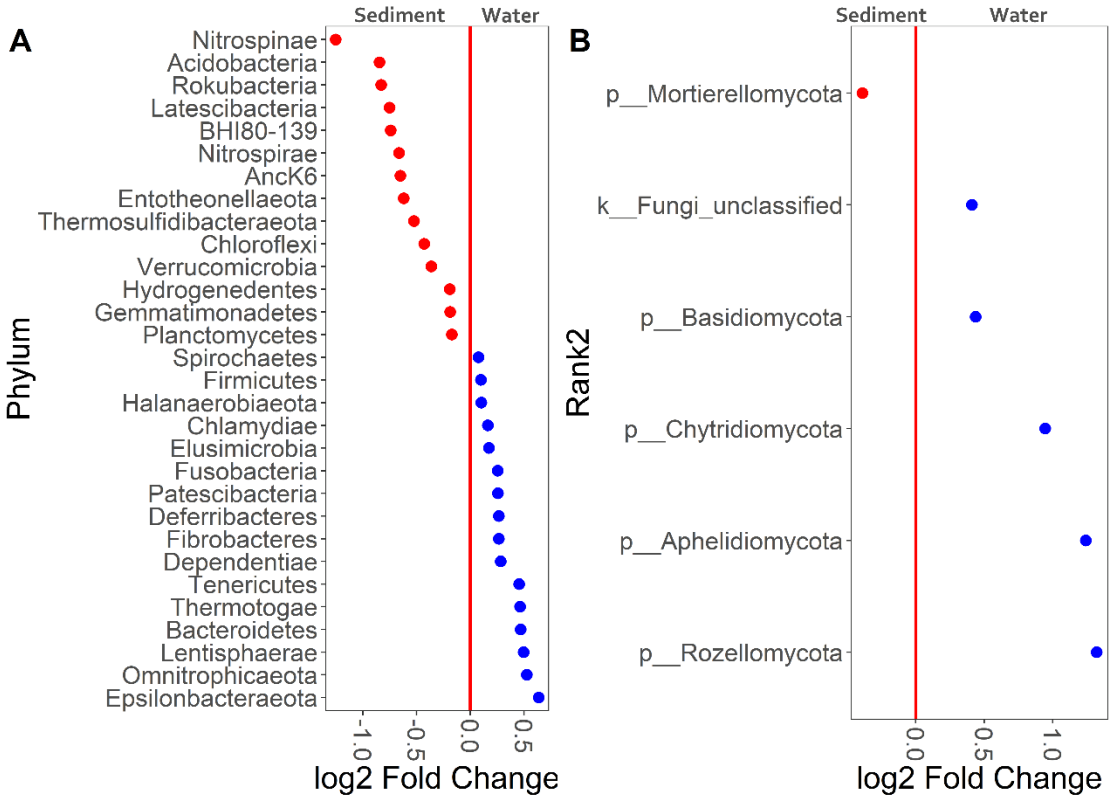

Figure 3.2 Binomial differential abundance of microbial phyla between (A) bacterial and (B) fungal communities of surface water samples collected in upstate NY from May to August 2017. Differential abundance was based on Wald test.

Differences in overall microbial communities' composition of sediment and water fractions were further investigated through beta diversity analysis using weighted UniFrac metric that measures dissimilarity between samples based on phylogenetic relatedness and relative abundance of microbial taxa identified in samples. The UniFrac distance matrix was used in the principal coordinate analysis (PCoA) to visualize similarities among sample microbial communities in a low-dimensional space. Separate PCoA plots generated for bacterial (Figure 3.3A) and fungal (Figure 3.3B) communities indicate that the composition of both bacterial and fungal communities was different in sediment fractions compared to water fractions, albeit the differences in bacterial communities were greater compared to fungal communities.

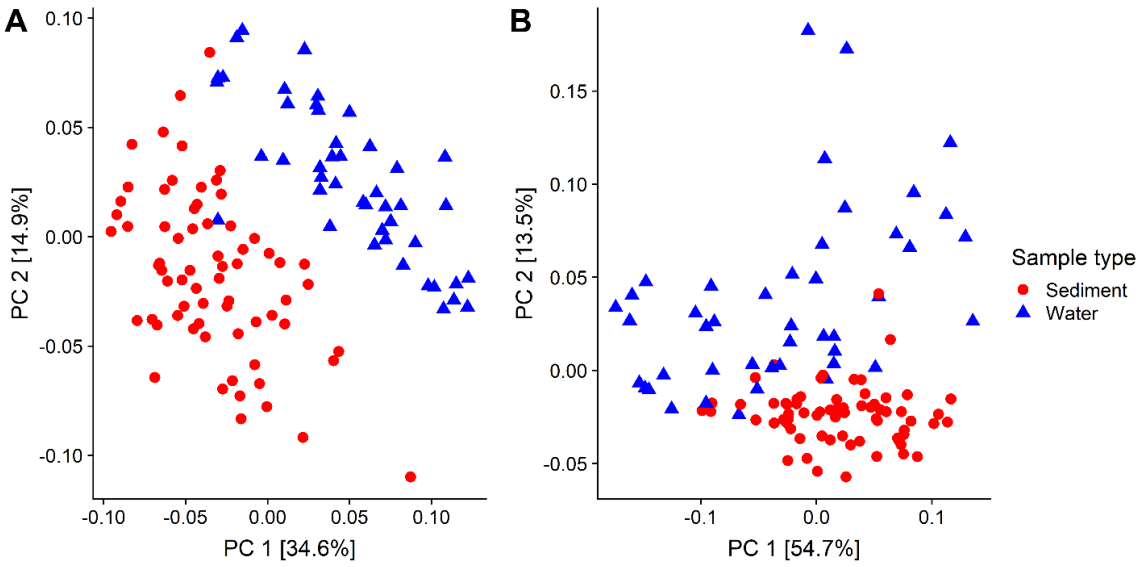

Figure 3.3 Principal Coordinate Analysis (PCoA) based on the UniFrac distances for (A) bacterial and (B) fungal families found in surface water samples. The analysis included sediment (n=68, red dots) and water (n=46, blue dots) fractions of collected water samples.

The differences between microbial communities of sediment and water fractions were further compared on a family level using PERMANOVA. The results of PERMANOVA analyses using samples collected across all 6 streams and all three months confirmed statistically significant differences in the composition of both bacterial ($p<0.001$) and fungal communities ($p<0.001$) between suspended sediment and water fractions (Supplemental Table A.2).

3.4.4 Microbial communities differed among samples collected from different streams

After confirming the distinct composition of sediment and water fraction microbiomes, we used statistical analyses to investigate whether samples collected from different streams have a distinct composition of microbial communities.

Microbial compositions of sediment and water samples were tested separately using pairwise PERMANOVA based on weighted UniFrac indices to test whether any of the 15 pairs of streams had a significantly different composition of bacterial or fungal OTUs at sampled locations. All 15 tested pairs of streams ($p < 0.05$), and 13 out of 15 tested pairs of streams ($p <0.05$) had a significantly different composition of bacterial and fungal communities in sediment fractions, respectively (Table 3.1). Furthermore, 10 out of 15 pairs of sites had a significantly different composition of bacterial communities in water fractions ($p < 0.05$), and 11 out of 15 pairs of sites for fungal communities in water fractions ($p<0.05$) (Table 3.1).

We further used descriptive analyses to investigate whether streams with a similar composition of microbial communities also share similar characteristics in terms of land use upstream of the sampling sites (Supplemental Materials Table S3). The predominant land

cover in all six watersheds was a natural cover, including forests, wetlands, shrubland or grassland. However, immediately upstream (i.e., 0-250 m from the sampling site) of streams A and D, the predominant land cover were developed areas, such as buildings, roads, and lawns. The predominant land cover upstream of the stream B was a natural cover, and upstream of streams C, E, and F pasture or hay. Furthermore, based on the observational data recorded during sampling, dairy and produce farms were visible from the sampling sites on streams C and F, and cattle herds had access to areas immediately upstream of the sampling location on the stream F. The sampling site on the stream E was adjacent to a small industrial complex, and dogs were observed roaming freely near the sampling site E. Streams B and D were near urban areas and stream D was located adjacent to a quarry. Noteworthy, the predominant bottom substrate in streams A, B, C, E, and F was either sand, gravel or cobble, while the bed of the stream D was predominantly organic matter, which was also found overlying the gravel bed in the stream C. These formal quantitative and informal descriptive observations were taken in consideration in the interpretation of the compositional differences in microbial communities found among different streams.

Table 3.1 Differences in bacterial and fungal communities' composition among different streams

Sampling site	Bacterial communities[a]		Fungal communities[a]	
	Soil	Water	Soil	Water
A vs B	**<0.015**	**<0.015**	**<0.015**	**<0.015**
A vs C	**<0.015**	**0.030**	**<0.015**	**0.030**
A vs E	**<0.015**	**<0.015**	0.060	**0.030**
A vs F	**<0.015**	0.060	**<0.015**	**0.045**
A vs D	**<0.015**	**0.045**	**<0.015**	**<0.015**
B vs C	**<0.015**	**0.045**	0.0750	**<0.015**
B vs E	**<0.015**	**<0.015**	**<0.015**	**0.030**
B vs F	**<0.015**	0.105	**<0.015**	0.075
B vs D	**<0.015**	**0.030**	**<0.015**	**<0.015**
C vs E	**<0.015**	0.885	**<0.015**	0.285
C vs F	**<0.015**	0.465	**<0.015**	1.000
C vs D	**<0.015**	**<0.015**	**<0.015**	**<0.015**
E vs F	**0.030**	0.660	**0.030**	1.000
E vs D	**<0.015**	**<0.015**	**<0.015**	**<0.015**
F vs D	**<0.015**	**<0.015**	**<0.015**	**<0.015**

[a]Bonferroni-corrected p-values obtained by pairwise PERMANOVA analyses of weighted UniFrac distances between pairs of sampling sites. Pairs of sites with significantly different microbial communities ($p < 0.05$) are indicated in bold type.

Kruskal-Wallis test confirmed that overall, bacterial communities in the sediment and water fractions differ by sampling site/stream consistently, regardless of which of the three alpha diversity indices were used in the comparative analyses ($p<0.05$) (Supplemental Materials Table A.4). Dunn's multiple comparison test that was conducted as a post-hoc test after the Kruskal-Wallis test indicated significant differences in alpha diversity between specific pairs of sampling streams and different sample types (Supplemental Materials Table A.4). The distribution of alpha diversity indices for bacterial and fungal communities of sediment and water fractions were plotted by sampling stream (Supplemental Materials Figure A.2). According to Dunn's multiple comparison test, alpha diversity of bacterial

communities found in sediment fractions from sampling sites A and D, B and D, C and E, D and F were significantly different ($p<0.05$). Furthermore, streams B and D had significantly different diversity of bacterial communities in water fractions. There were no other significant differences between the communities of fungal sediment communities, whereas, sampling site A and B, A and E had significantly different alpha diversity of fungal communities in water fractions.

3.4.5 A number of physicochemical factors were associated with composition of microbial communities

CCA results showed that different physicochemical factors had an effect on the composition of microbial communities (Figure 3.4). Conductivity ($p = 0.001$), pH ($p = 0.001$), and flow rate ($p=0.001$) were associated with the composition of bacterial communities in sediment samples (Figure 3.4A) but did not affect fungal communities nor bacterial communities in water fractions. Turbidity ($p = 0.023$) and conductivity ($p = 0.004$) were associated with fungal communities in water fractions (Figure 3.4B). Relationship between physicochemical properties and the abundance of the different phyla (Figure 3.5) identified with linear correlation. Dissolved and pH were mostly negatively associated with bacterial communities based on the Pearson's correlation coefficient ($p < 0.05$). Seven fungal phyla were significantly either positively or negatively correlated with dissolved oxygen in sediment samples, and four phyla were significantly positively correlated with the turbidity of the sediment samples.

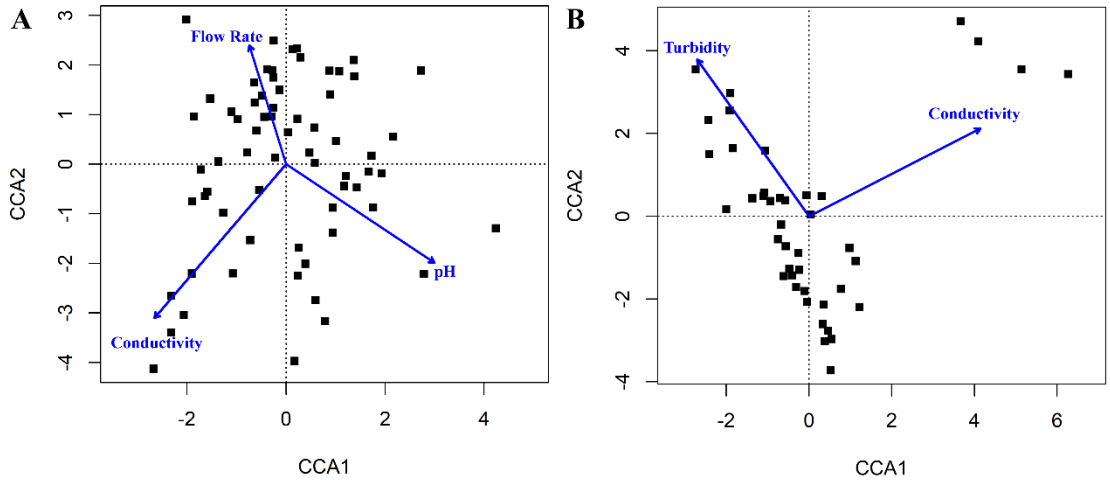

Figure 3.4 Canonical correspondence analysis (CCA) plot indicating physicochemical properties that had an effect on the composition of (A) bacterial communities found in sediments and (B) fungal communities found in water. The blue arrows and labels correspond to the physicochemical factors that were significantly associated with the composition of microbial communities in samples shown as black dots.

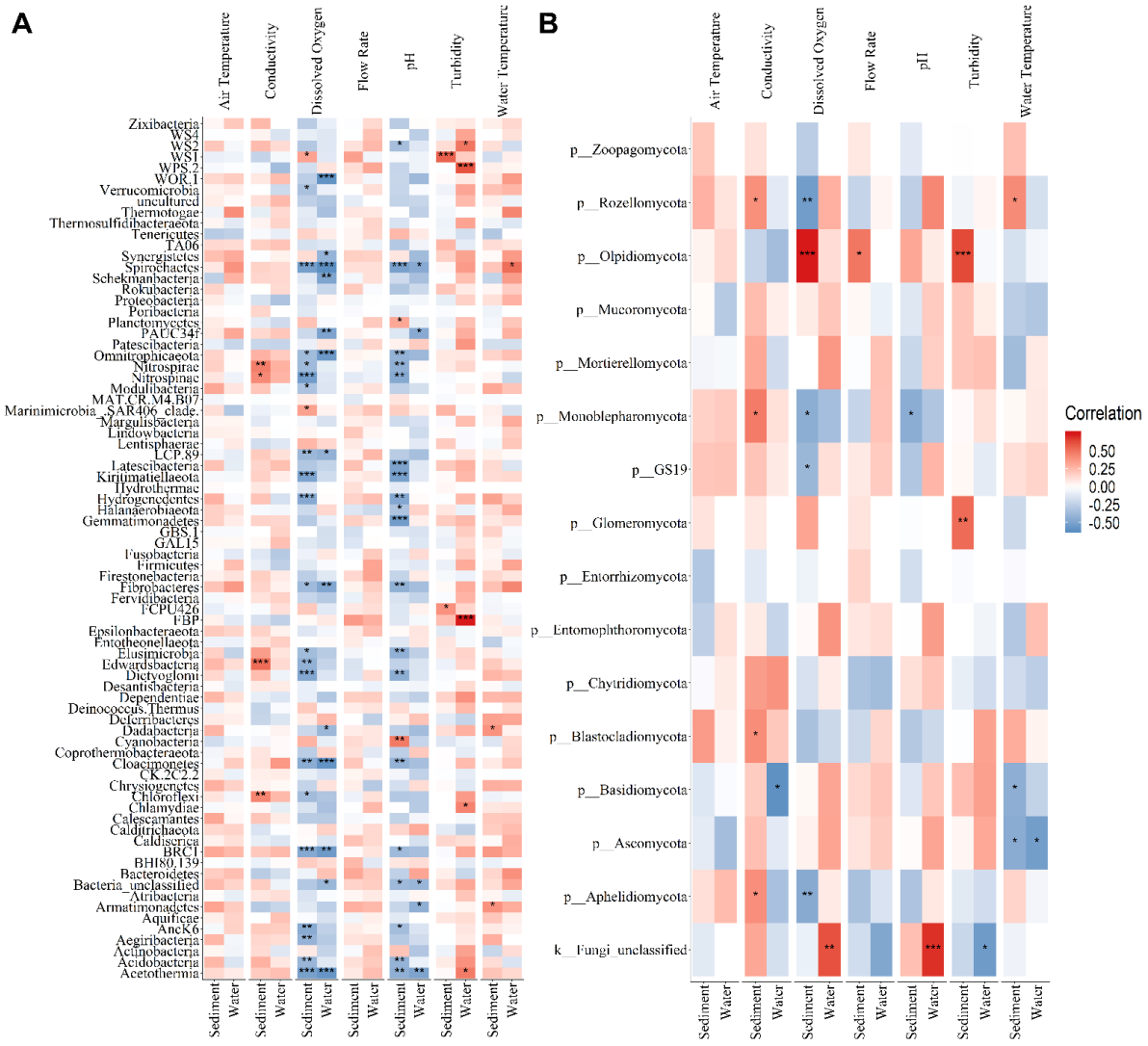

Figure 3.5 Correlation matrix between different environmental measurements and phylum level abundance of the (A) bacterial and (B) fungal communities. Samples are divided into two groups based on the samples type (i.e., sediment, water). Asterisks indicates p-value less than 0.05 (*), 0.01 (**), 0.001 (***) respectively.

3.4.6. The accuracy of predicting pathogen presence based on the microbiome composition is very low

The UniFrac-based PCoA clustering (Supplemental Materials Figure A.3) and PERMANOVA analyses (Supplemental Materials Table A.6) indicated a lack of significant association between microbial community composition and presence of cultured *Salmonella* and *L. monocytogenes* ($p > 0.05$). STEC detected by quantitative PCR showed a significant association, which might be due to the higher prevalence of presumptive positives. Hence, random forest (RF) classification was used to assess whether a single or a combination of microbial community classifiers (i.e., microbial families, also referred to as variables) can be used to predict the presence of foodborne pathogens. Classification variables were selected using the area under the curve (AUC) variable importance measure. Prediction of *Salmonella, L. monocytogenes,* and STEC presence was based on the random forest models that used 33 families and 10001 trees based on the parameter tuning process. Model accuracy and kappa statistic indicated that the overall prediction has low accuracy. Based on the model accuracy obtained after 10-fold cross-validation, all of the models predicting the presence of *Salmonella* and STEC in both bacterial and fungal communities of sediment and water fractions had 45 – 55 % accuracy with low kappa statistics close to 0, which indicates that classifications are not accurate. Moreover, even though the accuracy of *L. monocytogenes* predictions were 88 – 98 %, kappa statistics were close to 0, which also indicates poor accuracy of classification. The families identified as predictive of individual pathogens are shown in Figure 3.6**.**

RF was used to calculate the variable importance (VI) that was used for the prediction of pathogen presence. Measurement of variable importance was reported as normalized variable importance and normalized VI with a value higher than 0 was considered as informative. Seventeen bacterial families including Chlamydiaceae, Sericytochromatia, Chitinivibrionaceae, and WPS-2 (candidate taxa) were identified as the families with the highest ability to predict the presence of *Salmonella* in sediment fractions (Figure 3.6A). Peptococcaceae, VadinHA49, Clostridiales_vadinBB60_group, and HglApr721 (a candidate under Gammaproteobacteria) were identified as most predictive of *Salmonella* presence in water (Figure 3.6B). The RF for fungal communities identified fungal families Exobasidiaceae, Valsariaceae, and unidentified family under Helotiales (Helotiales_fam_Incertae_sedis) as predictive of *Salmonella* presence in sediment fractions (Figure 3.6C), and Halomycetaceae, Leotiaceae, and unidentified family under Rhizophydiales (Rhizophydiales_fam_Incertae_sedis) as predictive of *Salmonella* presence in water fractions (Figure 3.6D). There were no informative VI generated from RF-based prediction of STEC and *L. monocytogenes*, with exception of fungal communities in sediment samples, where Platygloeaceae and Dipodascaceae were identified as predictive of the presence of *L. monocytogenes* (Figure 3.6E).

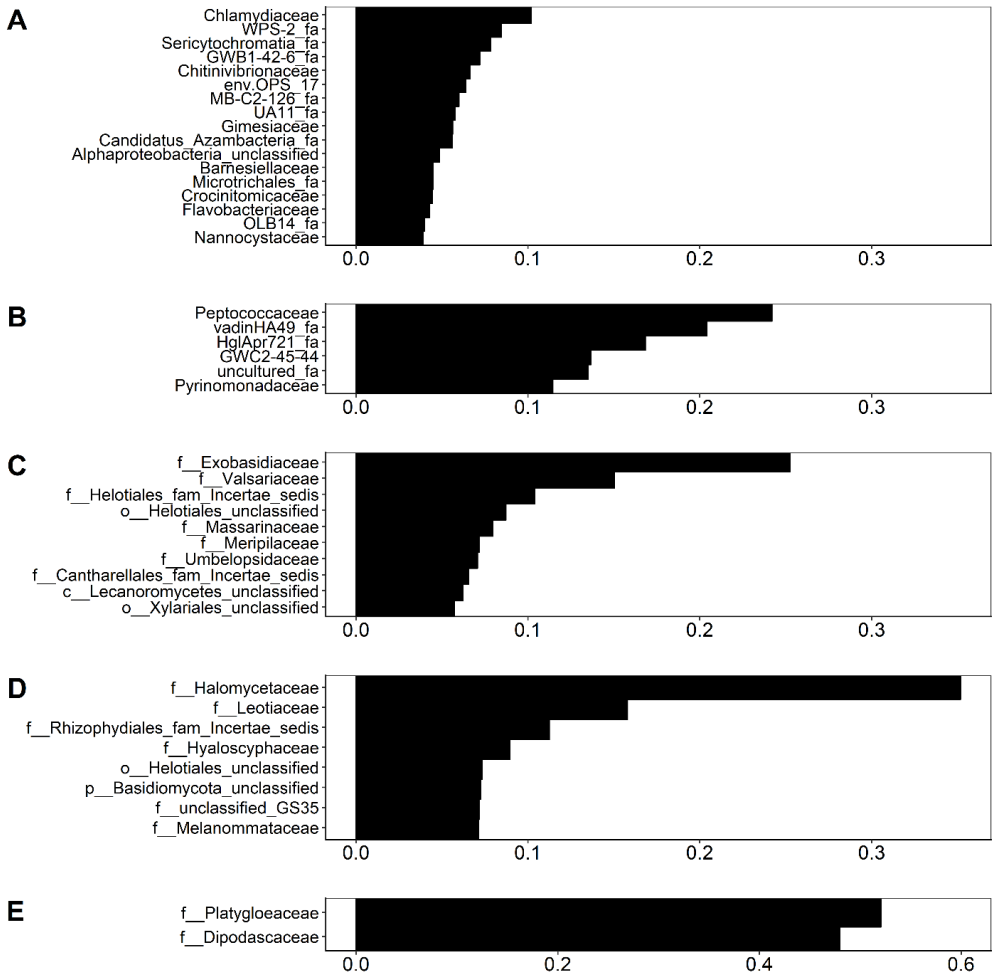

Figure 3.6 Conditional variable importance for the prediction of *Salmonella* and *Listeria monocytogenes* presence in water samples. Conditional variable importance was quantified based on the area under the curve (AUC) using random forest classification. (A) bacterial families predictive of *Salmonella* in sediment fractions (accuracy = 48%), (B) bacterial families predictive of *Salmonella* in water fractions (accuracy = 42%), (C) fungal families predictive of *Salmonella* in sediment fractions (accuracy = 42%), (D) fungal families predictive of *Salmonella* in water fractions (accuracy = 52%), (E) fungal families predictive of *Listeria monocytogenes* in sediment fractions (accuracy = 89%).

3.5 Discussion

Recent high-profile foodborne outbreaks linked to contaminated water used for irrigation of produce demonstrated the need for improved monitoring of microbiological hazards in agricultural waters. Current indirect methods used for the microbiological safety assessment of water rely on generic *E. coli* or other indicators that do not correlate well with the occurrence of foodborne pathogens under all relevant conditions (11,18,63). Hence, there is a need for better understanding of microbial ecology of surface waters that may help in identification of novel, complementary indicators that could enhance the utility of currently used indicators. In this study, we characterized microbial communities in water samples collected from six streams in the upstate New York to improve general understanding of the of the microbial diversity of surface water and to identify associated factors that may facilitate identification of novel indicators of microbiological food safety hazards. microbial diversity and associated factors that may facilitate identification of novel indicators of microbiological food safety hazards. The data presented here provide new insights into the temporal and spatial variability in microbial communities' composition across the streams, albeit from a single geographical region in the rural northeast U.S.

3.5.1 Asymptotic estimations of sample richness indicate sufficient sequencing depth

Samples included in our study were not sequenced at an equal sequencing depth. Equal sequencing depth across all samples, however, is not necessarily needed if samples with different microbial diversity are sequenced. For instance, a sample with low microbial diversity will not require the same sequencing depth as a sample with high microbial

diversity, in order to describe the same proportion of microbial diversity in a sample. To estimate the proportion of discovered diversity, we first plotted rarefaction curves. It was evident that rarefaction curves for some samples did not reach a plateau, indicating that the substantial number of additional unique OTUs would likely be discovered at a sequencing depth greater than that used in the present study. In order to quantitatively assess whether the sequencing depth that was achieved in our study was sufficient for our purpose, we carried out an asymptotic richness estimation and found that all samples except one reaches 70% of coverage, which was acceptable for our purpose.

3.5.2 Different normalization approaches have a significant effect on the alpha diversity

Two commonly encountered challenges with microbial community data analyses include variable library sizes and a large number of zero values due to rare taxa that are not universally present across all analyzed samples (37,40). The bias introduced by uneven sequencing depth across different libraries is typically addressed using OTU normalization, however, whether and how normalization should be carried out is debatable (37,40). We applied three broadly used normalization methods, assessed and compared their effect on the calculated alpha diversity prior to subsequent data analyses. The first method, proportional transformation, conserves low abundant taxa and is strongly influenced by the sequencing depth that increases the detected richness on account of rare species. This approach is not appropriate when libraries are not sequenced at even depth, as in our case. To minimize the effect of uneven sequencing depth, a broadly applied rarefying can be used (37,64,65), however, rarefying may not be appropriate since it discards potentially informative low-

abundant OTUs (40). As a compromise between proportional transformation and rarefying, a statistical modeling approach to data normalization RLE has been proposed and used in our study (39,64,66). Based on the three alpha diversity indices, rarefied or proportional transformed data showed relatively lower alpha diversity values and smaller variances among the samples than statistical normalization, which is likely a result of discarded low-abundant OTUs, potentially informative variables. Hence, all subsequent data analyses were based on a dataset normalized using a statistical RLE to avoid losing informative data through rarefying by random subsampling (39,66). To minimize the effect of low-abundant taxa and place a higher weight on abundant taxa, inverse Simpson index was used in further analyses (44).

3.5.3 Microbial communities in sediment fractions differed from those found in water fractions

Previous studies that investigated water microbiomes mostly used the membrane filter after filtration to capture the microorganisms, and have not separated the sediment particles from water fractions (67–72). Here, we used Moore (i.e., cotton strip) swabs, which allowed us to capture microbial communities that had passed a sampling location in a water column within 24 hours. With that, we also captured a large amount of sediment suspended in the streams due to frequent rain. Hence, we used centrifugation to pellet the sediments. The remaining water fraction was further filter to avoid overlooking any microorganisms potentially persisting in the supernatant. To the best of our knowledge, previous studies have not used this sampling approach to characterize water microbiomes within a specified period (71–73). We observed that specific microbial communities persisted in water fractions after the

centrifugation, and that they differed compared to communities found in sediment fractions. While this approach allowed us to analyze sediment and water-associated microbial communities separately, it did not allow us to infer the relative abundance of microbial taxa in the whole sample extracted from the Moore swab. Nevertheless, it is reasonable to assume that the majority of microbial biomass was collected in a sediment fraction after centrifugation. Interestingly, rarefaction curves and richness estimation showed similar levels of richness in water and sediment fractions, indicating that the sample separation step did not reduce microbial diversity in water fractions.

In contrast to sample richness, beta diversity of sediment and water fractions was significantly different as indicated by PCoA ordination plot and PERMANOVA analyses. Previous studies on Mississippi river and Baltic Sea (brackish water) microbiome also revealed that sediment and water communities were distinct in structure even though their approach to processing sediment and water samples was different, as they used multiple filtration steps to capture the particulate (73–75). These observations suggest a specific micro-scale heterogeneity exists in the microbial communities of water, which was also observed in our study. The micro-scale heterogeneity is expected to be dependent on environmental factors that facilitate the bed sediment suspension and sedimentation dynamics.

Previous studies reported Actinobacteria, Proteobacteria, and Bacteroidetes as commonly detected in high relative abundance in surface waters in both water fraction and suspended sediment, which is consistent with our results (70,71,73,76). Even though these and other expected phyla were detected in our samples, their abundance between sediment

and water fractions was different. Bacterial phyla Nitrospinae, Acidobacteria, Rokubacteria, and Latescibacteria, were more abundant in sediment fractions, while Epsilonbacteraeota, Omnitrophicaeota, Lentisphaerae, and Bacteroidetes were more abundant in water fractions. These phyla are widespread in various environments including water and soil (77–81). Some of the phyla have recently been detected and named based on the metagenomic sequencing but have not been characterized in terms of their phenotypic traits. This leads to the lack of understanding in respect to their biological characteristics and functionalities in microbial communities.

Hongchen at el. revealed that bacterial communities in sediment of lake Chaka in China were dominated by phylum with low guanine-cytosine (GC) content (82). In contrast to the previous study, according to the differential abundance test results from our study, phyla that are significantly more abundant in water fraction than sediment tend to show lower G-C content. Epsilonbacteraeota (formerly classified as Epsilonproteobacteria) (average of 38 % GC), Lentisphaerae (40.95 % GC), Bacteroidetes (46 % GC), Tenericutes (32.2 % GC), Chlamydiae (40.3 % GC), Firmicutes (43.1 % GC), and Spirochaetes (40.6 % GC) were previously reported to have lower GC content (46–49). On the other hand, bacterial phyla with higher relative abundance in the sediment samples than water fraction tend to have higher GC content. For example, Acidobacteria (60 % GC), Rokubacteria (66 – 71 % GC), Verrucomicrobia (52 – 73 % GC), Chloroflexi (48 – 65 % GC), and Planctomycetes show relatively higher GC contents compared to higher abundant phyla in water fractions(83–86). On the other hand, bacterial phyla with higher relative abundance in the sediment samples than water fraction tend to have higher GC content. For example, Acidobacteria (60 % GC), Rokubacteria (66 – 71 % GC), Verrucomicrobia (52 – 73 % GC),

Chloroflexi (48 – 65 % GC), and Planctomycetes show relatively higher GC contents compared to higher abundant phyla in water fractions (79,87,88). In bacteria, high GC contents implies broader tolerance to external stresses that are imposed on the microorganisms (89), however, GC contents by itself cannot predict the stress tolerance of an organism (89). Compositional differences between sediment and water fractions after sample centrifugation indicate that different microbial taxa have different sedimentation properties that may affect the speed of their sedimentation in a stream after a rain event. Hence, it may be necessary to re-assess the water sampling methods to ensure appropriate timing and strategies for water sampling that will improve the chances of foodborne pathogen detection. Recently published report of the California romaine lettuce-associated *E. coli* outbreak investigation suggested possible transmission of the outbreak-causing strain from sediment of the irrigation water source, further emphasizing this need (90).

3.5.4 Composition and diversity of microbial communities differ among water streams

Alpha diversity analyses of microbial communities in samples collected from different streams for both sediment and water fraction indicate that 9 out of 60 compared pairs of streams had significantly different bacterial and fungal alpha diversity. Streams B and D had significantly different alpha diversity of bacterial communities in both sediment and water fractions. Upstream land use within 0 – 250 meters from the sampling site at streams B and D was predominantly covered with developed urban areas, suggesting that the land use immediately adjacent to the sampling site may not explain these differences. However, stream B was covered by mostly natural areas (91.4% forest, grassland, shrubland or wetland) in a greater distance of up to 1000 m from the sampling site, whereas, stream D

had 40.3 % of upstream land in the radius of 1000 m covered with a dairy area and 40.3 % with the natural areas. It appears that the land use up to 1000 m upstream from the sampling site had an effect on the diversity of the microbial communities, although these are merely observational speculations that may guide further hypothesis testing in follow-up studies.

In terms of beta diversity, pairwise PERMANOVA confirmed that all pairs of samples from different streams had a significantly different bacterial composition in sediment fractions. Ten out of 15 pairs of sampling sites also had significant differences in composition of bacterial water communities. For fungal communities, 13 out of 15 for sediment, and 11 out of 15 pairs had significantly different composition in water fractions.

Overall, the composition of microbial communities found in samples collected from the stream F was most similar to those found in other streams. Communities found in stream F were significantly different only compared to those found in stream D. The stream D stood out in that its microbial communities were distinct compared to communities found in all other streams. The area up to 250 m upstream of the sampling site on stream D was covered by both urban and natural surfaces and had pastoral activity. On the other hand, 79.7 % of the land 500 m upstream from the sampling location on the stream F was mainly covered by the dairy-related areas. This different land use upstream from the sampling site might be one of the important factors shaping microbial communities in sampled streams, particularly during and after rainfall. However, data for only six streams and merely descriptive link between land use upstream of the sampling sites and microbial communities do not allow for solid and statistically valid conclusions. Further studies are needed to characterize samples

collected from a larger number of streams to investigate the effects of the land use on the composition of microbial communities.

3.5.4 Poor accuracy of foodborne pathogen presence prediction based on select microbial families

We utilized variable importance measures (VIMs) from random forest classification to identify specific microbial families predictive of the presence of foodborne pathogens. The 17 selected families were identified as predictive of the presence of *Salmonella* spp. in sediment and 6 families in water communities, but at a very low, insufficient accuracy. Two fungal families were identified as predictive of the *L. monocytogenes* presence in water samples, but with a low accuracy. Based on the results from the present dataset analyses, we can conclude that the presence of foodborne pathogens cannot be predicted with sufficient accuracy based on the microbial community composition. Further application of the random forest classifier on a larger set of samples collected from a larger number of streams is needed to provide more insight as to whether the prediction accuracy determined in this study is low due to the lack of biologically meaningful associations between predictive families and pathogens or due to the small dataset. Furthermore, the majority of the identified predictive families are not well characterized or are considered as uncultured or candidate families. Nevertheless, some of the identified families have been well documented in previous studies. Among these, species of Chlamydiaceae family have been reported to cause human and animal infections (91). The genus *Chlamydia* causes the preventable blindness and sexually transmitted disease (91). Typical hosts of the family Chlamydiaceae are mammals, birds, cats, guinea pigs, and humans (92). Barneslellaceae are commonly

found in the human gut microbiome, and Bressa et al. reported that it is significantly correlated with the percentage of body fat of individuals (93), and Flavobacteriaceae are found in a wide variety of environments, from marine, freshwater to soil (94). Flavobacteriaceae is the second most abundant family in the entire dataset analyzed in this study. Members of certain species in this family are pathogenic toward humans, birds, fish, and other animals (95). However, due to the taxonomic diversity of the family, higher resolution of the taxonomic profiles (i.e., species level classification) is needed to judge the pathogenic capacity of representatives found in water samples collected in our study (95). Overall, based on the results of the random forest data analyses, it was not possible to predict presence of pathogens of interest and further research is warranted to determine whether this is a result of the absence of pathogen biomarkers or a failure in identifying them due to the small sample size.

3.6 Conclusions

Microbial communities of surface waters were associated with sample fractions, environmental factors including physicochemical properties, and geospatial factors. Taxonomic composition of suspended sediments and water fractions of collected water samples were significantly different. This finding implies a specific heterogeneity exists in the microbial communities of water, which might need to be taken in consideration when assessing the microbial quality of the untreated surface waters. Microbial communities also differed significantly between streams. That differences might be due to the land use and anthropogenic activity upstream of the sampling site, and their distance from the sampling

site. Furthermore, physicochemical properties of the sample showed a significant effect on microbial community composition. That multi-factor complexity of microbial communities' composition is a major challenge to identify the microbial quality of water and its association with microbial community. In general, applying machine learning can be helpful in analyses of complex datasets to indicate which taxa are the most discriminating in predicting specific outcomes of interest. In our data, random forest classification revealed that microbial communities had low accuracy of prediction of foodborne pathogen presence; thus, further research is needed to disclose whether the low accuracy of prediction was due to the absence of pathogen biomarkers, or the insufficient sample size. This study tested a novel approach for microbial quality indicator discovery that combined metagenomic, statistical, and machine learning methods. Thus, it provided baseline data describing the relationships between microbial community composition and detected pathogen occurrence in the northeast U.S.

3.7 References

1. Molly A. Maupin, Joan F. Kenny, Susan S. Hutson, John K. Lovelace, Nancy L. Barber and KSL. Estimated Use of Water in the United States in 2010. 2014. https://pubs.usgs.gov/circ/1405/

2. Dieter CA, Maupin MA, Caldwell RR, Harris MA, Ivahnenko TI, Lovelace JK, et al. Estimated use of water in the United States in 2015. Circular. Reston, VA; 2018. http://pubs.er.usgs.gov/publication/cir1441

3. Markland SM, Ingram D, Kniel KE, Sharma M. Water for Agriculture: the

Convergence of Sustainability and Safety. Microbiol Spectr. 2017 May;5(3).

4. Doyle MP, Erickson MC. Summer meeting 2007 – the problems with fresh produce: an overview. J Appl Microbiol. 2008 Aug 1;105(2):317–30. https://doi.org/10.1111/j.1365-2672.2008.03746.x

5. Ceuppens S, Hessel CT, de Quadros Rodrigues R, Bartz S, Tondo EC, Uyttendaele M. Microbiological quality and safety assessment of lettuce production in Brazil. Int J Food Microbiol. 2014 Jul 2;181:67–76. https://www.sciencedirect.com/science/article/pii/S0168160514002013

6. Weller D, Wiedmann M, Strawn LK. Irrigation Is Significantly Associated with an Increased Prevalence of *Listeria monocytogenes* in Produce Production Environments in New York State. J Food Prot. 2015 Jun;78(6):1132–41. http://jfoodprotection.org/doi/10.4315/0362-028X.JFP-14-584

7. Jones LA, Worobo RW, Smart CD. Plant-Pathogenic Oomycetes, *Escherichia coli* Strains, and Salmonella spp. Frequently Found in Surface Water Used for Irrigation of Fruit and Vegetable Crops in New York State. Elkins CA, editor. Appl Environ Microbiol. 2014 Aug 15;80(16):4814 LP – 4820. http://aem.asm.org/content/80/16/4814.abstract

8. CDC, Multistate Outbreak of *E. coli* O157:H7 Infections Linked to Romaine Lettuce (Final Update). 2018. https://www.cdc.gov/ecoli/2018/o157h7-04-18/index.html

9. CDC. National Outbreak Reporting System (NORS). 2017. https://wwwn.cdc.gov/norsdashboard/

10. FDA. Investigation Summary: Factors Potentially Contributing to the Contamination of Romaine Lettuce Implicated in the Fall 2018 Multi-State Outbreak of *E. coli* O157:H7. 2019. https://www.fda.gov/Food/RecallsOutbreaksEmergencies/Outbreaks/ucm631243.htm

11. Draper AD, Doores S, Gourama H, LaBorde LF. Microbial Survey of Pennsylvania Surface Water Used for Irrigating Produce Crops. J Food Prot. 2016;79(6):902–12. http://www.ncbi.nlm.nih.gov/pubmed/27296593%0A

12. McEgan R, Mootian G, Goodridge LD, Schaffner DW, Danyluk MD. Predicting *Salmonella* Populations from Biological, Chemical, and Physical Indicators in Florida Surface Waters. Appl Environ Microbiol. 2013 Jul 1;79(13):4094–105. http://aem.asm.org/lookup/doi/10.1128/AEM.00777-13

13. FDA. FSMA Final Rule on Produce Safety. 2017. https://www.fda.gov/Food/GuidanceRegulation/FSMA/ucm334114.htm

14. FDA. United States Food and Drug Administration. Standards for the growing, harvesting, packing, and holding of produce for human consumption. 2013. http://www.fda.gov/downloads/Food/GuidanceRegulation/FSMA/UCM360734.pdf.

15. Rock CM, Brassill N, Dery JL, Carr D, McLain JE, Bright KR, et al. Review of water quality criteria for water reuse and risk-based implications for irrigated produce under the FDA Food Safety Modernization Act, produce safety rule. Environ Res. 2019;172:616–29. http://www.sciencedirect.com/science/article/pii/S0013935118306856

16. Truchado P, Hernandez N, Gil MI, Ivanek R, Allende A. Correlation between *E. coli* levels and the presence of foodborne pathogens in surface irrigation water: Establishment of a sampling program. Water Res. 2018 Jan 1;128:226–33. https://www.sciencedirect.com/science/article/pii/S0043135417308643

17. Topalcengiz Z, Strawn LK, Danyluk MD. Microbial quality of agricultural water in Central Florida. PLoS One. 2017;12(4):e0174889. 10.1371/journal.pone.0174889

18. Haley BJ, Cole DJ, Lipp EK. Distribution, diversity, and seasonality of waterborne salmonellae in a rural watershed. Appl Environ Microbiol. 2009 Mar;75(5):1248–55.

19. Pachepsky Y, Shelton D, Dorner S, Whelan G. Can *E. coli* or thermotolerant coliform concentrations predict pathogen presence or prevalence in irrigation waters. Crit Rev Microbiol. 2016 May 3;42(3):384–93. https://doi.org/10.3109/1040841X.2014.954524

20. Vierheilig J, Savio D, Ley RE, Mach RL, Farnleitner AH, Reischer GH. Potential applications of next generation DNA sequencing of 16S rRNA gene amplicons in microbial water quality monitoring. Water Sci Technol. 2015;72(11):1962–72. https://www.ncbi.nlm.nih.gov/pubmed/26606090

21. USDA Natural Resources Conservation Service. https://www.nrcs.usda.gov/wps/portal/nrcs/site/national/home/

22. Multi-Resolution Land Cover Characteristics (MRLC) Consortium. www.mrlc.gov

23. Early Warning and Environmental Monitoring Program (EWEM). https://earlywarning.usgs.gov/

24. Spatial Data Sets Available on the WRD NSDI Node. https://water.usgs.gov/lookup/getgislist

25. Walters W, Hyde ER, Berg-Lyons D, Ackermann G, Humphrey G, Parada A, et al. Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. Bik H, editor. mSystems. 2016;1(1).

26. Ihrmark K, Bodeker ITM, Cruz-Martinez K, Friberg H, Kubartova A, Schenck J, et al. New primers to amplify the fungal ITS2 region--evaluation by 454-sequencing of artificial and natural communities. FEMS Microbiol Ecol. 2012 Dec;82(3):666–77.

27. 16S Metagenomic Sequencing Library Preparation. https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

28. Daum C. iTag Sample Amplification QC. 2016. https://jgi.doe.gov/wp-content/uploads/2016/10/iTag-Sample-Amplification-QC-v1.4.pdf

29. 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. 2009 Dec;75(23):7537–41.

30. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Vol. 79, Applied and

Environmental Microbiology. 1752 N St., N.W., Washington, DC; 2013. p. 5112–20.

31. Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, et al. 25 years of serving the community with ribosomal RNA gene reference databases and tools. J Biotechnol. 2017;261:169–76.

32. Koljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, et al. Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol. 2013 Nov;22(21):5271–7.

33. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016 Oct 18;4:e2584. https://peerj.com/articles/2584

34. Schloss PD, Handelsman J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol. 2005;71(3):1501–6.

35. Wickham H. ggplot2 : elegant graphics for data analysis. Dordrecht; New York: Springer; 2009. http://ggplot2.org

36. Chao A, Ma KH, Hsieh TC. SpadeR: Species Prediction and Diversity Estimation with R. R package; 2015. http://chao.stat.nthu.edu.tw/wordpress/software_download/

37. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017 Mar;5(1):27.

38. Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre

Legendre D, McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos MHH, Stevens ES and HW. vegan: Community Ecology Package. R package. 2018. https://cran.r-project.org/package=vegan

39. Maza E. In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. Front Genet. 2016 Sep 16;7:164. https://www.ncbi.nlm.nih.gov/pubmed/27695478

40. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10. https://doi.org/10.1371/journal.pcbi.1003531

41. Chao A. Nonparametric Estimation of the Number of Classes in a Population. Scand J Stat. 1984;11(4):265–70. http://www.jstor.org/stable/4615964

42. Shannon CE. A Mathematical Theory of Communication. Bell Syst Tech J. 1948 Jul 1;27(3):379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

43. Simpson EH. Measurement of Diversity. Nature. 1949 Apr 30;163:688. http://dx.doi.org/10.1038/163688a0

44. Morris EK, Caruso T, Buscot F, Fischer M, Hancock C, Maier TS, et al. Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. Ecol Evol. 2014/08/28. 2014 Sep;4(18):3514–24. https://www.ncbi.nlm.nih.gov/pubmed/25478144

45. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive

Analysis and Graphics of Microbiome Census Data. PLoS One. 2013 Apr 22;8(4):e61217. https://doi.org/10.1371/journal.pone.0061217

46. Whittaker RH. Vegetation of the Siskiyou Mountains, Oregon and California. Ecol Monogr. 1960 Feb 1;30(3):279–338. https://doi.org/10.2307/1943563

47. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007;35(21):7188–96.

48. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2011 Feb;5(2):169–72.

49. Anderson MJ. Permutational Multivariate Analysis of Variance (PERMANOVA). Wiley StatsRef: Statistics Reference Online. 2017. (Major Reference Works). https://doi.org/10.1002/9781118445112.stat07841

50. Martinez Arbizu P. pairwiseAdonis: Pairwise multilevel comparison using adonis. 2019. https://github.com/pmartinezarbizu/pairwiseAdonis

51. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilit\`{a}. Pubbl del R Ist Super di Sci Econ e Commer di Firenze. 1936;8:3–62. citeulike-article-id:1778138

52. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8

53. Alfred Ssekagiri WTS, Ijaz UZ. microbiomeSeq: An R package for microbial community analysis in an environmental context. 2018. http://userweb.eng.gla.ac.uk/umer.ijaz/projects/microbiomeSeq_Tutorial.html

54. Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. 1995. p. 278–82 vol.1.

55. Breiman L. Random Forests. Mach Learn. 2001 Oct;45(1):5–32. https://doi.org/10.1023/A:1010933404324

56. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome. 2013;1(1):11. https://doi.org/10.1186/2049-2618-1-11

57. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Software; Vol 1, Issue 5. 2008; https://www.jstatsoft.org/v028/i05

58. Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Comput Stat Data Anal. 2009 Sep 1 [cited 2019 Mar 1];53(11):3735–45. https://www.sciencedirect.com/science/article/pii/S0167947309001601#!

59. SMEETON, C. N. Early history of the kappa statistic. Biometrics. 1985;41:795. http://ci.nii.ac.jp/naid/10030965590/en/

60. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9(1):307. https://doi.org/10.1186/1471-2105-9-307

61. Janitza S, Strobl C, Boulesteix A-L. An AUC-based permutation variable importance measure for random forests. BMC Bioinformatics. 2013;14(1):119. https://doi.org/10.1186/1471-2105-14-119

62. Jain YK, Bhandare SK. Min max normalization based data perturbation method for pMin max normalization based data perturbation method for privacy protectionrivacy protection. Int J Comput Commun Technol. 2011;2(8):45–50.

63. Henry R, Schang C, Kolotelo P, Coleman R, Rooney G, Schmidt J, et al. Effect of environmental parameters on pathogen and faecal indicator organism concentrations within an urban estuary. Estuar Coast Shelf Sci. 2016;174:18–26.

64. McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. Methods for normalizing microbiome data: An ecological perspective. Methods Ecol Evol. 2019 Mar 1;10(3):389–400. https://doi.org/10.1111/2041-210X.13115

65. Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol. 2016;26(5):330–5. http://www.sciencedirect.com/science/article/pii/S1047279716300722

66. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26. https://doi.org/10.1093/bioinformatics/btp616

67. Hassell N, Tinker KA, Moore T, Ottesen EA. Temporal and spatial dynamics in microbial community composition within a temperate stream network. Environ Microbiol. 2018 Oct 1;20(10):3560–72. https://doi.org/10.1111/1462-2920.14311

68. Teachey ME, McDonald JM, Ottesen EA. Rapid and stable microbial community assembly in the headwaters of third-order stream. Appl Environ Microbiol. 2019 Apr 5;AEM.00188-19. http://aem.asm.org/content/early/2019/04/01/AEM.00188-19.abstract

69. Cagle R, Ramachandran P, Reed E, Commichaux S, Mammel MK, Lacher DW, et al. Microbiota of the Hickey Run Tributary of the Anacostia River. Microbiol Resour Announc. 2019 Mar 21;8(12):e00123-19. https://www.ncbi.nlm.nih.gov/pubmed/30938701

70. 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. 2013 Aug 8;115(5):1147–58. https://doi.org/10.1111/jam.12323

71. Van Rossum T, Peabody MA, Uyaguari-Diaz MI, Cronin KI, Chan M, Slobodan JR, et al. Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality. Vol. 6, Frontiers in Microbiology. 2015. p. 1405. https://www.frontiersin.org/article/10.3389/fmicb.2015.01405

72. Ghaju Shrestha R, Tanaka Y, Malla B, Bhandari D, Tandukar S, Inoue D, et al. Next-generation sequencing identification of pathogenic bacterial genes and their relationship with fecal indicator bacteria in different water sources in the Kathmandu Valley, Nepal. Sci Total Environ. 2017 Dec;601–602:278–84. http://linkinghub.elsevier.com/retrieve/pii/S0048969717312056

73. Payne JT, Millar JJ, Jackson CR, Ochs CA. Patterns of variation in diversity of the Mississippi river microbiome over 1,300 kilometers. PLoS One. 2017 Mar 28;12(3):e0174890. https://doi.org/10.1371/journal.pone.0174890

74. Rieck A, Herlemann DPR, Jürgens K, Grossart H-P. Particle-Associated Differ from Free-Living Bacteria in Surface Waters of the Baltic Sea. Front Microbiol. 2015 Dec 1;6:1297. https://www.ncbi.nlm.nih.gov/pubmed/26648911

75. Jackson CR, Millar JJ, Payne JT, Ochs CA. Free-Living and Particle-Associated Bacterioplankton in Large Rivers of the Mississippi River Basin Demonstrate Biogeographic Patterns. Wommack KE, editor. Appl Environ Microbiol. 2014 Dec 1;80(23):7186 LP – 7195. http://aem.asm.org/content/80/23/7186.abstract

76. Wang P, Zhao J, Xiao H, Yang W, Yu X. Bacterial community composition shaped by water chemistry and geographic distance in an anthropogenically disturbed river. Sci Total Environ. 2019;655:61–9. http://www.sciencedirect.com/science/article/pii/S004896971834590X

77. Spieck E, Keuter S, Wenzel T, Bock E, Ludwig W. Characterization of a new marine nitrite oxidizing bacterium, Nitrospina watsonii sp. nov., a member of the newly proposed phylum “Nitrospinae.” Syst Appl Microbiol. 2014;37(3):170–6. http://www.sciencedirect.com/science/article/pii/S0723202014000186

78. Kielak AM, Barreto CC, Kowalchuk GA, van Veen JA, Kuramae EE. The Ecology of Acidobacteria: Moving beyond Genes and Genomes. Front Microbiol. 2016 May 31;7:744. https://www.ncbi.nlm.nih.gov/pubmed/27303369

79. Becraft ED, Woyke T, Jarett J, Ivanova N, Godoy-Vitorino F, Poulton N, et al. Rokubacteria: Genomic Giants among the Uncultured Bacterial Phyla . Vol. 8, Frontiers in Microbiology. 2017. p. 2264. https://www.frontiersin.org/article/10.3389/fmicb.2017.02264

80. Farag IF, Youssef NH, Elshahed MS. Global Distribution Patterns and Pangenomic Diversity of the Candidate Phylum WS3. Löffler FE, editor. Appl Environ Microbiol. 2017 May 15;83(10):e00521-17. http://aem.asm.org/content/83/10/e00521-17.abstract

81. Waite DW, Vanwonterghem I, Rinke C, Parks DH, Zhang Y, Takai K, et al. Comparative Genomic Analysis of the Class Epsilonproteobacteria and Proposed Reclassification to Epsilonbacteraeota (phyl. nov.). Vol. 8, Frontiers in Microbiology. 2017. p. 682. https://www.frontiersin.org/article/10.3389/fmicb.2017.00682

82. Jiang H, Dong H, Zhang G, Yu B, Chapman LR, Fields MW. Microbial diversity in water and sediment of Lake Chaka, an athalassohaline lake in northwestern China. Appl Environ Microbiol. 2006;72(6):3832–45.

83. Thrash JC, Cho J-C, Vergin KL, Morris RM, Giovannoni SJ. Genome sequence of Lentisphaera araneosa HTCC2155T, the type species of the order Lentisphaerales in the phylum Lentisphaerae. J Bacteriol. 2010;192(11):2938–9.

84. Bohlin J, Snipen L, Hardy SP, Kristoffersen AB, Lagesen K, Dønsvik T, et al. Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genomics. 2010;11(1):464.

85. Rekadwad BN, Khobragade CN. Determination of GC content of Thermotoga maritima, Thermotoga neapolitana and Thermotoga thermarum strains: A GC dataset for higher level hierarchical classification. Data Br. 2016;8:300.

86. Reichenberger ER, Rosen G, Hershberg U, Hershberg R. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. Genome Biol Evol. 2015;7(5):1380–9.

87. Wertz JT, Kim E, Breznak JA, Schmidt TM, Rodrigues JLM. Genomic and physiological characterization of the Verrucomicrobia isolate Diplosphaera colitermitum gen. nov., sp. nov., reveals microaerophily and nitrogen fixation genes. Appl Environ Microbiol. 2012;78(5):1544–55.

88. Ward LM, Hemp J, Shih PM, McGlynn SE, Fischer WW. Evolution of phototrophy in the Chloroflexi phylum driven by horizontal gene transfer. Front Microbiol. 2018;9:260.

89. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun. 2006;347(1):1–3.

90. FDA. FDA Continues Investigation into Source of *E. coli* O157:H7 Outbreak Linked to Romaine Lettuce Grown in CA; CDC Reports End to Associated Illnesses. 2019. https://www.fda.gov/Food/RecallsOutbreaksEmergencies/Outbreaks/ucm626330.htm

91. Brown MA, Potroz MG, Teh S-W, Cho N-J. Natural Products for the Treatment of Chlamydiaceae Infections. Microorganisms. 2016 Oct 16;4(4):39.

https://www.ncbi.nlm.nih.gov/pubmed/27754466

92. Bush RM, Everett KD. Molecular evolution of the Chlamydiaceae. Int J Syst Evol Microbiol. 2001;51(1):203–20.

93. Bressa C, Bailén-Andrino M, Pérez-Santiago J, González-Soltero R, Pérez M, Montalvo-Lominchar MG, et al. Differences in gut microbiota profile between women with active lifestyle and sedentary women. PLoS One. 2017 Feb 10;12(2):e0171352. https://doi.org/10.1371/journal.pone.0171352

94. McBride M.J. The Family Flavobacteriaceae. In: Rosenberg E., DeLong E.F., Lory S., Stackebrandt E. TF, editor. The Prokaryotes. Springer, Berlin, Heidelberg; 2014. https://link.springer.com/referenceworkentry/10.1007%2F978-3-642-38954-2_130#citeas

95. Bernardet J-F, Nakagawa Y. An introduction to the family Flavobacteriaceae. Prokaryotes Vol 7 Proteobacteria Delta, Epsil Subclass. 2006;455–80.

Appendix A

Chapter 3 Supplemental Materials

A.1 Supplemental Tables

Table A.1 Metadata for collected samples with pathogen detection results

ID	Site	swab placed	*Salmonella* spp.	STEC	*Listeria* spp.	*L. monocytogenes*
A1	A	6/5/2017	+	+	MS Lost	MS Lost
A3	A	6/6/2017	-	+	-	-
A5	A	6/7/2017	-	+	-	-
A7	A	6/8/2017	+	+	-	-
A9	A	6/9/2017	+	+	-	-
A27	A	7/17/2017	+	+	-	-
A29	A	7/18/2017	+	+	-	-
A31	A	7/19/2017	-	+	+	-
A33	A	7/20/2017	-	+	-	-
A43	A	8/7/2017	-	+	-	-
A45	A	8/8/2017	+	+	+	-
A47	A	8/9/2017	+	+	+	+
A49	A	8/10/2017	-	+	-	-
B15	B	7/17/2017	+	+	-	-
B17	B	7/18/2017	+	-	-	-
B19	B	7/19/2017	-	-	-	-
B21	B	7/20/2017	+	+	+	-
B23	B	7/21/2017	-	+	-	-
B37	B	8/1/2017	-	-	-	-
B34	B	8/2/2017	-	+	-	-
B36	B	8/3/2017	+	+	+	-
B25	B	7/31/2017	+	+	-	-
B40	B	8/21/2017	-	+	-	-
B42	B	8/22/2017	-	-	+	-
B44	B	8/23/2017	-	-	-	-
B46	B	8/24/2017	+	+	-	-
C8	C	6/5/2017	+	+	-	-
C10	C	6/6/2017	+	+	-	-
C12	C	6/7/2017	+	+	-	-
C14	C	6/8/2017	+	+	-	-

C25	C	7/17/2017	+	+	+	+
C27	C	7/18/2017	-	-	+	+
C29	C	7/19/2017	-	+	-	-
C31	C	7/20/2017	-	+	+	+
C33	C	7/21/2017	+	+	-	-
ID	Site	swab placed	*Salmonella* spp.	STEC [a]	*Listeria* spp.	*L. monocytogenes*
C45	C	8/2/2017	+	-	-	-
C47	C	8/3/2017	-	+	-	-
C44	C	8/1/2017	-	-	-	-
D8	D	6/5/2017	+	+	MS Lost	MS Lost
D10	D	6/6/2017	+	+	-	-
D12	D	6/7/2017	+	+	-	-
D14	D	6/8/2017	+	+	-	-
D16	D	6/9/2017	-	-	-	-
D27	D	7/17/2017	+	+	-	-
D29	D	7/18/2017	+	+	-	-
D31	D	7/19/2017	-	+	+	-
D33	D	7/20/2017	-	+	-	-
D36	D	8/7/2017	-	+	-	-
D38	D	8/8/2017	+	+	+	-
D40	D	8/9/2017	-	+	-	-
D42	D	8/10/2017	+	+	-	-
E1	E	6/6/2017	-	+	-	-
E4	E	6/7/2017	+	+	-	-
E6	E	6/8/2017	-	+	-	-
E9	E	7/17/2017	+	+	-	-
E11	E	7/18/2017	-	+	+	+
E13	E	7/19/2017	+	+	+	+
E15	E	7/20/2017	+	+	+	-
E17	E	7/21/2017	+	+	+	-
E20	E	8/7/2017	+	+	-	-
E22	E	8/8/2017	-	+	-	-
E24	E	8/9/2017	-	+	-	-
E26	E	8/10/2017	+	-	-	-
F8	F	6/5/2017	+	+	-	-
F10	F	6/6/2017	+	+	-	-
F12	F	6/7/2017	+	+	-	-
F14	F	6/8/2017	-	+	-	-
F25	F	7/17/2017	-	+	+	-

F27	F	7/18/2017	+	+	+	-
F31	F	7/20/2017	+	+	+	-
F34	F	8/7/2017	+	+	-	-
F36	F	8/8/2017	+	+	+	-
F38	F	8/9/2017	-	+	+	-
F40	F	8/10/2017	-	+	+	-

ID	Site	swab placed	*Salmonella* spp.	STEC [a]	*Listeria* spp.	*L. monocytogenes*

Table A.2 Metadata for collected samples with physicochemical properties

ID	site	pH	water temperature (°C)	Turbidity (NTU)	DO (mg/L)	conductivity (uS/cm)	air temperature (°C)	flow rate (m/s)
A1	A	7.610	15.000	22.850	7.355	510.000	18.400	0.500
A3	A	7.625	14.550	13.900	7.705	479.000	14.300	0.583
A5	A	7.655	14.950	13.200	7.685	487.500	15.700	0.350
A7	A	7.680	16.750	10.615	7.230	515.000	20.000	0.250
A9	A	7.715	18.150	10.515	7.970	536.500	21.500	0.300
A27	A	7.465	23.500	22.100	6.545	532.500	28.800	0.550
A29	A	7.550	24.400	20.000	6.455	530.500	31.400	0.550
A31	A	7.615	23.450	11.550	6.290	597.500	26.600	0.350
A33	A	7.605	23.650	9.250	6.115	623.500	25.900	0.300
A43	A	7.770	18.550	27.555	7.853	552.882	22.600	0.600
A45	A	7.700	18.000	29.000	7.445	546.802	22.700	0.750
A47	A	7.825	18.250	9.625	7.665	546.001	24.550	0.500
A49	A	7.870	18.550	8.655	7.855	546.314	22.450	0.350
B15	B	7.665	18.450	6.365	8.385	465.000	20.400	0.150
B17	B	7.760	18.800	7.610	8.470	468.500	18.800	0.150
B19	B	7.895	19.600	4.030	8.568	484.000	20.100	0.100
B21	B	7.740	20.000	10.620	8.483	472.000	21.550	0.100
B23	B	7.660	19.900	10.580	8.190	469.000	21.100	0.050
B37	B	8.135	19.150	2.975	8.668	490.516	19.950	0.050
B34	B	8.225	19.000	1.580	8.989	502.638	20.050	0.000
B36	B	8.190	19.300	1.620	8.839	519.856	20.800	0.000
B25	B	7.824	21.243	2.881	9.039	474.286	24.000	0.100
B40	B	8.170	22.450	2.465	8.802	486.157	24.300	0.000
B42	B	8.190	19.750	2.945	8.914	470.471	25.550	0.000
B44	B	8.220	18.700	2.420	9.114	450.460	20.850	0.050
B46	B	8.165	17.600	1.725	9.114	458.342	20.500	0.100
C8	C	8.180	15.200	41.695	9.500	316.000	19.100	0.950

ID	site	pH	water temperature (°C)	Turbidity (NTU)	DO (mg/L)	conductivity (uS/cm)	air temperature (°C)	flow rate (m/s)
C10	C	7.985	15.700	43.335	9.030	252.000	18.400	0.800
C12	C	8.250	18.050	6.450	9.545	278.000	19.300	0.850
C14	C	8.440	19.600	5.040	9.825	316.000	23.900	0.900
C25	C	7.600	18.200	8.410	8.165	329.500	19.400	0.850
C27	C	7.710	18.600	10.550	8.200	348.500	17.300	0.850
C29	C	7.775	19.625	6.420	8.115	363.500	20.200	0.950
C31	C	7.750	19.875	3.090	8.090	377.500	21.800	0.850
C33	C	7.770	19.950	3.950	7.930	387.500	19.100	0.650
C35	C	7.995	19.000	2.205	8.980	421.000	25.350	0.750
C45	C	8.345	19.400	1.420	8.796	463.300	25.000	0.750
C47	C	8.265	19.900	1.735	8.631	473.854	26.200	0.750
C44	C	8.440	20.871	1.860	9.254	429.486	26.300	0.800
D8	D	7.620	14.275	5.385	7.325	692.500	22.650	0.100
D10	D	7.630	14.600	4.890	7.550	690.000	17.800	0.050
D12	D	7.640	14.550	5.845	7.700	697.000	18.700	0.050
D14	D	7.620	15.225	11.760	7.300	711.500	21.550	0.100
D16	D	7.600	16.225	9.960	6.980	714.500	24.350	0.050
D27	D	7.525	22.200	13.055	7.155	673.000	32.700	0.050
D29	D	7.645	22.600	8.355	7.465	686.500	32.000	0.100
D31	D	7.625	21.400	10.335	7.275	719.000	27.900	0.100
D33	D	7.640	21.850	12.093	7.260	734.000	27.100	0.100
D36	D	8.010	16.450	5.940	8.149	657.585	20.600	0.100
D38	D	8.045	16.250	6.160	8.188	643.734	20.950	0.050
D40	D	7.995	16.300	6.415	8.413	645.751	23.300	0.050
D42	D	8.000	16.600	5.415	8.415	648.664	22.250	0.050
E1	E	8.100	16.850	2.670	8.870	288.500	19.750	0.250
E4	E	8.050	17.700	2.675	8.745	293.000	24.200	0.250
E6	E	8.070	18.700	2.530	8.670	298.000	25.800	0.250
E9	E	7.495	18.900	5.690	8.295	308.000	23.450	0.600
E11	E	7.565	19.450	8.000	8.300	304.000	21.850	0.550
E13	E	7.755	20.250	5.025	8.240	319.500	25.950	0.400
E15	E	7.605	20.100	4.050	8.255	310.500	26.600	0.400
E17	E	7.620	19.450	3.975	8.425	310.500	23.750	0.450
E20	E	8.200	15.800	2.390	9.191	350.145	17.200	0.200
E22	E	8.200	15.750	2.365	9.221	338.246	17.700	0.200
E24	E	8.190	15.650	1.900	9.130	369.635	19.950	0.150
E26	E	8.190	16.250	1.480	8.992	389.347	20.550	0.150

ID	site	pH	water temperature (°C)	Turbidity (NTU)	DO (mg/L)	conductivity (uS/cm)	air temperature (°C)	flow rate (m/s)
F8	F	7.890	15.350	378.300	14.265	314.500	20.450	1.100
F10	F	7.980	15.750	14.345	9.195	328.000	21.300	0.800
F12	F	8.075	16.900	8.520	9.070	364.500	23.750	0.700
F14	F	8.110	17.650	7.900	9.000	403.500	29.650	0.650
F25	F	8.050	16.700	1.840	9.295	661.500	24.250	0.600
F27	F	8.080	16.700	3.220	9.405	665.500	24.250	0.700
F31	F	8.070	18.100	5.680	9.240	623.500	24.450	0.400
F34	F	8.410	13.650	2.185	9.138	531.059	16.550	0.200
F36	F	8.435	13.450	1.730	9.085	531.916	16.000	0.400
F38	F	8.420	13.350	1.475	9.173	544.417	16.650	0.400
F40	F	8.400	13.750	1.315	9.072	570.067	17.900	0.400

Table A.3 Statistical differences in bacterial and fungal communities between sediment and water fractions comparing three different normalization approaches

Normalization method	Bacterial communities[b]	Fungal communities[b]
Rarefying	**0.0001**	**0.0001**
Proportional Transformation	**0.0001**	**0.0001**
edgeR RLE[a]	**0.0001**	**0.0001**

[a]Relative log expression.

[b]P values equal or less than 0.05 are indicated in bold type.

Table A.4 Differences in bacterial and fungal communities between suspended sediment and water fractions.

Communities[a]	df[b]	SS[c]	MS[d]	F.Model[e]	p-value
Bacterial					
Sample type	1	0.25108	0.25108	30.538	< **0.001**
Residuals	111	0.91264	0.00822		
Total	112	1.16373			
Fungal					
Sample type	1	0.172	0.172	16.402	< **0.001**
Residuals	110	1.1536	0.0104		
Total	111	1.3256			

[a]Sample type, suspended sediment or water fraction

[b]df, degrees of freedom.

[c]SS, sum of square.

[d]MS, mean square.

[e]Pseudo F statistic.

Table A.5 Percent of upstream watershed that was developed, natural (i.e., forest, grassland., shrubland or wetland), pasture and cropland for each watershed and distance class.

	Percent of Upstream Area Classified in Each Land Cover Class																
	0-250 m Upstream of Site[a]				0-500 m[b]				0-1000 m [c]				Whole Watershed [d]				
Stream	Developed	Natural	Pasture	Crop	Developed	Natural	Pasture	Crop	Developed	Natural	Pasture	Crop	Developed	Natural	Pasture	Crop	Total Area (km2)
A	35.6	30.0	34.4	0.0	17.3	25.8	37.2	19.7	6.2	25.0	28.1	40.3	3.1	41.4	21.8	33.7	22.3
B	41.2	58.8	0.0	0.0	25.1	74.9	0.0	0.0	8.6	91.4	0.0	0.0	3.9	63.8	23.0	9.1	176.1
C	6.5	0.0	56.9	36.6	3.6	20.9	40.9	34.6	8.8	24.4	38.3	25.2	4.1	42.8	31.6	21.2	143.2
D	47.8	46.3	5.7	0.0	16.1	37.7	14.3	0.0	6.2	40.3	22.7	14.7	8.1	37.0	18.3	35.3	40.5
E	16.4	17.6	66.1	0.0	28.7	15.4	48.5	6.7	14.1	44.3	26.2	15.0	3.4	77.4	6.9	11.8	25.4
F	9.7	3.6	79.7	6.9	11.8	5.1	60.1	22.9	6.7	15.2	33.9	44.2	2.7	71.4	16.9	7.0	19.5

[a] Open water and barren land comprised less than 2% of each watershed 0-250 m upstream of the sampling site.

[b] Open water and barren land comprised less than 2% of each watershed 0-250 m upstream of the sampling site, with the exception of Stream D. For Watershed D 0-500 m upstream for the sampling site, 6.6% was open water and 25.4% was barren land.

[c] Open water and barren land comprised less than 2% of each watershed 0-250 m upstream of the sampling site, with the exception of Stream D and C. For Watershed C 0-1000 m upstream for the sampling site, 3.3 % was open water and <1% was barren land. For Watershed D 0-1000 m upstream for the sampling site, 4.1% was open water and 12.1% was barren land.

[d] Open water and barren land comprised less than 2% of each watershed 0-250 m upstream of the sampling site.

Table A.6 Differences in alpha diversity in samples collected from different water streams

Alpha diversity index	Bacterial communities[a]		Fungal communities[a]	
	Sediment	Water	Sediment	Water
Chao 1	**0.0073**	**0.0021**	0.0768	0.6270
Shannon	**0.0026**	**0.0004**	0.6678	**0.0003**
Inversed Simpson	**0.0000**	**0.0003**	0.9168	**0.0003**

[a]P-values were obtained using Kruskal-Wallis. P values equal or less than 0.05 are depicted in bold type.

Table A.7 Pairwise comparison of microbial alpha diversity in samples collected from different sampling stream

Bacterial communities[a]							
Sediment				Water			
Sampling site	Chao1	Shannon	InvSimpson	Sampling site	Chao1	Shannon	InvSimpson
A vs B	**0.0076**	0.8626	1	A vs B	0.4155	1	0.3239
A vs C	**0.0304**	0.2226	1	A vs C	0.603	1	1
A vs D	0.4464	**0.0082**	**0.0185**	A vs D	**0.0137**	0.0902	1
A vs E	**0.033**	1	1	A vs E	**0.001**	1	0.7936
A vs F	0.1805	0.7237	1	A vs F	0.2035	1	1
B vs C	1	1	0.6782	B vs C	1	1	1
B vs D	1	1	**0.0076**	B vs D	1	**0.0002**	**0.0003**
B vs E	1	1	1	B vs E	0.917	1	1
B vs F	1	1	1	B vs F	1	0.1674	1
C vs D	1	1	1	C vs D	1	**0.014**	0.0723
C vs E	1	0.2847	**0.0134**	C vs E	0.834	1	1
C vs F	1	1	1	C vs F	1	1	1
D vs E	1	**0.012**	**0**	D vs E	1	**0.0285**	**0.0013**
D vs F	1	1	**0.0442**	D vs F	1	1	0.395
E vs F	1	0.8662	1	E vs F	1	1	1
Fungal communities[a]							
Sediment				Water			
Sampling site	Chao1	Shannon	InvSimpson	Sampling site	Chao1	Shannon	InvSimpson
A vs B	0.9921	1	1	A vs B	1	**0.0001**	**0**
A vs C	1	1	1	A vs C	1	0.3684	0.2276
A vs D	1	1	1	A vs D	1	0.2238	0.1091
A vs E	1	1	1	A vs E	1	**0.0031**	**0.0234**
A vs F	1	1	1	A vs F	1	0.2677	1
B vs C	1	1	1	B vs C	1	0.2868	0.2818
B vs D	0.7018	1	1	B vs D	1	0.7052	0.837
B vs E	1	1	1	B vs E	1	1	1
B vs F	0.0941	1	1	B vs F	1	1	0.4626
C vs D	1	1	1	C vs D	1	1	1
C vs E	1	1	1	C vs E	1	1	1
C vs F	0.476	1	1	C vs F	1	1	1
D vs E	1	1	1	D vs E	1	1	1
D vs F	1	1	1	D vs F	1	1	1
E vs F	1	1	1	E vs F	1	1	1

[a]P-values were obtained using Dunn's post-hoc test of Kruskal-Wallis analyses after Bonferroni-correction for multiple comparisons. P values equal or lower than 0.05 are indicated in bold type.

Table A.8 Statistical differences in microbial community composition among samples positive for individual pathogens.

Foodborne pathogen	Bacterial communities[b]		Fungal communities[b]	
	Sediment	Water	Sediment	Water
Salmonella spp.	0.442	0.25	0.444	0.953
[a]STEC	**0.029**	**0.004**	0.434	**0.011**
Listeria spp.	0.317	0.095	0.429	0.674
Listeria monocytogenes	0.448	0.167	0.622	0.325

[a]Shiga-toxin producing *E. coli.*

[b]P values were obtained using pairwise PERMANOVA. P values equal or less than 0.05 are indicated in bold type.

A.2 Supplemental Figures

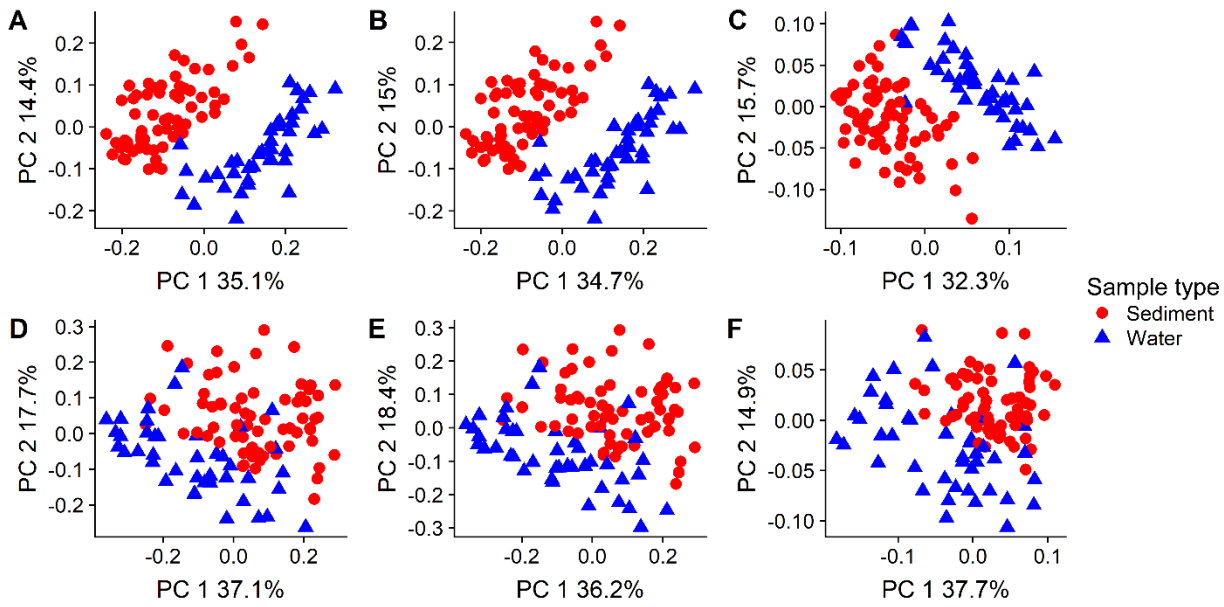

Figure A.2 Principal Coordinate Analysis (PCoA) based on the UniFrac distances for three different normalization approaches. Bacterial communities with (A) rarefied, (B) proportionally transformed, (C) normalized by RLE, and fungal communities with (D) rarefied, (E) proportionally transformed, and (F) normalized by RLE were compared respectively.

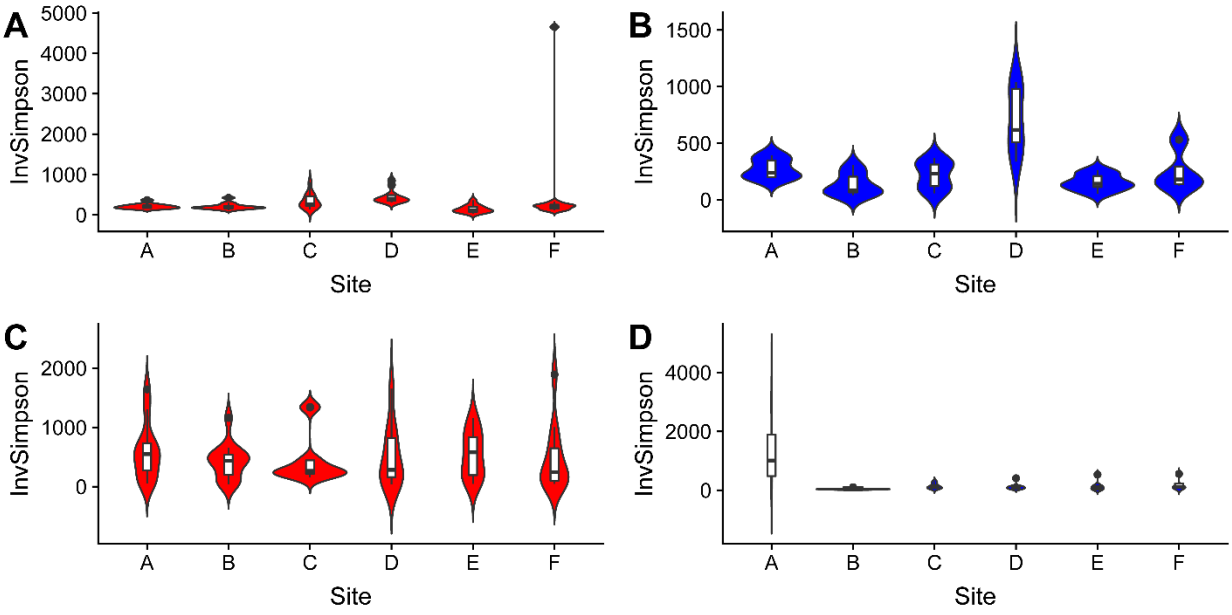

Figure A.3 Alpha diversity of (A, B) bacterial and (C, D) fungal communities found in samples collected from six different stream using inverse Simpson index. Analysis included sediment (n=68, red plots) and water (n=46, blue plots) fractions of collected water samples.

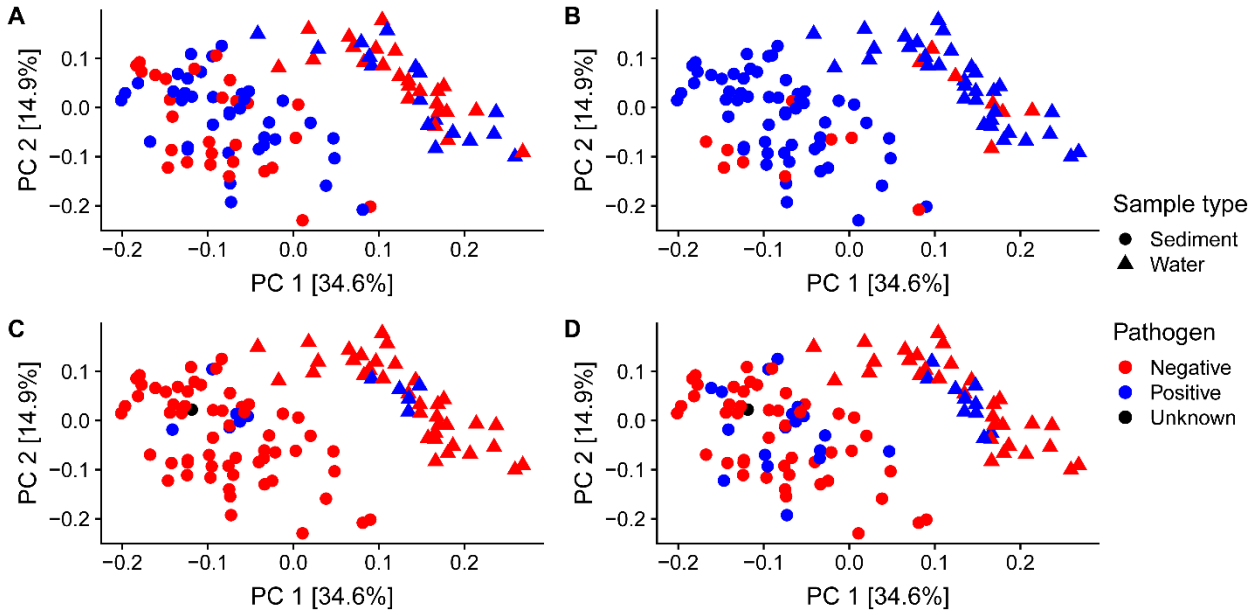

Figure A.4 Principal Coordinate Analysis (PCoA) based on the UniFrac distances for (A) *Salmonella* spp. (B) STEC (C) *Listeria monocytogenes* (D) *Listeria* spp. detection results. Blue dots indicate the presence of targeted microorganisms, red dots indicate the absence, and black dots indicate unknown due to missing swabs during sampling.

Appendix B

Optimization of PCR conditions for amplification of 16S rRNA and ITS

Objective of this study was to optimize the conditions of the polymerase chain reaction (PCR) for amplicon of 16S rRNA and ITS sequences to increase the PCR product yields and at the same time minimize the potential PCR bias due to stringent PCR annealing conditions.

Materials and methods

Preparation of template DNA and PCR master mix

Total DNA was extracted from environmental samples described in the previous chapter (Chapter 3.2.2) using DNeasy PowerWater DNA isolation kits or DNeasy PowerSoil DNA isolation kits (Qiagen), respectively. DNA concentration and purity were measured by Nanodrop UV-vis spectrophotometer (Thermo Fisher Scientific). PCR reaction mix without template DNA was prepared to contain PCR Ready Mix Kit (Kapa Biosystems), forward and reverse primer described in the previous chapter (Chapter 3.3.4), and PCR grade water (Table B1, B2).

Table B.1 Composition of PCR reaction mix for 16S rRNA amplification

Item	Original Concentration	Volumes for one reaction (µl)
2x KAPA HIFI HotStart Ready Mix	[1]2x	12.5
16S rRNA PCR Reverse Primer (µM)	10 µM	1
16S rRNA PCR Forward Primer (µM)	10 µM	1
PCR Grade Water	-	9.5
Total Volume	-	24

[1] 2x means that the concentration of the components in that solution is double their concentration in the final solution.

Table B.2 Composition of PCR reaction mix for ITS amplification

Item	Original Concentration	Volumes for one reaction (µl)
2x KAPA HIFI HotStart Ready Mix	2x	12.5
ITS PCR Reverse Primer (µM)	10 µM	1
ITS PCR Forward Primer (µM)	10 µM	1
PCR Grade Water	-	8.5
Total Volume	-	23

Annealing temperature optimization

Gradient PCR was performed to identify the optimal annealing temperature. A median annealing temperature was calculated based on the calculated melting temperature (T_m) of the primer (55.2°C for 16S rRNA V4 region and 51.6°C for ITS region). Ten different annealing temperatures (50°C to 59°C) were tested using the same PCR reaction mix as described above with same amount of template DNA. All other PCR thermal cycling conditions were consistent, except for the annealing temperature. Detailed thermal cycling conditions are described in the previous chapter (Chapter 3.2.4).

Template DNA concentration standardization

To find the optimal concentration of template DNA for PCR reaction, DNA was diluted in PCR-grade water to concentration within the range of 0-10 ng/µl, based on the concentration obtained through Nanodrop measurement. The diluted DNA concentration was determined by using the Qubit dsDNA High Sensitivity Assay Kit and Qubit 3. The concentration of the undiluted DNA was calculated by multiplying the measured concentration of diluted DNA by the dilution factor. PCR reaction was performed with standardized concentration (undiluted, 10ng/µl, 1ng/µl) of one randomly selected DNA sample extracted from a surface water sample (47.5 ng/µl). The reaction mix was prepared as described above.

Other factors tested in the PCR optimization

A total number of PCR cycles (20 – 30 cycles), initial denaturation temperature (98 to 95°C) and time (2 min – 5 min), extension time (15 sec – 60 sec), primer concentration (1 µM to 10 µM) were tested to test the conditions based on the recommended protocols (1,2).

Visualization and confirmation of PCR products

All tested PCR products were visualized by running electrophoresis on the 1% agarose gel. Additionally, a set of 12 pooled samples which were amplified before and after optimization was confirmed and compared using 2100 Bioanalyzer (Agilent) to assess the effect of optimization (3).

Results and Discussion

Different annealing temperature results in different amplification yield

The annealing temperature directly affects the results of amplification (4). If the annealing temperature is too low, primers may bind nonspecifically to the template, or annealing process will not work due to the high GC content of the PCR product (4). On the other hand, if the annealing temperature is too high, the yield of the desired product wil be reduced due to poor annealing of primers (4,5). Rychlik at el. developed the empirical formulation of annealing temperature based on the melting temperatures of primers (T_m) (6). Besides the complex equation, a typical annealing temperature is calculated as 5°C below the T_m of the primer (4). In our analyses, a temperature less than 53°C resulted in no or weak amplification or 16S rRNA, and higher temperature resulted in more smearing which might be due to the unspecific binding of primers (Figure B.1A). The optimal annealing temperature for 16S rRNA was determined as 54 - 56°C based on the intensity and clarity of the bands visualized on the agarose gel. For the ITS region, annealing temperature higher than 54°C showed weak or no amplification of the targeted region (Figure B.1B.). Temperatures between 52 – 54°C indicated weak amplification of one band that was otherwised produced when an annealing temperature of 50 – 51°C was used. Thus, the optimal temperature of the ITS region amplification was determined between 50 – 51°C. These results of protocol optimization were similar to the optimal theoretical annealing temperature that was 55.2°C for 16S rRNA V4 region primers and 51.6°C for ITS region primers.

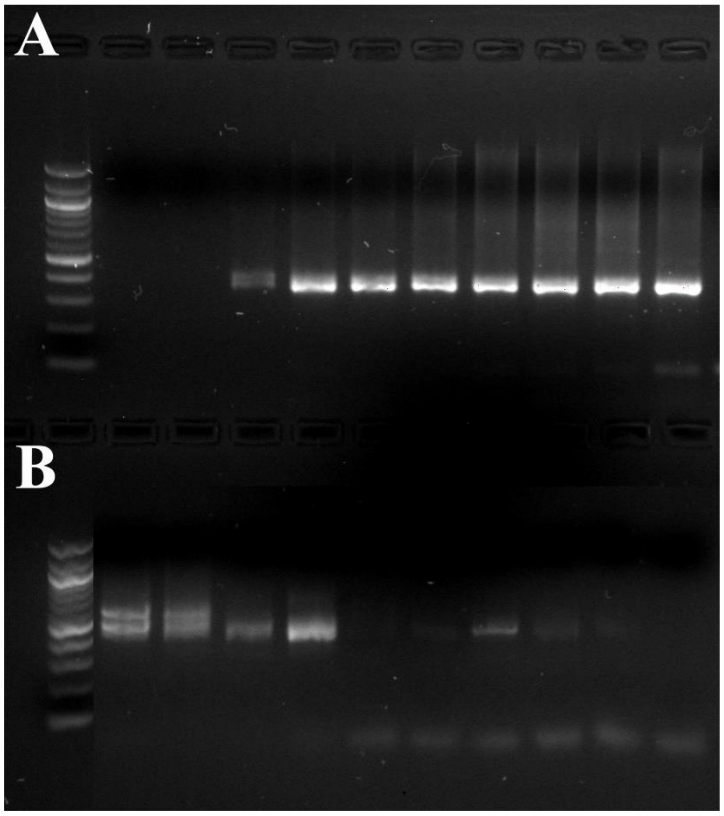

Figure B.1 Effect of annealing temperature during PCR on DNA amplification. Two different rows indicated (A) 16S rRNA V4 region and (B) ITS region amplification. Each column from the left after the DNA ladder represented the amplicons resulting from a gradual increase in annealing temperature from 50 to 59°C, with a 1°C increment.

The optimal concentration of template DNA was determined as 10 ng/µl

If the template concentration is too high, the polymerase can be inhibited due to carryover of inhibitors or inefficient denaturing due to the imbalance ratio between polymerase and template DNA concentration (4,5). In other ends, if the template concentration is too low, the yield of amplifying would reduce (4,5). Between 1 to 100 ng/µl is generally required; however, it depends on the purity of the extracted DNA sample (4,5). According to the visualization of agarose gel, 10 ng/µl was determined as the optimal concentration of template DNA for 16S rRNA and ITS amplification (Figure B.2).

Undiluted DNA showed the clear smearing on ITS amplification, and the lowest concentration (1 ng/µl) indicate weak amplification on 16S rRNA based on the visual intensity of the band.

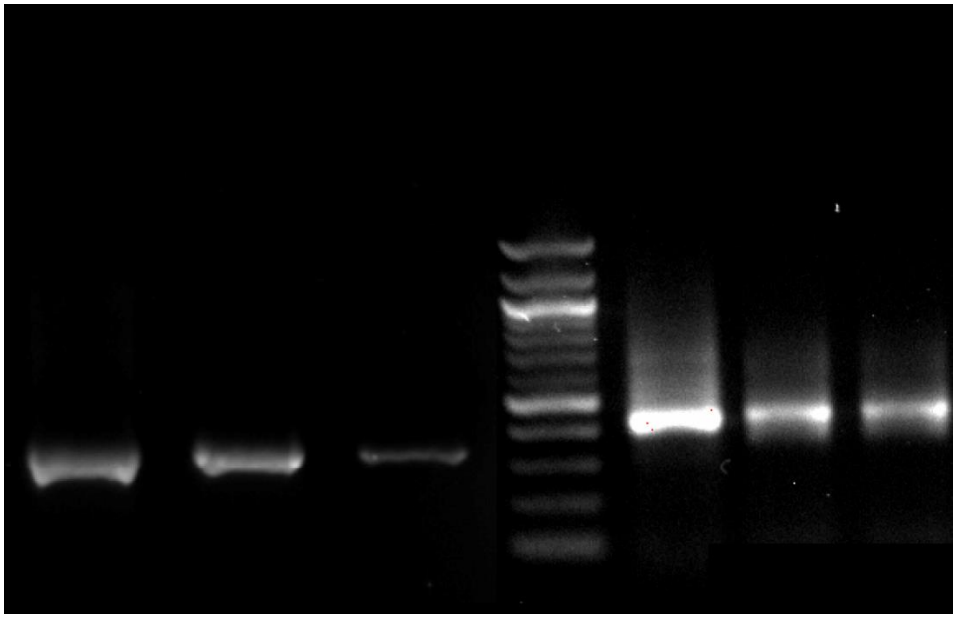

Figure B.2 Effect of template DNA concentration on DNA amplification. First three bands from the left showed 16S rRNA V4 region amplification with three different template DNA concentration in the order of undiluted, 10ng/µl, 1ng/µl. Last three bands indicated ITS region amplification with three different template DNA concentration in the order of undiluted, 10ng/µl, 1ng/µl.

Multiple factors tested to confirm the condition

Three notable problems were recognized in the PCR products (i.e., no band or faint band, nonspecific bands, or smeared bands). Those problems were due to the multiple factors, including PCR cycling condition (each time and temperature during the cycle) and PCR reaction components (concentration and purity of the reaction mix). Empirical optimization was completed to minimize those problems (data not shown). As a result, the optimal condition of PCR amplification of 16S rRNA and ITS were determined, respectively. For 16S rRNA, initial denaturation temperature was increased, and

denaturation, annealing, and extension times have reduced (Table B.3). For ITS, initial denaturation temperature was increased, extension time has increased, and the final extension was also increased compared with the original protocol (Table B.4). This optimized protocol was based on our specific sample type and may not be best when using a different set of samples with different characteristics. Thus, it is important to optimize PCR conditions based on the type of samples under investigation.

Table B.3 Thermo-cycling program for 16S rRNA

Step	Temperature	Time	
Initial Denaturation	98ºC	2 min	
Denaturation	95ºC	15 sec	25 Cycles
Annealing	55ºC	15 sec	
Extension	72ºC	15 sec	
Final Extension	72ºC	5 min	
Storage (Hold)	4ºC	Hold	

Table B.4 Thermo-cycling program for ITS

Step	Temperature	Time	
Initial Denaturation	98ºC	5 min	
Denaturation	95ºC	45 sec	30 Cycle
Annealing	50ºC	60 sec	
Extension	72ºC	60 sec	
Final Extension	72ºC	5 min	
Storage (Hold)	4ºC	Hold	

Bioanalyzer results indicate a notable effect of optimization on the PCR fragment size distribution

The Bioanalyzer (Agilent) utilizes capillary electrophoresis on a microchip device that is capable of rapidly sizing small DNA fragments (3). The data showed the migration-time plots to check the amplified bands' sizes, and artificial gel image based on the plots (3). Based on the results of pooled twelve samples before and after the optimization, notable differences between the two plots were observed (Figure B.3A, Figure B.3C). After optimization, the plots showed one clear peak that is constant throughout the samples. Those amplified fragments were represented of 16S rRNA V4 region; however, the plots before the optimization (Figure B.3C) showed unclear peak and unspecified size of the fragment, which was an evidence of weak amplification. Valid results of ITS are shown in Figure B.3B, whereas, it is hard to recognize the failure of amplification due to its nature of variation in the size of different fragment of ITS depending on the organisms. Thus, multiple peaks were concluded not to imply poor performance of PCR amplification.

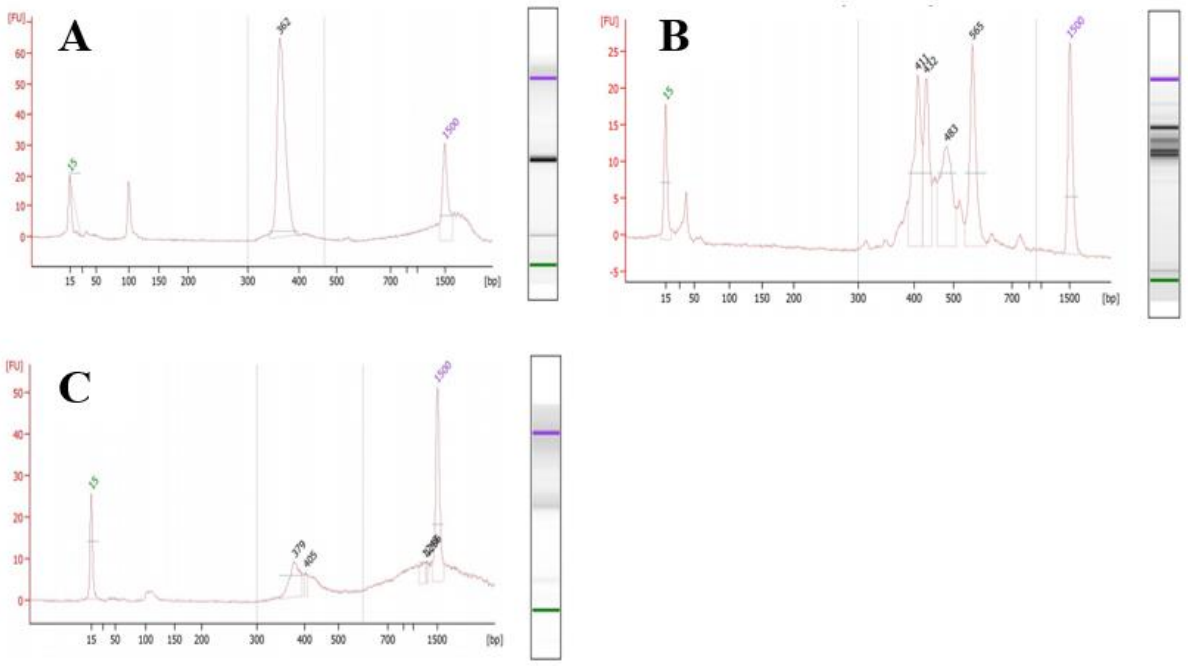

Figure B.3 Example of bioanalyzer results. (A) Valid result indicating successful amplification of 16S rRNA after optimization of PCR condition; (B) Valid result

indicating successful amplification of ITS after optimization of PCR condition; (C) Unsuccessful amplification of 16S rRNA before optimization. Unspecific or unclear peak indicates a failed PCR amplification.

References

1. 16S Metagenomic Sequencing Library Preparation. https://support.illumina.com/documents/documentation/chemistry_documentation/16S/16S-metagenomic-library-prep-guide-15044223-b.pdf
2. Daum C. iTag Sample Amplification QC. 2016. https://jgi.doe.gov/wp-content/uploads/2016/10/iTag-Sample-Amplification-QC-v1.4.pdf
3. Panaro NJ, Yuen PK, Sakazume T, Fortina P, Kricka LJ, Wilding P. Evaluation of DNA Fragment Sizing and Quantification by the Agilent 2100 Bioanalyzer. Clin Chem. 2000 Nov 1;46(11):1851 LP – 1853. http://clinchem.aaccjnls.org/content/46/11/1851.abstract
4. Innis MA, Gelfand DH, Sninsky JJ, White TJ. PCR protocols: a guide to methods and applications. Academic press; 2012.
5. Kramer MF, Coen DM. Enzymatic Amplification of DNA by PCR: Standard Procedures and Optimization. Curr Protoc Mol Biol. 2001 Oct 1;56(1):15.1.1-15.1.14. https://doi.org/10.1002/0471142727.mb1501s56
6. Rychlik W, Spencer WJ, Rhoads RE. Optimization of the annealing temperature for DNA amplification in vitro ; . Nucleic Acids Res. 1990 Nov 21;18(21):6409–12. https://doi.org/10.1093/nar/18.21.6409

Appendix C

Computational workflow of analysis of microbial communities

#Data analysis workflow
#Taejung Chung
#tuc289@psu.edu
#The Pennsylvania State University

#Sequence data analysis (Mothur v 1.40)
1. Bacterial communities analysis

From the raw sequencing files (.fastq)
Making '.files' which include list of samples and the sequences associated with those samples (paired R1, R2)
mothur > make.file(inputdir="" , type=fastq, prefix=16s)

Combine the paired-end reads together
Extract the sequence and quality score data from fastq files, create the contigs
mothur > make.contigs(file=16s.files)

Summary.seqs provide descriptive statistics of all sequences in .fasta file
It will summarize the quality of sequences (Start position, end position - aligned sequences only, Number of bases, ambiguous sequence, Polymer, and number of total sequences by percentile).
mothur > summary.seqs(fasta=16s.trim.contigs.fasta)

Clean up sequences
Remove any sequences with ambiguous bases ("N"), shorter than 292, longer than 294 - these arbitrary parameters are based on summary.seqs report
mothur > screen.seqs(fasta=16s.trim.contgis.fasta, group=16s.contigs.groups, summary=16s.trim.contigs.summary, minlength=292, maxlength=294, maxambig=0)

Collapse identical sequences. Representative sequence will be picked and stored in .fasta and corresponding sequences will be saved just as name of sequences to reduce the computational work
mothur > unique.seqs(fasta=16s.trim.contigs.good.fasta)

Create a count table of current unique sequences (name of representative sequences and abundance in sample sample)
mothur > count.seqs(name=16s.trim.contigs.good.names, group=16s.contigs.good.groups)

Customize database to targeted region of interest
Refine start and end position from whole 16s database
mothur > pcr.seqs(fasta=silva.nr_v132.align, start=11894, end=25319, keepdots=F)
Rename reference files to make it clear to put in command line
mothur > rename.file(input = silva.nr_v132.pcr.align, new= silva.v4.align)

Align sequence of 16S rRNA to Silva reference database
mothur > align.seqs(fasta=16s.trim.contigs.good.unique.fasta, reference=silva.v4.align, flip=T)

Screen sequences that are before or after the sites of alignment from the previous step, likely due to insertion or deletion at the terminal ends of the alignments. Define predominant start and stop sites based on summary.seqs from the previous step
SILVA dataset does not have more than 8 homopolymers (same base pair in a row), thus max homopolymer number to allow is 8
mothur > screen.seqs(fasta=16s.trim.contgis.good.unique.align, count=16s.trim.contigs.good.count_table, minlength=292, maxlength=294, maxhomop=8)

Eliminate overhangs to double-check analyses since paired-end sequencing already removed overhang during contig making step previously
Also filter the alignment characters that only consist of "-" (Vertical=T).
trump=. Allow to remove all the sequences contains '.'
mothur > filter.seqs(fasta=16s.trim.contigs.good.unique.good.align, vertical=T, trump=.)

Rerun unique.seqs in case new redundant sequences were created by filtering
mothur > unique.seqs(fasta=16s.trim.contigs.good.unique.good.filter.fasta, count = 16s.trim.contigs.good.count_table)

De-noise sequences. Abundant sequences are more likely to generate erroneous sequences than rare sequences (high possibility). Tak the current set of unique reads and collapse those that are similar to one another. Those that are within the threshold are merged with a larger sequence
diffs=2 means threshold of mismatch of the sequence. Sequences with less than 2 base pairs differences are merged
mothur > pre.cluster(fasta=16s.trim.contigs.good.unique.good.filter.unique.fasta, count=16s.trim.contigs.good.unique.good.filter.count_table, diffs=2)

Reads a fasta and count file to chimera sequences using UCHIME algorithm. The basic mechanism of detecting chimeric sequences is that the query sequence is divided into four non-overlapping segments, and used to search a reference database, which is considered as

chimera free. Comparison of results is calculated and detected chimeric sequences are based on differences from four segments
Dereplication is used when checking for chimeras by groups. If the parameter is set to false, if one group finds the sequence to be chimeric, then all groups find it to be chimeric
mothur >
chimera.uchime(fasta=16s.trim.contigs.good.unique.good.filter.unique.precluster.fasta, count=16s.trim.contigs.good.unique.good.filter.unique.precluster.count_table, dereplicate=t)

Remove sequences that have been detected as chimeric sequences
mothur >
remove.seqs(fasta=16s.trim.contigs.good.unique.good.filter.unique.precluster.fasta, accnos=16s.trim.contigs.good.unique.good.filter.unique.precluster.denovo.vsearch.accnos)

Assign the the taxonomy
mothur >
classify.seqs(fasta=16s.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta, count=16s.trim.contigs.good.unique.good.filter.unique.precluster.denovo.uchime.pick.count_table, reference=silva.nr_v132.align, taxonomy=silva.nr_v132.tax)

See if any undesired sequences have persisted in the dataset. Remove chloroplast, mitochondria that are only present in eukaryotic organisms
mothur >
remove.lineage(fasta=16s.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta, count=16s.trim.contigs.good.unique.good.filter.unique.precluster.denovo.uchime.pick.count_table,
taxonomy=16s.trim.contigs.good.unique.good.filter.unique.precluster.pick.nr_v132.wang.taxonomy, taxon=Chloroplast-Mitochondria-unknown-Eukaryota)

Calculate uncorrected pairwise distances between aligned DNA sequences. By default, a gap is only penalized once (string of gaps is considered as a single gap), and all distances are calculated. Cutoff value means distances larger than 0.03 will not be saved (97% similarity)
mothur >
dist.seqs(fasta=16s.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.fasta, cutoff=0.03)

Assign sequences to OTUs, Mothur provides three different methods of alignment. By default, opticlust method is used. Opticlust makes clusters of OTUs using metrics to determine the quality of clustering. Besides that, nearest neighbor, furthest neighbor, average neighbor method can be used.
mothur >
cluster(column=16s.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.dist, count
=16s.trim.contigs.good.unique.good.filter.unique.precluster.denovo.vsearch.pick.pick.count_table)

Determine how many sequences are in each OTU at the 0.03 cutoff level. Distribute OTUs into the groups
mothur >
make.shared(list=16s.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.opti_mcc.list,count=16s.trim.contigs.good.unique.good.filter.unique.precluster.denovo.vsearch.pick.pick.count_table, label=0.03)

Determine taxonomy for all OTUs. Outcome of this command will be used 'taxonomy file' which will be used for downstream analysis.
Outcome file will project: name of OTU, size (how many exist in a whole sample set), taxonomic profiles
mothur >
classify.otu(list=16s.trim.contigs.good.unique.good.filter.unique.precluster.pick.pick.opti_mcc.list,count=16s.trim.contigs.good.unique.good.filter.unique.precluster.denovo.vsearch.pick.pick.count_table,taxonomy=16s.trim.contigs.good.unique.good.filter.unique.precluster.pick.pds.wang.pick.taxonomy, label=0.03)

#2. Fungal communities analysis

From the raw sequencing files (.fastq)
Making .files which include list of samples and the sequences associated with those samples (paired R1, R2)
mothur > make.file(inputdir="" , type=fastq, prefix=ITS)

Combine the paired-end reads together
Extract the sequence and quality score data from fastq files, create the contigs
mothur > make.contigs(file=ITS.files)

Summary.seqs provides descriptive statistics of all sequences in .fasta file
Summarize the quality of sequences (start position, end position - aligned sequences only, number of bases, ambiguous sequence, polymers, and number of total sequences by percentile)
mothur > summary.seqs(fasta=ITS.trim.contigs.fasta)

Clean up sequences
Remove any sequences with ambiguous bases ("N"), longer than Y - these arbitrary parameter is based on summary.seqs report, for ITS, it is not recommended to set the minimum length due to its variability between organisms.
mothur > screen.seqs(fasta=ITS.trim.contigs.fasta, group=ITS.contigs.groups, maxlength=Y, summary=ITS.trim.contigs.summary, maxhomop=13, maxambig=0)

Collapse identical sequences. Representative sequence will be picked and stored in .fasta and corresponding sequences will be saved just as names of sequences to reduce computational work

mothur > unique.seqs(fasta=ITS.trim.contigs.good.fasta)

Create a count table of current unique sequences (name of representative sequences and abundance in sample sample)
mothur > count.seqs(name=ITS.trim.contigs.good.names, group=ITS.contigs.good.groups)

De-noise sequences. Abundant sequences are more likely to generate erroneous sequences than rare sequences (high possibility). Take the current set of unique reads and collapse those that are similar to one another. Those that are within the threshold are merged with the larger sequence
diffs=2 means threshold of mismatches of the sequence. Sequences with less than 2 base pairs differences are considered to be merged.
mothur > pre.cluster(fasta=ITS.trim.contigs.good.unique.fasta, count=ITS.trim.contigs.good.count_table, diffs=2)

Reads a fasta and count file to chimera sequences using UCHIME algorithm. The basic mechanism of detecting chimeric sequences is the query sequence is divided into four non-overlapping segments, and used to search a reference database, which is considered as chimera free. Comparison results are calculated and detected using chimeric sequences based on differences from four segments.
Dereplication is used when checking for chimeras by group. If the parameter is set to false, if one group finds the sequence to be chimeric, then all groups find it to be chimeric
mothur > chimera.uchime(fasta=ITS.trim.contigs.good.unique.precluster.fasta, count=ITS.trim.contigs.good.unique.precluster.count_table, dereplicate=t)

Remove sequences that have been detected as chimeric sequences
mothur > remove.seqs(fasta=ITS.trim.contigs.good.unique.precluster.fasta, accnos=ITS.trim.contigs.good.unique.precluster.denovo.vsearch.accnos)

Assign taxonomy
mothur > classify.seqs(fasta=ITS.trim.contigs.good.unique.precluster.pick.fasta, count=ITS.trim.contigs.good.unique.precluster.denovo.uchime.pick.count_table, reference=UNITEv6_sh_dynamic_s.fasta, taxonomy=UNITEv6_sh_dynamic_s.tax)

See if any undesired sequences have persisted in the dataset
mothur > remove.lineage(fasta=ITS.trim.contigs.good.unique.precluster.pick.fasta, count=ITS.trim.contigs.good.unique.precluster.denovo.uchime.pick.count_table, taxonomy=ITS.trim.contigs.good.unique.precluster.pick.UNITEv6_sh_dynamic_s.wang.taxonomy, taxon= Protozoa-Chromista-Eukaryota_kgd_Incertae_sedis-Bacteria-Animalia-Plantae-Plantae_unclassified-unknown-Protista-Fungi_unclassified-unclassified_Fungi)

Calculate uncorrected pairwise distances between aligned DNA sequences. By default, a gap is only penalized once (string of gaps is considered as a single gap), and all distances are calculated. Cutoff value means distances larger than 0.05 will not be saved (95% similarity)

mothur > dist.seqs(fasta=ITS.trim.contigs.good.unique.precluster.pick.pick.fasta, cutoff=0.05)

Assign sequences to OTUs, Mothur provides three different methods of alignment. By default, opticlust method is used. Opticlust makes clusters of OTUs using metrics to determine the quality of clustering. Besides that, nearest neighbor, furthest neighbor, average neighbor method can be used.
mothur > cluster(column=ITS.trim.contigs.good.unique.precluster.pick.pick.dist, count = ITS.trim.contigs.good.unique.precluster.denovo.vsearch.pick.pick.count_table)

Determine how many sequences are in each OTU at the 0.05 cutoff level. Distribute OTUs into the groups
mothur > make.shared(list=ITS.trim.contigs.good.unique.precluster.pick.pick.opti_mcc.list, count=ITS.trim.contigs.good.unique.precluster.denovo.vsearch.pick.pick.count_table, label=0.05)

Determine taxonomy for all OTUs. Outcome of this command will be used 'taxonomy file' which will be used for downstream analysis
Outcome file will project: name of OTU, size (how many exist in whole sample set), taxonomic profiles
mothur > classify.otu(list=ITS.trim.contigs.good.unique.precluster.pick.pick.opti_mcc.list, count=ITS.trim.contigs.good.unique.precluster.denovo.vsearch.pick.pick.count_table, taxonomy=ITS.trim.contigs.good.unique.precluster.pick.UNITEv6_sh_dynamic_s.wang.pick.taxonomy, label=0.03)

Downstream analayses
Carried out in R (version 3.5.2)

Import data
setwd - set directory with OTUs table, taxonomy table, and metadata
Install and load the required packages
source('http://bioconductor.org/biocLite.R')
biocLite('phyloseq')
library('phyloseq')
library('vegan')
library('ape')

Import .shared file (OTUs table from Mothur)
set.seed(336)
otus_16s <- import_mothur(mothur_shared_file="16S_WMP_OTU.shared")
otus_ITS <- import_mothur(mothur_shared_file="ITS_WMP_OTU.shared")

Convert an object into data frame, transpose the data
otus_16s_dataframe <- as.data.frame(otus_16s)

```
otus_ITS_dataframe <- as.data.frame(otus_ITS)
otus_16s_t <- t(otus_16s_dataframe)
otus_ITS_t <- t(otus_ITS_dataframe)

## Exclude the samples that has less than 9000 reads, rarefy by the minimum reads
otus.t.total_16s <- cbind(rowSums(otus_16s_t), otus_16s_t)
otus.t.sub_16s <- subset(otus.t.total_16s, otus.t.total_16s[,1]>=9000)
otus.t.sub_16s <- otus.t.sub_16s[,-1]
rarefy <- min(rowSums(otus.t.sub_16s))
otus.r_16s <- rrarefy(otus.t.sub_16s, rarefy)

otus.t.total_ITS <- cbind(rowSums(otus_ITS_t), otus_ITS_t)
otus.t.sub_ITS <- subset(otus.t.total_ITS, otus.t.total_ITS[,1]>=9000)
otus.t.sub_ITS <- otus.t.sub_ITS[,-1]
rarefy_ITS <- min(rowSums(otus.t.sub_ITS))
otus.r_ITS <- rrarefy(otus.t.sub_ITS, rarefy_ITS)

## Create a phyloseq object
OTU_16s = otu_table(otus.r_16s, taxa_are_rows=FALSE)
OTU_ITS = otu_table(otus.r_ITS, taxa_are_rows=FALSE)

taxon_16s <- import_mothur(mothur_constaxonomy_file="16S_WMP_TAX.taxonomy")
colnames(taxon_16s) <- c("Domain", "Phylum", "Class", "Order", "Family", "Genus")
TAX_16s = tax_table(as.matrix(taxon_16s))

taxon_ITS <- import_mothur(mothur_constaxonomy_file="ITS_WMP_TAX.taxonomy")
TAX_ITS = tax_table(as.matrix(taxon_ITS))

metadat_16s <- read.table("Newmeta_pathogen.csv", sep=",", header=T, row.names=1)
metadat_16s$pos_s <- as.factor(metadat_16s$pos_s)
metadat_16s$pos_stec <- as.factor(metadat_16s$pos_stec)
metadat_16s$lm <- as.factor(metadat_16s$lm)
metadat_16s$ls <- as.factor(metadat_16s$ls)
META_16s = sample_data(metadat_16s)

metadat_ITS <- read.table("Newmeta_pathogen_ITS.csv", sep=",", header=T,
row.names=1)
metadat_ITS$pos_s <- as.factor(metadat_ITS$pos_s)
metadat_ITS$pos_stec <- as.factor(metadat_ITS$pos_stec)
metadat_ITS$lm <- as.factor(metadat_ITS$lm)
metadat_ITS$ls <- as.factor(metadat_ITS$ls)
META_ITS = sample_data(metadat_ITS)

## Obtain phyloseq object for 16S rRNA
phyloseq1 = phyloseq(OTU_16s,TAX_16s,META_16s)
```

```
TREE_16s = rtree(ntaxa(phyloseq1), rooted=TRUE, tip.label = taxa_names(phyloseq1))
phyloseq1 = phyloseq(OTU_16s,TAX_16s,META_16s, TREE_16s)
phyloseq1

phyloseq1 %>%
  subset_taxa(Domain == "Bacteria" &
              Family != "mitochondria" &
              Class != "Chloroplast") -> phyloseq1

phyloseq1

## Obtain phyloseq object for ITS
phyloseq2 = phyloseq(OTU_ITS,TAX_ITS,META_ITS)
TREE_ITS = rtree(ntaxa(phyloseq2), rooted=TRUE, tip.label = taxa_names(phyloseq2))
phyloseq2 = phyloseq(OTU_ITS,TAX_ITS,META_ITS, TREE_ITS)
phyloseq2

phyloseq2 %>%
  subset_taxa(Rank1 == "k__Fungi" ) -> phyloseq2
phyloseq2

## Test different normalization approaches
## 1. rarefying
## Exclude samples that have less than 9000 reads (3% from total sample set)
otu_with_count1 <- rbind(colSums(OTU_16s), otus_16s)
otu_with_count1_t <- t(otu_with_count1)
otu_with_count_sub1 <- subset(otu_with_count1_t, otu_with_count1_t[,1]>=9000)
otu_with_count_sub1 <- otu_with_count_sub1[,-1]
otus.r_16s <- rrarefy(otu_with_count_sub1,min(rowSums(otu_with_count_sub1)))
OTU_16s_rare <- otu_table(otus.r_16s, taxa_are_rows=FALSE)
phyloseq1_rare <- phyloseq(OTU_16s_rare, TAX_16s, META_16s, TREE_16s)
phyloseq1_rare

otu_with_count2 <- rbind(colSums(OTU_ITS), otus_ITS)
otu_with_count2_t <- t(otu_with_count2)
otu_with_count_sub2 <- subset(otu_with_count2_t, otu_with_count2_t[,1]>=9000)
otu_with_count_sub2 <- otu_with_count_sub2[,-1]
otus.r_ITS <- rrarefy(otu_with_count_sub2,min(rowSums(otu_with_count_sub2)))
OTU_ITS_rare <- otu_table(otus.r_ITS, taxa_are_rows=FALSE)
phyloseq2_rare <- phyloseq(OTU_ITS_rare, TAX_ITS, META_ITS, TREE_ITS)
phyloseq2_rare

## 2. proportional transformation
otus_16s.perc<-t(OTU_16s)/colSums(t(OTU_16s))*100
OTU_16s_prop <- otu_table(otus_16s.perc, taxa_are_rows=TRUE)
```

```
phyloseq1_prop <- phyloseq(OTU_16s_prop, TAX_16s, TREE_16s, META_16s)
phyloseq1_prop

otus_ITS.perc <- t(OTU_ITS)/colSums(t(OTU_ITS))*100
OTU_ITS_prop <- otu_table(otus_ITS.perc, taxa_are_rows =TRUE)
phyloseq2_prop <- phyloseq(OTU_ITS_prop, TAX_ITS, TREE_ITS, META_ITS)
phyloseq2_prop

## 3. edgeR statistical model with RLE method
## Write a function for normalization
## norm.edgeR
biocLite("edgeR")
require(edgeR)
norm.edgeR =function(physeq, ...) {
   if (!taxa_are_rows(physeq)){
      physeq <- t(physeq)
}
x= as(otu_table(physeq), "matrix")
x=x+1
y=edgeR::DGEList(counts=x, remove.zeros=TRUE)
z=edgeR::calcNormFactors(y, ...)
return(z)

## Normalization
otus_16s_edgeR <- norm.edgeR(OTU_16s, method="RLE")
otus_ITS_edgeR <- norm.edgeR(OTU_ITS, method="RLE")

OTU_16s_edgeR <- otu_table(otus_16s_edgeR$counts, taxa_are_rows = TRUE)
OTU_ITS_edgeR <- otu_table(otus_ITS_edgeR$counts, taxa_are_rows = TRUE)

phyloseq1_edgeR <- phyloseq(OTU_16s_edgeR, TAX_16s, TREE_16s, META_16s)
phyloseq1_edgeR
phyloseq2_edgeR <- phyloseq(OTU_ITS_edgeR, TAX_ITS, TREE_ITS, META_ITS)
phyloseq2_edgeR

## Transform the OTUs into family level taxa
phyloseq1_family_rare <- phyloseq1_rare %>%
  tax_glom(taxrank = "Family")
phyloseq2_family_rare <- phyloseq2_rare %>%
  tax_glom(taxrank = "Rank5")

phyloseq1_family_prop <- phyloseq1_prop %>%
  tax_glom(taxrank = "Family")
phyloseq2_family_prop <- phyloseq2_prop %>%
  tax_glom(taxrank = "Rank5")
```

```
phyloseq1_family_edgeR <- phyloseq1_edgeR %>%
  tax_glom(taxrank = "Family")
phyloseq2_family_edgeR <- phyloseq2_edgeR %>%
  tax_glom(taxrank = "Rank5")


## Total PCoA plot at a family level using different normalization datasets
set.seed(336)
library(plyr)
PCoA = function(physeq, ...){
  require(phyloseq)
  require(ggplot2)
  a= ordinate(physeq,"PCoA", "Unifrac", weighted= TRUE)
  b= plot_ordination(physeq, a, color="Sampletype", shape="Sampletype")
  y= sub(".*\\[(.*)\\].*", "\\1", b$labels$y, perl=TRUE)
  x= sub(".*\\[(.*)\\].*", "\\1", b$labels$x, perl=TRUE)
  c= b+ geom_point(size=3)  + labs(x= paste("PC 1",x) , y= paste("PC 2",y)) +
scale_color_manual(values= c("Red", "Blue")) +theme(panel.grid.major = element_blank(),
panel.grid.minor = element_blank(),panel.background = element_blank(), axis.line =
element_line(colour = "black"), legend.position="none")
  c
}
library(cowplot)
legend_plot = PCoA_plot_ITS + geom_point(size=3) + scale_color_manual(values=
c("Red","Blue")) +theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),panel.background = element_blank(), axis.line = element_line(colour
="black"))
legend_plot$data$Sampletype <- revalue(legend_plot$data$Sampletype,
c("Sediment"="Sediment"))
legend <- get_legend(legend_plot)

## Plot all six PCoAs for comparison
library(cowplot)
plot <- plot_grid(PCoA(phyloseq1_family_rare), PCoA(phyloseq1_family_prop),
PCoA(phyloseq1_family_edgeR),PCoA(phyloseq2_family_rare),PCoA(phyloseq2_family_
prop), PCoA(phyloseq2_family_edgeR), ncol = 3, nrow =2,
labels=(c("A","B","C","D","E","F")),label_size =16)

plot_fin <- plot_grid(plot, legend,ncol=2, rel_widths = c(7, 1))
ggsave("FigureS1.tiff", device="tiff", dpi=600, width=10, height=5, units="in")

## Run PERMANOVA for different normalization effect on sample type
require(vegan)
set.seed(1)
```

```
PERMANOVA = function(physeq, ...){
  require(vegan)
  require(phyloseq)
  a= phyloseq::distance(physeq, method="unifrac", weighted=TRUE)
  b= data.frame(sample_data(physeq))
  perm = vegan::adonis(a~Sampletype,
                       data= b, permutations =9999)
  p = as.data.frame(perm$aov.tab$'Pr(>F)')[1,1]
  p
}

Bacterial<- c(PERMANOVA(phyloseq1_family_rare),
PERMANOVA(phyloseq1_family_prop), PERMANOVA(phyloseq1_family_edgeR))
set.seed(2)
Fungal<- c(PERMANOVA(phyloseq2_family_rare),
PERMANOVA(phyloseq2_family_prop), PERMANOVA(phyloseq2_family_edgeR))
fin <- data.frame(Bacterial, Fungal, row.names=c("Rarefied", "Proportion", "edgeR
(RLE)"))
fin

## Effect of normalization approach on alpha diversity
a = estimate_richness(phyloseq1_rare, measures=c("Chao1","Shannon","InvSimpson"))
b = estimate_richness(phyloseq1_prop, measures=c("Chao1","Shannon","InvSimpson"))
c = estimate_richness(phyloseq1_edgeR, measures=c("Chao1","Shannon","InvSimpson"))
write.csv(a, "a.csv")
write.csv(b, "b.csv")
write.csv(c, "c.csv")

d = estimate_richness(phyloseq2_rare, measures=c("Chao1","Shannon","InvSimpson"))
e = estimate_richness(phyloseq2_prop, measures=c("Chao1","Shannon","InvSimpson"))
f = estimate_richness(phyloseq2_edgeR, measures=c("Chao1","Shannon","InvSimpson"))
write.csv(d, "d.csv")
write.csv(e, "e.csv")
write.csv(f, "f.csv")


## Estimated total richness for each sample
library(SpadeR)
## Original script adopted from "https://cran.r-
project.org/web/packages/SpadeR/SpadeR.pdf"
## SpadeR::ChaoSpecies function is used for estimate richness of the communities

richness_estimate = function(otu,...) {
  options(warn = -1)
  b = data.frame(matrix(nrow=as.matrix(dim(otu))[2,], ncol=3))
```

```
  colnames(b) <- c("Chao1 Estimates","Observed OTUs", "%Covered Species")
  for (i in 1:as.matrix(dim(otu))[2,]) {
    a =SpadeR::ChaoSpecies(otu[,i], datatype="abundance", k=10, conf=0.95)
    b[i,1]= as.numeric(a$Species_table[3,1])
    b[i,2]= apply(as.data.frame(a$Basic_data_information),2,as.numeric)[2,2]
    b[i,3]= (b[i,2]/b[i,1])*100
    rownames(b) <- colnames(otu) }
print(b)
}

spadeR_16s_estimate <- richness_estimate(spadeR_16s)
spadeR_ITS_estimate <- richness_estimate(spadeR_ITS)

## Rarefaction curves
library(ggrare)
rarecurve_16s <- ggrare(phyloseq1, step = 100, se=TRUE, color="Sampletype")
rarecurve1 <- rarecurve_16s + facet_grid(Sampletype ~.) +
scale_color_manual(values=c("Red", "Blue"))+  theme(strip.background =
element_blank(),strip.text.y = element_blank(), legend.position="none")+xlab("Number of
OTUs") + ylab("Number of unique OTUs") +scale_x_continuous(expand = c(0, 0)) +
scale_y_continuous(expand = c(0, 0))+annotate("segment", x=-Inf, xend=Inf, y=-Inf, yend=-
Inf)+annotate("segment", x=-Inf, xend=-Inf, y=-Inf, yend=Inf)

rarecurve_ITS <- ggrare(phyloseq2, step = 100, se=TRUE, color="Sampletype")
rarecurve2 <- rarecurve_ITS + facet_grid(Sampletype ~., scales = "free_x")
+scale_color_manual(values= c("Red", "Blue"))+  theme(strip.background =
element_blank(),strip.text.y = element_blank())+xlab("Number of OTUs") + ylab("Number
of unique OTUs") + scale_x_continuous(expand = c(0, 0)) + scale_y_continuous(expand =
c(0, 0))+annotate("segment", x=-Inf, xend=Inf, y=-Inf, yend=-Inf)+annotate("segment", x=-
Inf, xend=-Inf, y=-Inf, yend=Inf)

rare1 <-plot_grid(rarecurve1, rarecurve2 + theme(legend.position="none"),
labels=c("A","B"),label_size = 20)
legend_rare <- get_legend(rarecurve2)
rare2 <-plot_grid(rare1, legend_rare, rel_widths = c(7,1))

ggsave("figure1.pdf", device="pdf", width=10, height=5, units="in", dpi=600)
ggsave("figure1.tiff", device="tiff", width=10, height=5, units="in", dpi=600)


## Use binomial test for assessment of significant differences of phyla in water and
sediment fractions
# Transform the OTUs to phyla
library(dplyr)
phyloseq1_phylum <- phyloseq1 %>%
```

```
  tax_glom(taxrank = "Phylum")
phyloseq2_phylum <- phyloseq2 %>%
  tax_glom(taxrank = "Rank2")
# Binomial test using DESeq2 package - bacteria
library('DESeq2')
sampletype <- phyloseq_to_deseq2(phyloseq1_phylum, ~ Sampletype)
gm_mean = function(x, na.rm=TRUE){
  exp(sum(log(x[x > 0]), na.rm=na.rm) / length(x))
}
geoMeans = apply(counts(sampletype), 1, gm_mean)
sampletype = estimateSizeFactors(sampletype, geoMeans = geoMeans)
sampletype = DESeq(sampletype, fitType="local")

install.packages("ggplot2")
library("ggplot2")

res1 <- results(sampletype, pAdjustMethod = "BH")
res1 <- res1[order(res1$padj, na.last=NA),]
alpha=0.05
sigtab1 = res1[(res1$padj < alpha),]
sigtab1 = cbind(as(sigtab1, "data.frame"), as(tax_table(phyloseq1)[rownames(sigtab1), ],
"matrix"))
sigtab1 <- sigtab1[order(sigtab1$log2FoldChange),]
sigtab1 = cbind(sigtab1,color1)

x = tapply(sigtab1$log2FoldChange, sigtab1$Phylum, function(x) max(x))
x = sort(x, TRUE)
sigtab1$Phylum = factor(as.character(sigtab1$Phylum), levels=names(x))
d1 <-ggplot(sigtab1, aes(y=Phylum, x=log2FoldChange, color=color1)) +
  geom_vline(xintercept = 0.0, color = "red", size = 1) +
  geom_point(size=3) +
  theme_set(theme_bw()) + scale_color_manual(values= c("Blue", "Red")) +
  labs(x ="log2 Fold Change") +
  theme(axis.text.x = element_text(angle = -90, hjust = 0, vjust=0.5, size=16),axis.text.y=
element_text(size=16), legend.position="none") +
  theme(axis.title= element_text(size=20)) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
d1

# Binomial test using DESeq2 package - fungi
sampletype2 <- phyloseq_to_deseq2(phyloseq2_phylum, ~ Sampletype)
gm_mean2 = function(x, na.rm=TRUE){
  exp(sum(log(x[x > 0]), na.rm=na.rm) / length(x))
}
geoMeans2 = apply(counts(sampletype2), 1, gm_mean2)
```

```
sampletype2 = estimateSizeFactors(sampletype2, geoMeans = geoMeans2)
sampletype2 = DESeq(sampletype2, fitType="local")

res2 <- results(sampletype2)
res2 <- res2[order(res2$padj, na.last=NA),]
alpha=0.05
sigtab2 = res2[(res2$padj < alpha),]
sigtab2 = cbind(as(sigtab2, "data.frame"), as(tax_table(phyloseq2)[rownames(sigtab2),],
"matrix"))
color2 <- c("Red","Blue","Blue","Blue","Blue","Blue")
sigtab2 <- sigtab2[order(sigtab2$log2FoldChange),]
sigtab2 = cbind(sigtab2,color2)
x2 = tapply(sigtab2$log2FoldChange, sigtab2$Rank2, function(x2) max(x2))
x2 = sort(x2, TRUE)
sigtab2$Rank2 = factor(as.character(sigtab2$Rank2), levels=names(x2))
d2 <-ggplot(sigtab2, aes(y=Rank2, x=log2FoldChange, color=color2)) +
  geom_vline(xintercept = 0.0, color = "red", size = 1) +
  geom_point(size=3) +
  theme_set(theme_bw()) + scale_color_manual(values= c("Blue", "Red")) +
  labs(x ="log2 Fold Change") +
  theme(axis.text.x = element_text(angle = -90, hjust = 0, vjust=0.5,
size=16),axis.text.y=element_text(size=16), legend.position="none") +
  theme(axis.title= element_text(size=20))+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
d2

library(cowplot)
dc <- plot_grid(d1+ theme(legend.position="none")
                , d2+ theme(legend.position="none")
                , labels=c("A","B"),label_size = 20)
dc
ggsave("Figure2.pdf", device="pdf", width= 10, height=7, dpi=600, units="in")
ggsave("Figure2.tiff", device="tiff", width= 10, height=7, dpi=600, units="in")

## PCoA Plot at a family level
## Transform the OTUs to family level
phyloseq1_family <- phyloseq1 %>%
  tax_glom(taxrank = "Family")

phyloseq2_family <- phyloseq2 %>%
  tax_glom(taxrank = "Rank5")
library(cowplot)
PCoA_total_16s = ordinate(phyloseq1_family, "PCoA", "Unifrac", weighted = TRUE)
PCoA_plot_16s = plot_ordination(phyloseq1_family, PCoA_total_16s,
color="Sampletype",shape="Sampletype")
```

```
PCoA_combined_16s = PCoA_plot_16s + geom_point(size=3) + labs(x= "PC 1 [34.6%]",
y= "PC 2 [14.9%]") + scale_color_manual(values= c("Red", "Blue"))
+theme(panel.grid.major = element_blank(), legend.position="none",panel.grid.minor =
element_blank() ,panel.background = element_blank(), axis.line = element_line(colour =
"black"))
set.seed(336)
colnames(sample_data(phyloseq2_family))[6] <- "Sampletype"

PCoA_total_ITS = ordinate(phyloseq2_family, "PCoA", "Unifrac", weighted = TRUE)
PCoA_plot_ITS = plot_ordination(phyloseq2_family, PCoA_total_ITS,
color="Sampletype",shape="Sampletype")
PCoA_combined_ITS = PCoA_plot_ITS + geom_point(size=3) + labs(x= "PC 1 [54.7%]",
y= "PC 2 [13.5%]")+ scale_color_manual(values= c("Red", "Blue"))
+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()
,panel.background = element_blank(), legend.position="none",axis.line =
element_line(colour = "black"))
legend = PCoA_plot_ITS + geom_point(size=3) +
  scale_color_manual(values= c("Red", "Blue"))
legend$labels$colour = "Sample type"
legend$labels$shape = "Sample type"
legend<- get_legend(legend)
p1 <- plot_grid(PCoA_combined_16s, PCoA_combined_ITS,labels=c("A","B"),label_size =
20 ,ncol=2)
p2 <- plot_grid(p1, legend, ncol=2, rel_widths = c(7, 1))
p2

ggsave("Figure3.pdf", device="pdf", width=10, height=5, units="in", dpi=600)
ggsave("Figure3.tiff", device="tiff", width=10, height=5, units="in", dpi=600)


## Run the PERMANOVA (vegan name : adonis) analyses per sample type (sediment vs.
water)
set.seed(1)
par(mfrow = c(1, 2)) # 2-by-2 grid of plots
par(oma = c(4, 4, 0, 0)) # make room (i.e. the 4's) for the overall x and y axis titles
par(mar = c(2, 2, 1, 1)) # make the plots be closer together

phyloseq1_dis_sampletype <- phyloseq::distance(phyloseq1_family,method="unifrac",
weighted=TRUE)
sample_permanova1_sampletype <- data.frame(sample_data(phyloseq1_family))
adonis_permanova1_sampletype <- adonis(phyloseq1_dis_sampletype~ Sampletype ,
data=sample_permanova1_sampletype, permutations =9999)
adonis_permanova1_sampletype
```

```
phyloseq2_dis_sampletype <- phyloseq::distance(phyloseq2_family,method="unifrac",
weighted=TRUE)
sample_permanova2_sampletype <- data.frame(sample_data(phyloseq2_family))
adonis_permanova2_sampletype <- adonis(phyloseq2_dis_sampletype~ Sampletype ,
data=sample_permanova2_sampletype, permutations =999)
adonis_permanova2_sampletype

## Subset data into two sample types
Sediment1 <- subset_samples(phyloseq1_family, Sampletype=="Sediment")
Water1 <- subset_samples(phyloseq1_family, Sampletype=="Water")
Sediment2 <- subset_samples(phyloseq2_family, Sampletype=="Sediment")
Water2 <- subset_samples(phyloseq2_family, Sampletype=="Water")

Sediment1_otu <- subset_samples(phyloseq1, Sampletype=="Sediment")
Water1_otu <- subset_samples(phyloseq1, Sampletype=="Water")
Sediment2_otu <- subset_samples(phyloseq2, Sampletype=="Sediment")
Water2_otu <- subset_samples(phyloseq2, Sampletype=="Water")

# Run pairwise PERMANOVA for site effect (OTUs level)
library(devtools)
library(pairwiseAdonis)

set.seed(10)
unifrac_Sediment1 <- phyloseq::distance(Sediment1_otu, method="unifrac",
weighted=TRUE)
sample_Sediment1 <- data.frame(sample_data(Sediment1))
Sediment1_pairadonis <- pairwise.adonis2(unifrac_Sediment1 ~ site,
data=sample_Sediment1)

set.seed(11)
unifrac_Water1 <- phyloseq::distance(Water1_otu, method="unifrac", weighted=TRUE)
sample_Water1 <- data.frame(sample_data(Water1))
Water1_pairadonis <- pairwise.adonis2(unifrac_Water1 ~ site, data=sample_Water1)

set.seed(12)
unifrac_Sediment2 <- phyloseq::distance(Sediment2_otu, method="unifrac",
weighted=TRUE)
sample_Sediment2 <- data.frame(sample_data(Sediment2))
Sediment2_pairadonis <- pairwise.adonis2(unifrac_Sediment2 ~ site,
data=sample_Sediment2)

set.seed(13)
unifrac_Water2 <- phyloseq::distance(Water2_otu, method="unifrac", weighted=TRUE)
sample_Water2 <- data.frame(sample_data(Water2))
Water2_pairadonis <- pairwise.adonis2(unifrac_Water2 ~ site, data=sample_Water2)
```

```
# Alpha diversity analyses visualization using violin plots for different indices
# Calculate alpha diversity using estimate_richness
alpha_S1 <- estimate_richness(Sediment1_otu, measures=c("Chao1","Shannon",
"InvSimpson"))
Arb_site1 = sample_data(Sediment1)$Arb_site
alpha_S1 <- cbind(alpha_S1, Arb_site1)

alpha_W1 <- estimate_richness(Water1_otu, measures=c("Chao1","Shannon",
"InvSimpson"))
Arb_site2 = sample_data(Water1)$Arb_site
alpha_W1 <- cbind(alpha_W1, Arb_site2)

alpha_S2 <- estimate_richness(Sediment2_otu, measures=c("Chao1","Shannon",
"InvSimpson"))
Arb_site3 = sample_data(Sediment2)$Arb_site
alpha_S2 <- cbind(alpha_S2, Arb_site3)

alpha_W2 <- estimate_richness(Water2_otu, measures=c("Chao1","Shannon",
"InvSimpson"))
Arb_site4 = sample_data(Water2)$Arb_site
alpha_W2 <- cbind(alpha_W2, Arb_site4)

# Kruskal-Wallis test
KW_Chao_S1 <- kruskal.test(Chao1 ~ sample_data(Sediment1_otu)$Arb_site, data =
alpha_S1)
KW_Shannon_S1 <- kruskal.test(Shannon ~ sample_data(Sediment1_otu)$Arb_site, data=
alpha_S1)
KW_Simpson_S1 <- kruskal.test(InvSimpson ~ sample_data(Sediment1_otu)$Arb_site,
data= alpha_S1)

KW_Chao_S2 <- kruskal.test(Chao1 ~ sample_data(Sediment2_otu)$Arb_site, data =
alpha_S2)
KW_Shannon_S2 <- kruskal.test(Shannon ~ sample_data(Sediment2_otu)$Arb_site, data=
alpha_S2)
KW_Simpson_S2 <- kruskal.test(InvSimpson ~ sample_data(Sediment2_otu)$Arb_site,
data= alpha_S2)

KW_Chao_W1 <- kruskal.test(Chao1 ~ sample_data(Water1_otu)$Arb_site, data =
alpha_W1)
KW_Shannon_W1 <- kruskal.test(Shannon ~ sample_data(Water1_otu)$Arb_site, data=
alpha_W1)
KW_Simpson_W1 <- kruskal.test(InvSimpson ~ sample_data(Water1_otu)$Arb_site, data=
alpha_W1)
```

```
KW_Chao_W2 <- kruskal.test(Chao1 ~ sample_data(Water2_otu)$Arb_site, data =
alpha_W2)
KW_Shannon_W2 <- kruskal.test(Shannon ~ sample_data(Water2_otu)$Arb_site, data=
alpha_W2)
KW_Simpson_W2 <- kruskal.test(InvSimpson ~ sample_data(Water2_otu)$Arb_site, data=
alpha_W2)

# Kruskal-Wallis multiple comparison test
install.packages("dunn.test")
library(dunn.test)
dunn_Chao_S1 <- dunn.test(alpha_S1$Chao1, sample_data(Sediment1)$Arb_site,
method="bonferroni",alt=TRUE)
dunn_Shannon_S1 <- dunn.test(alpha_S1$Shannon, sample_data(Sediment1)$Arb_site,
method="bonferroni",alt=TRUE)
dunn_InvSimpson_S1 <- dunn.test(alpha_S1$InvSimpson,
sample_data(Sediment1)$Arb_site, method="bonferroni",alt=TRUE)

dunn_Chao_W1 <- dunn.test(alpha_W1$Chao1, sample_data(Water1)$Arb_site,
method="bonferroni",alt=TRUE)
dunn_Shannon_W1 <- dunn.test(alpha_W1$Shannon, sample_data(Water1)$Arb_site,
method="bonferroni",alt=TRUE)
dunn_InvSimpson_W1 <- dunn.test(alpha_W1$InvSimpson,
sample_data(Water1)$Arb_site, method="bonferroni",alt=TRUE)

dunn_Chao_S2 <- dunn.test(alpha_S2$Chao1, sample_data(Sediment2)$Arb_site,
method="bonferroni",alt=TRUE)
dunn_Shannon_S2 <- dunn.test(alpha_S2$Shannon, sample_data(Sediment2)$Arb_site,
method="bonferroni",alt=TRUE)
dunn_InvSimpson_S2 <- dunn.test(alpha_S2$InvSimpson,
sample_data(Sediment2)$Arb_site, method="bonferroni",alt=TRUE)

dunn_Chao_W2 <- dunn.test(alpha_W2$Chao1, sample_data(Water2)$Arb_site,
method="bonferroni",alt=TRUE)
dunn_Shannon_W2 <- dunn.test(alpha_W2$Shannon, sample_data(Water2)$Arb_site,
method="bonferroni",alt=TRUE)
dunn_InvSimpson_W2 <- dunn.test(alpha_W2$InvSimpson,
sample_data(Water2)$Arb_site, method="bonferroni",alt=TRUE)

# Violin plots by different sampling stream (Inversed Simpson)
aS1 <- ggplot(alpha_S1, aes(x=Arb_site1, y=InvSimpson)) +
geom_violin(trim=FALSE,fill='Red') +
  geom_boxplot(width=0.1) +labs(x="Site", y = "InvSimpson")
aS1
aS2 <- ggplot(alpha_W1, aes(x=Arb_site2, y=InvSimpson)) +
geom_violin(trim=FALSE,fill='Blue') +
```

```
  geom_boxplot(width=0.1) +labs(x="Site", y = "InvSimpson")
aS2
aS3 <- ggplot(alpha_S2, aes(x=Arb_site3, y=InvSimpson)) +
geom_violin(trim=FALSE,fill='Red') +
  geom_boxplot(width=0.1) +labs(x="Site", y = "InvSimpson")
aS3
aS4 <- ggplot(alpha_W2, aes(x=Arb_site4, y=InvSimpson)) +
geom_violin(trim=FALSE,fill='Blue') +
  geom_boxplot(width=0.1) +labs(x="Site", y = "InvSimpson")
aS4

alpha_bac <- subset(alpha_2, amplicon=="16S")
alpha_fun <- subset(alpha_2, amplicon=="ITS")

inv1 <- ggplot(alpha_bac, aes(x=variable, y=value),) + geom_violin(aes(fill=SampleType),
trim=FALSE) +
  scale_fill_manual(values=c("red","blue")) +
  labs(x="Normalization method", y="Inverse Simpson") +
  theme(axis.text=element_text(size=14), axis.title=element_text(size=14),
        text=element_text(size=14, color="black"))+
  theme(panel.grid.major = element_blank(), legend.position="none",panel.grid.minor =
element_blank(),panel.background = element_blank(), axis.line = element_line(colour =
"black"))
inv1
inv2 <- ggplot(alpha_fun, aes(x=variable, y=value)) + geom_violin(aes(fill=SampleType),
trim=FALSE) +
  scale_fill_manual(values=c("red", "blue"))+
  labs(x="Normalization method", y="Inverse Simpson") +
  theme(axis.text=element_text(size=14), axis.title=element_text(size=14),
        text=element_text(size=14, color="black"))+
  theme(panel.grid.major = element_blank(), legend.position="none",panel.grid.minor =
element_blank(),panel.background = element_blank(), axis.line = element_line(colour =
"black"))
inv2

plot_grid(inv1, inv2, ncol=1, nrow=2, labels=c("A", "B"), label_size=16)
ggsave("alpha.pdf", dpi=600, device="pdf", width=10, height=5, units="in")

plot_grid(aS1,aS2,aS3,aS4, labels=c("A","B","C","D"), label_size=20)
ggsave("Figure4.pdf", dpi=600, device = "pdf", width=10, height=5, units="in")
ggsave("Figure4.tiff", dpi=600, device = "tiff", width=10, height=5, units="in")


#Environmental factors relatied to the microbial community diversity/composition
#Environmental variables in different file - load
```

```
metadat_16s_env <- read.table("Environmental_all_samples_16S.csv", sep=",", header=T,
row.names=1)
META_16s_env = sample_data(metadat_16s_env)

phyloseq1_env <- phyloseq(OTU_16s, TAX_16s, META_16s_env)
TREE_16s = rtree(ntaxa(phyloseq1_env), rooted=TRUE, tip.label =
taxa_names(phyloseq1_env))
phyloseq1_env = phyloseq(OTU_16s,TAX_16s,META_16s_env,TREE_16s)
phyloseq1_env

metadat_ITS_env <- read.table("Environmental_all_samples_ITS.csv", sep=",", header=T,
row.names=1)
META_ITS_env = sample_data(metadat_ITS_env)

phyloseq2_env <- phyloseq(OTU_ITS, TAX_ITS, META_ITS_env)
TREE_ITS = rtree(ntaxa(phyloseq2_env), rooted=TRUE, tip.label =
taxa_names(phyloseq2_env))
phyloseq2_env = phyloseq(OTU_ITS,TAX_ITS,META_ITS_env,TREE_ITS)
phyloseq2_env

##Normalization - edgeR (RLE)##
otus_16s_edgeR <- norm.edgeR(otu_table(phyloseq1_env), method="RLE")
otus_ITS_edgeR <- norm.edgeR(otu_table(phyloseq2_env), method="RLE")

OTU_16s_edgeR <- otu_table(otus_16s_edgeR$counts, taxa_are_rows = TRUE)
OTU_ITS_edgeR <- otu_table(otus_ITS_edgeR$counts, taxa_are_rows = TRUE)

phyloseq1_edgeR <- phyloseq(OTU_16s_edgeR, TAX_16s, TREE_16s, META_16s_env)
phyloseq1_edgeR
phyloseq2_edgeR <- phyloseq(OTU_ITS_edgeR, TAX_ITS, TREE_ITS, META_ITS_env)
phyloseq2_edgeR

phyloseq1_env <- phyloseq1_edgeR
phyloseq2_env <- phyloseq2_edgeR

install.packages("dplyr")
library(dplyr)
phyloseq1_env %>%
  subset_taxa(Domain == "Bacteria" &
              Family != "mitochondria" &
              Class != "Chloroplast") -> phyloseq1_env
phyloseq2_env %>%
  subset_taxa(Rank1 == "k__Fungi" ) -> phyloseq2_env

## Phyloseq object for the environmental effects analyses
```

```
phyloseq1_env_phylum <- taxa_level(phylose1_env, "Phylum")
phyloseq2_env_phylum <- taxa_level(phyloseq2_env, "Rank2")

## Canonical Correspondence Analysis
color <- c("#084594","#FFF5F0","#99000D","#FB6A4A","#737373","#E6E6FA")
Sediment_env <- subset_samples(phyloseq1_env_phylum, Sample.Type =="Sediment")
Water_env <- subset_samples(phyloseq1_env_phylum, Sample.Type =="Water")


library(vegan)
otu.table.cca1 <- data.frame(otu_table(Sediment_env))
data1 = data.frame(sample_data(Sediment_env))
cca.avg1 <- cca(otu.table.cca1~ av_ph + av_cond + av_a_t + av_w_t + av_flow + av_do +
av_turb, data=data1[,-c(1:10)])
cca.avg1

mod0<- cca(otu.table.cca1 ~ 1, data=data1[,-c(1:10)]) # create a null model
set.seed(9999)
mod <- step(mod0, scope = formula(cca.avg1), test = "perm", perm.max = 100)

cca.var <- cca(formula= otu.table.cca1 ~ av_cond + av_ph + av_flow, data=data1)
anova(cca.avg1)
anova(cca.var)
set.seed(9999)
anova(cca.var, by="axis", perm=100)
anova(cca.var, by="terms")

otu.table.cca2 <- data.frame(otu_table(Water_env))
data2 = data.frame(sample_data(Water_env))
cca.avg2 <- cca(otu.table.cca2~ av_ph + av_cond + av_a_t + av_w_t + av_flow + av_do +
av_turb, data=data2[,-c(1:10)])
cca.avg2

mod02<- cca(otu.table.cca2 ~ 1, data=data2[,-c(1:10)]) # create a null model
set.seed(998)
mod <- step(mod02, scope = formula(cca.avg2), test = "perm", perm.max = 100)
cca.var2 <- cca(formula= otu.table.cca1 ~ av_turb + av_w_t + av_flow, data=data1)
anova(cca.avg2)
anova(cca.var2)
set.seed(998)
anova(cca.var2, by="axis", perm=100)
anova(cca.var2, by="terms")
cca.var2 <- cca(formula = otu.table.cca2 ~ av_flow + av_w_t, data=data2)

## CCA - fungi
```

```
Sediment_env2 <- subset_samples(phyloseq2_env_phylum, Sampletype =="Sediment")
Water_env2 <- subset_samples(phyloseq2_env_phylum, Sampletype =="Water")

otu.table.cca3 <- data.frame(otu_table(Sediment_env2))
data3 = data.frame(sample_data(Sediment_env2))
cca.avg3 <- cca(otu.table.cca3~ av_ph + av_cond + av_a_t + av_w_t + av_flow + av_do +
av_turb, data=data3)
cca.avg3

mod03<- cca(otu.table.cca3 ~ 1, data=data3) # create a null model
set.seed(997)
mod <- step(mod03, scope = formula(cca.avg3), test = "perm", perm.max = 100)

cca.var3 <- cca(formula= otu.table.cca3 ~ av_flow + av_w_t + av_a_t + av_ph, data=data3)
anova(cca.avg3)
anova(cca.var3)
set.seed(900)
anova(cca.var3, by="axis", perm=100)
anova(cca.var3, by="terms")
cca.var3 <- cca(formula= otu.table.cca3 ~ av_flow +  av_a_t + av_ph, data=data3)

otu.table.cca4 <- data.frame(otu_table(Water_env2))
data4 = data.frame(sample_data(Water_env2))
cca.avg4 <- cca(otu.table.cca4~ av_ph + av_cond + av_a_t + av_w_t + av_flow + av_do +
av_turb, data=data4)
cca.avg4

mod04<- cca(otu.table.cca4 ~ 1, data=data4) # create a null model
set.seed(996)
mod <- step(mod04, scope = formula(cca.avg4), test = "perm", perm.max = 100)

cca.var4 <- cca(formula= otu.table.cca4 ~ av_cond + av_turb + av_a_t + av_w_t,
data=data4)
anova(cca.avg4)
anova(cca.var4)
set.seed(996)
anova(cca.var4, by="axis", perm=100)
anova(cca.var4, by="terms")
cca.var4 <- cca(formula= otu.table.cca4 ~ av_cond + av_turb, data=data4)

tiff("Plot1.tiff", width = 5, height = 5, units = 'in', res = 600)
plot(cca.var, type="n", display="sites")
points(cca.var, display = "sites", pch=c(15), cex = 0.9, font = 1)
points(cca.var, display = "bp", lwd = 2, col = "blue")
dev.off()
```

```
text(cca.var, display = "bp", col = "blue", font = 2, cex=1.5)

tiff("Plot2.tiff", width = 5, height = 5, units = 'in', res = 600)
plot(cca.var4, type="n", display="sites")
points(cca.var4, display = "sites", pch=c(15), cex = 0.9, font = 1)
points(cca.var4, display = "bp", lwd = 2, col = "blue")
text(cca.var4, display = "bp", col = "blue", font = 2, cex=1.5)
dev.off()


## Associations between microbial communities and environmental factors
library(devtools)
library(plyr)
options(devtools.install.args = c("--no-multiarch", "--no-test-load"))
install_github("umerijaz/microbiomeSeq")
library(microbiomeSeq)
library(extrafont)
loadfonts(device="win")

env.taxa.cor_avg <- taxa.env.correlation(phyloseq1_env_phylum, grouping_column =
"Sample.Type", method = "pearson",
                              pvalue.threshold = 0.05, padjust.method = "BH", adjustment = 5,
num.taxa = 78,
                              select.variables =
c("av_ph","av_w_t","av_turb","av_do","av_cond","av_a_t","av_flow"))
p1 <- plot_taxa_env(env.taxa.cor_avg)
p1$data$Env <- revalue(p1$data$Env, c(av_a_t ="Air Temperature",
                              av_cond = "Conductivity",
                              av_do = "Dissolved Oxygen",
                              av_flow = "Flow Rate",
                              av_ph ="pH",
                              av_turb ="Turbidity",
                              av_w_t ="Water Temperature"))
print(p1)


## Fungi
env.taxa.cor2 <- taxa.env.correlation(phyloseq2_env_phylum, grouping_column =
"Sampletype", method = "pearson",
                              pvalue.threshold = 0.05, padjust.method = "BH", adjustment = 5,
num.taxa = 16)
p2 <- plot_taxa_env(env.taxa.cor2)
print(p2)
```

```
env.taxa.cor_avg_2 <- taxa.env.correlation(phyloseq2_env_phylum, grouping_column =
"Sampletype", method = "pearson",
                                pvalue.threshold = 0.05, padjust.method = "BH", adjustment = 5,
num.taxa = 78,
                                select.variables =
c("av_ph","av_w_t","av_turb","av_do","av_cond","av_a_t","av_flow"))
p2 <- plot_taxa_env(env.taxa.cor_avg_2)
p2$data$Env <- revalue(p2$data$Env, c(av_a_t ="Air Temperature",
                              av_cond = "Conductivity",
                              av_do = "Dissolved Oxygen",
                              av_flow = "Flow Rate",
                              av_ph ="pH",
                              av_turb ="Turbidity",
                              av_w_t ="Water Temperature"))


print(p2)

library(cowplot)
p1 <- p1 + theme(legend.position = "none",strip.text.x = element_text(size=14, angle =90,
family="serif"),axis.text.x = element_text(size=16, family ="serif"), axis.text.y
=element_text(size=16, family="serif"))
legend_pp <- get_legend(p2)
p2 <- p2 + theme(legend.position = "none",strip.text.x = element_text(size=14, angle =90,
family="serif"), axis.text.x = element_text(size=16, family="serif"), axis.text.y
=element_text(size=16, family="serif"))
env_cor <- plot_grid(p1, p2, ncol =2,labels=c("A","B"), label_size = 30)
env_cor2 <- plot_grid(env_cor, legend_pp, rel_widths = c(7,1),  label_size =16)
env_cor2


ggsave("corr.tiff", device="tiff", width=20, height=30, units="in", dpi=600)

## PCoA for pathogen presence
# Bacteria
set.seed(420)
PCoA_total_16s = ordinate(phyloseq1_family, "PCoA", "Unifrac", weighted = TRUE)
PCoA_plot_16s_sal = plot_ordination(phyloseq1_family, PCoA_total_16s,
color="pos_s",shape="Sampletype")
sal <- PCoA_plot_16s_sal + geom_point(size=3) + labs(x= "PC 1 [34.6%]", y= "PC 2
[14.9%]") + scale_color_manual(values= c("Red", "Blue")) +theme(panel.grid.major =
element_blank(), panel.grid.minor = element_blank())
PCoA_plot_16s_stec = plot_ordination(phyloseq1_family, PCoA_total_16s,
color="pos_stec",shape="Sampletype")
```

```
stec <- PCoA_plot_16s_stec + geom_point(size=3) + labs(x= "PC 1 [34.6%]", y= "PC 2
[14.9%]") + scale_color_manual(values= c("Red", "Blue")) +theme(panel.grid.major =
element_blank(), panel.grid.minor = element_blank())
PCoA_plot_16s_lm = plot_ordination(phyloseq1_family, PCoA_total_16s,
color="lm",shape="Sampletype")
lm <- PCoA_plot_16s_lm + geom_point(size=3) + labs(x= "PC 1 [34.6%]", y= "PC 2
[14.9%]") + scale_color_manual(values= c("Red", "Blue","Black"))
+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
PCoA_plot_16s_ls = plot_ordination(phyloseq1_family, PCoA_total_16s,
color="ls",shape="Sampletype")
ls <- PCoA_plot_16s_ls + geom_point(size=3) + labs(x= "PC 1 [34.6%]", y= "PC 2
[14.9%]") + scale_color_manual(values= c("Red", "Blue","Black"))
+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())

pp1 <- plot_grid(sal+ theme(legend.position="none")
        , stec+ theme(legend.position="none")
        , lm+ theme(legend.position="none")
        , ls+ theme(legend.position="none"), labels=c("A","B","C","D"))

lm_legend <- lm +
  scale_color_manual(name="Pathogen", breaks=c("0","1","MS Lost"), labels=c("Negative",
"Positive", "Unknown"),values= c("Red", "Blue","Black")) +
  scale_shape_discrete(name= "Sample type", labels=c("Sediment", "Water"))
legend2 <- get_legend(lm_legend)

pp2 <- plot_grid(pp1, legend2, rel_widths = c(8,1))
pp2

ggsave("figureS2.tiff", device="tiff", width=10, height=5, dpi=600, units="in" )
ggsave("figureS2.pdf", device="pdf", width=10, height=5, dpi=600, units="in")

# Fungi
set.seed(421)
PCoA_total_ITS = ordinate(phyloseq2_family, "PCoA", "Unifrac", weighted = TRUE)
PCoA_plot_ITS_sal = plot_ordination(phyloseq2_family, PCoA_total_ITS,
color="pos_s",shape="Sampletype")
sal2 <- PCoA_plot_ITS_sal + geom_point(size=3) + labs(x= "PC 1 [34.6%]", y= "PC 2
[14.9%]") + scale_color_manual(values= c("Red", "Blue")) +theme(panel.grid.major =
element_blank(), panel.grid.minor = element_blank())
PCoA_plot_ITS_stec = plot_ordination(phyloseq2_family, PCoA_total_ITS,
color="pos_stec",shape="Sampletype")
stec2 <- PCoA_plot_ITS_stec + geom_point(size=3) + labs(x= "PC 1 [34.6%]", y= "PC 2
[14.9%]") + scale_color_manual(values= c("Red", "Blue")) +theme(panel.grid.major =
element_blank(), panel.grid.minor = element_blank())
```

```
PCoA_plot_ITS_lm = plot_ordination(phyloseq2_family, PCoA_total_ITS,
color="lm",shape="Sampletype")
lm2 <- PCoA_plot_ITS_lm + geom_point(size=3) + labs(x= "PC 1 [34.6%]", y= "PC 2
[14.9%]") + scale_color_manual(values= c("Red", "Blue","Black"))
+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())
PCoA_plot_ITS_ls = plot_ordination(phyloseq2_family, PCoA_total_ITS,
color="ls",shape="Sampletype")
ls2 <- PCoA_plot_ITS_ls + geom_point(size=3) + labs(x= "PC 1 [34.6%]", y= "PC 2
[14.9%]") + scale_color_manual(values= c("Red", "Blue","Black"))
+theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank())

pf1 <- plot_grid(sal2+ theme(legend.position="none")
                 , stec2+ theme(legend.position="none")
                 , lm2+ theme(legend.position="none")
                 , ls2+ theme(legend.position="none"), labels=c("A","B","C","D"))
lm_legend <- lm2
lm_legend <- lm2 +
  scale_color_manual(name="Pathogen", breaks=c("0","1","MS Lost"), labels=c("Negative",
"Positive", "MS Lost"),values= c("Red", "Blue","Black")) +
  scale_shape_discrete(name= "Sample type")
legend2 <- get_legend(lm_legend)

pf2 <- plot_grid(pf1, legend2, rel_widths = c(8,1))
pf2


# Run PERMANOVA per pathogen presence
set.seed(114)
adonis_sal_Sediment1 <-
adonis(phyloseq::distance(Sediment1,method="unifrac",weighted=TRUE) ~ pos_s
       , data= as(sample_data(Sediment1), "data.frame"))
adonis_sal_Sediment2 <-
adonis(phyloseq::distance(Sediment2,method="unifrac",weighted=TRUE) ~ pos_s
                    , data= as(sample_data(Sediment2), "data.frame"))
adonis_sal_Water1 <-
adonis(phyloseq::distance(Water1,method="unifrac",weighted=TRUE) ~ pos_s
                    , data= as(sample_data(Water1), "data.frame"))
adonis_sal_Water2 <-
adonis(phyloseq::distance(Water2,method="unifrac",weighted=TRUE) ~ pos_s
                    , data= as(sample_data(Water2), "data.frame"))

adonis_sal_Sediment1
adonis_sal_Sediment2
adonis_sal_Water1
adonis_sal_Water2
```

```
set.seed(115)
adonis_stec_Sediment1 <-
adonis(phyloseq::distance(Sediment1,method="unifrac",weighted=TRUE) ~ pos_stec
                    , data= as(sample_data(Sediment1), "data.frame"))
adonis_stec_Sediment2 <-
adonis(phyloseq::distance(Sediment2,method="unifrac",weighted=TRUE) ~ pos_stec
                    , data= as(sample_data(Sediment2), "data.frame"))
adonis_stec_Water1 <-
adonis(phyloseq::distance(Water1,method="unifrac",weighted=TRUE) ~ pos_stec
                     , data= as(sample_data(Water1), "data.frame"))
adonis_stec_Water2 <-
adonis(phyloseq::distance(Water2,method="unifrac",weighted=TRUE) ~ pos_stec
                     , data= as(sample_data(Water2), "data.frame"))

adonis_stec_Sediment1
adonis_stec_Sediment2
adonis_stec_Water1
adonis_stec_Water2

set.seed(116)
adonis_lm_Sediment1 <-
adonis(phyloseq::distance(Sediment1,method="unifrac",weighted=TRUE) ~ lm
                    , data= as(sample_data(Sediment1), "data.frame"))
adonis_lm_Sediment2 <-
adonis(phyloseq::distance(Sediment2,method="unifrac",weighted=TRUE) ~ lm
                    , data= as(sample_data(Sediment2), "data.frame"))
adonis_lm_Water1 <-
adonis(phyloseq::distance(Water1,method="unifrac",weighted=TRUE) ~ lm
                     , data= as(sample_data(Water1), "data.frame"))
adonis_lm_Water2 <-
adonis(phyloseq::distance(Water2,method="unifrac",weighted=TRUE) ~ lm
                     , data= as(sample_data(Water2), "data.frame"))

adonis_lm_Sediment1
adonis_lm_Sediment2
adonis_lm_Water1
adonis_lm_Water2

set.seed(117)
adonis_ls_Sediment1 <-
adonis(phyloseq::distance(Sediment1,method="unifrac",weighted=TRUE) ~ ls
                    , data= as(sample_data(Sediment1), "data.frame"))
adonis_ls_Sediment2 <-
adonis(phyloseq::distance(Sediment2,method="unifrac",weighted=TRUE) ~ ls
                    , data= as(sample_data(Sediment2), "data.frame"))
```

```
adonis_ls_Water1 <- adonis(phyloseq::distance(Water1,method="unifrac",weighted=TRUE)
~ ls
                        , data= as(sample_data(Water1), "data.frame"))
adonis_ls_Water2 <- adonis(phyloseq::distance(Water2,method="unifrac",weighted=TRUE)
~ ls
                        , data= as(sample_data(Water2), "data.frame"))

adonis_ls_Sediment1
adonis_ls_Sediment2
adonis_ls_Water1
adonis_ls_Water2


# Conditional variable importance measurement based on AUC
library(pdp)
library(party)
library(caret)
library(extrafont)
library(ggplot2)
library(forcats)

# [function] Change OTUs name into family name for rf variable importance
otu_to_family <- function(sig_table, physeq, bacteria = TRUE){
  a= as.data.frame(sig_table$predictors)
  a <- as.vector(a[,1])
  a <- prune_taxa(a, physeq)
  if (bacteria == TRUE){sig_table$Family <- factor(tax_table(a)[,"Family"])}
  else {sig_table$Family <- factor(tax_table(a)[,"Rank5"])}
  sig_table$Family <- factor(sig_table$Family,
levels=sig_table$Family[order(sig_table[,2])])
  print(sig_table)
}

# Salmonella
set.seed(47)
data.controls <-cforest_unbiased(ntree=10001, mtry=33)

# Bacteria, sediment
model_sal1 <- as.data.frame(cbind(otu_table(Sediment1), sample_data(Sediment1)[,9]))
model_sal1$pos_s <- as.factor(model_sal1$pos_s)

rf_sal1 <- cforest(pos_s~.,data=model_sal1,controls=data.controls)
varimp_sal1 <- data.frame(varimpAUC(rf_sal1, conditional=TRUE, OOB=TRUE))
varimp_sal1
```

```
oob4 = predict(rf_lm1, OOB=T)
table(model_lm1$lm, oob4)

names(varimp_sal1)[names(varimp_sal1) == 'X'] <- 'Variable'
names(varimp_sal1)<- 'varimp'

# Find "informative" variables that are greater than the threshold of abs(min(varimpTRUE))
varimp_sal1_sig<-subset(varimp_sal1,
varimp_sal1$varimp>abs(min(varimp_sal1$varimp)))
varimp_sal1_sig$normimp<-varimp_sal1_sig$varimp/sum(varimp_sal1_sig$varimp)
varimp_sal1_sig <- data.frame(predictors = rownames(varimp_sal1_sig),
varimp_sal1_sig$normimp)

varimp_sal1_sig$predictors <- factor(varimp_sal1_sig$predictors,
                             levels =
varimp_sal1_sig$predictors[order(varimp_sal1_sig$varimp_sal1_sig.normimp)])
varimp_sal1_sig <- otu_to_family(varimp_sal1_sig, Sediment1, bacteria=TRUE)


sal1_plot<-ggplot(varimp_sal1_sig, aes(x=Family, y = varimp_sal1_sig.normimp)) +
  geom_bar(stat = "identity",color='black', fill='black',width=1.2) +
  coord_flip()+
  xlab("") +
  ylim(0,.352)+
  ylab("") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"))+
  theme(axis.text=element_text(size=14), axis.title=element_text(size=14),
        text=element_text(size=14, color="black"))
sal1_plot

# Bacteria, water
model_sal2 <- cbind(t(otu_table(Water1)), sample_data(Water1)[,9])
model_sal2 <- as.data.frame(model_sal2)
model_sal2$pos_s <- as.factor(model_sal2$pos_s)

rf_sal2 <- cforest(pos_s~.,data=model_sal2,controls=data.controls)
varimp_sal2 <- data.frame(varimpAUC(rf_sal2, conditional=TRUE, OOB=TRUE))
varimp_sal2

names(varimp_sal2)<- 'varimp'
# Find "informativarimp_sal2ve" variables that are greater than the threshold of
abs(min(varimpTRUE))
varimp_sal2_sig<-subset(varimp_sal2,
varimp_sal2$varimp>abs(min(varimp_sal2$varimp)))
```

```
varimp_sal2_sig$normimp<-varimp_sal2_sig$varimp/sum(varimp_sal2_sig$varimp)
varimp_sal2_sig <- data.frame(predictors = rownames(varimp_sal2_sig),
varimp_sal2_sig$normimp)

varimp_sal2_sig$predictors <- factor(varimp_sal2_sig$predictors,
                             levels =
varimp_sal2_sig$predictors[order(varimp_sal2_sig$varimp_sal2_sig.normimp)])
varimp_sal2_sig <- otu_to_family(varimp_sal2_sig, Water1, bacteria=TRUE)

sal2_plot<-ggplot(varimp_sal2_sig, aes(x=Family,y = varimp_sal2_sig.normimp)) +
  geom_bar(stat = "identity",color='black', fill='black',width=1.2) +
  coord_flip()+
  xlab("") +
  ylim(0,.352)+
  ylab("") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"))+
  theme(axis.text=element_text(size=14), axis.title=element_text(size=14),
        text=element_text(size=14, color="black"))
sal2_plot

#fungal, sediment
model_sal3 <- cbind(t(otu_table(Sediment2)), sample_data(Sediment2)[,9])
model_sal3 <- as.data.frame(model_sal3)
model_sal3$pos_s <- as.factor(model_sal3$pos_s)

rf_sal3 <- cforest(pos_s~.,data=model_sal3,controls=data.controls)
varimp_sal3 <- data.frame(varimpAUC(rf_sal3, conditional=TRUE, OOB=TRUE))
varimp_sal3

names(varimp_sal3)<- 'varimp'
# Find "informativarimp_sal3ve" variables that are greater than the threshold of
abs(min(varimpTRUE))
varimp_sal3_sig<-subset(varimp_sal3,
varimp_sal3$varimp>abs(min(varimp_sal3$varimp)))
varimp_sal3_sig$normimp<-varimp_sal3_sig$varimp/sum(varimp_sal3_sig$varimp)
varimp_sal3_sig <- data.frame(predictors = rownames(varimp_sal3_sig),
varimp_sal3_sig$normimp)

varimp_sal3_sig$predictors <- factor(varimp_sal3_sig$predictors,
                             levels =
varimp_sal3_sig$predictors[order(varimp_sal3_sig$varimp_sal3_sig.normimp)])
varimp_sal3_sig <- otu_to_family(varimp_sal3_sig, Sediment2, bacteria=FALSE)

sal3_plot<-ggplot(varimp_sal3_sig, aes(x=Family,y = varimp_sal3_sig.normimp)) +
```

```
  geom_bar(stat = "identity",color='black', fill='black',width=1.2) +
  coord_flip()+
  xlab("") +
  ylim(0,.352)+
  ylab("") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"))+
  theme(axis.text=element_text(size=14), axis.title=element_text(size=14),
        text=element_text(size=14, color="black"))
sal3_plot

# Fungi, sediment
model_sal4 <- cbind(t(otu_table(Water2)), sample_data(Water2)[,9])
model_sal4 <- as.data.frame(model_sal4)
model_sal4$pos_s <- as.factor(model_sal4$pos_s)

rf_sal4 <- cforest(pos_s~.,data=model_sal4,controls=data.controls)
varimp_sal4 <- data.frame(varimpAUC(rf_sal4, conditional=TRUE, OOB=TRUE))
varimp_sal4

names(varimp_sal4)<- 'varimp'
# Find "informativarimp_sal4ve" variables that are greater than the threshold of
abs(min(varimpTRUE))
varimp_sal4_sig<-subset(varimp_sal4,
varimp_sal4$varimp>abs(min(varimp_sal4$varimp)))
varimp_sal4_sig$normimp<-varimp_sal4_sig$varimp/sum(varimp_sal4_sig$varimp)
varimp_sal4_sig <- data.frame(predictors = rownames(varimp_sal4_sig),
varimp_sal4_sig$normimp)

varimp_sal4_sig$predictors <- factor(varimp_sal4_sig$predictors,
                             levels =
varimp_sal4_sig$predictors[order(varimp_sal4_sig$varimp_sal4_sig.normimp)])
varimp_sal4_sig <- otu_to_family(varimp_sal4_sig, Water2, bacteria=FALSE)

sal4_plot<-ggplot(varimp_sal4_sig, aes(x=Family,y = varimp_sal4_sig.normimp)) +
  geom_bar(stat = "identity",color='black', fill='black', width=1.2) +
  coord_flip()+
  xlab("") +
  ylim(0,.352)+
  ylab("") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"))+
  theme(axis.text=element_text(size=14), axis.title=element_text(size=14),
        text=element_text(size=14, color="black"))
sal4_plot
```

```
# L. monocytogenes
set.seed(49)

# Fungi, sediment
model_lm3 <- cbind(t(otu_table(Sediment2)), sample_data(Sediment2)[,12])
model_lm3 <- as.data.frame(model_lm3)
model_lm3$lm <- as.factor(model_lm3$lm)

rf_lm3 <- cforest(lm~.,data=model_lm3,controls=data.controls)
varimp_lm3 <- data.frame(varimpAUC(rf_lm3, conditional=TRUE, OOB=TRUE))
varimp_lm3

names(varimp_lm3)<- 'varimp'
# Find "informativarimp_lm3ve" variables that are greater than the threshold of
abs(min(varimpTRUE))
varimp_lm3_sig<-subset(varimp_lm3,
varimp_lm3$varimp>abs(min(varimp_lm3$varimp)))
varimp_lm3_sig$normimp<-varimp_lm3_sig$varimp/sum(varimp_lm3_sig$varimp)
varimp_lm3_sig <- data.frame(predictors = rownames(varimp_lm3_sig),
varimp_lm3_sig$normimp)

varimp_lm3_sig$predictors <- factor(varimp_lm3_sig$predictors,
                                    levels =
varimp_lm3_sig$predictors[order(varimp_lm3_sig$varimp_lm3_sig.normimp)])
varimp_lm3_sig <- otu_to_family(varimp_lm3_sig, Sediment2, bacteria=FALSE)

lm3_plot<-ggplot(varimp_lm3_sig, aes(x=Family,y = varimp_lm3_sig.normimp)) +
  geom_bar(stat = "identity",color='black', fill='black', width=1.2) +
  coord_flip()+
  xlab("") +
  ylim(0,0.6)+
  ylab("") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"))+
  theme(axis.text=element_text(size=14), axis.title=element_text(size=14),
        text=element_text(size=14, color="black"))
lm3_plot

install.packages("extrafont")
library(extrafont)
font_import()
loadfonts(device="win")
fonts()
```

```
rf_plots <- plot_grid(sal1_plot+ theme(axis.text.y = element_text(size= 12),axis.text.x =
element_text(size = 13)+theme(text=element_text(family="Times New Roman",
face="bold", size=13))),
        sal2_plot+ theme(axis.text.y = element_text(size= 12),axis.text.x = element_text(size
= 13)+theme(text=element_text(family="Times New Roman", face="bold", size=13))),
        sal3_plot+ theme(axis.text.y = element_text(size= 12),axis.text.x = element_text(size
= 13)+theme(text=element_text(family="Times New Roman", face="bold", size=13))),
        sal4_plot+ theme(axis.text.y = element_text(size= 12),axis.text.x = element_text(size
= 13)+theme(text=element_text(family="Times New Roman", face="bold", size=13))),
        lm3_plot+ theme(axis.text.y = element_text(size= 12),axis.text.x = element_text(size
= 13)+theme(text=element_text(family="Times New Roman", face="bold", size=13))),
        nrow=5, labels = c("A","B","C","D","E"),label_size = 20, align="v", rel_heights =
c(2,1,1.5,1.5,0.75))


rf_plots

ggsave("figure6.pdf", plot=rf_plots, width = 10, height =10, units="in", dpi=600)
ggsave("figure6.tiff",device="tiff" ,width = 10, height =10, units="in", dpi=600)
=10, units="in", dpi=600)
```