METAGENOMICS APPROACHES FOR IMPROVED WATER QUALITY MONITORING, MICROBIAL SOURCE TRACKING, AND PUBLIC HEALTH RISK ASSESSMENT

A Dissertation
Presented to
The Academic Faculty

by

Brittany Suttner

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
December 2020

METAGENOMICS APPROACHES FOR IMPROVED WATER QUALITY MONITORING, MICROBIAL SOURCE TRACKING, AND PUBLIC HEALTH RISK ASSESSMENT

Approved by:

Dr. Konstantinos T. Konstantinidis, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Thomas DiChristina
School of Biological Sciences
Georgia Institute of Technology

Dr. Spyros G. Pavlostathis
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Amy E. Kirby
Waterborne Disease Prevention Branch
Centers for Disease Control and Prevention

Dr. Joseph M. Brown
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: November 24, 2020

To my grandpa

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Kostas Konstantinidis for his guidance and support through the last five years. I'm deeply grateful for the opportunity to work in his lab. He always encourages his students to strive for excellence and I think that has made me more successful as a PhD student. I don't think I could have made it this far in my academic career without his mentorship, kindness, patience, and empathy. I would also like to thank Dr. Joe Brown for his collaboration on various projects I worked on during the course of my PhD, as it was instrumental in shaping my thesis. I also would like to express my gratitude to my committee members Dr. Spyros Pavlostathis and Dr. Thomas DiChristina for their valuable input and feedback. I especially enjoyed the coursework I took under them. I must also give a huge thank you to those who helped me with the labor intensive mesocosm projects: Minjae Kim, Kevin Zhu, and Blake Lindner. I am further and eternally grateful for Minjae and Blake driving to Athens with me to collect cow and pig poop (on multiple occasions). Eric, Coto and Miguel were of immense help when I was completely new to the CLI. Additionally, I would like to thank all the members of the Kostas lab (especially Minjae, Smruthie, Carlos, Roth, Blake, and Angela) who have been there to help me with various bioinformatics challenges along the way. Many thanks also to Dr. Janet Hatt whose expertise was invaluable for my wet-lab work. It was a pleasure and an honor to work with such remarkable and talented people. I'm forever grateful for all the friends I made along the way and for the Zouk Atlanta dance community for helping me stay active and de-stress after a long day in the lab. I would be remiss if I did not thank Dr. Emily Latch and the COMPASS scholarship program at UWM. Being the first person

in my family to go to college, I had no idea how to navigate the grad school application process; Dr. Latch was very helpful through all of this. I am also grateful for my husband Flavio, who was my pillar of support through this journey (especially during the rough times when I felt like giving up). I must also thank my grandfather, Carroll, who taught me so much and got me first interested in science! His passing last year was devastating for me and my family and I regret that he can't be here to see this, but I know how proud he would be if he were. Finally, and most of all I would like to thank my father, Richard. I know it wasn't easy for him to raise four kids on his own, so I appreciate how hard he worked to provide for me and my brothers. His love and support has made this journey so meaningful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

MST	Microbial Source Tracking
FIB	Fecal Indicator Bacteria
VBNC	Viable But Not Culturable
qPCR	Quantitative Polymerase Chain Reaction
MAG	Metagenome Assembled Genome
NGS	Next-Generation Sequencing
OTU	Operational Taxonomic Unit
ARG	Antibiotic Resistance Gene
TAD	Truncated Average sequencing Depth
QMRA	Quantitative Microbial Risk Assessment
LOD	Limit of Detection
ANI	Average Nucleotide Identity
AAI	Average Amino acid Identity
DEG	Differentially Expressed Gene

SUMMARY

Fecal contamination in waters is a primary source of waterborne pathogens and is one of the most common impairments of water quality, affecting over a billion people worldwide. The impact and burden of fecal-contaminated waters have had significant public health and economic reach and is the main cause of death in children under five. Because it is not practical to directly monitor the numerous pathogens that cause waterborne diseases, water quality and public health risk are assessed using fecal indicator bacteria (FIB) as proxies for pathogens. However, there are many known limitations associated with current FIB-based methods that lead to inaccurate water quality and risk assessments. For example, the commensal enteric bacteria, *Enterococcus faecalis* is considered the "gold standard" FIB for water quality monitoring and previous epidemiological studies show that its presence correlates well with gastrointestinal illness cases at recreational beaches. However, decades of research show that "naturalized" populations of this organism are found in extraenteric environments and can enter a viable but non culturable (VBNC) state when under survival stress. The extent that these naturalized and/or VBNC populations confound water quality testing is unclear. Motivated by these uncertainties, in **chapter 2** we investigate how environmental stress affects survival and metabolic response in enteric versus environmental isolates of *E. faecalis* in laboratory mesocosms simulating an oligotrophic freshwater habitat. For this, we developed a 16S rRNA/rDNA viability assay for *E. faecalis* in order to elucidate how this organism regulates rRNA levels under environmental stress. We also describe how

currently used methods to enumerate *E. faecalis* (i.e., quantitative polymerase chain reaction [qPCR] and culturing) are confounded by the VBNC state.

Because *E. faecalis* is found in the guts of most animals, it is not useful for distinguishing between different hosts and thus, identifying the source of fecal pollution such as municipal sewage, livestock, wildlife or pets. Accordingly, recent efforts have focused on finding new host-specific FIB as targets for more robust qPCR assays. Among these microbial source tracking (MST) targets are members of the *Bacteroides* genus or their bacteriophages. Some 16S rRNA gene amplicon-based surveys have shown this genus is host-specific and there are many published qPCR assays targeting various hosts (humans, ruminants, gulls, dogs, etc.). However, most of these assays are not sufficiently host-specific or -sensitive (e.g., not all members of a host type carry the marker) and there is little epidemiological data linking them to waterborne disease risk. One of the guiding hypotheses of this thesis is that whole genomes and/or functional genes related to host-microbe interactions may provide more robust markers with higher resolution than the current MST assays targeting the 16S rRNA gene. To this end, in **chapter 3** we use shotgun metagenomic sequencing to compare the decay kinetics of fecal microbes from three different hosts (cow, pigs, and humans) in freshwater mesocosms simulating a pollution event. We identified several host-specific metagenome assembled genomes (MAGs) and functional genes as putative targets for more robust water quality monitoring and demonstrated the advantages of metagenomic methods over traditional qPCR and culture-based tests. Notably, the identified MAGs differ from the most commonly used FIB both taxonomically and functionally.

An important aspect when developing new MST markers and methodologies is to test them *in situ* with data from natural environments. Thus, in **chapter 4** we used time series metagenomics and newly established bioinformatics pipelines to determine the potential effects of cattle ranching in agricultural creek sediments. Our results revealed that these sediment communities are extremely diverse and robust against inputs from agricultural surface runoff and cattle ranching. In summary, this thesis critically assessed the advantages and limitations of meta-omics for MST biomarker discovery and public health risk assessment. Further, we provide evidence that FIB cannot be effectively distinguished from their naturalized counterparts based on our mesocosm incubations, an important limitation that is not applicable to the newly proposed MAGs. We also developed novel methods and bioinformatic protocols that should be useful for future studies to apply metagenomic techniques for MST.

CHAPTER 1. INTRODUCTION

1.1 The indicator paradigm and microbial source tracking (MST)

Fecal contamination is a primary source of pathogens that cause waterborne disease and is one of the most common impairments of water quality, affecting over a billion people worldwide. The impact and burden of fecal-contaminated waters have had significant public health and economic reach (Eisenberg, Bartram, and Wade 2016). Because it not practical to monitor the full spectrum of infectious agents associated with fecal contamination, water quality and public health risk are assessed using fecal indicator bacteria (FIB) as proxies for pathogens. Densities of commensal enteric organisms, *Enterococcus* spp. and *E. coli,* are considered the "gold standard" for water quality monitoring and are used worldwide (WHO 2003; USEPA 2012). However, since these organisms can be found in most vertebrate gastrointestinal tracts, they provide no information on the source of fecal contamination, such as combined sewer overflows, faulty septic systems, pets, wildlife, or farms (Field and Samadpour 2007). Contamination from human sewage is considered the most serious threat to human health because pathogenic viruses tend to be host specific. Yet, current regulations treat all fecal contamination as equally hazardous to human health (USEPA 2012). The pathogen source in the majority of drinking water and recreational water outbreaks reported to the CDC are unknown; however, roughly 18% of recreational water outbreaks from 1970-2000 and 14% of drinking water outbreaks from 1999-2004 were possibly animal-related (WHO 2004; USEPA 2009). From a public health perspective, more information is needed on the risk of exposure to animal-contaminated water as more recent studies suggest that risk from

exposure to water impacted by non-human feces of particular host types may be similar or equal to that of human feces (Soller et al. 2010; Probert, Miller, and Ledin 2017). Furthermore, FIB do not correlate well with enteric viruses, one of the dominant etiological agents of waterborne disease (Sinclair, Jones, and Gerba 2009). This is because enteric viruses react to waste water treatment in different ways and are not eliminated as successfully as bacteria (Carducci et al. 2009; Schmitz et al. 2016). Therefore, in recent years, many studies have been published trying to evaluate existing FIB tests and/or develop new host-specific markers for microbial source tracking (MST).

1.2 Limitations of culture-based monitoring: persistence and naturalized populations

Epidemiological data support a correlation between elevated FIB concentration in water and increased incidence of disease (Wade et al. 2003; Arnold et al. 2016). However, several studies have also shown FIB levels exceeding regulation standards do not always correlate to the presence of pathogens (Harwood et al. 2014). In addition to their lack of host specificity, these organisms can persist outside the host and even grow in the environment long after risk of disease from fecal pathogens has dissipated (M. N. Byappanahalli et al. 2012; J. Jang et al. 2017). Further, endemic populations have been found to occur naturally in the environment (heretofore referred to as environmental strains), which could possibly confound the use of these bacteria as FIB (Ishii and Sadowsky 2008; Luo et al. 2011; Weigand et al. 2014). These environmental strains are sometimes phylogenetically and phenotypically indistinguishable (**Figure 1-1**) from their enteric relatives based on current FIB testing methods so their recovery during a routine test may not indicate recent fecal contamination (that is, false positive signal) and may be

the cause for poor correlations of FIB and pathogens (Mote et al. 2012). Moreover, unique *E. coli* populations have shown differential survival/growth ability in freshwater mesocosms (Anderson, Whitlock, and Harwood 2005) and in temperate soils (Ishii et al. 2010). Comparative transcriptome analysis suggests that environmental *E. coli* isolates are better adapted to low-nutrient conditions than their enteric counterparts (Vital et al. 2015). The long-term survival and growth of FIB in the environment may also have important public health implications. Previous studies have reported multidrug-resistant (Dhanji et al. 2011; Walsh et al. 2011; Jeonghwan Jang et al. 2013) and potentially pathogenic (Muruleedhara N. Byappanahalli et al. 2015; Q. Zhang et al. 2016) *E. coli* strains are found, and even grow, in the natural environment. Similarly, recent whole genome comparisons of environmental and enteric isolates of *Enterococcus faecalis*, a commonly used FIB for MST, revealed distinct habitat-specific genetic signatures such as genes associated with metabolism of sugars that are often abundant in the gut, as well as antibiotic resistance and virulence genes, to be specific or highly enriched in the enteric genomes. In contrast, nickel and cobalt transport systems are overrepresented in the environmental genomes (**Figure 1-2**; Weigand 2014, Cesare 2014, He et al. 2018). These results suggest that the accessory gene content may contribute to differential survival and adaptation in different habitats and thus, it is possibly to distinguish between enteric and naturalized populations based on genomic means. However, this hypothesis has not been rigorously tested yet for this important model FIB. Hence, in **Chapter 2**, we explore this further as described below.

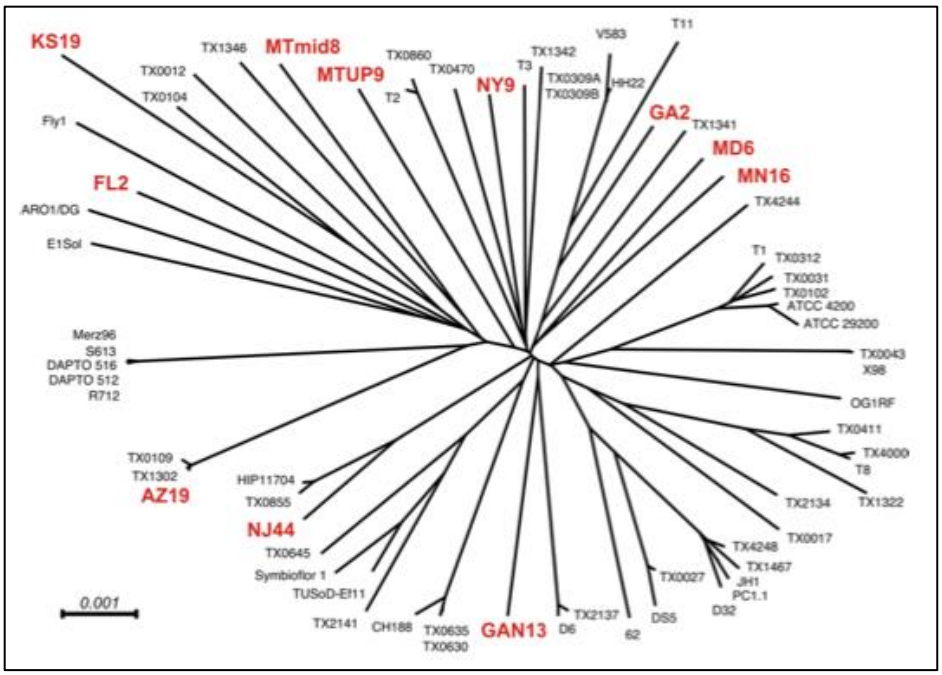


Figure 1-1: Core genome phylogeny of 70 *E. faecalis* genomes. The tree is based on nucleotide alignment of 1000 shared orthologous genes. The 11 environmental genomes (highlighted in red) are deep-branching and dispersed throughout the tree indicating that these isolates are phylogenetically indistinguishable from enteric strains. Adapted from Weigand et al. 2012.

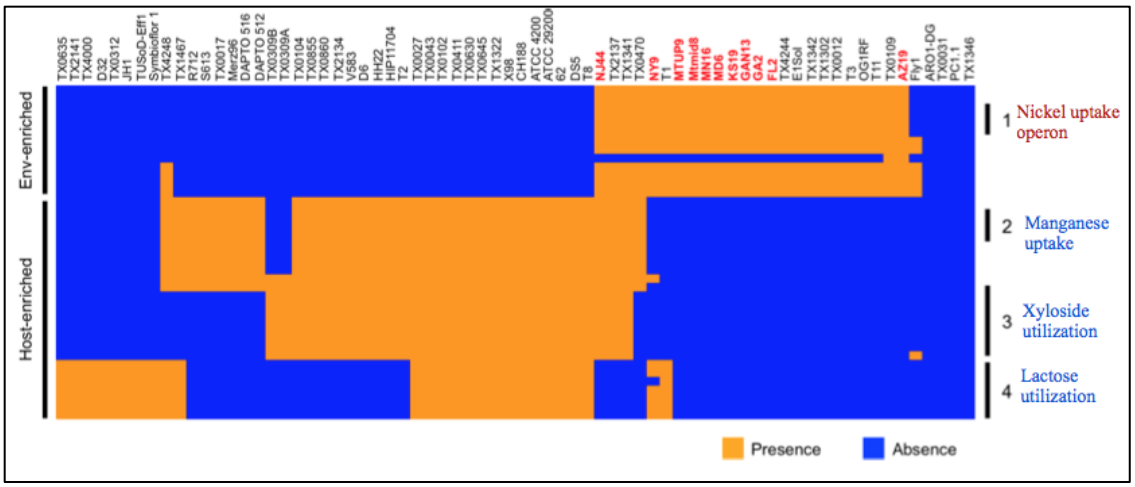


Figure 1-2: Gene signatures over-represented among enteric or environmental genomes. Each column is an *E. faecalis* isolate genome with red and black labels indicating environmental or enteric isolates, respectively. Each row represents a gene that was differentially enriched between the two isolate types. Although the core gene phylogeny is conserved (Fig. 1-1), there is evidence that the accessory gene content may encode habitat-specific functions such as sugar utilization in the gut (i.e., xyloside and lactose) and resource scavenging in the environment (i.e. nickel uptake), which can contribute to differential survival in the different habitats. Adapted from Weigand et al. 2012.

Additionally, *E. faecalis* is known to enter a viable but non-culturable (VBNC) state as a survival response to environmental stress such as those found in oligotrophic aquatic habitats (e.g. low temps, light, nutrient limitation), which can also lead to inaccurate assessments of water quality (e.g., false negative signal). VBNC cells are viable in that they preserve membrane integrity and low levels of gene expression, but they typically do not form colonies using traditional culture-based methods and can be resuscitated upon

return to favorable conditions (del Mar Lleò, Tafi, and Canepari 1998). The protein expression and membrane changes that occur when *E. faecalis* enters the VBNC state have been well characterized in the laboratory (Heim et al. 2002). However, the importance of this state for improved water quality monitoring has been largely unexplored.

To test the significance of the results by Weigand and colleagues (2014) and others (Luo et al. 2011, He et al. 2018) for distinguishing enteric *E. faecalis* isolates from their environmental counterparts, we incubated 9 enteric and 9 environmental isolates whose genomes carried the characteristic gene signatures identified in **Figure 1-2** (e.g. the nickel uptake operon *nik(MN)QO*) in sterilized lake water mesocosms. We found these oligotrophic growth conditions induced the VBNC state invariably for both isolate types and there was no clear difference in survival using traditional culture and qPCR-based tests (**Figure 1-3**). However, qPCR did not distinguish between live, dead, or VBNC cells. Therefore, in **Chapter 2**, we used a viability assay based on the ratio of the transcript (rRNA) vs. DNA (rDNA) copy number of the 16S rRNA gene to better detect physiological differences between the two groups of isolates. The rRNA/rDNA ratio (mostly of the small submit ribosomal RNA gene or 16S rRNA) has been used to detect growing and/or metabolically active cells in the environment (Kemp et al. 1993; Kerkhof and Ward 1993; Muttray and Mohn 2001; Kamke et al. 2010). Since cell death is a spectrum that can occur before cell lysis (**Figure 1-4**), looking at levels of rRNA is a more accurate assessment of the state of cellular activity than techniques based on membrane permeability (e.g. PMA-qPCR, live-dead staining microscopy). The idea behind using rRNA/rDNA ratio is that VBNC cells are still expressing genes, and thus synthesizing ribosomes, so they should have higher levels of ribosomal RNA relative to DNA copies of the ribosomal genes

compared to a population that contains mostly dead cells. Although both enteric and environmental isolates entered the VBNC state, the relative activity and regulation of rRNA levels may be different between the different isolates and reflect adaptations to the different habitats. Our work described in **Chapter 2** directly tested this hypothesis.

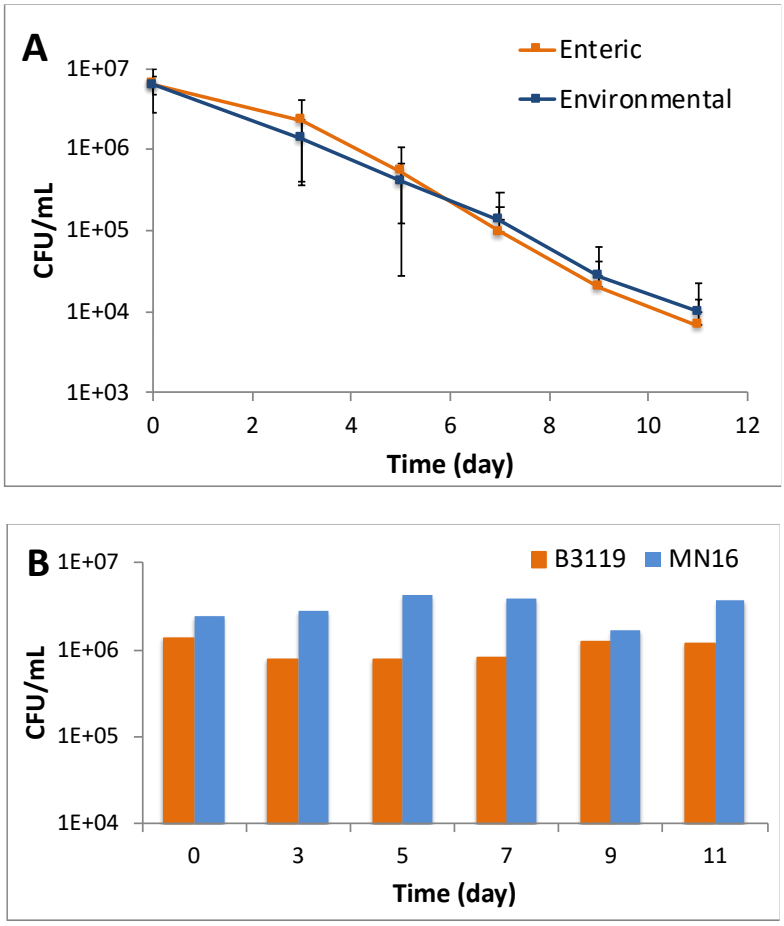


Figure 1-3: Monitoring decay in sterilized lake water glass bottle mesocosms with (A) culturing and (B) qPCR with an E. faecalis 16S assay. Panel A shows the average viable cell counts decreases over time while in panel B, the qPCR-based cell counts are maintained at the starting concentration throughout the duration of the experiment. Only one enteric (B3119) and one environmental (MN16) isolate is shown in part B for demonstration purposes, the trend was consistent among all 18 isolates evaluated. Hence,

in subsequent experiments (Chapter 2) we used an rRNA/rDNA assay to detect differences between isolates with more precision (Suttner et al., in preparation).

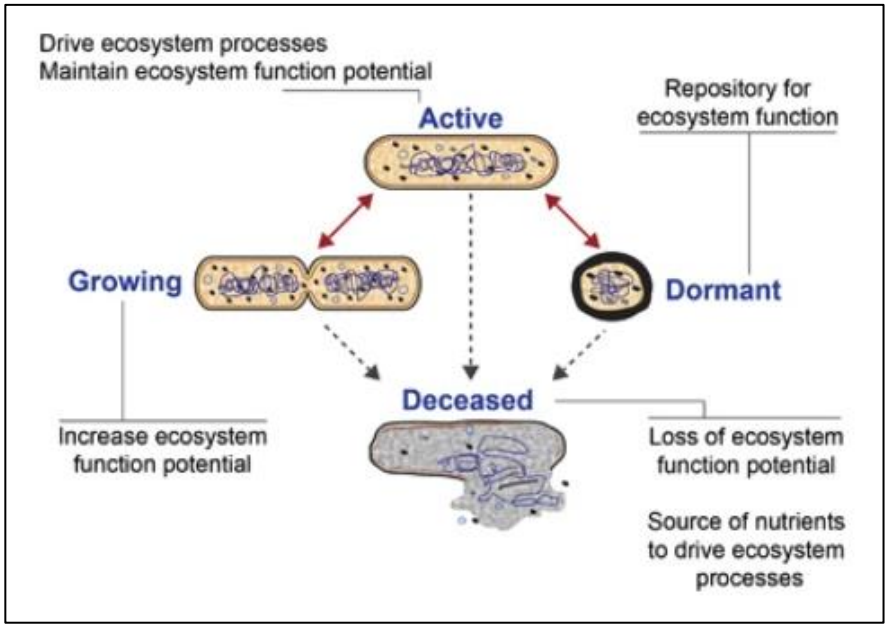


Figure 1-4: Metabolic states relevant for relative microbial activity assessment. Viable microorganisms exist in one of three general metabolic states that are all subject to mortality. Definitions of terms: *Growing*—cells are actively dividing, *Active*—cells are measurably metabolizing (catabolic and/or anabolic processes) but are not necessarily dividing, *Dormant*—cells are not measurably dividing or metabolizing, *Deceased*—cells are not metabolically active or capable of becoming metabolically active in the future, but intact macromolecules may persist. Viable but non culturable (VBNC) *E. faecalis* are likely somewhere between the dormant and fully active states. Adapted from Blazewicz et al. 2013.

1.3 Current culture-independent MST methods and limitations of the "on marker one assay" design of qPCR tests

Culture-based testing is problematic for timely management decisions because results typically take more than 24 hours to obtain. Recent efforts have focused on rapid culture-independent methods such as qPCR and looking for more robust markers of fecal contamination (Bernhard and Field 2000, Bernhard et al. 2003; Seurinck et al. 2005; Kildare et al. 2007; Field and Samadpour 2007; Bae and Wuertz 2009; Converse et al. 2009; Mieszkin et al. 2009, Liu et al. 2015, Fisher et al. 2015, (Stachler et al. 2017), García-Aljaro et al. 2017; Cinek et al. 2018; Liang et al. 2018). In addition to traditional FIB, fecal anaerobes have emerged as targets for new alternative markers, which were not as amenable to culture-based methods compared to traditional FIB like *E. coli* and *E. faecalis*. However, recent advances in genomic techniques during the last two decades have allowed the development of molecular assays bypassing the need to isolate these organisms in culture (Haugland et al. 2010). Several decades of research show that the genus *Bacteroides* tends to co-evolve with the host and are particularly suitable for MST because they are the most abundant genera in stool, have a narrow host range exclusive to warm-blooded mammals, and presumably have short-term survival rates in water because they are strict anaerobes (Ahmed, Hughes, and Harwood 2016). Hence, their presence in the aerobic aquatic environment should be indicative of recent fecal inputs. However, there is evidence suggesting *Bacteroides* can persist and even grow under some environmental conditions (Green et al. 2011; Weidhaas et al. 2015). More recently, CrAssphage, a DNA bacteriophage named after the metagenomic data mining technique used to discover it (Dutilh et al. 2014), has emerged as a promising new MST biomarker because it is one of the most abundant phages in some human gut populations (Stachler and Bibby 2014). Further, CrAssphage is thought to have *Bacteroides* as its preferred host. There are now

several published qPCR assays targeting CrAssphage for MST (Stachler et al. 2017, García-Aljaro et al. 2017; Cinek et al. 2018; Liang et al. 2018) though most of these assays have not been tested in environmental systems yet. Several recent studies report that these markers have high concentrations in sewage and sewage-impacted waters but have some cross-reactivity with other non-human hosts (Ahmed et al. March 2018; Stachler et al. 2018; Ahmed et al. Aug 2018) and may be too abundant in sewage for monitoring highly polluted waters (Stachler et al. 2018). Clearly, more research is needed on CrAssphage concentration and persistence in the environment, including how well these phages correlate to risk of infection with enteric pathogens. Nevertheless, the end goal when identifying new MST biomarkers, regardless of how advanced or sophisticated the method used to discover them may be, is usually to design a single qPCR assay for routine water quality monitoring. This framework has many known limitations as described below.

Many studies assessed MST qPCR marker performance under various experimental or environmental conditions and sampling techniques (Anderson et al. 2005; Bae and Wuertz 2009; Chern et al. 2009; Haugland et al. 2010; Dick et al. 2010; Green et a. 2011; Bae and Wuertz 2015; Liu et al. 2015; Li et al. 2016; Zhang et al. 2016; Cloutier and McLellan 2017; Mantha et al. 2017; Mattioli et al. 2017; Rothenheber an Jones 2018; Korajkic et al. 2018; Ballesté et al. 2018). However, methodological issues pose a substantial challenge for the field application of qPCR assays for regulatory purposes, as there are no standardized methods for sample collection, processing, DNA extraction, etc. Further, DNA markers are detectable by qPCR long after the living FIB (and presumably pathogens) have been inactivated (Bae and Wuertz 2009). Robust qPCR assays require small amplicon sizes (~100-200bp) so even highly degraded DNA from dead cells can be

detected by this method. Furthermore, little has been documented on the recovery efficiency of these markers from complex environmental matrices and the effect of PCR inhibitors carried over from the extraction, which makes it challenging to reliably compare the performance of different assays (Ahmed et al. 2016). In an attempt to address these uncertainties, a recent study involving 27 labs evaluated the performance of 41 MST markers and found none to be perfectly sensitive or specific for their target host. However, the human specific *Bacteroides* qPCR marker, HF183, was the best performing host specific marker overall (Boehm et al. 2013). Accordingly, the EPA has recently released Method 1696 targeting human specific fecal pollution with the HF183 qPCR assay (USEPA 2019). Although the HF183 assay is among the best performing qPCR assays available today, several studies have shown that it has poor correlation with pathogens and disease risk (Harwood et al. 2014), and has often low host sensitivity (e.g., not all humans carry the marker; **Chapter 3**). For example, a geographically expansive study evaluated human and ruminant specific markers in 16 different countries and 6 continents. While ruminant specific markers were globally suitable, the human associated markers (including HF183) were less prevalent and stable in some regions of the world (Reischer et al. 2013).

There is also a need for studies that examine persistence of MST markers over time in the environment and how they vary between different hosts. Fecal pollution of surface waters is often the result of a complex mixture of multiple inputs further complicated by environmental dispersion and deposition. Differential decay characteristics of DNA markers confound the interpretation of MST results for public health risk assessment and accurately attributing relative concentrations of different pollution sources (Cloutier and McLellan 2017). Although an absolute gene count can be obtained via qPCR, estimates of

the relative abundance of each marker and the relative contribution of various fecal sources in the natural environment cannot be quantitative without this decay information. More studies are also needed on the geographical and temporal stability of non-human host gut microbiomes (i.e. if host-associated bacteria show substantial degree of biogeography and do not apply well globally or across large geographic distances). The use of FIB to assess water quality has undoubtedly helped to reduce human health risk; however, the currently used approaches are not ideal in several aspects, as discussed above. In **Chapter 3** we tracked the decay of cow, pig, and human fecal communities in a freshwater lake over time using dialysis bag mesocosms simulating a pollution event coupled with shotgun metagenomics in order to discover new, host-specific biomarkers and evaluate their decay rates during the mesocosm incubation time. We also compared the metagenomic results against traditional FIB and MST methods (i.e., culturing and qPCR). Consistent with the previous findings described above, the human-specific HF183 assay was not detectable in two of the three human samples used in the mesocosms and the ruminant-specific assay had perfect host sensitivity among the six cows included in our study.

1.4 Application of Next-Generation Sequencing (NGS) technologies for MST

1.4.1 A brief overview of NGS and metagenomics

Next-generation sequencing (NGS) technologies have revolutionized microbial ecology research by bypassing the need to isolate microbes in pure culture and allowing us to study the "uncultivable majority" directly from an environmental sample. The most common NGS technologies are based on high-throughput sequencing of short DNA fragments (150- 350bp) that are either PCR amplicons for targeted deep sequencing of a

specific gene (e.g., the 16S rRNA), or a random census of all DNA fragments present in a microbial community (i.e., shotgun metagenomics). While 16S rRNA amplicon sequencing is useful for characterizing the taxonomic composition (Woese and Fox 1977), it does not provide any information on the functional potential of microbial populations. Metagenomics, however, is able to shed light on both "who" is there (i.e. what taxa) and what metabolic functions they are doing (Handelsman et al. 2007) since both functional and 16S rRNA gene fragments are recovered as part of a metagenomic dataset. However, shotgun metagenomics have a higher materials and infrastructure cost than gene-amplicon sequencing and bioinformatics tools for analysis are not as standardized as for amplicons. Consequently, gene-amplicon-based datasets and studies have prevailed.

A critical first step in metagenomic data analysis is piecing the small DNA sequences (or "reads") back together into larger fragments with a process called assembly. Assembly algorithms work by aligning short DNA reads at overlapping regions to piece together a longer consensus sequence, or contig (Miller et al. 2010, Rodriguez-R and Konstantinidis 2014, Sczyrba et al. 2017). If the metagenomic assembly is of sufficient quality and sequencing depth, the contigs can be further grouped (or "binned") into putative microbial genome populations (akin to a jigsaw puzzle). The end result of this binning process is a collection of typically fragmented (i.e., not a single, closed circular genome) metagenome-assembled genomes (MAGs) that represent the distinct microbial genomes that were originally present in the environmental sample (Tyson et al. 2004). Recently, there has been an effort to produce longer read sequences based on the PacBio and Oxford Nanopore technologies (English et al. 2012, Amarasinghe et al. 2020) in order to merge contigs of the same organisms or population and thus, close the genome. However, the

application of these technologies to environmental samples remains challenging; mostly due to the requirement to obtain high quality and molecular weight DNA for sequencing (Quince et al. 2017). Hence, the majority of environmental studies is currently based on short read shotgun metagenomics or amplicon sequencing.

1.4.2 16S rRNA gene amplicon-based sequencing for MST

Most research efforts utilizing NGS technologies for MST thus far have focused on 16S rRNA gene amplicon sequencing (Unno et al. 2018). However, this gene is too conserved to distinguish between closely related, yet distinct species. Accordingly, the most common way that the method is applied, i.e., clustering of 16S rRNA gene sequences into Operational Taxonomic Units (or OTUs; a proxy for species) based on an identity threshold (e.g. 97%) can miss environmentally or ecologically relevant groups. A few studies have attempted to address this limitation using oligotyping, a novel computational method that classifies closely related sequences based on minimum entropy decomposition and can distinguish groups at the resolution of a single nucleotide, which would otherwise fall into a single OTU (Eren et al. 2013). A recent study surveying sewage influent from 71 cities in the U.S. found 27 oligotypes that were common to all samples and were highly abundant. Interestingly, the structure and distribution of the human-fecal community predicted whether samples were from lean or obese populations with 81-89% accuracy (Newton et al. 2015). This result was somewhat surprising considering individual human gut microbiomes are highly variable and do not have a conserved "core" taxa (Huttenhower et al. 2012). The potential for oligotyping to capture traits associated with human health and demographics has important implications for public health and MST. The presence of conserved oligotypes suggests that although sewage represents an amalgam of individual

gut microbiomes there might be some highly (but probably not universally) shared taxa among human individuals that could serve as a signature of municipal sewage and distinguish it from other sources of pollution. Fisher et al. 2015 expanded on these findings and used oligotyping to develop host-specific biomarkers using sewage as a proxy for humans. They identified 99 oligotypes that were specific to human sewage that were also found in sewage from Spain and Brazil (Fisher et al. 2015), which suggests these oligotypes have potential as global alternative indicators. However, it remains unclear whether these biomarkers will work well in areas that lack centralized sewer and sanitation systems. More recently, this technique has shown to lack the resolution power to differentiate between fecal sources with similar bacterial communities, such as sewage effluent and lake water (Brown et al. 2019). Another study found that oligotypes for the bloom-forming cyanobacterium, *Microcyctis*, did not correlate with toxin genes and could not be used for inferring toxic ecotypes (Berry et al. 2017). These results are consistent with our doubts that single nucleotide variations in the 16S rRNA gene (i.e., oligotyping) can reflect ecologically or phylogenetically significant host-associated microbial populations. Clearly, more information is needed about the gene repertoires that facilitate host associations and specializations and if they are stable across members of the same host type (e.g., human individuals).

1.4.3 Improving MST and public health risk assessments with metagenomics

Deciphering the functional basis for host-associated microbiome patterns requires moving beyond 16S rRNA gene amplicon data to a broader sequencing approach (i.e.,

shotgun metagenomics) to characterize any selective, adaptive, and evolutionary processes underlying host-specific bacterial signatures (i.e., ecotypes and genes) at the whole genome level. Microbial source tracking and water quality assessments could be significantly improved with metagenomics, yet this technique has seen little application in MST field to date (Sharma and Sharma 2020). In **Chapter 3**, we established metagenomic and bioinformatic techniques as tools for water quality monitoring and identify host-specific taxa and functional genes. MST marker development is often a compromise between sensitivity (detected in all members of the target host type) and specificity (not detected in any non-target hosts). Efforts to design new MST biomarkers can overcome imperfect specificity by targeting a marker that is significantly more abundant in the target host compared to any other cross-reactivity; because a marker that is more abundant (i.e. higher sensitivity) may be useful even if it is not 100% specific to the target host. This reasoning provides support for metagenomics for biomarker discovery because metagenomes are biased towards the most abundant sequences in the population (whereas 16S rRNA gene amplicon sequencing also captures more rare community members). Comparing metagenomes from different host gut microbiomes focuses on the most abundant (and presumably more important) members for the host gut community. Moreover, those that are most abundant will be the easiest to detect in environmental matrices. It is also likely that metagenomic methods can be combined with conventional MST methods to obtain more accurate measures of fecal pollution in watersheds, since qPCR has generally a lower limit of detection than metagenome shotgun sequencing (**Chapter 4**).

Metagenomics has clear advantages for identifying new host-specific biomarkers, but it can also be useful for water quality monitoring and source tracking in environmental

samples (e.g., to assess multiple species found in a sample, including pathogens, simultaneously). In **Chapter 4**, the metagenomic techniques developed as part of the mesocosm incubations (**Chapters 2** and **3**) were applied and validated with field samples. Specifically, Southwestern California is one of the most productive agricultural regions in the U.S. and is associated with many foodborne *E. coli* O157:H7 outbreaks; nearly half of the major produce outbreaks in the U.S. between 1995-2006 have been traced to spinach or lettuce grown in these areas (M. Cooley et al. 2007). Produce contamination can be caused by exposure to contaminated irrigation or flood water, deposition of feces by livestock, or in the field application of manure as fertilizer (Mantha et al. 2017). Not only a source of pathogens, animal feces from farms is an emerging public health issue because of the current antibiotic practices (WHO 2014). Antibiotics are regularly administered to livestock at prophylactic concentrations to prevent infection and food animal production is responsible for a significant proportion of total antibiotic use (Landers et al. 2012). This is known to contribute to prevalence of antibiotic resistance genes (ARGs) in the environment (Jechalke et al. 2013; Zhu et al. 2013; Karkman, Pärnänen, and Larsson 2018), which can spread rapidly to other taxa on mobile genetic elements, including human pathogens of clinical importance (Walsh et al. 2011). Surprisingly, there is very little regulation of antibiotic use in the livestock industry, even though these operations can be major contributors of fecal pollution and spreading of ARGs to the environment (Durso and Cook 2014; Berendonk et al. 2015). In **Chapter 4** we investigated whether the impact of cattle ranching can be detected in sediment communities using metagenomics and if community structure is correlated to pathogenic *E. coli* detected by traditional culture-based methods.

CHAPTER 2. TRANSCRIPTOMIC AND RRNA/RDNA SIGNATURES OF ENVIRONMENTAL VS. ENTERIC *ENTEROCOCCUS FAECALIS* ISOLATES UNDER OLIGOTROPHIC FRESHWATER CONDITIONS

Brittany Suttner, Minjae Kim, Eric R. Johnston, Luis H. Orellana, Carlos A. Ruiz-Perez, Luis M Rodriguez-R, Janet K. Hatt, Joe Brown, Konstantinos T. Konstantinidis

2.1 Abstract

Enterococcus faecalis is used worldwide as an indicator for fecal contamination in water but the efficacy of this organism for risk assessment has been brought into question by recent studies showing the existence of "naturalized" populations of *E. faecalis* in the extraenteric environment in a viable but not culturable (VBNC) state. The extent to which these naturalized or VBNC *E. faecalis* can confound water quality monitoring is unclear. Here, we compared the decay patterns of three *E. faecalis* isolates from both the natural environment (environmental strains) and the human gut (enteric strains) in laboratory mesocosms that simulated well an oligotrophic, aerobic freshwater environment in order to determine if strains isolated from different habitats would display different survival strategies and responses. For this, we applied both the traditional culture-based and qPCR tests as well as a new rRNA/rDNA viability assay and metatranscriptomics. Our results

showed similar decay rates between isolates from the two habitat types based on viable plate and qPCR counts, yet a distinct spike in the rRNA/rDNA viability assay was observed for enteric vs. environmental isolates between day 1 and day 3. Despite this significant result for the viability assay, there was no strong evidence of differential gene expression or habitat adaptation in the metatranscriptomes from the mesocosm RNA. Overall, our results indicated that enteric strains may exhibit a different physiological response upon introduction into a nutrient-limiting environment. However, this difference may not be substantial or consistent enough for integration in water quality monitoring.

2.2 Introduction

Enterococcus faecalis is used worldwide as fecal indicator bacteria based on the assumption that *E. faecalis* is found only in the intestinal systems of animal hosts and dies off quickly upon release to the natural environment. However, "naturalized" populations of *Enterococcus* spp. are known to exist in freshwater environments with no sign of recent fecal inputs (heretofore referred to as "environmental" strains) (Byappanahalli et al. 2012, Devane et al. 2020) These environmental strains are phenotypically and phylogenetically indistinguishable from their enteric relatives based on standard selective media so their recovery during a water quality test by conventional methods would be considered a positive indicator of fecal contamination (Mote et al. 2012, Weigand et al. 2014). Whole genome comparisons of environmental and enteric strains revealed distinct habitat-specific genetic signatures such as genes associated with metabolism of sugars, as well as antibiotic resistance and virulence genes to be specific or highly enriched in the enteric genomes while, nickel and cobalt transport systems are overrepresented in the environmental genomes (Weigand et al. 2014, Cesare et al. 2014, He et al. 2018). These results suggest

that the accessory gene content contributes to differential survival and adaptation in different habitats despite the genetic relatedness, measured by genome-aggregate average nucleotide identity (ANI) or another metric, being indistinguishable between enteric and environmental *E. faecalis* strains (Weigand et al. 2014, He et al. 2018). However, the practical application and use of these alternative gene markers to distinguish innocuous, naturally-occurring from enteric strains that indicate risk to public health have not been tested yet.

Furthermore, *E. faecalis* is known to enter a viable but non-culturable (VBNC) state as a survival response mechanism to environmental stressors, such as introduction into an extraenteric environment. VBNC cells are viable in that they preserve membrane integrity and low levels of gene expression, but typically do not form colonies using traditional culture-based methods and typically have distinct proteomic signatures compared to non-VBNC cells (del Mar Lleò et al. 2000; Signoretto et al. 2000; Heim et al. 2002). However, VBNC cells can be resuscitated and grow upon return to favorable conditions (del Mar Lleò et al. 1998, Desmarais et al. 2002) and thus, represent risk to public health. Accordingly, culture-based approaches can also lead to inaccurate assessments of health risks due to VBNC (false negatives) or natural reservoirs of enterococci (false positives). Elucidating the extent to which naturalized populations and/or VBNC state cells may confound water quality monitoring is therefore critical for robust public health risk assessment.

Several studies have used cellular ribosomal RNA levels, often expressed as the copy number ratio of 16S rRNA transcripts to 16S rRNA genes (i.e. rRNA/rDNA ratio), to detect active and/or growing microbes (Kemp et al. 1993; Kerkhof and Ward 1993;

Muttray et al. 2001; Kamke et al. 2010). This is based on the assumption that the levels of ribosomal RNA are much higher in actively growing and metabolizing cells relative to dormant or dying cells. Although this has been shown to be true in several bacterial genera, the relationship between rRNA/rDNA ratio and growth rate varies significantly between taxa and some studies have even reported an indirect relationship (e.g., higher rRNA levels observed during low growth states) between rRNA concentrations and growth rate (Flärdh et al. 1992; Worden and Binder 2003; Sukenik et al. 2012). Furthermore, more information is needed on the relationship between rRNA levels and non-growth activities (e.g. VBNC state). Since these ratios are taxa-specific, baseline data on rRNA/rDNA levels in *E. faecalis* during different stages of activity and decay are needed in order to determine if it can be used as a viability assay for water quality monitoring to distinguish environmentally adapted from enteric strains. Cell death is a spectrum that can occur before cell lysis; thus, looking at levels of rRNA is a more accurate assessment of the state of cellular activity than techniques based on membrane permeability (e.g. PMA-qPCR, live-dead staining microscopy). VBNC cells are still expressing genes, and thus synthesizing ribosomes, so they should have higher levels of ribosomal RNA relative to DNA copies of the ribosomal genes compared to a population that contains mostly dead cells. Furthermore, environmentally adapted *E. faecalis* strains (if such strains exist) have higher rRNA/rDNA ratios in surface water environment compared to enteric strains because the former strains are better able to survive environmental stressors like O_2, sunlight, and nutrient limitation. In contrast, enteric strains, if they are able to persist in that same environment, are expected to be in a lower activity state.

The guiding hypothesis of this study is that the strains associated with different habitats (i.e., enteric vs. environmental) have distinct genetic and/or physiological adaptations that cause differential survival in freshwater ecosystems, and this can be detected and quantified for more accurate public health risk assessment based on rRNA/rDNA gene copy number ratios and gene expression profiles. To test this hypothesis, we performed laboratory incubations that simulated well the natural freshwater environment and were spiked in with three environmental and three enteric isolates (in separate mesocosm) that were reported previously to be phylogenetically and phenotypically indistinguishable from one another (Weigand et al. 2014). The change in viable cell counts (i.e. plate counts) and rRNA/rDNA ratios were monitored over two weeks. Therefore, this study provided important baseline information on the regulation of rRNA levels in *E. faecalis* under different growth conditions and new insights into the use of rRNA for improved water quality monitoring.

2.3 Methods

2.3.1 Compare rRNA/rDNA ratios in enteric vs. environmental E. faecalis isolates in dialysis bag mesocosms

2.3.1.1 Dialysis bag mesocosm set-up:

Lake water was collected from Lake Lanier (Georgia, USA) in acid-washed 10 L carboys and transported immediately back to the lab for mesocosm set-up the following day. Lake water to be used for inoculating with the *E. faecalis* strains was first filtered through 0.2 um sterivex filters as described previously (Tsementzi et al. 2014) while the remaining, unfiltered water was used to fill 10-gallon aquarium tanks where the dialysis bags would

be suspended during the incubations, as described below. Frozen glycerol stocks of the *E. faecalis* isolates (Table 2-1) were streaked for single colonies onto tryptic soy agar (TSA) plates and grown overnight at 37 ^{0}C. A single colony from each isolate was then inoculated into four mL of tryptic soy broth (TSB) and incubated at 37 ^{0}C with shaking at 150 rpm for 14 hours. One mL from each overnight culture was washed once with phosphate buffered saline (PBS) before inoculating into filtered lake water to a final concentration of ~10^6 CFU/mL. The initial concentration for each overnight culture was also determined by plate counts on TSA. The dialysis bags (6-8 kDa molecular weight cutoff) were filled to a total volume of 110mL (~21 cm length of dialysis tube) and closed on both ends using polypropylene Spectra/Por clamps (Spectrum Laboratories). Enough dialysis bags were filled to sample each isolate in triplicate at four time points, plus four filtered lake water negative control bags. The dialysis bags were then transferred to 10-gallon aquarium tanks filled with unfiltered lake water and stored in environmentally controlled rooms at 22 ^{0}C in the dark. A small water pump was included in each tank for aeration and nutrient distribution. A small headspace of air was left in each bag when sealing with the clamps so that they could float freely in the tanks.

2.3.1.2 Mesocosm sampling:

Destructive sampling of the dialysis bags occurred at days 1, 3, 8 and 11 after the initial set up day and each time point included triplicate biological replicates per isolate and a single lake water negative control. Fifty mL from each dialysis bag were filtered onto 0.45 um polycarbonate membranes then transferred into 2mL screw-cap tubes that had been pre-filled with 0.8 mL Qiagen buffer RLT (with 1% beta-mercaptoethanol) and 100 mg of acid-washed 0.1 mm beads. Bead tubes were stored at -80 ^{0}C until ready for extraction.

Additionally, water from each bag was diluted serially 10-fold with PBS for culture-based enumeration on TSA and mEnterococcus agar. All dilutions yielding measurements within the acceptable range of quantification were averaged to estimate CFUs/mL of each isolate.

2.3.1.3 Total nucleic acid extraction:

The frozen filters were defrosted on ice before the cells were mechanically lysed using a BioSpec BeadBeater and four 1-minute intervals with icing in between to prevent the samples from excessive heating and to protect the integrity of the RNA. Nucleic acids were extracted from cell lysates using the Qiagen AllPrep DNA and RNA extraction kit following the manufacturer's protocol for animal tissue. Contaminating DNA was removed from RNA samples by digestion (1-2 times depending on the sample concentration) with the Ambion TURBO DNase kit and following the manufacturer's protocol. RNA integrity was assessed with an Agilent 2100 Bioanalyzer instrument and the Agilent RNA 6000 Pico kit.

2.3.1.4 Assessment of quality of the RNA and DNA extractions:

Elimination of DNA from RNA samples was confirmed by end-point PCR amplification with the same primers used for the *E. faecalis* specific 16S rRNA qPCR assay (Santo-Domingo et al. 2013). Two uL of undiluted RNA was used as template in 20uL PCR reactions with 0.5 uM primers, 200uM dNTPs, 0.025 units/uL TaKara TAQ polymerase and 1x TaKara PCR buffer. The thermocycling conditions are as follows: 1 minute at 95 ^{0}C then 30 cycles of 95 ^{0}C for 15 seconds and 61 ^{0}C for 30 seconds followed by 72 ^{0}C for 1 minute. The PCR products were visualized with gel electrophoresis and the absence of

any detectable bands in the gel indicated that there was no significant DNA contamination in the RNA samples.

The absence of PCR inhibitors in the RNA and DNA samples was confirmed by "poison control" reactions where undiluted RNA or DNA was used as template in end-point PCR reactions as above, except a known amount (~10^7 copies) of standard plasmid was spiked into the PCR reaction mix. The same PCR master mix and thermocycling conditions were used as above except the primers were targeting the nickel uptake gene in the standard plasmid (not published). The PCR products were run on a 1% agarose gel and the presence of a single band at the expected size of the PCR amplicon in the standard plasmid confirmed the absence of any PCR inhibitors.

2.3.1.5 Quantification of 16S rRNA and rDNA using reverse transcriptase quantitative PCR (RT-qPCR) and quantitative PCR (qPCR):

DNA and RNA concentrations were quantified using the Qubit High Sensitivity DNA and RNA kits (Thermo Fisher Scientific), respectively, and Qubit 2.0 flourometer. Template nucleic acids were then diluted to below 0.5ng/uL before amplification using an *E. faecalis* specific 16S rRNA gene assay (Santo-Domingo et al. 2003). The standard plasmid used for absolute quantification was an *E. faecalis* 16S rRNA gene ligated into a pCR™2.1-TOPO® TA vector and cloned using One Shot® Chemically Competent TOP10 *Escherichia coli* and the TOPO®-TA cloning kit (Invitrogen), following manufacturer's instructions. The standard plasmid was isolated using the QIAprep Spin Miniprep Kit (Qiagen) following the manufacturer's instructions and quantified using the Qubit HS DNA kit. Eight, 10-fold serial dilutions (10^8 to 10^1 copies per reaction) of qPCR standard

plasmids were run in triplicate on every 96-well plate. All reactions were performed on the Applied Biosystems 7500Fast machine and Bio-Rad Universal Probes reagents following the manufacturers protocol. Reactions were performed in triplicate in a total volume of 20uL that included 2uL of the template or standard plasmid and 250 nM primer and TaqMan (5' hydrolysis) probe (the RT-qPCR reactions also included 0.5uL of iScript reverse transcriptase). Thermocycle conditions for qPCR consisted of an initial 50 ^{0}C step for 2 minutes followed by 95 ^{0}C for 10 minutes, then 40 cycles of 95 ^{0}C for 15 seconds and 60 ^{0}C for 60 seconds. The RT-qPCR thermocycle conditions were the same except for the initial step of 50 ^{0}C for 10 minutes followed by 95 ^{0}C for 2 minutes. The calibration curve from each plate was used to calculate rRNA and rDNA copy numbers in each sample which were averaged among technical replicates, multiplied by elution volume (200 or 50 uL for DNA and RNA, respectively), and then divided by the filter volume (50mL) total copies per milliliter.

2.3.2 Metatranscriptome sequencing and analysis of total RNA from dialysis bag mesocosms

2.3.2.1 Metatranscriptome library preparation and sequencing:

The triplicate RNA extractions from each isolate at each of the four time points were pooled together in order to obtain enough high-quality RNA for metatranscriptomic sequencing, and cDNA libraries were prepared using the ScriptSeq v2 RNA-Seq Library Preparation kit (Illumina) following the manufacturer's instructions except a half ng (~1% of total library size) of a luciferase internal RNA standard was included during the RNA fragmentation (step 3A) for absolute quantification of transcript copy numbers as described

below. The quality and insert size of each cDNA library was inspected using the Agilent High Sensitivity DNA kit and Agilent 2100 Bioanalyzer instrument. Library concentrations were determined using the Qubit HS DNA kit and diluted to ~5 nM before loading into the flow cell and sequencing on the Illumina HiSEQ 2500 instrument as described previously (Johnston et al. 2019).

2.3.2.2 Luciferase internal RNA standard preparation:

The Promega pGEM®-luc plasmid vector (accession number X65316) containing a 1094 nucleotide fragment of the firefly luciferase gene was digested with SphI-HF restriction enzyme (New England Biosystems) at 37 ^{0}C for 1 hour followed by clean up with the Qiagen PCR clean-up kit. The digested DNA was gel-purified using 1.5% low melt agarose gel and the MO BIO UltraClean® 15 DNA purification kit followed by end repair with the Thermo Scientific Fast DNA End Repair kit and another clean up with the Qiagen PCR clean-up kit but with a 30 uL elution volume. The DNA was concentrated by ethanol precipitation before transcribing to RNA with the Promega Riboprobe® in vitro Transcription T7 System and following the manufacturers protocol 4.F for synthesis of large amounts of RNA. The RNA standard quantity and quality were determined using the Qubit HS RNA kit and Agilent Bioanalyzer as described above.

2.3.2.3 Transcriptome sequence analysis:

All transcriptomic reads were quality filtered and trimmed as described previously (Kim et al. 2018). Trimmed reads were filtered to remove rRNA sequences using SortMeRNA v2.1 (Kopylova et al. 2012) with all rRNA databases in the program and the following options: --blast 1 --num_alignments 1 -v -m 8336. The internal luciferase standard sequences were

identified by blastn search against the 1094 bp length nucleotide luciferase reference sequence carried on the pGEM®-luc plasmid vector (accession number X65316). Matches were filtered for best match using a threshold of 97% identity and alignment length that is 80% of the query read length, and resulting matches were subsequently removed from the transcriptomic datasets (for the reference luciferase matches). The number of internal standard sequences recovered were used to estimate the absolute number of mRNA transcripts in the sample and sample sequencing depth (actual number of mRNA reads in sequenced in metatranscriptome divided by the absolute number of mRNA transcripts in the sample) as described in Satinsky et al. 2013.

Reference genome assemblies for the four isolates that were used as inocula in the mesocosms (Table 2-1) were downloaded from NCBI. Prodigal v2.6.1 (Hyatt et al. 2010) was used to predict genes from the assemblies, which were then annotated against the Swiss-Prot (downloaded March 2019; UniProt Consortium 2017) using blastp (options: e_value 1E-6 and max_target_seqs 10). Matches to the reference Swiss-Prot sequences were filtered for best matching, using 40% identity and 40% query cover alignment length as threshold. All genes that had no match to the Swiss-Prot database were annotated against the TrEMBL database (downloaded May 2018; Uniprot Consortium 2017) using the same match filtering cut-off. Non-rRNA metatranscriptomic reads (i.e. after removing internal standard sequences) were mapped against predicted genes for the corresponding isolate that was used as inocula in that sample using MegaBLAST (Camacho et al. 2009) and matches with <97% identity and <50 bp alignment length were removed from further analysis. Read count tables against predicted genes were generated using custom scripts and were used as the input for DESeq2 (Anders and Huber 2010) along with the sample

sequencing depth as determined from the internal standard for the estimate size factors step. Differentially expressed genes (with P_{adj} <0.05) between enteric and environmental isolates were determined using the Likelihood Ratio Test as implemented in DESeq2.

2.3.3 rRNA/rDNA ratio in E. faecalis in pure culture under standard laboratory conditions

2.3.3.1 Batch culture growth conditions and sampling:

A frozen glycerol stock of *E. faecalis* strain MTUP9 (Table 2-1) was streaked for single colonies onto a TSA plate and grown overnight at 37 ^{0}C. A single colony was then inoculated into 4 mL of TSB and incubated at 37 ^{0}C with shaking at 150 rpm for 14 hours. 100uL of the overnight liquid culture was inoculated into 60mL fresh TSB in triplicate to start the growth curve experiment ($O.D._{600}$ < 0.1 at time 0) and incubated at 37 ^{0}C with shaking at 150rpm. Each triplicate culture was sampled at 14 time points over 73 hours to capture the different growth phases. At each sampling point, 1 mL of each triplicate culture was collected for $O.D._{600}$ reading, 0.1 mL was serially diluted 10-fold in PBS for plate counts on TSA, and 0.5-1 mL of the culture was collected for nucleic acid extraction by centrifuging at 10,000 rpm for 5 minutes and decanting the supernatant. Cell pellets were re-suspended in 600uL buffer RLT (Qiagen) with 1% beta-mercaptoethanol added and stored at -80 ^{0}C until ready for extraction. The re-suspended cell pellets were defrosted on ice and transferred to 2mL screw-cap tubes pre-filled with 100 mg of acid-washed 0.1 mm beads. Total nucleic acids were extracted and used for rRNA/rDNA analysis following the same protocol for filters as described above.

2.4 Results

2.4.1 rRNA/rDNA ratio of enteric vs. environmental E. faecalis isolates in dialysis bag mesocosms simulating an oligotrophic freshwater habitat

The ratio of the transcript (rRNA) vs. DNA (rDNA) copy number of 16S ribosomal RNA gene has been used as a viability assay to distinguish live, VBNC, or dead cell populations with more precision compared to traditional qPCR approaches (Poulsen et al. 1993, Campbell et al. 2011, Gaidos et al. 2011, Simister et al. 2012). We used the rRNA/rDNA ratio to compare the physiological response of human enteric versus environmental *E. faecalis* isolates in laboratory mesocosms simulating an oligotrophic freshwater environment. Oligotrophic growth conditions have been shown to induce the VBNC state invariably for both isolate types in our previous, unpublished pilot experiment using glass bottle mesocosms and traditional qPCR methods only (Figure 1-3). However, in our previous pilot experiment, we were not able to detect any difference using traditional qPCR only and instead present the results here for dialysis bag mesocosms and the rRNA/rDNA assay. Laboratory mesocosms consisted of 100mL dialysis bags filled with filtered lake water and inoculated with individual *E. faecalis* pure cultures (3 enteric and 3 environmental; Table2-1) to a final concentration of ~10^6 CFU/mL and suspended in aquarium tanks filled with unfiltered lake water. The dialysis bags have a pore size that allows passage of small molecules and ions but prevents passage of molecules larger than 6-8 kDa (e.g. bacteria and viral particles). These six isolates were selected because previous comparative genomic analysis showed that they contained the putative habitat-specific genes signatures identified by Weigand and colleagues (Weigand et al. 2014). Mesocosm sampling occurred in triplicate for each strain on days 1, 3, 8 and 11 (D1, D3, D8, and D11) and included: plating for viable cell counts and filtering for total nucleic acid extraction to

determine rRNA/rDNA ratios. Although the experiments started out with filter-sterilized lake water, some growth was observed in the negative control bags on TSA plates by D8. However, no growth was observed on the *Enterococcus*-specific media (data not shown). This result suggested that the integrity of the dialysis bags started to break down over time and that some of the microbes from the non-sterilized lake water in the tanks outside of the bags were able to pass through the dialysis membranes. Hence, we primarily focused our analysis and interrelations on the first three sampling points.

Table 2-1: *E. faecalis* isolates used in the dialysis bag mesocosm experiments. Total RNA from the mesocosm samples were also analyzed with metatranscriptomics for the isolates in bold. Isolation source describes whether the strains were isolated from the human gut (enteric) or an extra-enteric environment with no sign of recent fecal inputs (environmental) according to Weigand et al. 2014.

Isolate Name	**Isolation source**	**GenBank Accession**
MMH594	enteric	AJDZ01000001.1
ERV62	enteric	ALZQ01000001.1
TX0104	enteric	ACGL01000001.1
MTUP9	environmental	AYOJ01000001.1
MTmid8	environmental	AYKU01000001.1
AZ19	environmental	AYLU01000001.1

All strains exhibited a decrease in viable cell counts over time, as expected, and were detectable by culture for the duration of the experiment (i.e., until D11; Figure 2-1A). Moreover, decay rates based on plate counts were not significantly different between the two isolate types (paired T-test P=0.082), consistent with our previous pilot experiment (Figure 1-3). The average rRNA/rDNA ratios in the three environmental and one of the enteric isolates were relatively stable from D1 to D3 (0.7 to 1.5-fold change in ratio). However, two of the enteric strains (ERV62 and MMH594) had ~6-fold increase in ratio from D1 to D3 (Figure 2-1B). By D8 and D11, the average ratios decreased and approached zero consistently for all six isolates. When comparing average rRNA/rDNA ratios overall (i.e. enteric vs. environmental across all time points), the enteric isolates were not significantly different from environmental isolates (Wilcoxon Rank Sum P=0.149). When looking at the habitat types separately over time, the average rRNA/rDNA ratios between D1 and D3 were not significantly different for environmental strains but were significant for the enteric strains (paired Wilcoxian P= 1.0 and 0.014, respectively). This result might suggest that the enteric and environmental isolates show different gene expression responses to environmental stress (e.g. nutrient limitation), which we examined more fully with metatranscriptomics below.

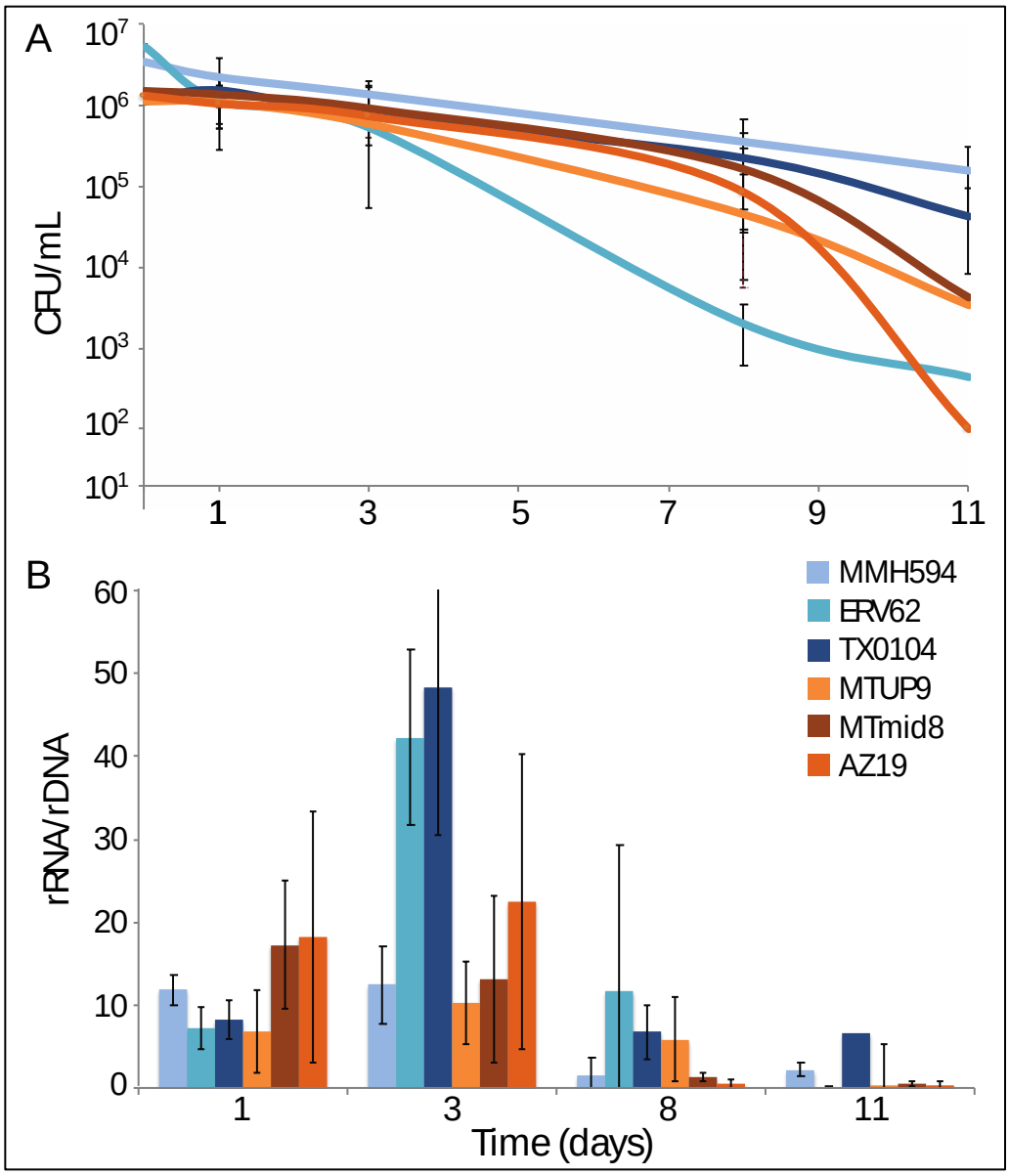


Figure 2-1: Comparing changes in (A) viable cell counts and (B) rRNA/rDNA ratios of enteric versus environmental E. faecalis isolates over time in dialysis bag mesocosms. Three enteric and three environmental isolates are represented by different shades of orange and blue, respectively. Error bars are standard deviation among three technical replicates.

2.4.2 Comparative metatranscriptomics of enteric and environmental isolates

The 16S rRNA/rDNA information alone does not provide information about specific gene functions and differences in mRNA expression levels that may serve as better biomarkers for the response to oligotrophic freshwater conditions. Thus, we used metatranscriptome sequencing profiles of the dialysis bag mesocosms to identify specific metabolic pathways that may underlie habitat adaptation and represent more reliable targets for improved FIB assays. We selected a subset of the mesocosm samples (two environmental and two enteric strains; Table 2-1) for total community RNA sequencing with an internal spiked control for absolute transcript quantification. Since the RNA extraction protocol used was designed for rRNA analysis and was not optimized for metatranscriptomic sequencing, we were not able to get enough mRNA for ribo-subtracted libraries. Thus, total RNA sequencing was used instead. The resulting metatranscriptome libraries had on average 3.2×10^7 ($\pm$ 9.7×10^6) reads per sample and ~95.7% of those reads were rRNA, on average. The internal RNA standard recovery in each metatranscriptome ranged from 0.02 to 0.13% of the original amount that was spiked in. The internal standard %recovery was used to estimate the absolute number of mRNA reads per ng RNA sequenced ($5.9 \times 10^7 \pm 3.5 \times 10^7$ on average in each sample) following the methods described by Satinsky and colleagues (Satinksy et al. 2013).

Reference genome sequences of the isolates were previously determined (Table 2-1) and were used for read mapping and to identify genes with significantly different expression between the two isolate types. Overall, there were no differentially expressed genes (DEGs) between enteric and environmental isolates across all time points. When controlling for the effect of time, there were only 31 strain specific DEGs observed between

D1 and D3, with 24 and 8 genes being more expressed in the environmental and enteric isolates, respectively (Figure 2-2). However, none of these genes were among the habitat-specific genes identified by the previous comparative genomic studies, such as the nickel uptake operon, *nik(MN)QO* (Weigand 2014, Cesare 2014, He et al. 2018). The DEGs found in the environmental isolates were mostly related to housekeeping genes such as ribosomal and transcription-related proteins (e.g. tRNA ligase and elongation factor T; Figure 2-2). While genes potentially related to cellular stress response, such as a putative transcription repressor (*niaR*) and DNA replication and repair gene (*recF*), had higher expression in the enteric isolates.

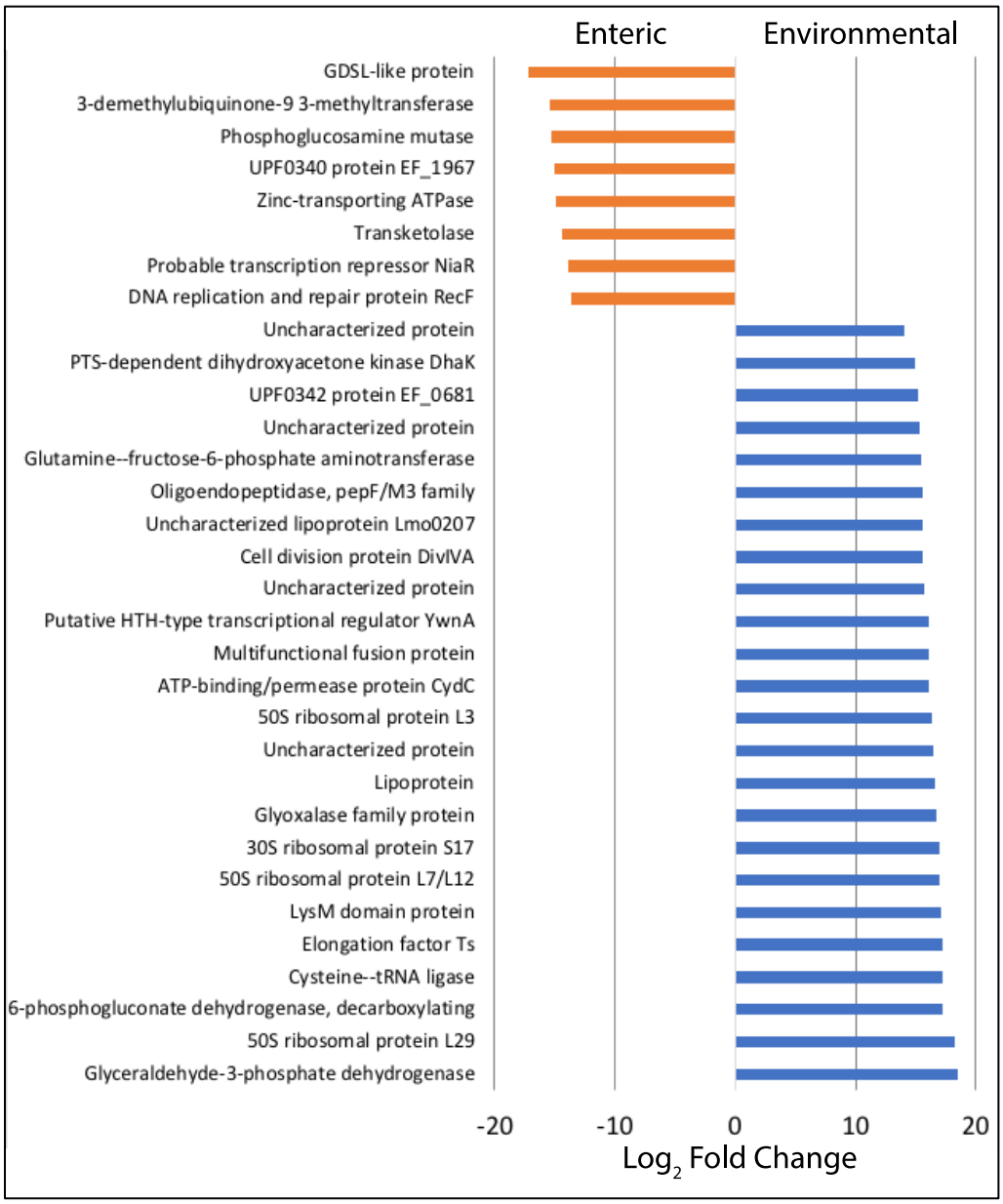


Figure 2-2: Differentially expressed genes between enteric and environmental *E. faecalis* isolates between days 1 and 3. Orange and blue bars indicate genes that were more expressed in enteric or environmental isolates, respectively. Functional gene annotation is based on the UniProt database and differentially expressed genes were identify by DeSeq2 analysis as described in section 2.3.2.3.

2.4.3 rRNA/rDNA ratios over the standard growth curve in pure culture

Since the relationship between rRNA/rDNA ratios and growth rate are taxa-specific (Blazewicz et al. 2013), we also collected baseline data on rRNA/rDNA levels in pure cultures of *E. faecalis* under standard laboratory conditions, which has not been examined previously for this species. For this, an *E. faecalis* strain (MTUP9) was grown in triplicate batch cultures and sampled over time to assess changes in the rRNA/rDNA ratio at the different growth and death phases (Figure 2-3). A typical bacterium growth curve was observed where the exponential growth phase lasted ~10 hours and maximum cell density ($1.5x10^9$ CFU/mL) was observed at 12.5 hours. Cell density remained relatively stable until the next measurement at 25 hours, where cell density was still around $1.1x10^9$ CFU/mL (Figure 2-3A). The rRNA/rDNA ratios ranged from 5.5 to 372, with the lowest ratios being observed during early exponential growth phase (i.e., during the first 5 hours), after which point the ratio started to increase but there was a high level of variation between biological replicates (Figure 2-3B). The highest levels of rRNA/rDNA ratios were observed in the early stationary phase (~hour 12; average ratio = 372), after which the ratios started to decrease during stationary and death phases.

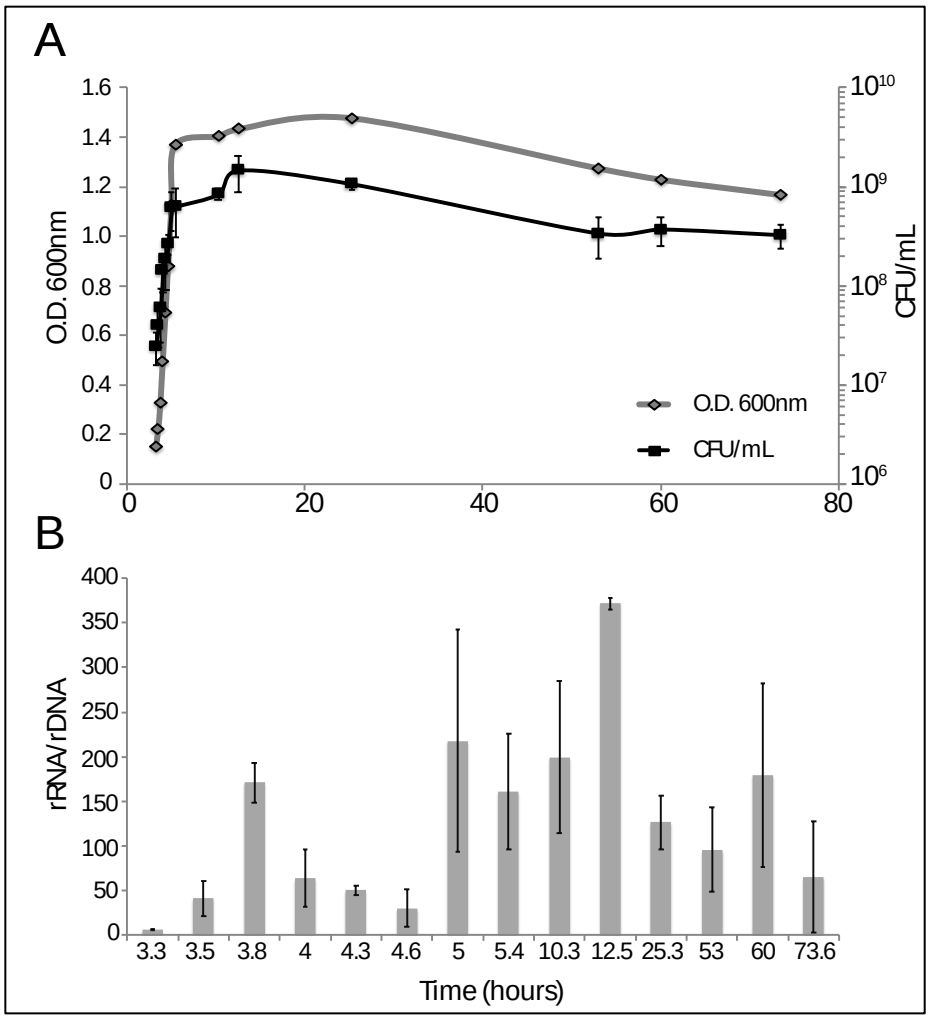


Figure 2-3: (A) Cellular abundance and (B) rRNA/rDNA ratios for E. faecalis MTUP9 in triplicate batch pure culture conditions. Error bars are standard deviation of biological and technical replicates.

2.5 Discussion

E. faecalis is still one of the most commonly used FIB for routine water quality monitoring despite several known limitations such as natural extraenteric populations and the ability to persist in the environment in a dormant, VBNC state (Byappanahalli et al. 2012). Several studies have assessed the use of 16S rRNA levels in the cell (normalized to

16S rDNA copy number) as a metric for distinguishing relative levels of cellular activity (Blazewicz et al. 2013); however, no study to date has investigated whether rRNA can be used in *E. faecalis* for improved environmental water quality monitoring. In this study, there was high variability in rRNA/rDNA ratios among biological replicates under oligotrophic mesocosm growth conditions (Figure 2-1B). Notably, the ratios under these conditions were, on average, roughly two orders of magnitude lower than those observed under standard lab conditions in pure culture (Figure 2-3B). These results are consistent with another study, which showed a high standard deviation in ratios and that copiotrophs have much lower ratios in oligotrophic systems relative to rich media (Lankiewicz et al. 2015). Notably, two of the three enteric isolates showed a six-fold increase in their rRNA/rDNA ratios from D1 to D3 (Figure 2-1B) and the ratios on D3 for these two isolates (~45 rRNA/rDNA) was similar to lower end of average values observed for *E. faecalis* in pure culture (e.g., during early exponential phase). However, there is no evidence to suggest that these isolates are actively growing or replicating in the mesocosms during this time based on the viable cell counts, and the incubation conditions are remarkably different. Hence the trends observed in the two experiments (i.e. lake water mesocosm vs. pure batch culture) are presumably the result of different biological factors.

A potential explanation for increasing rRNA/rDNA ratio coupled with decreasing cells observed in the oligotrophic mesocosm conditions is that the enteric isolates are increasing gene expression for pathways related to non-growth activities, such as environmental stress or cell homeostasis, that results in more ribosomes (and thus, more rRNA). Accumulating or maintaining high rRNA levels during periods of low activity may confer a competitive advantage upon return to favorable conditions, especially in copiotrophic environments

that favor fast growers that can respond quickly to nutrient stimuli (Roller et al. 2016). Enteric isolates maintaining high cellular rRNA levels through D3 could indicate an adaptive strategy for high nutrient environments like the gut, whereas the environmental isolates are not "evolutionarily primed" to expect high nutrient influxes and don't devote as much energy to maintain high rRNA levels.

Previous studies in other copiotrophs under balanced growth conditions in pure culture have shown that cellular rRNA concentration correlates well with growth rate (Neidhart and Magasanik 1960, Kerkhof and Ward 1993, Wagner 1994). As such, we expected to see the highest rRNA/rDNA ratios for *E. faecalis* during the exponential phase in pure culture. However, the highest ratios were observed around hour 12 when growth was beginning to reach stationary phase (Figure 2-3). The relationship between RNA levels and growth is not linear or consistent between different taxa, especially environmental oligotrophic bacteria (Binder and Liu 1998, Worden and Binder 2003). Therefore, this result is not necessarily surprising, but it does suggest that the regulation of cellular RNA levels in *E. faecalis* may be more complicated and not linearly correlated to growth. One possible explanation for the trends observed is that during early exponential phase, the cells are rapidly replicating their genomes and may have multiple genome copies per cell as a result of rolling replication, resulting in the observed lower rRNA/rDNA ratios. As nutrients in the batch culture start to become depleted and cell growth slows, there is a lag in the ribosome transcription feedback loop around hour 12 where ribosome concentration briefly exceeds cell demand for rapid growth and results in the observed higher ratio. The high amount of variation between biological replicates observed in both experiments also suggests that this rRNA/rDNA assay should be tested in more isolates in order to confirm

these trends and the amount of natural variability in rRNA/rDNA ratios between isolates, as well as provide more support for the putative explanations given above.

Although there was some difference between rRNA/rDNA ratios observed in the enteric and environmental isolates, our results are not conclusive with respect to whether or not this assay is suitable for distinguishing isolate types in water quality monitoring applications because the differences where not large enough and were strain-specific (as opposed to habitat-type-specific). However, it may be useful to provide information on the age of the pollution event. All six isolates had significantly higher rRNA/rDNA ratios on D1 and D4 compared to D8 and D11, with overall average ratios of 11.7, 24.8, 4.7, and 1.8, respectively. That is, the rRNA/rDNA ratio were substantially higher in the early stages, and this could serve as a sign of recent fecal pollution. Specifically, higher ratios (> 6 or 7) could indicate a more recent pollution event, whereas lower ratios (< 5) could indicate that the public health risk from exposure to pathogens is not as high. Nevertheless, RNA is generally more difficult and expensive to work with compared to DNA and requires more technical expertise and higher sterility (e.g., often requires a -80 ^{0}C freezer, RNase-free consumables, etc.), making this approach not practical for local municipalities with limited laboratory resources.

Yet, the viable cell counts indicated that the abundance of *E. faecalis* was still exceeding the EPA recreational water quality criteria of 36 CFU/100mL for all isolates on D8 (~10^5 CFU/mL; Figure 3-1A), thus these lake water samples would still be considered a public health risk according to current EPA standards. However, our findings that the rRNA/rDNA ratio is decreased after D4 suggests that these cells have largely become inactive (e.g. enter VBNC) and/or likely have started dying by D4 and hence, represent a

lower risk compared to D1. Consistent with these interpretations, a recent quantitative microbial risk assessment (QMRA) analysis of sewage pollution suggested that the risk of exposure to pathogens is not significant after three days (Boehm et al. 2018). In water bodies that consistently exceed EPA regulations for *Enterococcus*, it could be useful to investigate whether this is the result of a natural reservoir (i.e. no pathogen risk) or chronic pollution (pathogen risk) and techniques like the rRNA/rDNA assay presented here could be useful to help inform appropriate monitoring, management, and/or mitigation strategies.

Metatranscriptomics revealed that housekeeping genes such as ribosomal and transcription-related proteins were significantly more expressed in the environmental isolates (Figure 2-2), which may indicate better survival because they are able to maintain general gene expression without a strong signal of environmental stress. In contrast, the enteric isolates had fewer DEGs but these included several that potentially reflect a stronger stress response compared to environmental isolates. However, the number of DEGs overall was small (only 31 genes) and these results may be spurious as about half of these DEGs detected could be due to chance based on the false discovery rate predicted by the DESeq2 analysis (i.e. expected ~17 DEGs by chance). Moreover, the RNA extraction protocol used in this experiment was originally optimized for the rRNA/rDNA assay (i.e., simultaneous and consistent extraction of both DNA and RNA from a single filter and to ensure that the same amount of starting material is used each time) and resulted RNA samples with concentrations too low for ribo-subtracted libraries. Accordingly, this protocol resulted in poor mRNA recoveries in the metatranscriptomes (< 5% mRNA; typical for non-ribo-subtracted libraries) and some of the DEG signal could have been lost as a result of this.

Future studies should include separate RNA extractions for the rRNA/rDNA ratio assay and metatranscriptomic sequencing.

Furthermore, we acknowledge that the methods employed in this study may also limit our ability to distinguish isolates from the two habitat types. Previous starvation experiments show that in some taxa, growing cells at maximum or medium growth rates before starvation can affect whether high rRNA levels are sustained even when cell activity decreases (Sobek et al. 1966, Oda et al. 2000) and suggests an organism's response to an event (e.g., introduction to extra-enteric environment through fecal shedding) can be determined by the conditions it was exposed to before that event. In our dialysis bag mesocosm experiment, we spiked pure cultures from rich media into lake water, which may not accurately reflect the life histories of environmental or enteric *E. faecalis* isolates and thus, different ratios may be observed *in situ* relative to our mesocosm condition. For example, an enteric cell can be first introduced into a sewage or septic system, which may not be nutrient limiting but have other stressors like oxidation or predation, before reaching a surface water body. It is also possible that the habitat-specific genes previously identified such as the nickel and cobalt transport systems in the environmental genomes (Weigand et al. 2014, Cesare et al. 2014, He et al. 2018) are tuned for different conditions or stimuli than the mesocosm conditions used here and this accounted for the lack of their differential expression in our datasets. Although mesocosm studies are helpful for comparing *E. faecalis* survival in a more controlled environment, they cannot simulate all of the complex biotic and abiotic factors that occur in aquatic habitats. Inspecting the ratios in extractions directly from known, natural extraenteric reservoirs of *Enterococcus* such as in algal mats (Whitman et al. 2003) could help to get a better understanding of how rRNA levels are

regulated in isolates that have been (presumably) under nutrient limitation for a longer period of time.

This work provides preliminary information on rRNA/rDNA ratios in *E. faecalis* isolates under both standard lab and *in situ*-like conditions, which to our knowledge, has not been investigated for this genus. Our results suggest there may be evidence for different habitat adaptations between environmental and enteric strains but the difference may be too subtle or not consistent enough to be used in water quality monitoring. Clearly, our preliminary results require testing with more strains and growth conditions to allow for more robust conclusions to emerge. Furthermore, this study provides new insights on the relationship between rRNA levels and non-growth activities, such as in VBNC cells, for an important FIB taxon.

2.6 Acknowledgements

This work was supported by the US National Science Foundation, award numbers 1511825 (to J.M.B and K.T.K) and 1831582 (K.T.K.), and the US National Science Foundation Graduate Research Fellowship under grant number DGE-1650044 (to B.S). The funding agencies had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CHAPTER 3. METAGENOME-BASED COMPARISONS OF DECAY RATES AND HOST-SPECIFICITY OF FECAL MICROBIAL COMMUNITIES FOR IMPROVED MICROBIAL SOURCE TRACKING

Brittany Suttner, Blake G. Lindner, Minjae Kim, Roth E. Conrad, Luis M. Rodriguez-R, Luis H. Orellana, Eric R. Johnston, Janet K. Hatt, Kevin J. Zhu, Joe Brown, and Konstantinos T. Konstantinidis

3.1 Abstract

Fecal material in natural environments and water distribution systems is a primary source of pathogens that cause waterborne diseases and affect over a billion people worldwide. Most microbial source tracking (MST) efforts to attribute fecal contamination are based on 16S rRNA gene amplicon sequencing but these single gene-based assays do not always provide the resolution needed. In this work, we used dialysis bag mesocosms simulating a natural freshwater environment that were spiked separately with cow, pig, or human feces to monitor the decay of host-specific fecal signals over time with metagenomics and traditional qPCR and culture-based methods. Our sequencing of the raw fecal communities used as inocula recovered 79 non-redundant metagenome-assembled genomes (MAGs) whose abundance patterns over time suggested little health risk after about four days of incubation. Several MAGs showed high host specificity and thus, represent good candidates for biomarkers for their respective host type. Although all of

these MAGs were fermentative anaerobes, functions related to biofilm formation, biotin metabolism, and transport of various metabolites distinguished MAGs from different host types. Traditional qPCR methods varied in their correlation with MAG decay kinetics. Notably, the human-specific *Bacteroidales* assay, HF183, consistently under-estimated fecal pollution due to not being present in all inocula and/or primer mismatches. This work provides new insights on the persistence and decay kinetics of host-specific gut microbes in the environment and identifies several MAGs as putative biomarkers for improved MST.

3.2 Introduction

Fecal-contaminated waters have caused significant public health and economic burdens around the world (Eisenberg, Bartram, and Wade 2016). Because it is not practical to monitor the full spectrum of pathogens associated with fecal contamination, water quality and public health risk are assessed using fecal indicator bacteria (FIB) as proxies. Accordingly, the historical emphasis on monitoring indicators has resulted in water quality regulations focused primarily on reducing FIB instead of controlling pathogens and protecting public health. Because culture-based efforts to count FIB are laborious and ineffective for timely water management decisions, recent efforts have focused on rapid culture-independent methods such as qPCR targeting new biomarkers that are not easily amendable to culture-based approaches such as anaerobes within the order *Bacteroidales* (Kildare et al. 2007, Haugland et al. 2010) and other organisms known only by culture-independent genomic approaches (McLellan and Eren 2014). Several decades of research show that the genus *Bacteroides* tends to co-evolve with the host and are particularly suitable for MST because they are among the most abundant genera in stool, have a narrow host range exclusive to warm-blooded mammals, and generally have poor survival rates

outside their host (Ahmed, Hughes, and Harwood 2016). However, recent evidence suggests the potential for *Bacteroides* to persist, and even grow under some environmental conditions (Green et al. 2011; Weidhaas et al. 2015), which brings the assumptions about their persistence outside of the host into question. More recently, CrAssphage, a DNA bacteriophage named after the metagenomic data mining technique used to discover it (Dutilh et al. 2014) has emerged as promising new MST biomarker because it is one of the most abundant phages in some human gut populations (Stachler and Bibby 2014). There are now several published qPCR assays targeting CrAssphage for MST (García-Aljaro et al. 2017; Cinek et al. 2018; Liang et al. 2018), however most of these assays have not been tested in environmental systems. Several studies report these markers have high concentrations in sewage and sewage-impacted waters but have some cross-reactivity with other (non-human) hosts (Ahmed et al. *Wat. Res.* 2018; Stachler et al. 2018; Ahmed et al. *Appl. Micro. & Biotech.* 2018) and may be too abundant in sewage for monitoring highly polluted waters (Stachler et al. 2018). Clearly, more research is needed on CrAssphage concentration and persistence in the environment including how well these phages correlate to risk of infection with enteric pathogens. Nevertheless, the end goal when identifying new MST biomarkers, regardless of how advanced or sophisticated the method used to discover them, is usually to design a single qPCR assay for routine water quality monitoring. However, this framework has many known limitations as described below (Savichtcheva and Okabe 2006).

Rapid culture-independent methods such as qPCR allow for same day results and there is a plethora of studies focused on designing more robust qPCR markers for MST (Bernhard et al. 2003; Seurinck et al. 2005; Kildare et al. 2007; Field and Samadpour 2007;

Bae and Wuertz 2009; Converse et al. 2009; Mieszkin et al. 2009, Liu et al. 2015) and assessing marker performance under various experimental or environmental conditions and sampling techniques (Anderson et al. 2005; Bae and Wuertz 2009; Chern et al. 2009; Haugland et al. 2010; Dick et al. 2010; Green et a. 2011; Bae and Wuertz 2015; Liu et al. 2015; Li et al. 2016; Zhang et al. 2016; Cloutier and McLellan 2017; Mantha et al. 2017; Mattioli et al. 2017; Rothenheber and Jones 2018; Korajkic et al. 2018; Ballesté et al. 2018). However, methodological issues pose a substantial challenge for the field application of qPCR markers for regulatory purposes, as there are no standardized methods for sample collection, processing, DNA extraction, internal controls, etc. Further, limited documentation exists on the recovery of these markers from complex environmental matrices and the effect of PCR inhibitors carried over from the extraction, which makes it challenging to reliably compare the performance of different assays across different studies (Ahmed et al. 2016). In an attempt to address these uncertainties, Boehm et al. 2013 evaluated the performance of 41 MST markers and found none to be perfectly sensitive or specific for their target host. Notably, the human specific *Bacteroides* marker, HF183, was the best performing host specific marker according to this study (Boehm et al. 2013). However, this and other human associated markers, are not prevalent in human populations worldwide (Reischer et al. 2013), which suggests no single qPCR marker is likely to be universally suitable for detecting human fecal contamination. More information is also needed on the geographical and temporal stability of gut microbiomes in human and other host groups in order to design biomarkers with improved within and between-host resolution. Moreover, fecal pollution of surface waters is often the result of a complex mixture of multiple inputs further complicated by environmental dispersion and deposition.

The decay characteristics of different DNA markers could also confound the interpretation of MST results (Cloutier and McLellan 2017). Although an absolute gene count can be obtained via qPCR, estimates of the relative abundance of each marker and the relative contribution of various fecal sources in the natural environment cannot be quantitative without this decay information. The use of FIB and MST to manage water quality has undoubtedly helped to reduce public health risks; however, the currently used approaches are not ideal in several areas. Hence, new, more comprehensive methods, such as metagenomics (Handelsman et al. 2007), are still needed to help improve biomarker discovery and overcome some of the limitations described above.

Most research efforts utilizing metagenomics and next-generation sequencing (NGS) technologies thus far have focused on 16S rRNA gene amplicon sequencing to develop new biomarkers (Unno et al. 2018). However, the 16S rRNA gene is highly conserved across *Bacteria* and *Archaea* and this method of clustering gene amplicon sequences in operational taxonomic units (or OTUs) based on an identity threshold (e.g. 97%) can lead to groups that are too broad to be environmentally or ecologically meaningful. As such, cross-reactivity with non-target hosts is common for all assays targeting even the most variable regions of the 16S rRNA gene (Harwood et al. 2012; Ahmed et al. 2016). Functional, protein-coding genes that are specific to a host's unique gut physiology (e.g. host-microbe interactions) are likely more suitable targets for host-specific markers. However, deciphering the functional basis for host-associated microbiome patterns requires moving beyond 16S rRNA amplicon data to a broader sequencing approach (i.e., shotgun metagenomics) to characterize any selective, adaptive, and evolutionary processes underlying host-specific bacterial signatures (i.e., ecotypes and genes) at the whole genome

level. It is also likely that metagenomic methods can be combined with conventional MST methods to obtain more accurate measures of fecal pollution in watersheds since qPCR has generally a lower limited of detection that metagenome shotgun sequencing (Suttner et al. 2020, Hong et al. 2020). Microbial source tracking and water quality assessments could be significantly improved with metagenomics, yet this technique has seen little application in MST field to date (Sharma and Sharma 2020). More research is needed to establish the best meta-omics and bioinformatic techniques as tools for water quality monitoring and public health risk assessment and identify host-specific taxa and their genes.

In this study, we used dialysis bag mesocosms simulating a fecal pollution event in a freshwater habitat and time-series metagenomics to track the decay of metagenome assembled genomes (MAGs) from human, cow, and pig fecal inputs over time. Additionally, we used traditional culture and qPCR-based MST markers and included a universal 16S rRNA gene qPCR assay for absolute quantification in order to compare marker concentrations determined by the traditional and metagenomics methods. Using the time-series abundance and cross-reactivity information, we were able to identify ~12 MAGs as putative MST biomarkers and compared their functional gene content in order to establish host-specific genomes and genes for improved water quality monitoring assays.

3.3 Materials and Methods

3.3.1 Mesocosm sample collection, set-up, and sampling

3.3.1.1 Lake water and fecal sample collection:

Lake water samples were collected from Lake Lanier (Georgia, USA) in acid washed 10L carboys and transported immediately back to the lab and stored in the dark at 4 °C until mesocosm set up the following day (within 24 hours). Human fecal samples were collected from human volunteers who had not taken any antibiotics within the past one month before sample collection. All human subjects in the study provided informed consent and the study was approved by the Georgia Institute of Technology institutional review board (IRB) and carried out in accordance with the relevant guidelines and regulations. Remel fecal collection kits (ThermoFisher) were provided to human volunteers, who were instructed to store their sample at 4 °C and return within two days after fecal collection. Cow and pig fecal samples were collected within six hours of defecation from animals at the University of Georgia Athens Department of Animal and Dairy Sciences farms. A portion of each freshly-excreted fecal sample was preserved for DNA extraction by adding 1:1 by volume feces into two mL sterile distilled water and stored at 4 °C until processing in the lab. Five mL of lysis buffer (Qiagen PowerBead solution) was added to the water:feces mixture, vortexed for 30 seconds, then spun for three minutes at 1500rpm to create a homogenized fecal slurry. One to two mL of the slurry was pipetted into two mL cryo-vials and stored at -80 °C until ready for DNA extraction. Another portion of each fecal sample was persevered in 1:1 by volume of 15 mL sterile Cary-Blair media and stored at 4 °C until ready for inoculation into mesocosms (within two days).

3.3.1.2 Mesocosm set-up :

Sterile glass bottles were filled with 1.6 L of lake water and inoculated with feces to a final concentration of 2.5 g/L and shaken well to thoroughly mix the feces:lake water mixture before dispensing into dialysis bags. The dialysis bags (6-8 kDa molecular weight cutoff)

were filled to a total volume of 110 mL (~21 cm length of dialysis tube) with the feces:lake water mixture, un-inoculated lake water, or sterile milliQ water negative controls and closed on both ends using polypropylene Spectra/Por clamps (Spectrum Laboratories). Enough dialysis bags were filled to sample each biological replicate in triplicate at each time point, i.e., 36 dialysis bags per host type (three technical replicates per three biological replicates at 4 sampling time points). Additionally, four uninoculated lake water and two sterile milliQ water negative control bags were included for both of the two mesocosm experiment batches. The dialysis bags were then transferred to ten-gallon aquarium tanks filled with lake water and stored in environmentally controlled rooms at 22 °C in the dark. A small water pump was included in each tank for aeration and nutrient distribution. A small headspace of air was left in each bag when sealing with the clamps so that they could float freely in the tanks.

3.3.1.3 Mesocosm sampling:

On the day of mesocosm set up, initial day zero (D0) reference community lake water samples were collected by filtering five separate 250mL aliquots of uninoculated lake water onto 0.45 um poly-carbonate (PC) membranes, three of which were stored at -80 °C in PowerFecal (Qiagen) two mL screw-cap bead tubes until ready for DNA extraction and analysis with MST qPCR assays, while the other two were stored at -80 °C in sterile two mL screw-cap tubes filled with acid-washed 0.1 mm glass beads until ready for analysis following the EPA Method 1611 (USEPA 2012). Finally, 100 ml of the lake water was filtered and cultured on mEI media (in triplicate) following EPA Method 1600 for culture-based enumeration of *Enterococcus* (USEPA 2002). Additionally, the feces:lake water slurry mixtures remaining after filling the dialysis bags were sampled following the same

protocol for the un-inoculated lake water except using a 25 mL filter volume and 10-fold serial dilutions in phosphate-buffered saline (PBS) for culture-based enumeration EPA Method 1600 (USEPA 2002). All dilutions yielding measurements within the acceptable range of quantification were averaged to estimate CFUs/100mL of each biological replicate.

Destructive sampling of the dialysis bags occurred at days 1, 4, 7 and 14 (D0, D1, D4, D7, and D14) after the initial set up day and included each biological replicate (i.e., triplicate biological replicates per sampling point) and a single lake water negative control. The milliQ sterile lake water negative control bags were sampled on D7 and D14 to test dialysis bag integrity over time. Three aliquots of 25 mL from each dialysis bag were filtered onto 0.45um polycarbonate (PC) membranes and two were stored as described above for DNA extraction and the EPA Method 1611 (USEPA 2012). The third filter was saved at -80 °C as an archive filter. Additionally, water from each bag was 10-fold diluted and assayed according the EPA Method 1600 as described above.

3.3.1.4 DNA extraction from feces and filters:

DNA was extracted from homogenized fecal slurries for the quantification of MST qPCR markers and metagenome sequencing. The Qiagen PowerSoil kit was used for the cow, pig, and human fecal slurries following a modified Human Microbiome Project protocol for stool samples described previously (Wesolowska-Andersen et al. 2014). Briefly, ~0.25 g of the fecal slurry was added to 500uL Bead Solution (Qiagen) then heated at 65 ^{0}C with spinning for ten minutes followed by heating at 95 ^{0}C for ten minutes. This solution was

transferred to the Qiagen PowerSoil bead tube and DNA was extracted following the manufacturer's protocol.

Two separate DNA extraction protocols were used for the PC filters (i.e., two filters per dialysis bag sample were used for DNA extraction with two different methods). One PC filter was used for quantification of MST qPCR markers and metagenome sequencing and DNA was extracted using the Qiagen PowerFecal kit following the manufacturer's instructions except mechanical cell lysis was performed by bead beating in two 1-minute intervals using the Biospec BeadBeater. The other PC filter was used for DNA extraction and enumeration of total *Enterococcus* following the EPA Method 1611 (USEPA 2012). This method was designed for rapid and simple water quality monitoring and does not include any chemical precipitation or clean-up steps. Therefore, the method results in a more "crude" DNA extraction that is not suitable for most qPCR and metagenomic methods. That is also the reason that two different DNA extraction methods were used.

3.3.2 qPCR for common MST markers

The qPCR markers used in this study are described in Table 3-1 and included the human-specific *Bacteroidales* HF183/BFDRev (hereafter HF183; Haugland et al. 2010), a ruminant-specific *Bacteroidetes* BacR (hereafter RumBac; Reischer et al. 2006), human mitochondrial DNA (hereafter HUMmt; Caldwell et al. 2007), *Enterococcus faecalis* 16S rRNA gene (hereafter EF16S; Santo-Domingo et al. 2003), and the standard EPA Method 1611 assay targeting total *Enterococcus* (hereafter EPA1611**)**. We also included a universal 16S rRNA gene qPCR assay for total cell counts (hereafter GenBac16S; Ritalahti et al. 2006**)**. All qPCR reactions were run using an Applied Biosystems 7500Fast thermocycler

machine and the thermal cycling parameters were as follows: 2 minutes at 50 °C, 10 minutes at 95 °C, then 40 cycles of 15 seconds at 95 °C and 60 seconds at 60 °C . The EPA Method 1611 assay was run using the standard calibrator cells and the ΔΔCt quantification method following the protocol described in (USEPA 2012)**.** All other qPCR assay reactions used two ul of extracted DNA as template in 20 uL qPCR reactions with the TaqMan Universal PCR Master Mix (Applied Biosystems). Template DNAs were ran undiluted or diluted 5-fold depending on the expected marker concentration and quality of each sample. The Taqman (i.e., 5' hydrolysis) probe and primer concentrations for each assay are listed in Appendix A, Table A 3 and only the RumBac assay reactions included 10 ug bovine serum albumin (BSA)**.** All samples were run in triplicate on 96-well plates and each run included triplicate no template controls (NTC). The amplification thresholds for the GenBac16S, EF16S, and HUMmt and the HF183 and RumBac assays were set to 0.02 and 0.03 ΔRn units, respectively.

Standard plasmids were used for absolute quantification. Target sequences were ligated into a pCR™2.1-TOPO® TA vector and cloned using One Shot® Chemically Competent TOP10 *Escherichia coli* and the TOPO®-TA cloning kit (Invitrogen), following manufacturer's instructions. The qPCR amplicon sequences were used as the target sequence in the standard plasmids for the HF183, RumBac, and HUMmt assays, while an *Enterococcus* faecalis 16S rRNA gene was used in the standard plasmid for the EF16S and GenBac16S assays. Genomic DNA from *Bacteroides* sp., Strain 1_1_6, HM-23D (obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project) was used at the template DNA for generating the HF183 qPCR amplicon for ligation with standard plasmid. Standard plasmids were isolated using the QIAprep Spin Miniprep Kit (Qiagen)

following the manufacturer's instructions and quantified using the Qubit HS DNA kit. Seven, 10-fold serial dilutions (10^6 to 10^0 copies per reaction) of qPCR standard plasmids were run in triplicate on every 96-well plate. Marker concentrations were determined using the corresponding calibration curve from each plate because no more than four plates were ran per assay. Details on the standard curves and average assay efficiencies for each assay are provided in Appendix A, Table A 3.

3.3.2.1 qPCR marker copy number calculations and detection limits

Marker copy number per qPCR reaction was calculated for all samples using the linear fit of log-transformed standard copy number versus threshold cycle (Ct). A marker was considered not detected (ND) for any sample that did not return a Ct value in two of the three triplicates and was considered detectable but not quantifiable (DNQ) if it returned an average Ct value that was above the average Ct value of the lowest concentration in the standard curve, which was ~2.1 counts/uL DNA for the HF183, RumBac, EF16S, and HUMmt assays and 168 counts/uL DNA for the GenBac16S assay. The number of genes copies in each detectable sample was averaged, normalized by the volume of DNA per reaction, multiplied by the DNA elution volume (100 uL) and then divided by the total water filter volume (mL) or the mass of feces used in the initial DNA extraction (mg). Gene counts were converted to cell counts where applicable as described below.

3.3.3 Bioinformatic analysis of metagenomic data sets.

3.3.3.1 Metagenome library sequencing, quality assessment and analyses:

Metagenome sequencing libraries were prepared using the Illumina Nextera XT kit and sequenced on the HiSEQ 2500 instrument as described previously (Johnston et al. 2019). Three additional negative control libraries that were not included in the initial HiSEQ 2500 run were sequenced later on the NovaSEQ 6000 S4 platform (i.e. animal_LL_D1, D4, and D7) The two D0 negative control libraries (i.e. human_LL_D0 and animal_LL_D0) were also re-sequenced on the NovaSEQ for quality control comparisons to HiSEQ data. Short reads were pass through quality filtering, trimming and assembly using MiGA (Rodriguez-R et al. *Nucl. Acids Res.* 2018**)** with the default settings (PHRED score cutoff of 20, only retain read pairs with both sisters ≥ 50bp after trimming, and assembly with IDBA-UD [Peng et al. 2012] using kmer values ranging from 20 to 80). MASH v1.0.2 (options: -s 100000; Ondov et al. 2016) was used to determine whole-community similarity between metagenomes in a reference-independent approach. The MASH distances were used for ANOSIM, ADONIS (compared by host type and sampling day), and non-metric multidimensional scaling (NMDS; number of dimensions =4) with the metaMDS function in the R package vegan v2.5-6. Average community coverage and diversity were estimated using Nonpareil v3.0 (Rodriguez-R et al. *mSystems* 2018) with kmer kernel and default parameters. Average genome size and genome sequencing depth (i.e. average sequencing depth of single copy genes) were estimated in each metagenomic sample using MicrobeCensus v1.0.6 with default parameters (Nayfach and Pollard 2015).

3.3.3.2 Gene functional annotation and determination of differentially abundant gene functions in host fecal and D7 metagenome assemblies:

Open reading frame (ORF) prediction from assembled contigs was performed using Prodigal (Hyatt et al. 2010) as implemented in MiGA (Rodriguez-R et al. *Nucl. Acids Res.*

2018). Resulting amino acid sequences were searched against the KEGG ortholog profile database using KoFamScan v1.2.0 (Aramaki et al. 2019) with the 'prokaryote' database and using the parameter '-f mapper' to provide only the most confident annotations (i.e., ORFs assigned an individual KO). Orthologies were matched to their corresponding functional annotations using a parsed version of the KEGG orthology table ('ko00001.keg'; https://github.com/edgraham/GhostKoalaParser). Sequence coverage of each gene was determined by mapping metagenomic short reads against the corresponding ORFs for each sample using Magic-BLAST v1.4.0 (Boratyn et al. 2019). The Magic-BLAST outputs were filtered for best match, 90% query cover alignment length, and a minimum read length of 50 bp. These read counts were used to determine DA functional annotations in samples grouped by host type (i.e. pairwise comparisons of human, cow, and pig fecal samples) and all host fecal metagenomic samples vs. D7 mesocosm metagenomes using the negative binomial test and false discovery rate (P_{adj} <0.05) as implemented in DESeq2 v1.4.5 (Anders and Huber 2010). DA functional annotations with Log_2 fold change (L2FC)>|3| for the host only and L2FC>|6| for the host vs. D7 comparisons were summarized into several hierarchical ranks including metabolic pathways and individual protein families based on the KEGG classification system (Chapter 3 Supplementary Data files S2 and S3). A larger L2FC cutoff was used for host fecal vs. D7 comparison so that the number of DA functions retained were feasible for manual inspection (i.e. <600 functions). Read counts for each summarized functional category were converted to genome equivalents (GE) by dividing by the average genome sequencing depth as determined using MicrobeCensus v1.0.6 (Nayfach and Pollard 2015).

Each category was divided by the average GE across all samples to provide unbiased counts for visualization purposes.

3.3.3.3 Binning of fecal and D7 metagenomes and dereplications of MAGs:

The host fecal and D7 dialysis bag metagenome assemblies were used for population genome binning with MaxBin 2.2.4 (Wu et al. 2014) and MetaBat 2.12.1 (Kang et al. 2019) with default settings. Only contigs longer than one kbp were used for binning. Fecal MAGs from the two algorithms within each host type were combined and de-replicated with DAS Tool 1.1.0 (Sieber et al. 2018), and only resulting MAGs with contamination <5% or MiGA quality score >50 were retained for further analysis. This collection of high-quality host fecal MAGs was also dereplicated against each other (i.e., across each host type), along with a collection of 477 Lake Lanier (LL) MAGs (Rodriguez-R et al. 2019), and the non-aggregated set of D7 MAGs (i.e. the total set of MAGs resulting from both MetaBat and MaxBin) using the MiGA derep workflow (Rodriguez-R et al. *Nucl. Acids Res.* 2018). That is, in cases where two MAGs shared >95%ANI, only the higher quality MAG was retained for further analysis. The average amino acid identities (AAI) calculated by MiGA were used to generate heatmaps with the seaborn library in python3.

3.3.3.4 Taxonomic and functional annotation of MAGs:

Taxonomy was assigned to the MAGs using the MiGA assign_taxonomy workflow and the RefSeq type material database. MAG phenotypes (aerobe, anaerobe, fermentation type) were assigned using Traitar v1.0.4 and the prediction results from the phypat classifier model only (Weimann et al. 2016). Functional annotations were assigned to genes of MAGs identified as potential biomarkers (see following section) using KoFamScan v1.2.0

(Aramaki et al. 2019) and the Swiss-Prot (Uniprot Consortium 2017) database as implemented in MicrobeAnnotator (Ruiz-P et al. *in review*; https://github.com/cruizperez/MicrobeAnnotator) with the -m sword and --light options.

3.3.3.5 Tracking abundance of MAGs, FIB and MST reference genomes in dialysis bag mesocosm metagenomes:

Magic-BLAST v1.4.0 (Boratyn et al. 2019; options: -no_unaligned -splice F -outfmt tabular -parse_deflines T) was used to map metagenomic short reads to MAG contigs to determine MAG abundance over time in the mesocosms expressed as average sequencing depth (base pairs recruited/genome length). Read matches were filtered for single best alignments, using a minimum 90% query cover alignment length and 95% nucleotide identity. In order to remove biases from highly conserved regions and contig edges, the 80% central truncated average of sequencing depth of all bases (TAD80; Rodriguez-R et al. 2020) was determined using custom scripts available at: https://github.com/rotheconrad/00_in-situ_GeneCoverage. MAG abundance in each metagenomic dataset (as % of total community) was calculated as the quotient of the MAG's TAD80 value and the average genome sequencing depth from MicrobeCensus (Nayfach and Pollard 2015). In addition to the MAGs recovered from this study, several common MST reference genomes were also queried against the time-series short reads and included the reference genomes associated with the MST qPCR assays described in Table 3-1, as well as the common commensal organism, *Escherichia coli* HS (accession: NC_009800.1) and CrAssphage (accession: JQ995537). The reference genomes were download from NCBI and their relative abundance in the time-series metagenomic datasets was determined as described above.

3.3.3.6 Absolute abundance and limit of detection (LOD) estimation of qPCR markers, reference genomes, and MAGs:

The absolute abundance (cells/mL) of a *Bacteroides dorei* reference genome (Table 3-1) was estimated in the human mesocosm metagenomes using TAD80 at 95% average nucleotide identity of reads mapping against the reference genomes (ANIr) divided by MicrobeCensus average genome sequencing depth as described above. This value was then multiplied by total cell density per mL as determined by the GenBac16S qPCR assay described above in order to estimate the total number of *Bacteroides* cells per mL. Absolute abundances for fecal MAGs in the mesocosm metagenomes was also calculated following the same method (i.e. TAD80 at 95% ANIr divided by genome sequencing depth then multiplied by total cell density) The RumBac qPCR assay is not associated with a known reference genome; therefore, we searched the cow fecal metagenomic contigs against the assay oligos to find contigs for a perfect match. Contigs were searched against the NR database using the NCBI blastn server to check for chimeras and that the best match was to a *Bacteroides* 16S rRNA gene. One contig was selected as the proxy reference genome and its abundance was determined as described above for genomes except no truncation was used when estimating sequencing depth (i.e. TAD100) at 99% identity instead of 95% due to the high sequence conservation of the 16S rRNA gene relative to the rest of the genome (99% identity at the 16S level represents within genus diversity and only species or closely-related species are captured [Yarza et al. 2014; Konstantinidis et al. 2017]). The resulting sequencing depth value was also normalized by 16S rRNA gene copy number for *Bacteroides* (n = 7) in addition to average genome sequencing depth from MicrobeCensus (Nayfach and Pollard 2015) and then multiplied by total cell density per mL as described

above in order to estimate total number of *Bacteroides* cells per mL. The same protocol used for *B. dorei* was followed for the reference human mitochondrial genome (Table 3-1) except sequencing depth was normalized only by metagenome dataset size (in Gbp).

3.4 Results

3.4.1 Performance and decay of traditional culture-based and qPCR markers

Dialysis bag mesocosms simulating a natural freshwater environment were spiked (in separate mesocosms) with cow, pig, or human feces to represent a pollution event and monitored over time with general FIB (i.e., *Enterococcus* spp.) and host-specific MST qPCR assays (Table 3-1) and the culture-based EPA Method 1600 for total enterococci (USEPA 2002). Since human gut microbiomes are known to vary geographically, only samples from within the state of Georgia (US) were used. Laboratory mesocosms consisted of 100mL dialysis bags filled with unfiltered lake water spiked with freshly collected fecal material. The dialysis bags have a pore size that allows passage of small molecules and ions but prevents passage of molecules larger than six to eight kDa (e.g., bacterial and viral cells). Three biological replicate fecal samples were used per host and are referred to hereafter as hum1, hum2, hum3, cow7, cow8, cow9, pig7, pig8, pig9 to indicate the specific individual host fecal sample that was used for DNA extraction and inoculation into the lake water mesocosms, whereas H1, H2, H3, C7, C8, C9, P7, P8, and P9 hereafter refer to the feces:lake water mesocosm sample for each individual host (e.g. H1 refers to lake water mesocosms spiked with feces from hum1). The host fecal mesocosm experiments were performed on two separate dates because of limited space and resources (e.g., available clamps for dialysis bags) in the lab. The cow and pig fecal mesocosms were performed

concurrently in fall of 2017 and the human experiments were ran later in the summer of 2018 (Appendix A, Table A 2). In both experiments, mesocosm sampling occurred in triplicate on days 0, 1, 4, 7 and 14 (hereafter, D0, D1, D4, D7, and D14) which included: culturing, DNA extraction and qPCR analysis using the markers described in Table 3-1.

The qPCR markers were first tested in the host fecal DNA samples used as inocula to compare their sensitivity and specificity. All of the MST markers had good host-specificity as they were not detected (ND) in any non-target hosts and none were quantifiable in the uninoculated lake water negative controls (Table 3-2). However, the human-specific HF183 marker was not detected in one of the three human fecal samples (hum2). Since the EPA1611 assay is designed for filtered water samples, it was not tested in the raw fecal extractions. The *E. faecalis* 16S rRNA gene marker (EF16S) was ND in the human feces and had very low abundance in the cow and pig feces (~150 copies/mg). Accordingly, EF16S was also ND upon feces dilution in the lake water (Table 3-2). However, the EPA Method 1600 culture-based test for *Enterococcus* showed that the dialysis bag mesocosms were exceeding the EPA's recreational water quality criteria (RWQC) of 36 CFU/100mL throughout the entire duration of the cow and pig experiment and in all of the human timepoints except on D14 (Figure 3-1A**)**. The EPA Method 1611 qPCR-based test for *Enterococcus* showed that all the time-series samples that could be quantified exceeded the EPA RWQC of 10^3 calibrator cell equivalents (CCE) per 100mL (Figure 3-1B). However, this assay was not detectable in the cow and pig samples by D14 and was only quantifiable in two of the three human samples on D1 and was not detected in the rest. Overall the concentration of *Enterococcus* spp. was similar based on the EPA1600 and EPA1611 assays for cows and pigs but was not consistent for the human

mesocosms because the culture-based counts were greater than the qPCR-based (Figure 3-1A-B).

When tested in the time-series dialysis bag mesocosm samples, the qPCR counts for all of the host-specific MST assays decreased with time and returned to very near or below assay detection limits by D14 (Figure 3-1C). Therefore, only samples from D0 through D7 were used for metagenomic sequencing. Consistent with the hum2 fecal DNA results, the HF183 marker was ND in any of the H2 mesocosm samples. Abundance of the HF183 marker in H3 mesocosms was four to two orders of magnitude larger than the abundances observed in H1 mesocosms on D0, D1 and D4, after which the HF183 marker became ND in the H1 samples (Figure 3-1C). Accordingly, only the H3 samples on D0 and D1 exceeded the quantitative microbial risk assessment (QMRA)-based water quality threshold of 41 copies/mL for HF183 (Boehm et al. 2018). The average counts/mL were consistent across the three biological replicates and were detectable until D14 for both the HUMmt and RumBac assays in human and cow mesocosms, respectively (Figure 3-1C).

Table 3-1: qPCR markers used in this study and associated reference genomes.

Host-specific MST markers include HF183, RumBac, and HUMmt; general FIB markers are EF16S and EPA1611. The GenBac16S assay was used for absolute quantification and LOD estimation for reference genomes in the metagenomes.

Marker	**Target**	**Reference**	**Reference genome**
HF183	Human *Bacteroides* 16S	Haugland et al. 2010	*Bacteroides dorei* CL03T12C01
RumBac	Ruminant *Bacteroides* 16S	Reischer et al. 2006	n/a
HUMmt	Human mtDNA NADH dehydrogenase subunit 5	Caldwell et al. 2007	Human mitochondrion genome
EF16S	*E. faecalis* 16S	Santo-Domingo et al. 2003	*E. faecalis* ATCC29212
EPA1611	*Enterococcus* 23S	EPA Method 1611	*E. faecalis* ATCC29212
GenBac16S	Universal 16S rRNA	Ritalahti et al. 2006	n/a

Table 3-2: Detection of MST qPCR markers in feces (inocula material) or un-inoculated lake water (negative control). Values are reported as average number of copies per mg or mL for fecal and lake water samples, respectively. ND = Not Detected, which indicates that sample did not return a Ct value for all biological replicates. DNQ = Detectable but not Quantifiable, which indicates that sample returned a Ct value that was higher than the lowest concentration in the standard curve.

	Pig feces	**Cow feces**	**Human feces**	**Lake water**
RumBac	ND	$4.1x10^6 \pm 6.1x10^5$	ND	ND
HF183	ND	ND	$8.9x10^4 \pm 1740$	DNQ
HUMmt	ND	ND	1645 ± 316	DNQ
EF16S	141 ± 51	153 ± 59	ND	ND

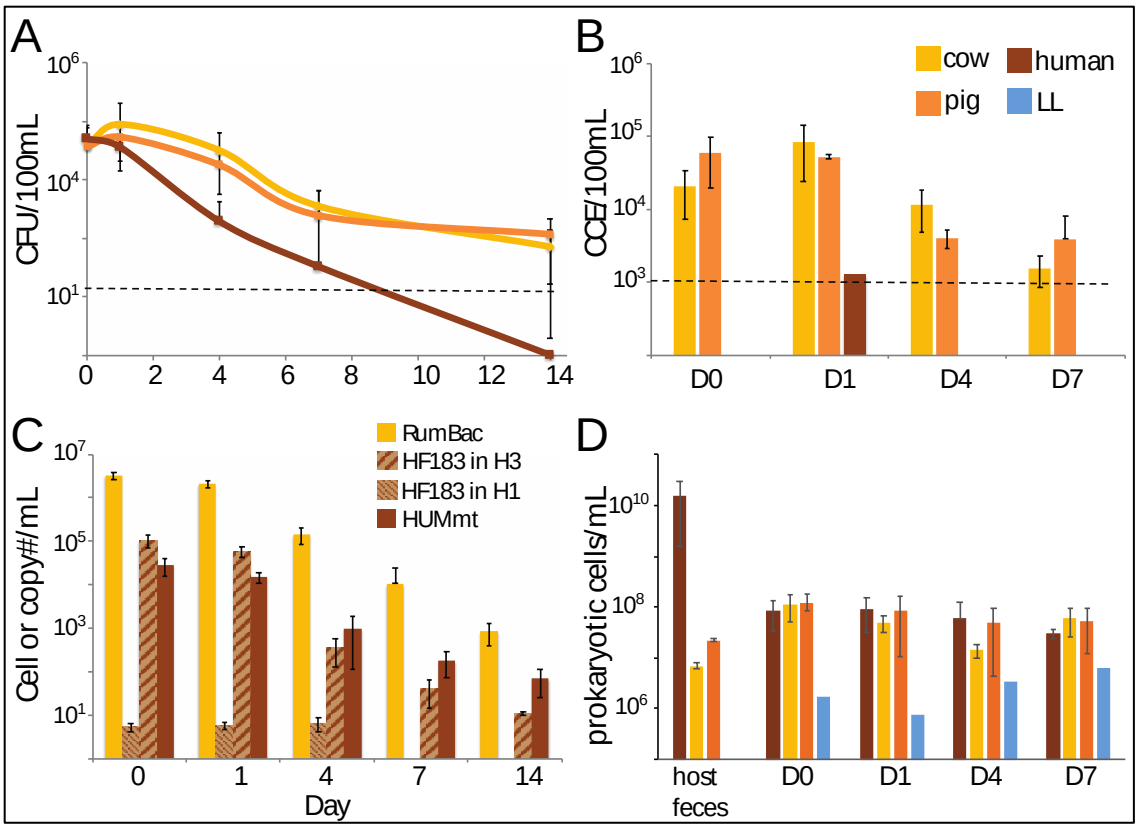


Figure 3-1: Traditional FIB, MST marker, and total bacterial cell abundances during the mesocosm incubations. (A) EPA Method 1600 culture-based enumeration of *Enterococcus*. (B) EPA Method 1611 qPCR-based enumeration of *Enterococcus*. Black dotted lines show the EPA's recreational water quality criteria (RWQC) limit for impaired waters for each assay (CFU= colony forming units; CCE= calibrator cell equivalents). (C) Host-specific MST qPCR assays that could be detected in the dialysis bag mesocosms. The HUMmt is reported as #copies/mL and the rest are reported as #cells/mL. (D) Prokaryotic cell density in the mesocosms over time based on a universal 16S qPCR assay (GenBac16S). In all figures, error bars are the standard deviation for averages that had more than three data points.

3.4.2 General description of metagenome samples and community coverage

A total of 56 metagenomic samples, ranging in size from 3.2 to 35.4 million reads (0.4 to 37 Gbp) were recovered from 36 dialysis bag mesocosm (fecal:lake water mixture), 5 lake water negative control, and 15 host fecal (inocula) samples (Appendix A, Table A 2). Fecal metagenomes from three other cow and pig individuals were also included in this study (in addition to the three cow and pig fecal samples used as inocula in the mesocosms) because considerably less fecal metagenomes have been sequenced for these hosts compared to humans and less is known about their gut microbiome diversity. The average total community covered by our sequencing efforts as determined by Nonpareil analysis (Rodriguez-R et al. *mSystems* 2018) was consistent across the biological replicates for the animal fecal metagenomes (54.8 ±8.7%, 70.9 ± 2.1%, and 84.5 ± 3.1% for the cow, pig, and human fecal samples, respectively). Therefore, it appeared that the cow samples were the most diverse on average (nonpareil diversity = 20.62; Appendix A, Table A 1), followed by pigs and then humans (19.4 and 17.6, respectively). Nonetheless, the sequencing coverage of these samples was within suitable range for whole community comparison at the individual gene/function or genome level (Rodriguez-R and Konstantinidis 2014).

The total cell density in the mesocosms based on the universal 16S qPCR assay (GenBac16S; Table 3-1) was ~10^8 cells/mL at the start of all mesocosm incubations and tended to decrease with time, reaching ~$1.5x10^7$ cells/mL by D7. The opposite trend was observed in the negative control bags, which started at ~10^6 cells/mL and increased by nearly an order of magnitude on D7 (Figure 3-1D). NMDS analysis based on MASH distances (Ondov et al. 2016) showed that samples tended to cluster by host type and time,

and that the later time points (e.g., D7) did not return to the natural community composition present in the lake water at D0 (Figure 3-2). Furthermore, ANOSIM analysis of the MASH distances showed that the samples were significantly different (P=0.001) by host type and sampling day (R=0.54 and 0.44, respectively) and ADONIS analysis predicted that these two variables explained (R^2) 44.0% and 41.3% of the variation in the MASH distances, respectively (P=0.001). These results indicated potential bottle effects during our incubations, which were assessed more fully by population genome binning of the D7 samples as described below.

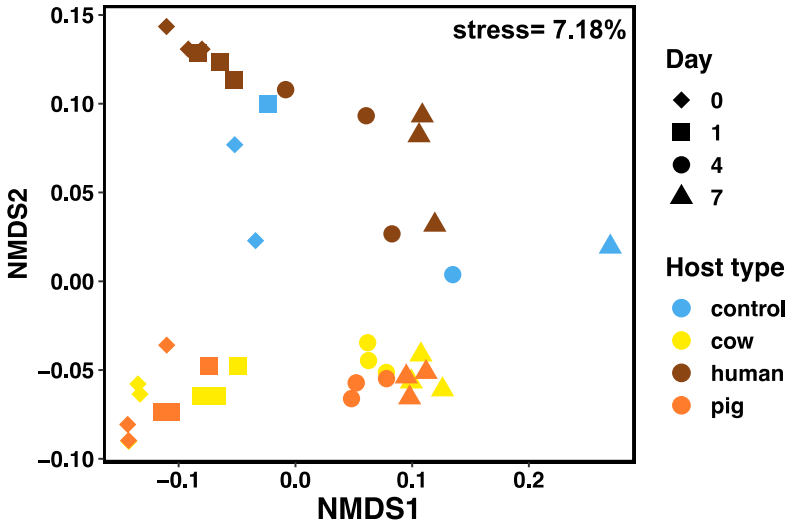


Figure 3-2: Similarity among the sequenced communities during the mesocosm incubations. Graph represents the non-metric multidimensional scaling (NMDS) of the whole-community MASH distances (i.e., overall kmer similarity of microbial communities). Each point represents a metagenome sample and samples from the same host type (or negative control) are denoted by the same color. Samples that are more similar are grouped closer together.

3.4.3 Taxonomic and phenotypic description of host fecal MAGs

The host fecal reads were assembled into contigs with total length and N50 values ranging from $2.5x10^7$ to $1.3x10^8$ and 1,913 to 19,034 base pairs, respectively (Appendix A, Table A 4). Contig binning resulted in an initial set of 30 cow, 13 human, and 82 pig high quality MAGs. The MAGs were first dereplicated at 95% nucleotide identity within each host and resulted in a new set of 18 cow, 13 human, and 50 pig MAGs, which were subsequently de-replicated against the high quality MAGs from all other host as well as the collection of 477 Lake Lanier (LL) MAGs (Rodriguez-Rojas et al. 2019) at 95% nucleotide identity to order to identify any MAGs that are non-host specific and/or found in the natural environment. This resulted in a final set of 17 cow, 13 human, and 49 pig HQ MAGs whose IDs are provided in Chapter 3 Supplementary data file S1. MAGs were named according to the individual fecal sample from which they were originally assembled followed by the closest relative of the MAG and the lowest taxonomic rank the two share according to the MiGA TypeMAT/NCBI database ($p<0.1$ threshold), i.e., P:phylum, C:class, O:order, F:family. G:Genus, S:Species. For instance, we use "cow4_20_Treponema_F" means MAG #20 assembled from cow4 fecal metagenome had a *Treponema* sp. as the closest relative and was classified (at the lowest level with statistical confidence) to the family *Spirochaetaceae*. Overall the MAGs were highly host specific at the species level (ANI >95%) and there were only two instances (cow4_20_Prevotella_F and pig7_9_Tolumonas_C) where a cow and pig MAG had ANI >95% with each other and were dereplicated into a single genomospecies. These MAGs were used for calculating decay in the mesocosms over time but were not considered as potential biomarkers due to the lack of host specificity. There was more overlap when evaluating the average amino

acid identity values (%AAI; Figure 3-3) among MAGs, revealing that these MAGs likely represent distinct but closely related species found in different hosts.

The taxonomy of the MAGs was determined using MiGA against the TypeMAT/NCBI database (Chapter 3 Supplementary data file S1; Rodriguez-R et al. *Nucl. Acids Res.* 2018). In all three host types, the majority of MAGs were classified at the class-level as *Bacteroidia* (41%, 46%, and 33% for cows, humans, and pigs, respectively) followed by *Clostridia* (24%, 23%, and 31% for cow, humans and pigs, respectively). Although MAGs from different hosts shared similar taxonomies at the class-level, there were several differences between the hosts at lower classification levels. In humans, the *Bacteroidia* MAGs were primarily assigned to the family *Bacteroidaceae*, whereas the cow MAGs were primarily *Prevotella*. The majority of the pig MAGs could not be classified well below class-level, i.e. they represented novel families (Chapter 3 Supplementary data file S1). Notably, although most of the cow and pig MAGs are *Bacteroidia* and *Clostridia*, two of the best putative cow biomarkers (see below; cow4_001_Treponema_F and cow8_3_Treponema_F) were actually classified in the family *Spirochaetaceae* while an *Actinobacteria* (pig4_16_Cellulomonas_C) and the archaeal phylum *Euryarchaeota* (pig4_38_Methanoplasma_F) were among the pig biomarker MAGs.

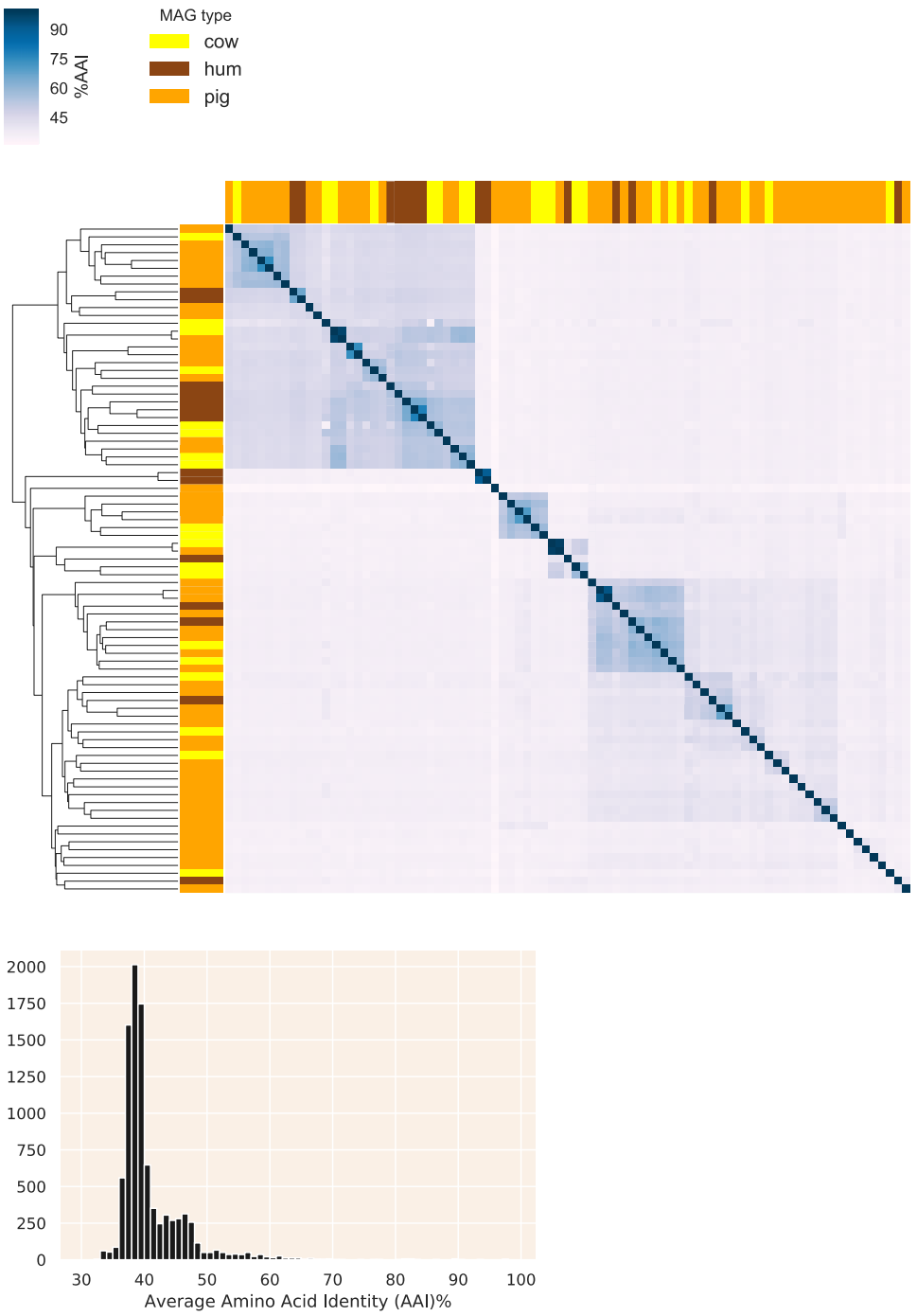


Figure 3-3: Genetic relatedness among the fecal MAG recovered by our study. (Upper) Heatmap comparing average amino acid (%AAI) of the MAGs assembled from pig, cow, and human fecal metagenomes. (Lower) Histogram of %AAI values.

None of the host fecal MAGs were phenotyped as aerobes using Traitar (Weimann et al. 2016) and the majority of MAGs were predicted to be anaerobes (100% human , 96% pigs, 82% cow; Appendix A, Figure A 4, A 5, A 6). The oxidative stress enzyme catalase was not found in any of the cow or pig MAGs but was detected in two of the human MAGs (hum1_013 _Akkermansia_G and hum2_003 _Rubritalea_C). Glucose fermentation was the most common energy yielding pathway in MAGs from all three host types (59% of cow, 71% of pig, 100% of human MAGs). In addition to glucose fermentation, 44, 15, and 15 unique sugar substrates for growth were identified in the pig, cow, and human MAGs, respectively, with lactose being the most common in the pig and human MAGs (76% and 85% of total MAGs, respectively) and maltose being most common in the cow MAGs (82%). These results were also consistent with the DESeq2 analysis at the individual gene level (see below).

3.4.4 Decay kinetics of host fecal MAGs in the mesocosms

The mesocosm metagenomic shorts reads were searched against the dereplicated set of 79 high quality host fecal MAGs from this study and the collection of 477 Lake Lanier (LL) MAGs (i.e., the native microbes present in the source lake water representing six years of surface water samples from the lake; Rodriguez-R et al. 2020) using Magic-BLAST (Boratyn et al. 2019) to assess MAG abundance dynamics over the incubation time. All 13 dereplicated human MAGs were detected in at least one human mesocosm, while only 13 out of the 17 total cow and 41 out of 49 total pig MAGs were detected in the cow and pig mesocosms, respectively. In general, most of the fecal MAGs decreased over time and were not detectable by D7 based on the TAD80 metric (Figures 3-4 and 3-5). There were a few MAGs that increased in abundance from D0 to D1, but this trend was

never consistent across all three biological replicates of a host type (e.g. hum1_001_ Bacteroides_S, cow4_001_ Treponema_F, cow7_4 _Kineothrix_F, and pig7_1 _Methylobacterium_C). The MAGs were highly host specific and none of the host MAGs were detected in any of the uninoculated lake water negative control metagenomes or mesocosm metagenomes from other hosts. Only a single cow MAG (cow4_10 _Prevotella_S) was detected in all three pig mesocosm metagenomes (Figure 3-5), the other cow MAG that was detected in the pig mesocosms (cow4_20 _Prevotella_F) shared >95% ANI with a pig MAG of lower quality. Similarly, the single pig MAG detected in all three cow mesocosm metagenomes (pig7_9_Tolumonas_C; Figure 3-4B) also had a close relative (>95%ANI) with a cow MAG that was removed during dereplication in the previous step described above. Thus, these two MAGs were expected to be detected in the both cow and pig mesocosms since they are highly related to each other. None of the non-human fecal MAGs were detected in any human mesocosm. The human MAGs showed high individual host specificity, i.e., MAGs assembled from an individual human fecal metagenome were always the most abundant in the mesocosms spiked with the feces from that individual and showed much lower abundances in the other two biological replicates (Figure 3-4A). In particular, among the hum2 MAGs, none were present in the H3 mesocosms and only two were detected in the H1 mesocosms; thus, none of the hum2 MAGs were selected as putative biomarkers (see below). Of the 477 LL MAGs, 59 (average 22 MAGs per mesocosm) and 139 were detected in the host fecal (Figure 3-6) and negative control (Appendix A, Figure A 8) mesocosms, respectively. The different host fecal inocula (and their associated communities) did not have a consistent effect on the abundance of the native LL MAGs over the time-series (Figure 3-6). Overall, the LL

MAGs showed much lower abundance than the host fecal MAGs in the mesocosms (<0.7%) and varied, more or less randomly, in their relative abundances with time, consistent with the assumption that they represent autochthonous taxa present in the source water used in the incubations.

Based on the decay and host specificity results, we identified putative targets for MST biomarkers as MAGs that were present in all three biological replicates of the same host type, were highly abundant on D0 (>0.1%) and were not detected after D4. Notably, the time of disappearance of these MAGs (i.e., D4) coincided with a previous quantitative microbial risk assessment (QMRA) analysis suggesting that health risk from sewage contamination is significantly reduced after three days (Boehm et al. 2018). Based on those criteria, we identified five cow, three human, and six pig MAGs that we investigated further as potential biomarkers for MST. These were: pig4_10_Paraprevotella_C, pig4_16_Cellulomonas_C , pig4_22_Emergencia_F, pig4_38_Methanoplasma_F , pig7_006_Acetobacteroides_C , pig8_10_Acetobacteroides_C , cow4_001_ Treponema_F, cow4_3_Acetobacteroides_C, cow5_2_Pseudonocardia_C , cow8_11_Phascolarctobacterium_F, cow8_3_Treponema_F, hum1_001_Bactreoides_S, human3_002_Bacteroides_S, and human3_3_Eubacterium_F (Figures 3-4 and 3-5; Chapter 3 Supplementary data file S1).

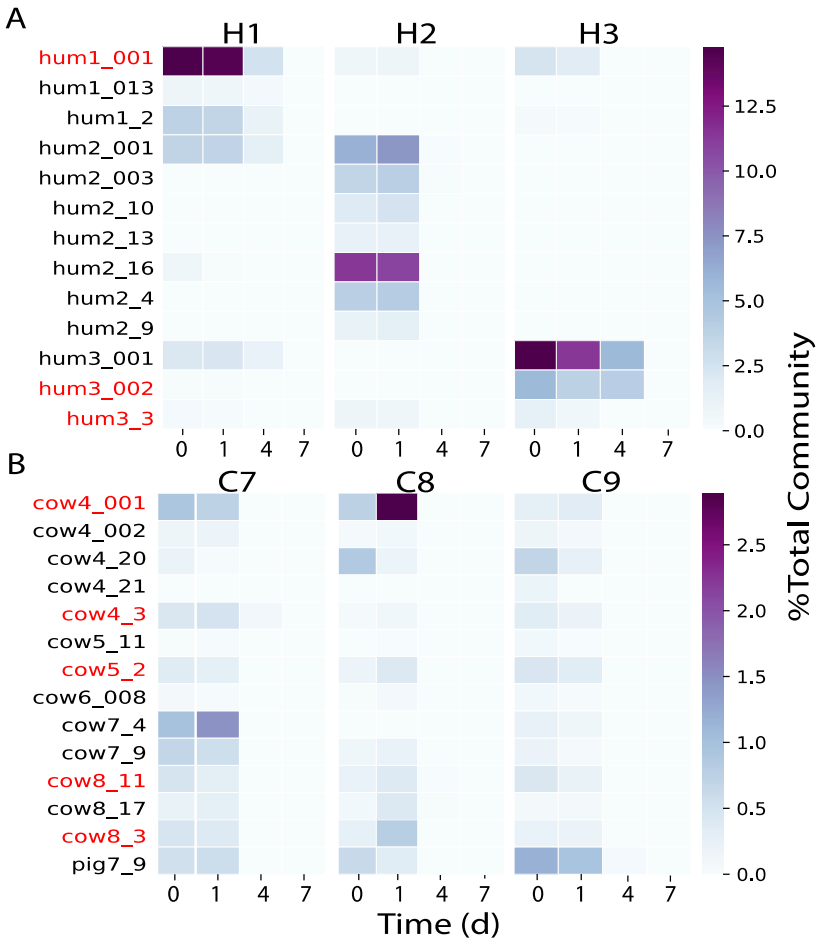


Figure 3-4: Decay kinetics of all host fecal MAGs (rows) that could be detected in the (A) human and (B) cow mesocosms (columns). Note that no non-target host MAG was detected in any of the human mesocosms; the single pig MAG shown was clustered into a single genomospecies at >95% ANI with another cow MAG that is not shown here because the pig MAG was of higher quality and was used in all downstream analyses. H1, H2, H3 and C7, C8, C9 are the different biological replicate mesocosm metagenomes for each host type. Abundances are reported as % of total microbial community (i.e. TAD80 divided by average genome sequencing depth). The MAGs identified as potential MST biomarkers have red labels. The MiGA TypeMAT/NCBI taxonomic identifications appending the MAG names as described in the main text are not included here due to space limitations.

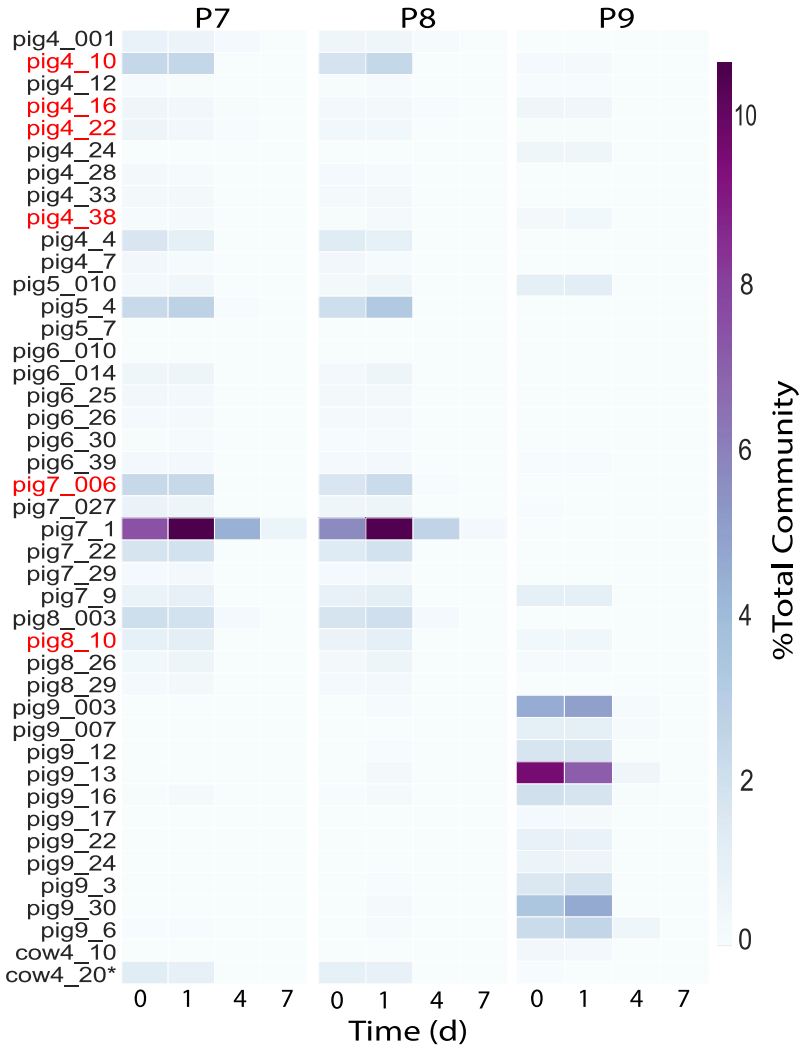


Figure 3-5: Decay kinetics of all host fecal MAGs (rows) that could be detected in the pig mesocosms (columns). The MiGA TypeMAT/NCBI taxonomic identifications of the MAG names (Chapter 3 Supplementary data file S1) are not included due to space limitations. Cow4_20 clustered into a single genomospecies at >95% ANI with a pig MAG that is not shown here because MAG cow4_20 was of higher quality and was used in all downstream analyses instead of the MAG obtained from the pig metagenome. P7, P8, and P9 are the different biological replicate mesocosm metagenomes. Abundances are reported as % of total microbial community (i.e. TAD80 divided by average genome sequencing depth). The MAGs identified as potential MST biomarkers have red labels.

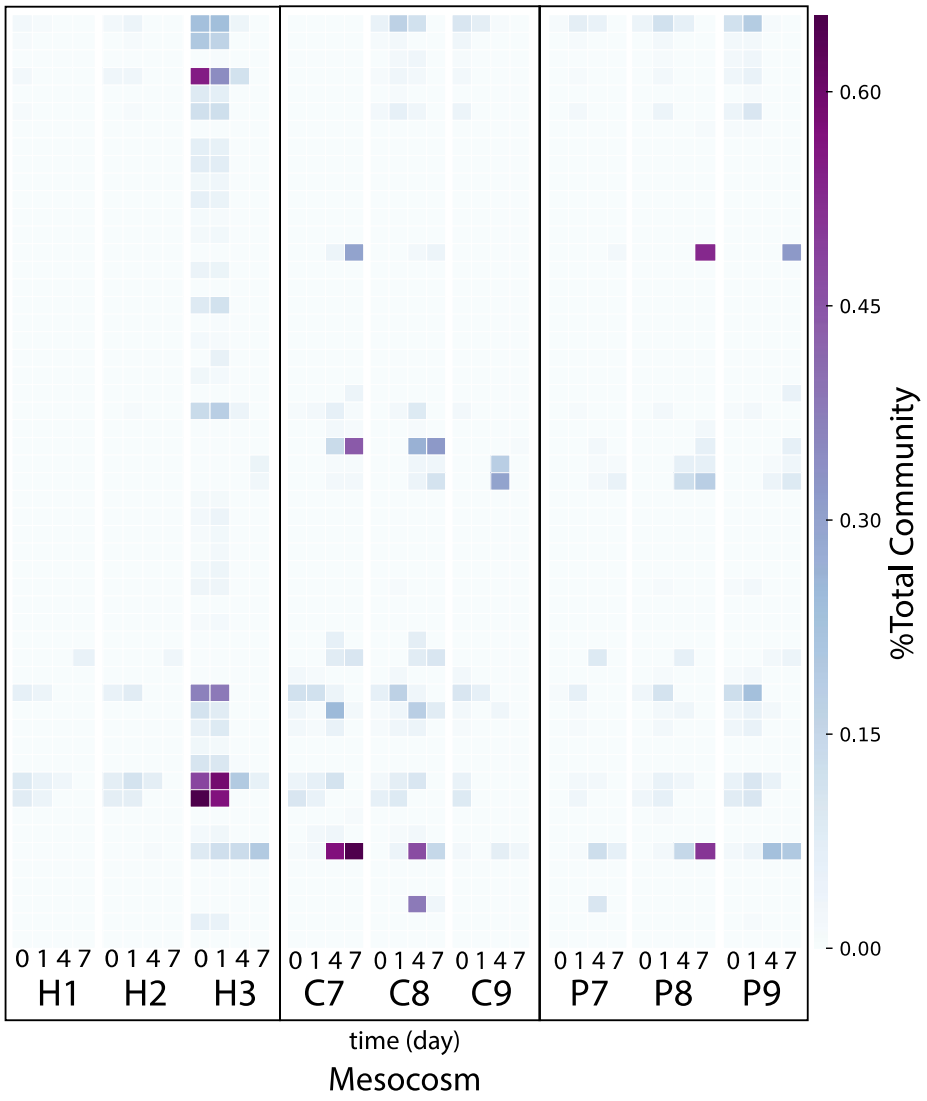


Figure 3-6: Abundance kinetics of Lake Lanier (LL) MAGs in the fecal mesocosm samples over time. The collection of 477 LL reference MAGs from Rodrigues-R 2019 was searched against the time-series mesocosm metagenomes and the abundance for 59 MAGs that could be detected are shown as individual rows in the heatmap. Each column is a mesocosm MG sample and the 0, 1, 4, or 7 refers to the sample time in days while the H, C, or P refers the human, cow or pig biological replicate mesocosm (e.g. the "0" column above H1 is hum1 mesocosm at day 0). Abundance is expressed as % of total bacterial community (i.e. TAD80 divided by average genome sequencing depth).

3.4.5 Functional annotation for MAGs identified as potential biomarkers and differentially abundant (DA) functions between host fecal metagenomes

The 14 MAGs identified as potential markers based on their host sensitivity, abundance, and decay kinetics in the mesocosms were functionally annotated and summarized into KEGG modules using MicrobeAnnotator (Ruiz-P et al. *in review*). There was an average of 2,175 genes in each MAG, of which roughly 48% could be annotated to the KEGG ontology database (2019-04-09 release; Aramaki et al 2019). Most of the KEGG modules identified in the MAGs were related to carbon and amino acid catabolism and the reductive pentose phosphate pathway for carbon fixation (Appendix A, Figure A 6**),** suggesting the potential for primary production in addition to fermenting carbon sources derived from host's diet. Common fermentation-related genes were present such as: fumarate reductase and succinate dehydrogenase, as well as modules for methanogenesis from methanol and methylamine, but there was no clear clustering of the MAGs by host type when examining the KEGG modules overall and no modules were clearly unique to a single host type (Appendix A, Figure A 6). Thus, DESeq2 differential abundance analysis (Anders and Huber 2010) was used at the gene level to identify specific functions that are DA in the host fecal metagenomic assemblies, which allowed for more comprehensive and sensitive functional comparisons of the host fecal communities (i.e. not restrict the analysis to only the functions that were binned into a MAG). Furthermore, the number of host fecal metagenomic reads that mapped to the collection of high quality fecal MAGs was greater than 2X different between samples (data not shown), which indicated high likelihood of false-positive results at the MAG level (i.e., finding DA functions by chance due to

differences in coverage; Rodriguez-R and Konstantinidis 2014). The gene level analysis circumvents this limitation.

Predicted ORFs from the assembled contigs of the host fecal metagenomes were annotated against the KEGG database similarly to the MAGs. There was an average of 107,588 ORFs per assembly, of which ~30% could be annotated with functions other than hypothetical or unknown. The metagenome short reads were mapped against the predicted ORFs to determine their abundances in each metagenome. Of the 2,080 total KEGG functions identified, 177 were significantly DA with $P_{adj} < 0.05$ and $\log_2$ fold change (L2FC) > 3 using pairwise comparisons between human, cow, and pig fecal samples. These 177 functions were manually grouped into 39 broader functional categories (Chapter 3 Supplementary data file S2) for visualization (Figure 3-7). Overall, it was not common for a single function to be highly abundant in one host gut and completely absent in the other two; rather there were functions that appeared to be significantly more abundant in a single or two host types. More specifically, the cow fecal assemblies were more abundant in genes related to biofilm formation, starch and sucrose metabolism, and maltose, urea, and putrescine transport. The pig fecal assemblies were more abundant in genes for amino acid (particularly lysine) degradation, ribosomal proteins and a ribonucleoside-diphosphate reductase gene (NrdB). Human fecal assemblies were enriched in ribose transport, biotin metabolism, and quorum sensing genes. The cow and pig assemblies were more abundant in biosynthesis pathways for amino acids and secondary metabolites, metabolism of cofactors and vitamins, and particularly eight genes for a type IV secretion system (T4SS) related to conjugation (TrbBCDEFGIL), which were absent in the all human fecal assemblies, except for TrbL, which was only identified in hum3 at low relative abundance

(94 reads matching; Chapter 3 Supplementary data file S2). Notably, the cow and pig assemblies were also enriched in methanogenesis genes associated with the CO_2 pathway (Tas et al. 2018) (FmdE; formylmethanofuran dehydrogenase subunit E) and the acetate pathway (AcsD; acetyl-CoA decarbonylase/synthase complex subunit delta) that were absent in the human assemblies. Human and cow assemblies had more genes related to zinc transport, lipid metabolism, and the T6SS secreted protein VgrG compared to pigs, whereas human and pig assemblies were more enriched for nitrogen, ascorbate and aldarate metabolism compared to cows (Figure 3-7). Despite being strictly an anaerobic environment, the cow and human samples were more abundant for catalase, which may indicate that these guts are more prone to aerobiosis e.g., from rapid biomass growth or infected epithelial tissues in the GI tract (Brioukhanov and Netrusov 2007) compared to the pig gut.

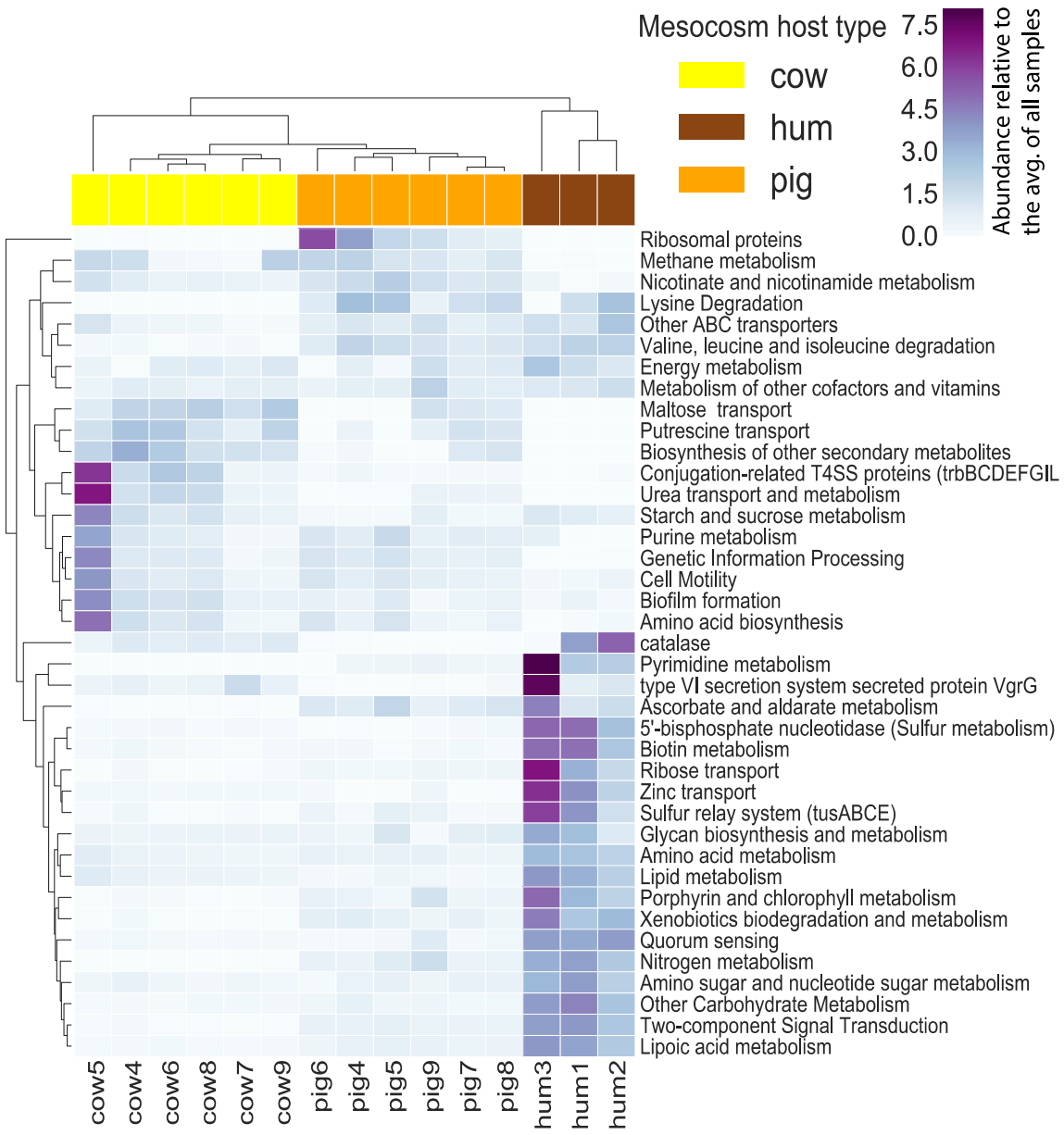


Figure 3-7: Gene functions enriched in the cow, pig, or human fecal metagenomes. The heatmap shows the KEGG functions (rows) that were differentially abundant between the different host types (columns) with $P_{adj} < 0.05$ as determined by DESeq2 analysis. Color scale indicates the abundance relative to the average across all metagenome samples.

Of the 177 significantly DA KEGG genes identified by DESeq2, 137 were also present in at least one of the putative biomarker MAGs described above. An average of 27 ±9, 23 ±18, and 10 ±9 DA genes were found in the human, cow, and pig putative biomarker MAGs, respectively. However only seven of these were found in all three human MAGs and none were found in all five cow or six pig MAGs, indicating that the host-specific functions are not shared consistently among the corresponding, host-specific MAGs. Alternatively, incomplete MAGs could account for these findings, e.g., these MAGs ranged between 70-95% in their completeness estimate (Chapter 3 Supplementary data file S1). Furthermore, two of the pig MAGs (pig7_006_Acetobacteroides_C and pig8_10_Acetobacteroides_C) only had only two DA KEGG genes, one of which was a glycine dehydrogenase subunit I gene (present in 4/6 MAGs). The 7 DA KEGGs that were found in all 3 human MAGs included a lactaldehyde reductase, hydroxylamine reductase, transaldolase, and several genes related to vitamin B12 synthesis. The most common KEGG gene present in the cow MAGs was pullulananse, which was present in 4 of 5 cow MAGs. Together, these results reinforce that most of the putative host-specific MAGs are robust targets because they contain genes that were also DA at the assembly level.

3.4.6 Comparisons to the reference MST marker genomes

3.4.6.1 Decay of potential biomarker MAGs and reference FIB and MST genomes in the mesocosms over time:

The absolute abundances (cells or viral particles/mL) of common FIB and MST biomarkers over time were also checked against the MAGs identified as potential host-specific biomarkers. The latter biomarkers included reference genomes associated with the qPCR

assays used in this study (Table 3-1) as well as genomes of *Escherichia coli* (NC_009800.1) and CrAssphage (JQ995537). None of the reference genomes used here were detectable in any of the LL negative controls and consistent with the EF16S qPCR results, the *E. faecalis* reference genome had TAD80 of zero (ND) in all of the mesocosm metagenomes sequenced in this study (data not shown). The *E. coli* genome had poor host specificity as it was detected in all host mesocosm metagenomes (Appendix A, Figure A 12A) and maintained higher abundances over time compared to the fecal MAGs (Figure 3-8). The *B. dorei* and CrAssphage genomes showed good host specificity as they were not detected in any cow or pig mesocosm metagenomes and they also had similar decay profiles to the human fecal MAGs (Figure 3-8 A&B), except the CrAssphage genome abundance increased from D0 to D1, whereas the *B. dorei* genome (and fecal MAGs) abundance consistently decreased with time, which could possibly indicate the predatory relationship between these two microbes (i.e. the CrAssphage is predating on the *Bacteroides*) (Figure 3-8 A&B**)**. Further, consistent with the qPCR results, these genomes had imperfect host sensitivity as neither were detected in any of the H2 mesocosm metagenomes. *Bacteroides* abundance based on the HF183 qPCR assay tended to be lower than the MAGs and B. dorei reference genome (see below). The human mitochondrial genome (mtGenome) was detected in all three human mesocosm metagenomes until D7 and showed a steady decay in abundance with time (Fig 3-8 A-B and Appendix A, Figure A 12B), consistent with the HUMmt assay, which was detectable by qPCR until D14 (Figure 3-1 C). The cow fecal MAGs were all ND by D7 and decayed faster compared to the *E. coli* reference genome and *Bacteroides* abundance based on the RumBac qPCR assay (Figure 3-8 C). Overall the cow biomarker MAGs showed decay kinetics that were similar

to those observed for other known strict anaerobes (i.e. *Bacteroides*) and is consistent with the functional annotation results.

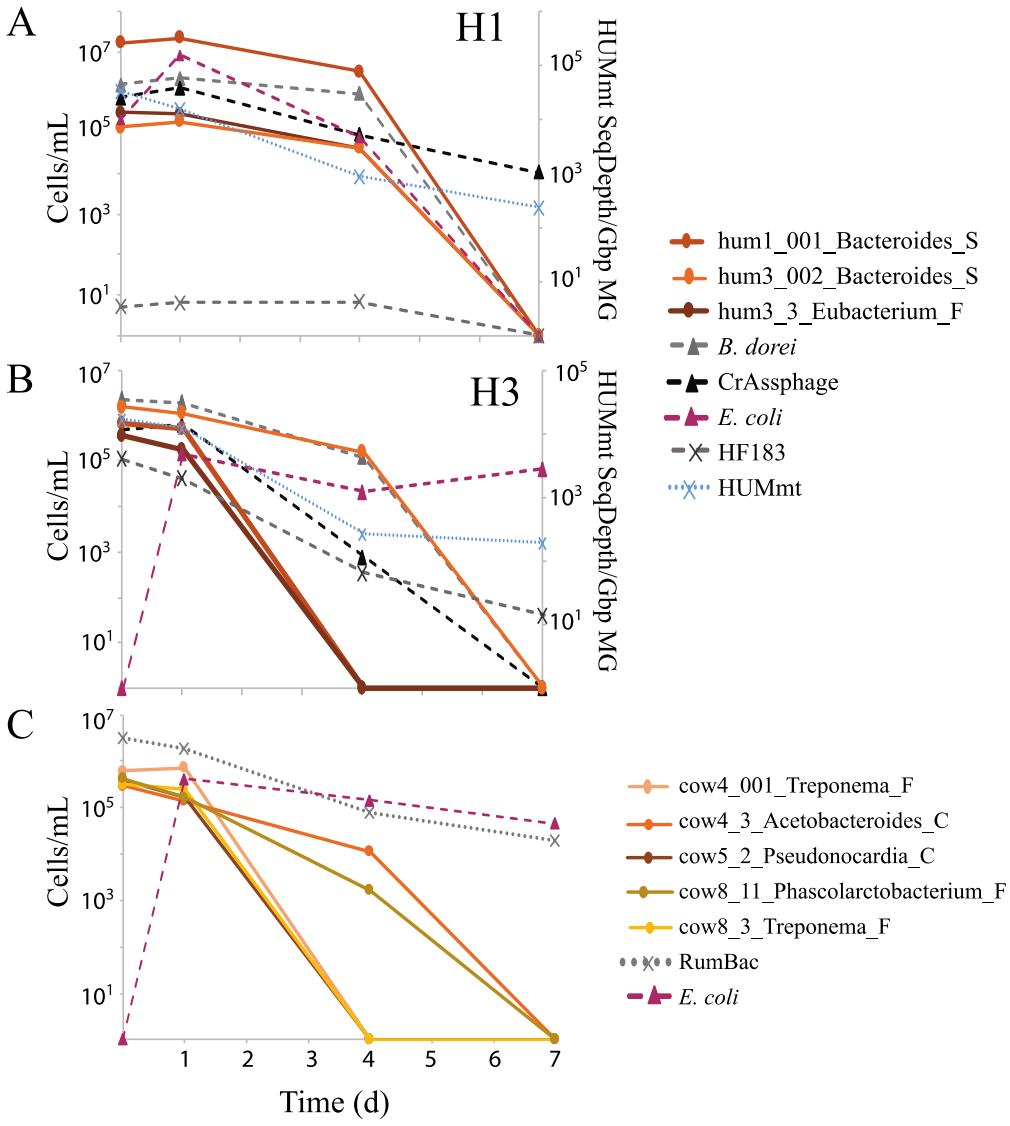


Figure 3-8: Compare absolute abundances of putative biomarker MAGs, traditional FIB and MST qPCR markers in (A) H1 mesocosms, (B) H3 mesocosms, and (C) the average of all 3 biological replicates of the cow fecal mesocosms. Absolute abundances (cells or viral particles per mL) were determined for all targets except for the human mitochondrial assay (HUMmt), which is expressed as relative abundance (SeqDepth/Gbp

MG is mtGenome sequencing depth per Gbp metagenome) and is shown on the secondary axis for (A) and (B). MAGs are represented by solid lines with circle markers. Reference genomes are represented by dashed lines with triangle markers and include *Bacteroides dorei*, CrAssphage, *Escherichia coli*, and a human mtGenome. The qPCR assays are represented by dotted lines with X markers and included the human-specific and ruminant specific *Bacteroides* assays (HF183 and RumBac, respectively) and the human mtDNA assay. The human mesocosms are plotted separately because they were more variable among each other compared to the cows and also because neither *B. dorei*, CrAssphage, or HF183 were detected in any of the H2 mesocosms. Thus, H2 is not shown here.

3.4.6.2 Correlation of MST qPCR markers to their metagenome counterpart:

In order to evaluate the performance of the qPCR assays against the metagenome-based results, we compared the abundances of MST targets against metagenomic abundances of the same target. Absolute abundance (expressed as cells/mL) of the RumBac and HF183 *Bacteroidetes* assays were compared to the abundance of the corresponding reference genome in the mesocosm metagenomes. Specifically, for HF183, the absolute abundance (cells/mL) of the *B. dorei* reference genome (Table 3-1) was estimated in the human mesocosm metagenomes by multiplying *B. dorei* sequencing depth and total cell density per metagenome (from the GenBac16S assay) as described in the Materials and Methods section. Since there is no known reference genome for the RumBac assay, a 317bp contig from a cow fecal metagenome with a perfect match to the assay oligos (cow5_scaffold246842) was used as a proxy to estimate the absolute abundance of

Bacteroides over time in the cow mesocosm metagenomes. The correlation between *Bacteroides* abundances based on qPCR counts and metagenomes was not consistent between the two assays (Figure 3-9 A&B; R^2=0.18 and 0.76 for HF183 and RumBac, respectively). The RumBac qPCR assay tended to give higher abundance estimates (linear regression slope = 0.16) than its metagenome counterpart, especially at the earlier time points (D0 and D1), but estimates became more similar to each other by D7, when both counts were lower (Figure 3-9 B). The HF183 qPCR assay consistently under-estimated the abundance of *Bacteroides* in the human mesocosms (linear regression slope= 10.26), especially in H1, in which the HF183 qPCR assay estimated only about 6 *Bacteroides* cells/mL in the mesocosms on D0, D1, and D4, well below the theoretical LOD for *B. dorei* in the metagenomes (~$3x10^4$ cells/mL; see Materials and Methods for LOD estimation**)**. However, the *B. dorei* reference genome was well above this concentration based on metagenome abundance (Figure 3-9 A). Further investigation showed that this was presumably caused by mismatches of the forward HF183 primer to the dominant *Bacteroides* strains present in the host fecal inocula (Appendix A, Figure A 11). Specifically, the short reads from the fecal inoculum were searched against the 16S rRNA gene of the reference *B. dorei* strain (which contains a perfect match to the HF183 assay primers and probe) to calculate its 99% identity truncated average sequencing depth (TAD80). For both hum1 and hum3 fecal metagenomes (there was no detection in hum2), the sequencing depths of the probe and reverse primer were similar to the overall average sequencing depth for the entire 16S rRNA gene (at about 42.0 and 247.0 for hum1 and hum3, respectively). However, the sequencing depth of the forward primer region was 0 in hum1 and ~40 (6x less than the average) in hum3. Furthermore, we manually checked the

metagenomic reads for perfect matches to the HF183 forward primer and found none in hum1 and only 17 in hum3, suggesting that this region is not present in the dominant *Bacteroides* strain that was assemble-able from each host.

Since it was not possible to estimate the human mtGenome copy number per cell, the HUMmt qPCR assay counts per mL were compared to the relative abundance of a reference human mtGenome (Table 3-1) expressed as sequencing depth per metagenome size (in Gbp) and a weak, but significant correlation (R^2=0.29) was observed (Figure 3-9 C). Together the results for the human-specific assays were consistent with known limitations for interpreting qPCR results, namely, host-sensitivity issues for HF183 and lack of absolute abundance estimates for HUMmt.

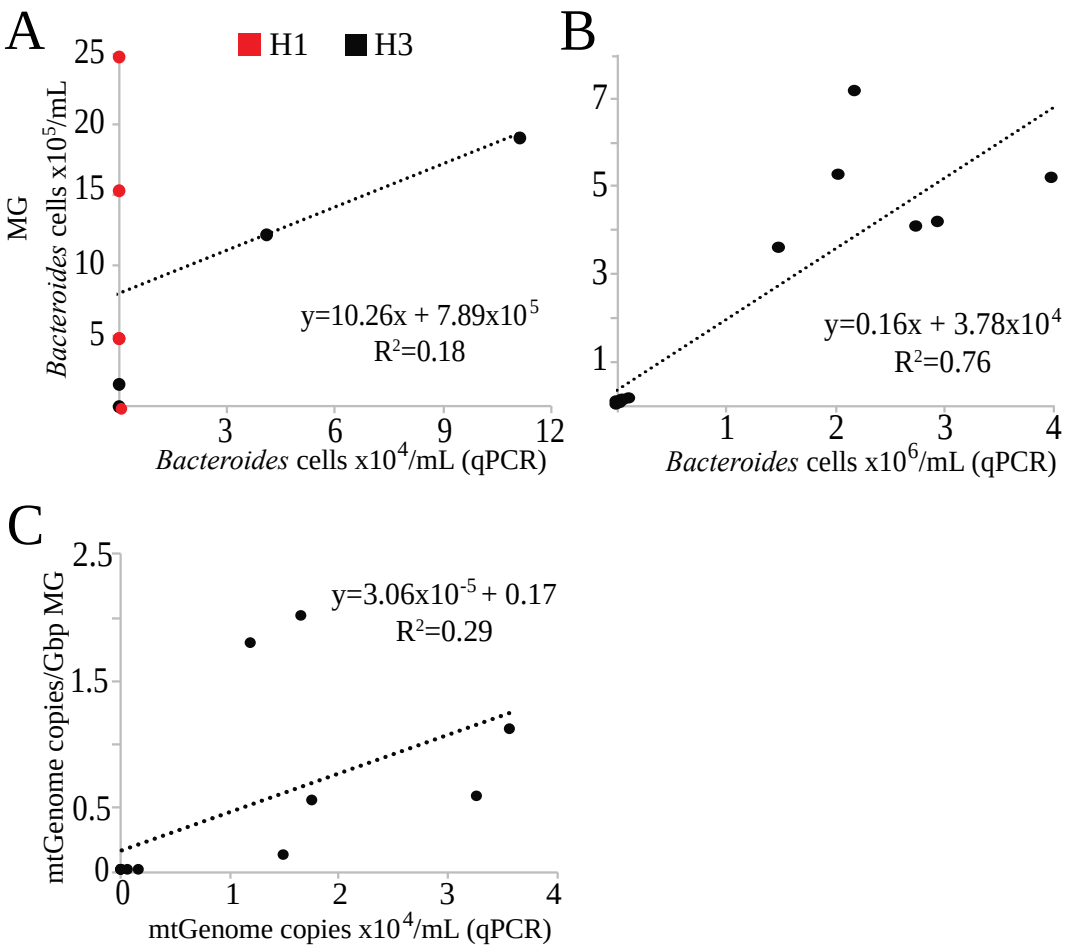


Figure 3-9: Correlation between qPCR and metagenome-based abundance estimates of MST markers and their reference genome counterparts. (A) Human-specific *Bacteroides* 16S (HF183) versus the absolute abundance of the reference genome *B. dorei* in the human mesocosm metagenomes. (B) Ruminant-specific *Bacteroides* 16S (RumBac) versus the absolute abundance of a contig recovered from the cow fecal inocula metagenomes carrying a perfect match to the RumBac assay in the cow mesocosm metagenomes. Absolute abundances for (A) and (B) are expressed as the number of *Bacteroides* cells/mL. (C) Human mtDNA (HUMmt) expressed as the number of gene copies per mL versus the relative abundance of a reference human mtGenome in the metagenome (expressed as TAD80 divided by library size in Gbp).

3.4.7 Bottle effect in mesocosms on D7

There was evidence of a bottle effect in the dialysis bag mesocosms that became apparent at the D7 time point. Nonpareil results (Rodriguez-R et al. *mSystems* 2018) showed an increase in overall community coverage (and thus a decrease in community diversity) over time for the dialysis bag samples that reached ~76% on D7 for all samples, including the negative controls (Appendix A, Table A 1). Furthermore, the average genome size as determined by MicrobeCensus (Nayfach and Pollard 2015) also showed an increase with time (Appendix A, Figure A 1). The high coverage in the D7 mesocosm samples indicated that it was possible to recover some MAGs from these samples (as opposed to the earlier time points, which were characterized by too high diversity based on Nonpareil to expect good assemblies or MAGs). Contig binning from the D7 samples resulted in 39 high quality MAGs that dereplicated into 17 genomospecies at 95% ANI. These MAGs appeared to be highly different from those assembled from the host fecal communities based on pairwise AAI comparison (Appendix A, Figure A 9**)** that showed none of the D7 MAGs formed a genomospecies (ANI >95%) with any of the host fecal or Lake Lanier (LL) MAGs. Taxonomic classification of these MAGs using MiGA and the TypeMAT/NCBI database showed that the closest relative to three of these MAGs (AAI ~47%) was *Methylobacterium platani* and the closest relative to four other MAGs (AAI ~66%) was *Cellvibrio japonicus*; two additional MAGs were classified in the family *Cytophagaceae* (Chapter 3 Supplementary data file S1). Therefore, these D7 MAGs were, in general, more related to each other than the host MAGs. Furthermore, none of the D7 MAGs were classified as class *Bacteroidia* or *Clostridia*; the most common taxa among the fecal MAGs. When looking at the abundance of the D7 MAGs overtime in the

mesocosms, they were not detectable in any of the mesocosm samples at D0 or D1 and only began to be detected in a few samples by D4 (Appendix A, Figure A 7**).** However, on D7, they increase to 30-50% of the total community in some cases and, accordingly, the majority of the metagenome short reads from D7 samples also mapped to these MAGs (Appendix A, Figure A 2). The LL MAGs in the negative control dialysis bags tended to decay over time (Appendix A, Figure A 8A) and were mostly undetectable by D7, when the D7 MAGs started to increase (Appendix A, Figure A 8B). These results suggest that the bottle effect in the D7 samples was likely consistent across all mesocosms and was likely not very substantial before D4.

The results of the D7 MAGs were further confirmed when comparing the relative abundances of the genes recovered in the D7 assemblies against the genes of the host fecal assemblies using DESeq2. Out of 2,906 total KEGG functions detected, 582 were significantly DA with $P_{adj} < 0.05$ and L2FC > 6 that were manually grouped into broader functional categories as described in (Chapter 3 Supplementary data file S3). There were many more significantly DA genes when comparing the D7 samples to all of the animal host fecal metagenomes and the differences were much greater and more distinct than in the animal host only comparisons described above. The D7 samples were particularly enriched for genes related to aerobic processes (Appendix A, Figure A 10 and Chapter 3 Supplementary data file S3) such as cytochromes and sulfide oxidation genes. Furthermore, the D7 samples were enriched for genes related to photosynthesis and porphyrin and chlorophyll metabolism and biofilm formation. In contrast, the host fecal metagenomes were more enriched in genes for anaerobic processes like methanogenesis, acetylaldehyde/alcohol dehydrogenase, sulfur and fumarate reductases, nitrogenase

(NifH), glycerol dehydrogenase, butyrate kinase, and phosphate butyryltransferase as well as ABC transporters for sugars and amino acid and cobalamin biosynthesis genes. These results suggested that the populations arising by D7 are likely "weed" species from the rare biosphere of the lake water used in the mesocosms that were able to form biofilms on the material of the dialysis bags and perform aerobic metabolism.

3.5 Discussion

In this study we used metagenomics to track the decay of cow, pig, and human fecal MAGs in laboratory mesocosms simulating a fecal pollution event in a freshwater habitat. For the conditions tested here, the majority of fecal MAGs from all three hosts were not detectable in the mesocosm metagenomes after D4 (Figures 3-4 & 3-5), which implies that the associated risk from fecal pollution was also presumably significantly diminished by D4. This result was consistent with a previous quantitative microbial risk assessment (QMRA) analysis that predicted that the gastrointestinal infection risk from sewage contamination in surface waters is not significant (<3% chance of infection) after 3.3 days (Boehm et al. 2018) in accordance with the EPA risk threshold for bathing water (USEPA RWQC 2012). Therefore, it appears recent fecal pollution events can be reliably detected up to 3-4 days post-event, and the risk for public health may not be high after four days, at least for natural freshwater environments like the conditions tested here.

Although the persistence of the fecal MAGs was similar across all three hosts, the MAGs were predominantly host-specific at the species level (i.e., >95% ANI) and we were able to identify several MAGs from each host as potential biomarkers for MST based on the host-specificity and abundance profiles during mesocosm incubation (Figures 3-4 & 3-

5). Notably, most of these MAGs represent novel species of existing genera (Chapter 3 Supplementary data file S1) and will be important to study in more detail in the future. The genome sequences recovered here should facilitate such studies. Notably, there was some overlap among the different host fecal MAGs at the genus level (>65% AAI; Figure 3-3), which could account for the cross-reactivity commonly observed for the various 16S qPCR assays targeting *Bacteroidales* at above the species level (Boehm et al. 2013; Harwood et al. 2012; Ahmed et al. 2016).

Consistently, the majority of MAGs in all hosts, including those identified as putative biomarkers, were within the classes *Bacteroidia* and *Clostridia*, however a few of the biomarker MAGs were classified to other taxa that were unique to cows or pigs. For example, two of the four cow putative biomarker MAGs (cow4_001_Treponema_F and cow8_3_Treponema_F) were from the *Spirochaetia*, while an *Actinobacteria* (pig4_16_Cellulomonas_C) and *Euryarchaeota* (pig4_38_Methanoplasma_F) were among the six pig biomarker MAGs (Chapter 3 Supplementary data file S1). Together, these results suggest that better and perhaps more host-specific biomarkers may be found in novel taxa that have not yet been considered for MST. Moreover, phenotype classification using Traitar (Weimann et al. 2016) showed that none of the potential biomarker MAGs were aerobes and all had mostly anaerobic phenotypes related to carbohydrate fermentation (Appendix A, Figure A 3, A 4, A 5). Accordingly, the best gene targets for MST assay development will likely be related to anaerobic functions specific to the different host types rather than the 16S rRNA gene, which has primarily been the target of most MST research to date.

E. faecalis and *E. coli*, despite being the "gold standard" FIB, performed worse than the MST markers assessed here. *E. faecalis* was not detected in any human feces or mesocosm samples by qPCR or metagenome-based methods and had too low abundance in the cow and pig feces to be detected upon dilution in lake water for the mesocosms. Hence, this organism would not be able to detect fecal contamination for any of the hosts in this study. *E. coli* was detected in all of the host mesocosms and persisted for ~1 week (Figure 3-8 A), longer than the presumed fecal contamination risk of 4 days described above. The longer persistence under the oxic conditions of our mesocosms is presumably consistent with the different physiologies of *E. coli* (facultative anaerobe) relative to the MAGs (mostly fermenters and methanogens) as revealed by the MAG sequences. Although it is well known that these organisms are not host specific and thus, not suitable for MST, these results confirmed our expectations and provided further evidence against the use of these organisms as FIB and the need for improved standard indicators.

Furthermore, we compared traditional qPCR-based abundances of common MST markers to metagenome-based methods leveraging total cell densities from GenBac16S qPCR counts to estimate absolute abundances (i.e., cells/mL) of the corresponding MST reference genome in the metagenomes. The ruminant-specific *Bacteroidetes* 16S assay, RumBac, consistently over-estimated the abundance compared to the metagenome-based methods. This could be due, at least partly, to the small amplicon size (118 bp), which is typical for most qPCR assays and could be detectable even in highly degraded DNA from dead cells (Bae and Wuertz 2009). In contrast to RumBac results, the human-specific HF183 assay was poor correlated and consistently under-estimated the abundance of *Bacteroides* compared to MG-based methods. This was most obvious in the H1 mesocosms

where qPCR-based estimates reported < 7 *Bacteroide*s cells/mL whereas metagenome-based estimates showed 0.5-3x10^6 *Bacteroides* cells/mL in D0, D1 and D4 samples. Our further investigation revealed a single contig in each of the hum1 and hum3 fecal assemblies that carried a perfect match to the HF183 assay reverse primer and probe. However, both of these contigs had mismatches in the forward primer region (Appendix A, Figure A 11). Neither of these contigs were binned into any of the fecal MAGs, which is not surprising because 16S-carrying contigs are often problematic for the abundance-based binning methods used here. When examining the short reads, there were no perfect matches in the hum1 fecal metagenome and only 17 reads carried a perfect match to the forward HF183 primer in the hum3 fecal metagenome. Together, these results suggest that the dominant *Bacteroides* populations in hum1 and hum3 do not contain a perfect HF183 forward primer match and presumably there is a population in hum3 with a perfect match whose abundance was too low to show up in any of the assembled contigs. These findings most likely accounted, at least in part, for the low HF183 qPCR-based *Bacteroides* abundance estimates compared to the metagenome-based estimates (i.e. less or no exponential doubling because PCR amplification is only happening at the reverse primer).

Both cases mentioned above exemplified the known limitations of using qPCR assays, which require small amplicon sizes (~100-200 bp), and how whole genome-based metagenome estimates for MST markers can be advantageous with this respect. However, targeting a single marker (whether it be a whole genome or qPCR assay) for MST can still be inadequate for water quality monitoring because of the high inter-person variability observed in the human gut. Neither the HF183 assay nor the *B. dorei* reference genome were detected in any of the H2 mesocosms (or the hum2 feces), thus these water samples,

although highly polluted with human feces in our experimental set up, would have been considered safe for public health if using only this *Bacteroides* marker. Evidence for inter-person variability/specificity was also evident when looking at the human fecal MAG abundances in the mesocosms (Figure 3-4 A). The most abundant MAGs in each human mesocosm corresponded to the human fecal samples that was used as inocula (e.g., hum1 MAGs were the most abundant in H1 mesocosms). Typically, the most abundant MAG in each mesocosm was ~10x more abundant than MAGs from the other two human fecal samples on average. Although both metagenomics and qPCR have unique benefits and limitations, the results presented here suggest that using a single marker only to assess human fecal pollution was inadequate for either method. This result is not so surprising considering previous studies of human gut communities have revealed extensive diversity among individuals (Costello et al. 2009; Garud et al. 2019) and intra-person temporal variability within the human gut microbiome, which suggests that no core human gut microbiome exists for the abundant taxa in each microbiome (Caporaso et al. 2011). Although the relative abundance and taxa can vary overtime, significant functional redundancy has been observed previously (Moya and Ferrer 2016, Li et al. 2014) as well as among our human MAGs (Figure 3). Therefore, the use of functional genes as opposed to the 16S rRNA gene or individual taxa (e.g., MAGs) for biomarkers may be more robust due to the high prevalence of some gene functions among individuals of the same host type, presumably driven by the host-specific gut physiology. Metagenomics, as shown in this study, can be useful in identifying these genes to be used as novel targets for more robust qPCR assays as well as elucidating the fecal signal in environmental (or mesocosm) samples relative to uncontaminated samples.

Although it was not common for any gene function to be highly abundant in one host and completely absent from the rest, there were several functions that were significantly enriched in one host compared to the others and could be targets for biomarker development. These patterns, and the accompanying high host-specificity of the MAGs recovered, are presumably driven, at least to some extent, by the different selection pressures prevailing in the gut of each animal, as also indicated by the type of fermenters present in the different hosts. Most notably, seven genes for a type IV secretion system (T4SS) protein (TrbBCDEFGI) were absent in the all human samples and present in both the cow and pig samples, with the highest abundance found in the cows (Figure 3-7). Evidence has shown that T4SS proteins are important for shaping community composition in the gut (Verster et al. 2017), which suggests these proteins could be viable targets for host-specific markers (especially in cows). Furthermore, some of the DA KEGG functions offered new insight on the fermentation pathways that distinguish cows and pigs. Fumarate reductase subunit D (FrdD), which is associated with the primitive electron transport chain (ETC) of some fermenters (Besten 2013), was more abundant in cows. The pig samples were instead enriched for two genes associated with butyrate-producing (AtoA; acetoacetate CoA-transferase beta subunit, Bcd; butyryl-CoA dehydrogenase) as well as H_2-producing (porD; pyruvate ferredoxin oxidoreductase delta subunit) fermentations (Chapter 3 Supplementary data file S2). These results indicated that fermenting microbes inhabiting the cow and pig gut carry out different strategies to sink excess reducing equivalents (the primitive ETC or H_2, respectively). However, these trends were not discernable for the human samples as fewer genes overall tended to be significantly more abundant in human inocula, which could be the result of sampling limitation (only 3 human

fecal samples were compared against 6 cow and 6 pig samples) and the higher inter-person diversity described above.

There was a bottle effect in our mesocosm incubations that resulted in an increase of some "weed" species by D7. The bottle effect is not apparent until D7; that is, after the point when most of the fecal organisms apparently had died off (i.e., D4). Therefore, this bottle effect is not expected to have a major impact on patterns reported here (e.g., host MAG dynamics) or our conclusions because the latter were primarily based on MAG abundance dynamics during the first four days. This is consistent with other similar studies such as in Ahmed et al. 2018, where sewage OTUs were not detected after four days, and found that the mesocosms did not return to the initial community composition even after 50 days (Ahmed et al. *Appl. Micro. & Biotech*. 2018). Consistently, we did not see a return to starting community within the duration of our experiments (7 days). Mattioli et al. 2018 also reported bottle effects from dialysis bag mesocosms (e.g., lower nutrient concentrations and higher chlorophyll a) that did not arise until after five days post-perturbation (Mattioli et al. 2018). These results confirm that delayed bag effects are a common limitation of the dialysis bag mesocosm method and that experiments need not be carried out for longer than 5 days or transferred to new bags. Further, it would be important to test these findings with field samples from recent pollution events to further corroborate the abovementioned conclusions.

Considering the high individual host variability, especially among human hosts, more work is needed to characterize the geographic stability of the putative biomarkers of human or animal hosts reported here and the degree of their biogeography. Most importantly, whether the biomarkers reported here may be universally applicable as opposed to only

locally useful (geographically). Many recent studies have made considerable effort to sequence metagenomes and/or assemble MAGs from cow rumen (Wilkinson et al. 2020, Almeida et al. 2020, Wang et. Al 2019, Stewart et al. 2019) as well a pig (Xiao et al. 2016, Wang et al. 2019) and chicken guts (Gilroy et al. 2020). However, this information has not yet been synthesized together for MST marker development. Future work should leverage these datasets to improve comparative functional gene analysis along with decay information to data mine for better DNA markers. Furthermore, as high-throughput sequencing becomes more affordable and routine, it may be possible to directly assess MST markers (and even pathogens) in environmental metagenomes. In order to make regulatory standards based on metagenome data, calculating absolute abundances of indicators (or pathogens) will be necessary. The methodologies proposed here should be helpful in these directions.

3.6 Acknowledgements

The authors would like to thank Dr. Robert Dove from the University of Georgia-Athens for providing the cow and pig fecal samples. This work was supported by the US National Science Foundation, award numbers 1511825 (to J.M.B and K.T.K) and 1831582 (K.T.K.), and the US National Science Foundation Graduate Research Fellowship under grant number DGE-1650044 (to B.S). The funding agencies had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CHAPTER 4. METAGENOMICS AS A PUBLIC HEALTH RISK ASSESSMENT TOOL IN A STUDY OF NATURAL CREEK SEDIMENTS INFLUENCED BY AGRICULTURAL AND LIVESTOCK RUNOFF: POTENTIAL AND LIMITATIONS

Brittany J. Suttner, Diana Carychao, Eric R. Johnston, Luis H. Orellana, Luis M. Rodriguez-R, Janet K. Hatt, Michelle Q. Carter, Michael B. Cooley, and Konstantinos T. Konstantinidis

Originally published in AEM on March 20th 2020, doi: 10.1128/AEM.02525-19

4.1 Abstract

Little is known about the public health risks associated with natural creek sediments that are affected by runoff and fecal pollution from agricultural and livestock practices. For instance, the persistence of foodborne pathogens originating from these practices such as Shiga Toxin-producing *Escherichia coli* (STEC) remains poorly quantified. Towards closing these knowledge gaps, the water-sediment interface of two creeks in the Salinas River Valley of California was sampled over a nine-month period using metagenomics and traditional culture-based tests for STEC. Our results revealed that these sediment communities are extremely diverse and comparable to the functional and taxonomic diversity observed in soils. With our sequencing effort (~4Gbp per library), we were unable to detect any pathogenic *E. coli* in the metagenomes of 11 samples that had tested positive

using culture-based methods, apparently due to relatively low abundance. Further, there were no significant differences in the abundance of human- or cow-specific gut microbiome sequences in the downstream, impacted sites compared to upstream, more pristine (control) sites, indicating natural dilution of anthropogenic inputs. Notably, a high number of metagenomic reads carrying antibiotic resistance genes (ARGs) was found in all samples that was significantly higher compared to ARG reads in other available freshwater and soil metagenomes, suggesting that these communities may be natural reservoirs of ARGs. The work presented here should serve as guide for sampling volumes, amount of sequencing to apply, and what bioinformatics analyses to perform when using metagenomics for public health risk studies of environmental samples such as sediments.

4.2 Importance

Current agricultural and livestock practices contribute to fecal contamination in the environment and the spread of food and water-borne disease and antibiotic resistance genes (ARGs). Traditionally, the level of pollution and risk to public health is assessed by culture-based tests for the intestinal bacterium, *E. coli*. However, the accuracy of these traditional methods (e.g., low accuracy in quantification, and false positive signal when PCR-based) and their suitability for sediments remains unclear. We collected sediments for a time series metagenomics study from one of the most highly productive agricultural regions in the U.S. in order to assess how agricultural runoff affects the native microbial communities and if the presence of STEC in sediment samples can be detected directly by sequencing. Our study provided important information on the potential for using metagenomics as a tool for assessment of public health risk in natural environments.

4.3 Introduction

Nearly half of the major produce-associated *Escherichia coli* O157:H7 outbreaks in the U.S. between 1995-2006 have been traced to spinach or lettuce grown in the Salinas Valley of California (M. Cooley et al. 2007). Fecal contamination of produce can be caused by exposure to contaminated irrigation or flood water, deposition of feces by wildlife or livestock, or during field application of manure as fertilizer (Mantha et al. 2017; Jay et al. 2007). From a public health perspective, more information is needed on the risk of exposure to animal fecal contamination as recent studies suggest that exposure to water impacted by cow feces may present public health risks that are similar or equal to human fecal contamination. For example, cattle are a reservoir of the major foodborne pathogen, Shiga Toxin-producing *E. coli* (STEC) (Soller et al. 2010; Probert, Miller, and Ledin 2017). Environmental contamination by animal feces from farms is an emerging public health issue not only as a source of pathogens but also as a source of antibiotic resistance genes (ARGs) (WHO 2014). Antibiotics are regularly administered to livestock at prophylactic concentrations to prevent infection, and food animal production is responsible for a significant proportion of total antibiotic use (Landers et al. 2012). Such practices are known to contribute to the prevalence of ARGs in the environment (Jechalke et al. 2013; Zhu et al. 2013; Karkman, Pärnänen, and Larsson 2018), which can spread rapidly to other microbes via horizontal gene transfer, including to human pathogens of clinical importance (Walsh et al. 2011; Maal-Bared et al. 2013). Surprisingly, there is very little regulation of antibiotic use in the U.S. livestock industry, even though these operations can be major contributors to fecal pollution and the spread of ARGs in the environment (Durso and Cook 2014; Berendonk et al. 2015).

Our previous culture- and PCR-based surveys of the Salinas watershed, and particularly Gabilan and Towne Creeks (heretofore called GABOSR and TOWOSR, respectively), indicated persistent presence of STEC in water and sediments (M. B. Cooley et al. 2013; 2014) and a potentially significant public health risk. Continued prevalence of STEC in both GABOSR and TOWOSR sites is hypothesized to be linked to the presence of cattle upstream. For instance, in several cases, STEC strains isolated from cattle fecal samples were identical to those found in water and sediment based on Multi-Locus Variable number tandem repeat Analysis (MLVA) typing. Indeed, the prevalence of STEC was strongly correlated with runoff due to rainfall (M. Cooley et al. 2007; M. B. Cooley et al. 2014). However, hydrologic modeling and surveys indicated that pathogen levels in streams were not only due to overland flow, but also to contributions from sediment (Dorner et al. 2006; Petit et al. 2017). These observations were further supported by several examples of identical MLVA types isolated from both water and sediment at the same location or downstream during periods of drought (M. Cooley et al. 2007; M. B. Cooley et al. 2013). Further, the levels of pathogen in the water column and sediment are difficult to measure and are generally underestimated when using culture-based tests due to the predominance of biofilms and viable but not culturable (VBNC) bacteria (M. B. Cooley, Carychao, and Gorski 2018). Determining accurate pathogen levels is also problematic when using culture-independent qPCR tests because these tests may detect small fragments of highly degraded DNA long after the living microbe and pathogens have been inactivated (Bae and Wuertz 2009). Furthermore, PCR methods do not give the complete picture of total functional and/or taxonomic shifts occurring in the sampled microbial communities. Therefore, metagenomic characterization of the creek sediments should provide

independent quantitative insights into the effect of agricultural practices on the surrounding environment.

River and creek sediments are among the most diverse communities sequenced to date and are largely under-sampled (Gibbons et al. 2014; Abia et al. 2018). Moreover, the sediments studied to date are exclusively from highly and/or historically polluted environments with varying industrial or sewage inputs and thus, each sediment is characterized by its unique properties in terms of flow dynamics, chemical environment, climatic conditions and anthropogenic inputs (Abia et al. 2018; Bowen 2011; Xu et al. 2014; Costa et al. 2015; Graves et al. 2016; Negi and Lal 2017; Huber et al. 2018). Accordingly, previous studies on the effect of anthropogenic inputs on sediments in lotic (free-flowing) aquatic systems have yielded mixed results on how surrounding land use practices impact sediment communities or were not directly relevant. Furthermore, in order to properly quantify the effect of anthropogenic antibiotic inputs, appropriate controls (e.g., pristine sampling sites) are needed to determine baseline levels of ARGs and other genes (Durso and Cook 2014; D'Costa et al. 2011).

In this study, we examined the effect of agricultural runoff on microbial communities from creek sediments in the Salinas watershed and whether community structure correlated with precipitation or culture-based detection of STEC. We sampled upstream sites with reduced human and cattle presence as a baseline to compare the abundance of anthropogenic signals (i.e. human and cow gut microbiome and ARGs) observed in the downstream sites that receive inputs from cattle ranches and produce farms. By combining culture-based STEC data with metagenome-based ARGs and animal host microbiome signal, we assessed the effect of cattle ranching run-off on the creek sediments

at multiple, independent levels, providing for more robust conclusions and interpretations. Furthermore, we compared these sites to other publicly-available sediment, soil, and river water metagenomes from both highly pristine and polluted environments in order to validate our results and assess anthropogenic pollution levels relative to other similar habitats.

4.4 Results

4.4.1 Description of sampling sites

Six sites from three creeks in the Salinas River valley in California were included in this study. Two of the sites (collectively referred to as the "downstream" samples/sites) are impacted by cattle ranching but vary in the level of agricultural activities in the directly surrounding area. Cattle have direct access to creeks at both locations and no effort is made to exclude them. At GABOSR, the cattle have access 2.38 km upstream from the sampling location and cattle access for TOWOSR is 0.68 km upstream. The creeks are isolated at the sampling locations but converge further downstream before emptying into the Salinas River. Gabilan (GABOSR) is directly downstream of organic strawberry produce fields that use both green and poultry manure fertilizer and has cattle ranching upstream of the strawberry farm. The second site, Towne Creek (TOWOSR), is roughly 2 Km north of GABOSR but does not have any abutting agricultural fields directly upstream and only receives input from cattle ranches. Ten samples from each of the two downstream sites, GABOSR and TOWOSR, collected over a 9-month period from September 2013 through June 2014 were selected for metagenome sequencing based on precipitation levels and detection of pathogenic *E. coli* via enrichment culture (Table 4-2). An additional seven

samples from four upstream sites (collectively referred to as the “upstream” samples/sites), were included to serve as upstream controls for metagenomic comparison (Table 4-2 and Figure 4-1). The samples from these locations included: three samples collected ~10 km upstream from Gabilan (“GABOSR Control”) on March 2016 (GC1-3); two samples collected ~3 km upstream from Towne Creek (“TOWOSR Control”) on April 2017 (TC1 and TC2); and finally, one sample from each of two sites on the west side of the Salinas River (“West Salinas”), ~60 km and 110 km southeast from the downstream sites collected in May 2017 (WS1 and WS2, respectively). The latter two samples are not upstream of GABOSR or TOWOSR but were included because they are more pristine sites with no known history of cattle impact, as opposed to the GC and TC samples, which may have had minimal inputs from previous cattle grazing.

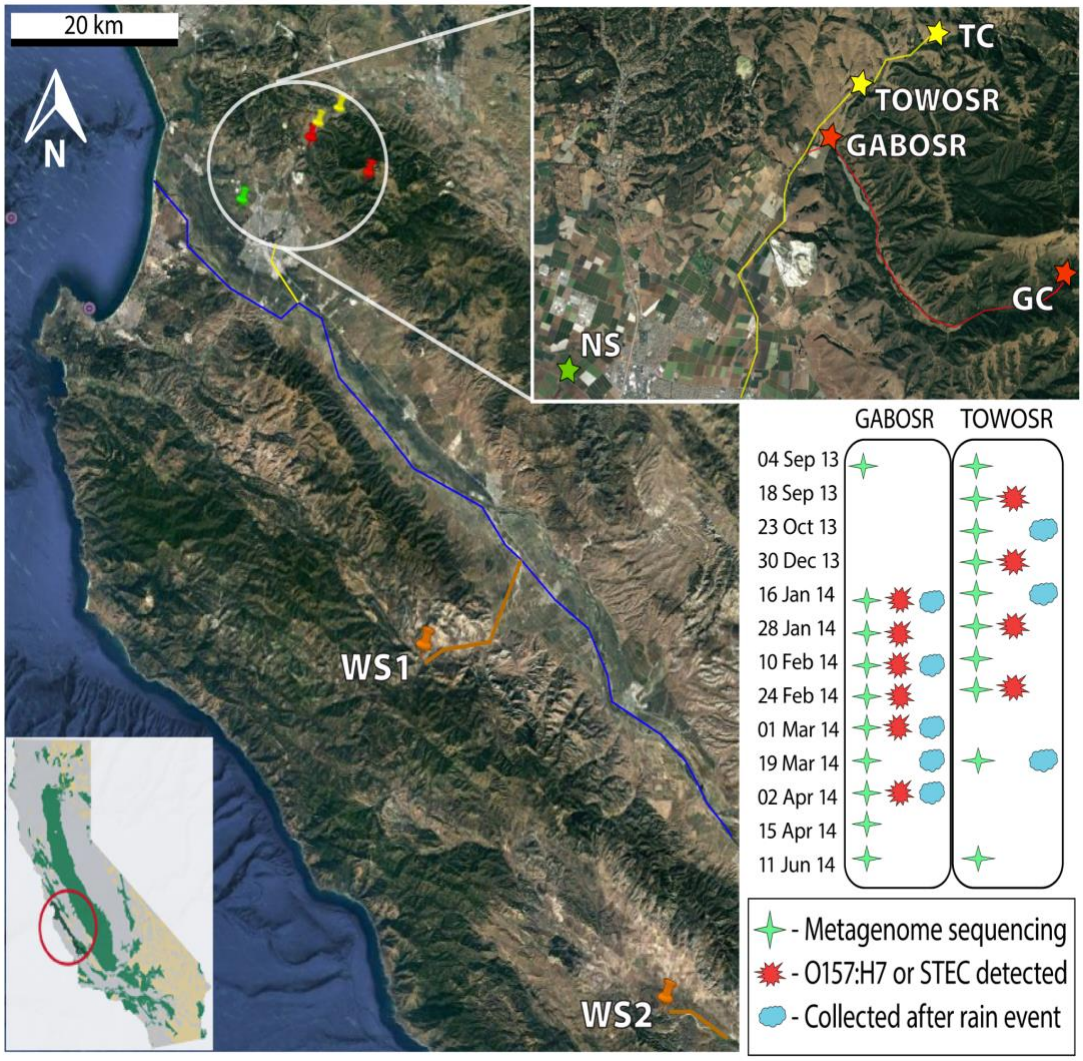


Figure 4-1: Location of sampling sites in the Salinas Valley, CA, and sampling scheme for time series metagenomics. Sampling site for Gabilan (GABOSR in red) and Towne Creek (TOWOSR in yellow). The upstream controls for Gabilan (GC) and Towne Creek (TC) are also indicated by the same colors. The red line shows the flow of the creek from GC to GABOSR, the yellow line shows the flow from TC to TOWOSR and the confluence of the two creeks before flowing into the Salinas river (blue line). Orange pins mark the West Salinas sites (WS1 and WS2) included as less agriculturally-impacted controls. Orange lines show the flow of these creeks from the sampling point to the Salinas River,

except for WS2, whose confluence point with the Salinas River is 70 kilometers upstream from where WS1 creek intersects and is not shown in the map. The North Salinas weather station (NS; green star) is approximately 11km SE of GABOSR and was the closest weather monitoring station to all samples shown in the subset map. GPS coordinates for all sampling locations are provided in Table 4-1. Inset: location of the Salinas Valley in the state of California.

Table 4-1: GPS coordinates for sampling sites and the weather station used in this study.

Location	GIS-Latitude (°N)	GIS-Longitude (°W)
North Salinas weather station	36.7168	-121.6806
GABOSR	36.7803	-121.5849
TOWOSR	36.7962	-121.5753
GC	36.7426	-121.5035
TC	36.8101	-121.5452
WS1	36.2575	-121.4269
WS2	35.8964	-121.0889

Table 4-2: Culture-based detection of STEC and precipitation (Precip) data reported in inches. [a] Samples in which STEC was detected by PCR of enrichment cultures are listed as either positive (+) or negative (-). [b]Copy number of the shiga toxin gene (*stx2*) was determined via ddPCR. [c, d]Precipitation levels (in inches) for the day of sample collection and the sum of precipitation levels for five days prior to the sampling day were obtained from the California Irrigation Management Information System database (http://ipm.ucanr.edu/calludt.cgi) for North Salinas weather station (the closest monitoring station to the downstream sites).

Sample ID[a]	Date collected (mo/day/yr)	STEC[b]	No. of copies stx_2/μg DNA[c]	Precipitation (in)[d]	
				Day 1	5-day sum
GABOSR					
G130904	9/4/13	−	8.1	0	0
G140116	1/16/14	+	8	0	0.01
G140128	1/28/14	+	0	0	0
G140210	2/10/14	+	4.4	0.01	1.1
G140224	2/24/14	+	1.8	0	0
G140301	3/1/14	+	1.5	0.33	2.01
G140319	3/19/14	−	0	0	0.01
G140402	4/2/14	+	1.4	0.03	1.04
G140415	4/15/14	−	0	0	0
G140611	6/11/14	−	2.4	0	0
TOWOSR					
T130904	9/4/13	−	14.2	0	0
T130918	9/18/13	+	15.3	0	0
T131023	10/23/13	−	0	0	0
T131230	12/30/13	+	3.9	0	0
T140116	1/16/14	−	0	0	0.01
T140128	1/28/14	+	0	0	0
T140210	2/10/14	−	1.7	0.01	1.1
T140224	2/24/14	−	1.5	0	0
T140319	3/19/14	−	0	0	0.01
T140611	6/11/14	−	0	0	0
Upstream GABOSR control					
GC1	3/9/16	−	0	0	2.84
GC2	3/9/16	+	0	0	2.84
GC3	3/9/16	−	0	0	2.84
Upstream TOWOSR control					
TC1	4/19/17	+	0	0	0.45
TC2	4/19/17	−	0	0	0.45
West Salinas					
WS1	5/4/17	−	0	0	0
WS2	5/4/17	−	0	0	0

4.4.2 Description of metagenomes and sequence coverage of microbial community

A total of 27 metagenomic samples, ranging in size from 8.7 to 20.1 million reads (2.5 to 5 Gbp) after trimming, were recovered from the six locations (Appendix B, Table B 1). For all samples, less than 28% of the total community (average 18.6%) was covered by our sequencing efforts as determined by Nonpareil analysis (Appendix B, Figure B 1). Consequently, the assembly of the metagenomes was limiting (e.g., the N50 values were poor; see Appendix B, Table B 1), consistent with our previous analysis of soil and sediment communities (Rodriguez-R and Konstantinidis 2014) and those of a few other metagenomic studies of river sediments. Thus, an un-assembled short read-based strategy was used for all subsequent analyses (paired-end, non-overlapping reads with an average length of 132-145 bp per dataset), unless noted otherwise. A total of $7.2x10^8$ protein sequences were predicted from the short reads, with an average of $2.7x10^7$ sequences per sample. The number of protein sequences that could be annotated to the Swiss-Prot database in each sample ranged between 10 and 16% (average 14.5%) of the total sequences.

4.4.3 OTU characterization and alpha diversity assessment

A total of 466,421 reads encoding fragments of the 16S- or 18S-rRNA gene were detected in all 27 metagenomes with an average of 601 (+/- 55) reads per million reads. All datasets were dominated by bacteria, with only 0.6% and 3.0% of the total rRNA reads, on average, having archaeal or eukaryotic origin, respectively. Closed-reference OTU picking at 97% nucleotide identity threshold resulted in a total of 25,764 OTUs from 349,886 reads for all 27 samples and an average of 4,465 OTUs per sample. Since the coverage was similar for

all datasets, the number of OTUs shared between all samples were compared without any further normalization. Only 138 OTUs (0.5%) were shared among all 27 samples, while 9,500 (36.9%) of the OTUs were present in only one sample. The OTU rarefaction plot showed that diversity was not saturated (Figure 4-2A), which agreed with the low number of shared OTUs and the Nonpareil estimates on the shotgun data reported above (Appendix B, Figure B 1).

Alpha diversity observed in the California samples was compared to three publicly-available river sediment metagenomes from Montana that had similar land use inputs (i.e. agricultural or small towns) and were the most appropriate data for comparison among lotic sediment metagenomes currently available (Gibbons et al. 2014). Species richness and diversity in Montana samples were significantly less than California samples ($P= 2.3x10^{-4}$ and 0.006, respectively; Figure 4-2). Within California sites, diversity and evenness were similar; however, average species richness in GABOSR was significantly lower than TOWOSR and the upstream samples ($P= 0.034$ and $4.1x10^{-4}$, respectively).

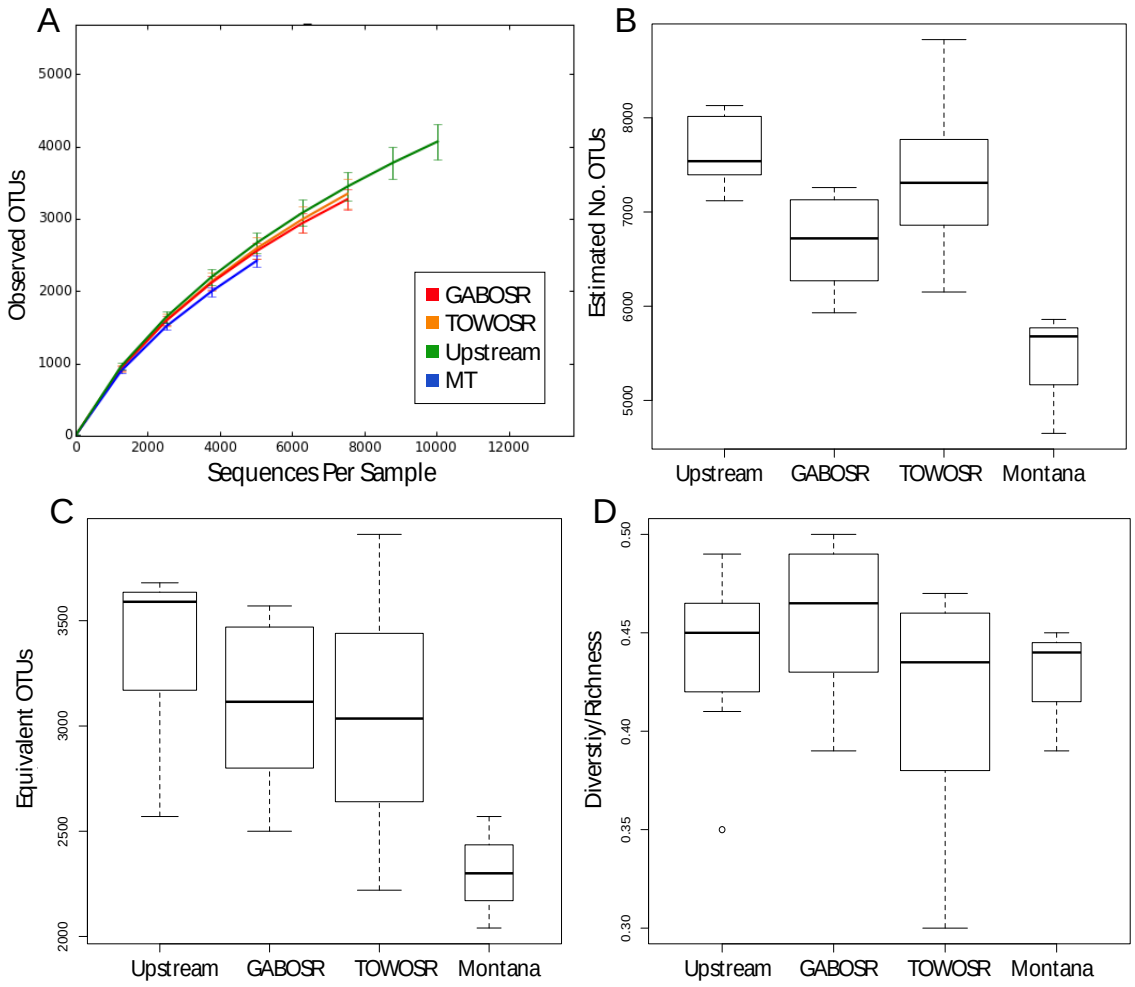


Figure 4-2: Taxonomic diversity of microbial communities in California (CA) creek sediments. Alpha diversity (based on 16S rRNA gene OTUs) of the 27 samples included in this study were compared to 3 sediment metagenomes from a river in Montana. (A) OTU Rarefaction plot: Multiple rarefactions were performed on OTU tables as implemented in MacQiime v.1.9.1. The rarefaction plots show that the diversity was not saturated by sequencing, which agrees with Nonpareil estimates (Fig. S1) and with the diversity observed in communities from similar habitats (i.e., the MT river sediments). (B-D) Alpha diversity indices: Comparisons between groups were conducted using two-sided t-tests and only the comparisons that yielded significant P-values are reported, as follows. (B) Species

richness was analyzed with the Chao 1 index. (C) Diversity was evaluated using the true diversity of order one and the Chao-Shen correction for unobserved species. (D) Evenness was calculated from the estimated values of diversity divided by richness.

4.4.4 Taxonomic composition and functional diversity of water-sediment microbial communities

OTUs were analyzed further to characterize the taxonomic profile of the communities sampled. *Proteobacteria* and *Bacteroidetes* were the most abundant phyla across most samples. However, some of the upstream samples had a higher abundance of *Actinobacteria* (Appendix B, Figure B 2A). Class level taxonomic distributions were consistent over time for GABOSR samples and revealed the high abundance of *Betaproteobacteria* (>19-24% of total sequences). TOWOSR samples varied more over time; five samples (T130918,T131230, T140128, T140210, T140611) had a higher abundance of *Deltaproteobacteria* and *Bacteroidia*, and one sample (T140116) had a higher abundance of *Cyanobacteria*. The upstream samples also showed a similar community composition and had higher relative abundance of *Alphaproteobacteria* (11-17%) compared to the downstream samples (Appendix B, Figure B 2B). These results were consistent with the TrEMBL taxonomic classification of protein-coding metagenomic reads, which were dominated by *Bacteria* (~95.2% per sample; Appendix B, Figure B 3).

4.4.5 Microbial community structure and dynamics in Salinas River valley creeks

Location was the strongest factor affecting clustering patterns observed in PCA ordinations of all distance matrices analyzed (Appendix B, Figure B 4). ADONIS analysis in the R package vegan (using location as a categorical variable) yielded $P<0.001$ and $R^2= 0.44$,

0.67, 0.41, and 0.56 for MASH, functional gene, OTUs Bray-Curtis (16S-BC) and OTUs weighted UniFrac (16S-WUF), respectively. This result was confirmed by correlation analysis of the NMDS ordinations to all metadata variables using the envfit function in vegan. After Bonferroni correction for multiple comparisons, location had the strongest correlation to all ordinations (MASH: P=0.001, R^2=0.879; Functional gene: P=0.001, R^2=0.845; 16S-BC: P=0.001, R^2=0.0.787; 16S-WUF: P=0.001, R^2=0.726), and was the only significant variable for MASH (Figure 4-3) and 16S rRNA gene-based measures of beta-diversity (Appendix B, Figure B 5, panels B and C) among those parameters evaluated. The functional gene ordination was also correlated, albeit weakly, to total 5-day precipitation (P=0.028, R^2=0.359; Appendix B, Figure B 5A). In order to control for spatial variance, a more rigorous db-RDA (Legendre and Anderson 1999) was used on constrained NMDS ordinations, which allows the influence of a matrix of conditioning variables (i.e., location) to be "removed" prior to analysis. No significant associations (P>0.05) were found in the functional gene and OTU Bray-Curtis ordinations, however, the MASH and OTU weighted UniFrac distances were significantly associated with sampling time (ANOVA: F=1.274, P=0.031; F=2.174, P=0.04, respectively).

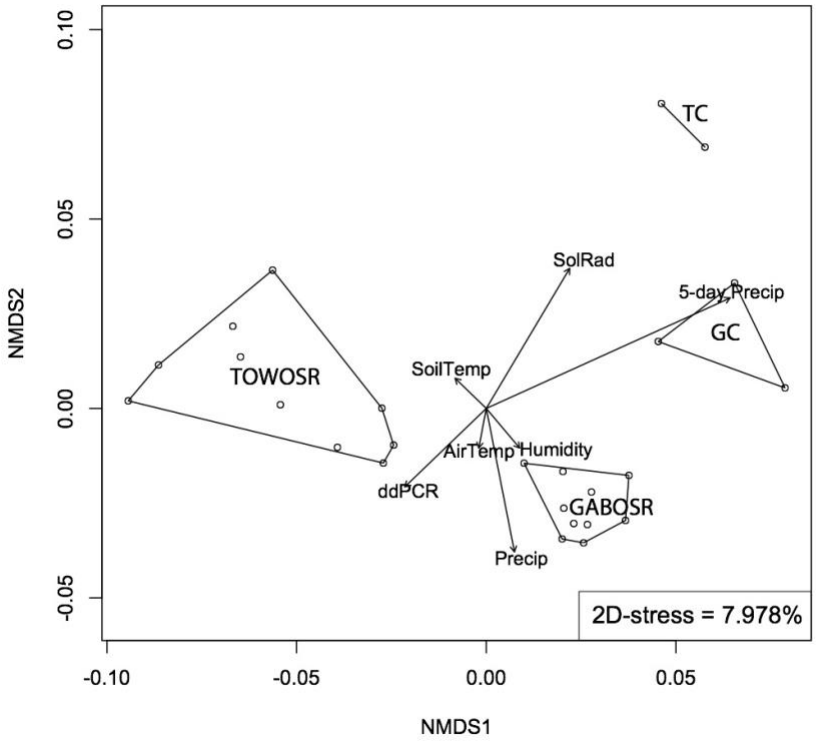


Figure 4-3: The effect of environmental parameters on microbial community structure. The graph shows non-metric multidimensional scaling (NMDS) of the sequenced communities based on whole-community MASH distances. Each dot represents a metagenome sample and metagenomes from the same location are connected by lines. Location (i.e., the polygons or lines) was the only variable that significantly correlated to the ordination. Arrowed vectors indicate correlation to other variables, none of which reached statistical significance, however. SoilTemp, AirTemp, SolRad, Precip, 5-day Precip, and ddPCR represent soil temperature, air temperature, solar radiation, day-of precipitation, 5-day precipitation, and digital droplet PCR counts for STEC, respectively.

4.4.6 Detection of E. coli by culture but not metagenomes

The abundance of reads annotated as *E. coli* in the metagenomes based on Blastn (nucleotide level) search against an STEC reference genome was low for all samples (~0.002% of total reads). Samples with the highest relative abundance of metagenomic reads matching to *E. coli* were negative for all culture-based tests (Table 4-3), which indicated spurious *in-silico* results (e.g., reads from non-*E. coli* genomes matching to conserved genes such as the rRNA operon). In addition, when using imGLAD (Castro et al. 2018) to predict the probability that *E. coli* was present in the metagenomes, a tool developed by our team to deal with spurious matches, all samples yielded a P-value of 1 (i.e., 0 probability of presence), which suggested that any *E. coli* populations (including STEC) were below the imGLAD estimated limit of detection for the metagenomic datasets in hand (i.e., 3% coverage of the *E. coli* genome at a minimum of 0.12X sequencing depth). The absolute abundance of the STEC based on ddPCR was also low (in the order of ~1 in 10^8 cells, assuming average molecular weight of a bp of DNA is 660g/mol, 5 Mb genome size, and 1 copy *stx*/genome) or absent in all samples, which supports our bioinformatic-derived conclusions that *E. coli* was probably too low in abundance to be detected by our metagenomic sequencing effort (Table 4-3).

Table 4-3: Culture-Based versus in-silico E. coli detection. [a] Culture-based methods to detect *E. coli* in resuspended sediment/water samples included an enrichment culturing step followed by shiga toxin (stx) PCR procedures to detect specific virulence genes and genotypes as described in the Materials and Methods section. Detection of *E. coli* O157 (EcO157) was determined using ELIZA serotyping and a sample was positive for STECs if PCR and/or ELIZA data yielded a positive result. [b] *In Silico* methods included a blastn

search of metagenomic reads against an STEC reference genome with a 95% identity and 97% read coverage cutoff for a read match, which was then normalized by dividing by the total number of reads per metagenome. The two samples with highest relative abundance of reads matching the STEC reference genome are shown in bold.

		Detection method			
		Culture based[a]		*In silico*[b]	
Location	Sample ID	EcO157	STEC	No. reads matching	% relative abundance
GABOSR	**G130904**	**−**	**−**	**1,652**	**0.0062**
	G140116	−	+	576	0.0019
	G140128	−	+	406	0.0020
	G140210	−	+	936	0.0032
	G140224	−	+	644	0.0024
	G140301	−	+	886	0.0026
	G140319	−	−	866	0.0028
	G140402	−	+	1,112	0.0030
	G140415	−	−	711	0.0022
	G140611	−	−	1,050	0.0029
TOWOSR	T130904	−	−	516	0.0022
	T130918	−	+	255	0.0011
	T131023	−	−	606	0.0023
	T131230	−	+	379	0.0013
	T140116	−	−	505	0.0019
	T140128	−	+	367	0.0016
	T140210	−	−	607	0.0035
	T140224	+	−	780	0.0026
	T140319	−	−	459	0.0016
	T140611	**−**	**−**	**1,495**	**0.0037**
Upstream GABOSR	GC1	−	−	411	0.0016
	GC2	−	+	419	0.0017
	GC3	−	−	423	0.0015
Upstream TOWOSR	TC1	−	+	478	0.0016
	TC2	−	−	626	0.0017
West Salinas	WS1	−	−	958	0.0032
	WS2	−	−	601	0.0022

4.4.7 Differentially abundant (DA) functions and taxa between locations

Of the 1,105 SEED subsystems (pathways) and 1806 taxonomic groups identified, 911 and 408 were significantly DA with $P_{adj} < 0.05$ for subsystems and taxa, respectively. Using

pairwise comparisons between GABOSR, TOWOSR, and the 7 upstream samples, 184 SEED subsystems had Log_2 fold change (L2FC) > 1, while 273 taxa had L2FC > 2, which were grouped into 36 and 35 broader functional and taxonomic categories, respectively (as described in the supplementary data files). The magnitude of the L2FC differences were somewhat low overall, with an average L2FC of 1.82 and 3.71 for DA functional genes and taxa, respectively. Still, this analysis revealed several notable trends that were consistent between the functional SEED and taxonomy results (Figures 4-4 and Appendix B, Figure B 6). More specifically, iron acquisition genes appeared to more abundant in the upstream samples, particularly in the samples collected upstream of TOWOSR (TC1 and TC2. Plant-associated and photosynthesis genes were more abundant in the more pristine samples (WS1 and WS2). Consistently, members of the phyla, *Alphaproteobacteria* (e.g. *Rhizobiales*; see Chapter 4 Supplementary data file S2), were more abundant upstream. Additional taxa that were more abundant in the upstream sites included those that are typically associated with soil and aquatic habitats (e.g., *Gemmatimonadetes* and *Armatimonadetes)*, which indicated that these sites may indeed receive less anthropogenic inputs.

Sample T140116 was enriched for both cyanobacteria based on OTU analysis (Appendix B, Figure B 6) and photosynthesis genes (Figure 4-4). TOWOSR appeared to be significantly more abundant in genes for anaerobic processes like anoxygenic photosynthesis and methanogenesis, along with genes related to archaeal DNA, RNA, and protein metabolism (all organisms known to carry out methanogenesis are *Archaea*). Consistently, the two TOWOSR samples (T140128 and T140210), which were most abundant in archaeal and methanogenesis genes, were also the most abundant in *Archaea*

and methanotrophs from the order *Methylococcales*, relative to the other sites. Other genes associated with anaerobic metabolisms, such as anoxygenic photosynthesis and sulfur metabolism genes (Figure 4-4), were congruent with taxonomic results that showed anoxygenic photosynthetic phyla *Chlorobi* (Green sulfur bacteria), *Chloroflexi* (Green non-sulfur), and the family *Chromatiaceae*, as well as known sulfur-metabolizing and anaerobic groups (e.g. *Thiobacillus* and *Clostridia*) to be more prevalent in the TOWOSR samples (Appendix B, Figure B 6). Additionally, the TOWOSR samples, in general, were more abundant in the *Firmicutes* and *Bacteroidetes*, which include gut-associated in addition to environmental members. Sample T140210 from TOWOSR was particularly enriched in specific enteric taxa: *Endomicrobia* and *Fibrobacteres*, which are rumen bacteria associated with cellulous degradation.

Collectively, these results indicated that our annotation and grouping methods were robust, e.g., archaeal taxa identified as more abundant in TOWOSR samples were consistent with an increased frequency of archaeal functional genes such as methanogenesis in these samples. These results also suggested that TOWOSR samples might be more anaerobic, which could potentially indicate an effect of runoff and eutrophication as a result of human activity at this location. It could also be that this is the result of natural factors that we did not test here and so we tried to look at specific DNA signals for anthropogenic pollution such as human and cow gut microbiome signal (see below). Also, *Actinobacteria* (i.e., common soil microbes and antibiotic producers) were all significantly more abundant in the upstream sites, which provides further evidence in support of this system being a natural source of ARGs (see below).

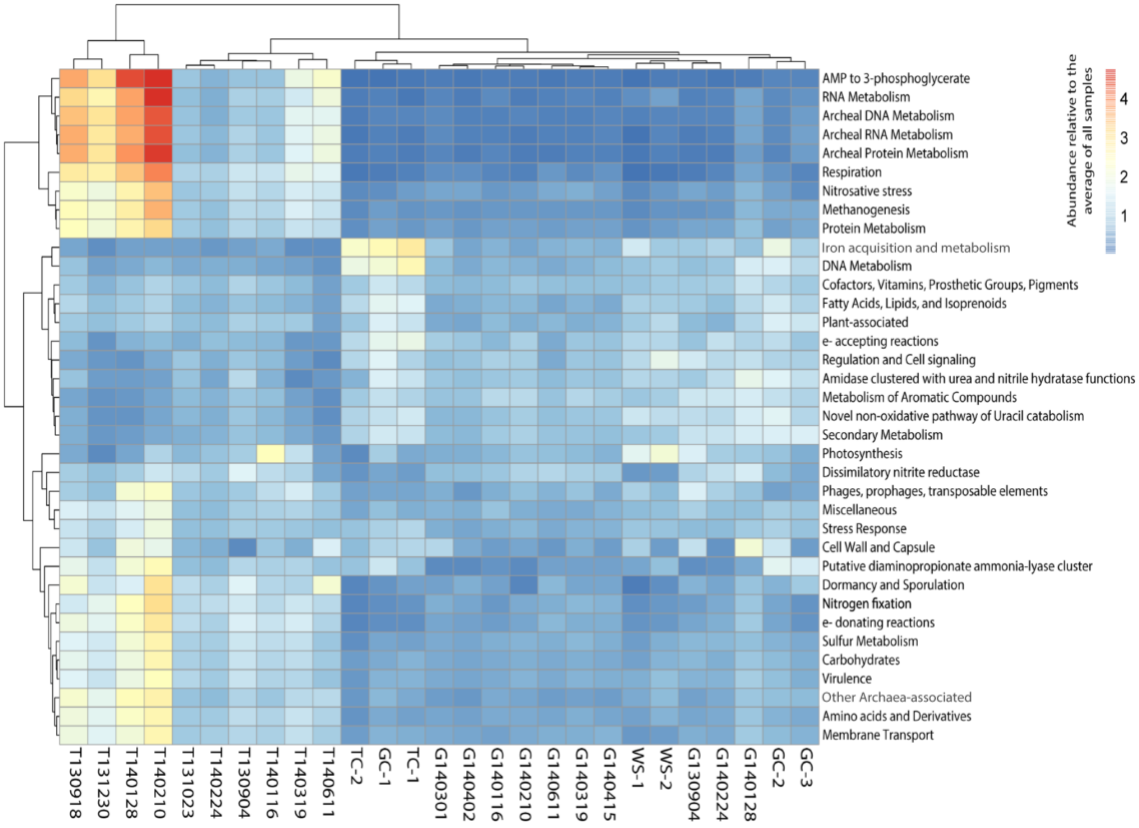


Figure 4-4: Functional profiles of creek sediment microbial communities. The heatmap shows SEED subsystems that were differentially abundant between locations (TOWOSR, GABOSR, and the upstream controls) with $P_{adj} < 0.05$. Color scale indicates the abundance relative to the average of all samples (increasing from blue to red). Letters T or G and date in the column names represent the sample site (TOWOSR or GABOSR) and collection date, respectively. TC, GC and WS represent the upstream TOWOSR Control, GABOSR Control, and West Salinas, respectively.

4.4.8 Quantifying anthropogenic and agricultural inputs

4.4.8.1 ARGs are more abundant in California samples compared to other similar environments:

The abundance of ARGs in each dataset was determined by blastp search against the Comprehensive Antibiotic Resistance gene Database (CARD; (McArthur et al. 2013)). The most abundant ARGs detected are shown in Appendix B, Figure B 7. A comparison of selected metagenomic datasets that included metagenomes from agricultural sediments from Montana (MT) and soils from Illinois (Urb, Hav), more pristine/remote samples from the Kalamas River (Kal) and Alaskan permafrost (AK), as well as a highly polluted sample from the Ganges River (Agra), was performed in order to benchmark the level of anthropogenic signal observed in the Salinas Valley against other environments. The abundance of ARGs in the California samples were significantly greater compared to the other environmental metagenomes included here (Kruskal-Wallis χ^2 =19.44, P = 0.0002; Figure 4-5A).

4.4.8.2 Abundance of genes associated with antibiotics used in cattle:

In order to better assess the impact (if any) of ARGs related to cattle ranching, we built ROCker models, a more accurate approach for finding metagenomic reads encoding a target gene of interest compared to simple homology searches (Orellana, Rodriguez-R, and Konstantinidis 2017), targeting tetracycline resistance (*tetM*) and production gene (*oxyT*) since tetracyclines are among the most common antibiotics used in livestock (US-FDA 2015). We also built a model targeting ketosynthase alpha subunit genes (*KSα*), which are involved in the synthesis of many antibiotics, including tetracyclines (Morlon et al. 2015). The antibiotic production genes were quantified in order to the test the hypothesis that if (the high abundance of) ARGs is naturally occurring (as opposed to being human-induced) then their abundance should correlate with that of the antibiotic production genes. To exclude the effect of potentially confounding variables, only the California samples were

used for linear regression analysis of the abundances of antibiotic production and resistance genes, and gene abundance was expressed as genome equivalents (GE), or the fraction of total genomes encoding the target gene of interest assuming the gene is single-copy -as it is usually the case for bacterial genes. In cases where the genes are in multiple copies, the GE will likely be >1 and would indicate genes per cell and not the fraction of genomes per total genomes. However, we did not observe cases of GEs >1, which indicated that our assumption was generally robust. ROCker analysis showed an abnormally high abundance of *tetM* in sample TC1 (Figure 4-6, left panel), which was thus considered an outlier and excluded from the linear regression analysis. The high abundance in TC1 was presumably attributed to the fact that *tetM* has the widest host range of all tetracycline resistance (*tet*) genes due to its association with highly mobile conjugative transposons that behave similarly to plasmids and have several antirestriction systems (Salyers et al. 1995; Roberts 2005). *OxyT* did not significantly correlate to *tetM* abundance (r^2 =0.031); however, *KSα* showed a moderate correlation to *tetM* (r^2=0.280) (Figure 4-6, right panel).

4.4.8.3 Abundance of cow and human gut (HG) microbiomes:

The abundance of cow- or human gut reads in the California creek and reference metagenomes from other environments was determined by Blastn search against a custom cow gut database and the Integrated Gene Catalog (IGC) of human gut microbiome genes (MetaHIT Consortium et al. 2014), respectively. The IGC is referred to as the Human Gut Database (HG) hereafter for clarity. The signal from the Ganges River (Agra) sample greatly exceeded all other samples in both the absolute number (Table 4-4) and relative abundance expressed as genome equivalents (GE), i.e., the fraction of total genomes encoding human gut genes assuming a single-copy of each gene per genome (33.5 GE; 8-

100x more abundant than all other samples; Figure 4-5B). There was a significant difference between the HG abundance averages observed in California metagenomes and the 8 metagenomes from 5 other habitats evaluated here (Kruskal-Wallis P=0.015). However, after correcting for multiple comparisons, none of the groups were significantly different (Wilcoxon Rank Sum P >0.1). Within California samples, there was no significant difference, overall, between abundances observed in the downstream samples and the average abundances of the upstream control samples (Kruskal-Wallis P=0.169).

The abundance of different cow gut genes had a similar trend to the human gut data (Table 4-4). However, two samples from TOWOSR (T140210 and T140611) showed an elevated signal for cow sequences (Figure 4-5C). Despite these two samples from TOWOSR with a higher level of cow gut signal, the average gene abundances were similar for California samples overall, and no significant difference was detected between the means compared to the other environmental metagenomes and the seven upstream control samples (Kruskal-Wallis P=0.090; Figure 4-5C).

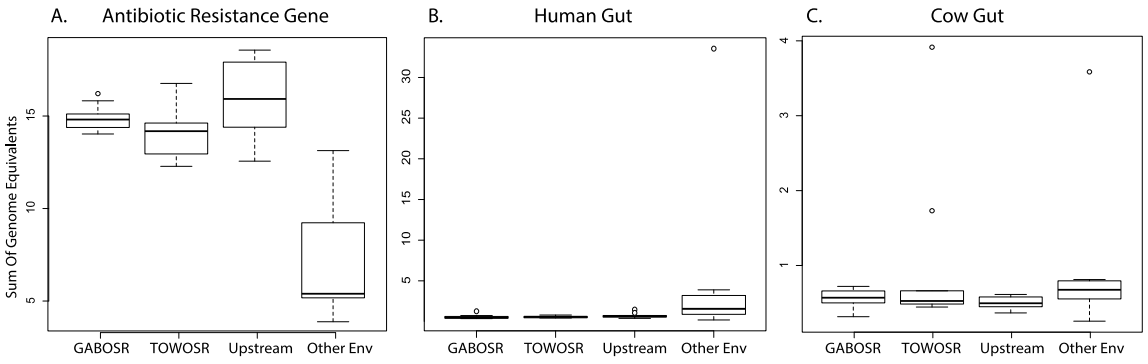


Figure 4-5: Abundance of antibiotic resistance genes, human gut, and cow gut sequences in the Salinas Valley metagenomes compared to other environmental metagenomes. The box and whisker plots show the interquartile range for the abundances with open dots indicating samples that exceeded 1.5x the interquartile range. The "Upstream" metagenomes represent the seven more pristine control samples, and included: three samples collected upstream from GABOSR, two collected upstream from TOWOSR, and two sites on the west side of the Salinas River that were farthest upstream from the rest of the sites (for more details, see main text and Figure 4-1). The other environmental metagenomes (Other Env) included: 3 river sediments, 2 agricultural soils, 1 permafrost soil, and 2 river water samples from the Kalamas and Ganges Rivers.

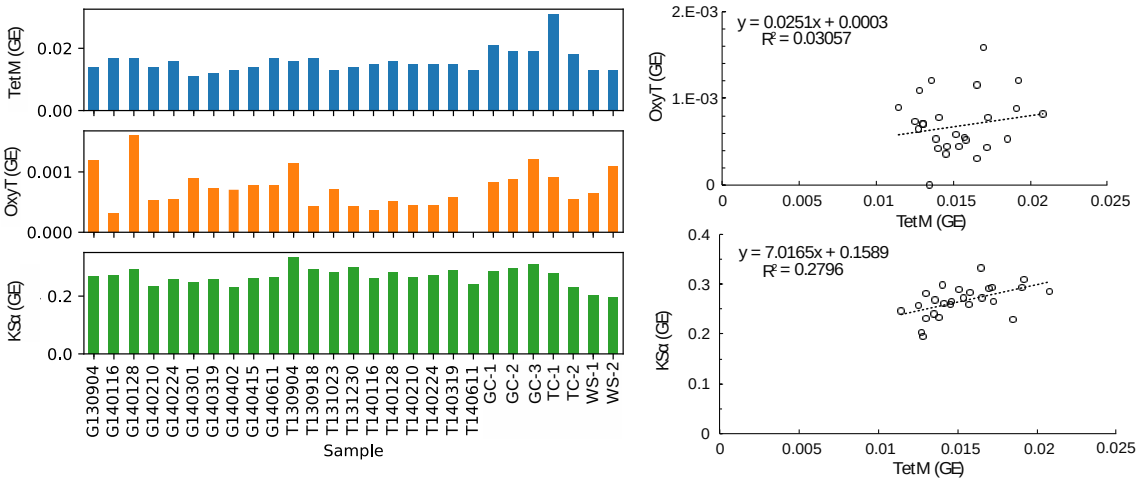


Figure 4-6: Abundances of selected antibiotic resistance and production genes in the Salinas Valley metagenomes. (LEFT) Abundance (expressed as genome equivalents) of *tetM*, *oxyT*, and *KSα* genes for the 27 sites included in this study. (RIGHT) Linear

regression of *tetM* versus *oxyT* or *KSα* gene abundances. TC1 was an outlier for *tetM* abundance and was removed from this analysis.

Table 4-4: Number of unique reference genes detected from the Comprehensive Antibiotic Resistance gene Database (CARD), Human and Cow Gut databases.

Unique genes detected in:	#samples	ARG	Human	Cow
all samples	35	1,776	167,481	15,497
TOWOSR	10	693	1,192	1,704
GABOSR	10	983	1,356	124
Upstream Controls	7	760	1,135	136
MT sediments	3	441	1,522	116
Agricultural Soils	2	722	9,877	270
AK permafrost	1	642	245	50
Kalamas River	1	475	3,952	554
Ganges River (Agra)	1	827	137,409	5,900
Total reference genes in database		2,820	9,879,896	459,176

4.5 Discussion

Analyses of planktonic microbial communities in rivers over time and land use have shown that these communities vary by average genome size, location, amount of sunlight, and nutrient concentrations (Van Rossum et al. 2015) as well as by sampling time more so than space (Meziti et al. 2016). However, the results presented here suggested that community composition of Salinas Valley creek sediments are structured primarily by spatial separation, and the local weather parameters tested here did not have a significant effect (Figure 4-3). More detailed *in-situ* metadata than those obtained here such as nutrient concentrations (e.g., organic carbon and biological oxygen demand) are needed in order to

discern the processes that are driving community diversity and structure within each Salinas Valley site. For example, anaerobic taxa and processes related to methane and sulfur metabolism and anoxygenic photosynthesis were significantly more abundant in TOWOSR (Figure 4-4 and supplemental material), which could indicate higher influence from agricultural run-off, lower permeability of the corresponding sediments by oxygen, or some other environmental factor that was not reflected by the local weather parameters measured here. It should be mentioned however, that we did not observe any significant differences in the type of sediment sampled (e.g., percent of fine sand) between the different sampling sites. Hence, the lower oxygen permeability appears to be a less plausible explanation for the functional differences observed compared to higher eutrophication (or another reason).

We compared abundances of metagenomic reads annotated as ARG, human or cow gut microbiome in order to assess levels of anthropogenic impacts on Salinas Valley creek sediment communities. No significant difference was detected between the downstream samples and the upstream controls for any of the three anthropogenic indicators (Figure 4-5), which suggested that the land use practices surrounding the creeks does not have a major or lasting impact on the natural community and the inputs are likely diluted or attenuated faster than the intervals sampled here. To gain further quantitative insights, we then benchmarked abundances observed in the creek sediments from this study against metagenomes from other environments. These included agricultural sediments and soils, permafrost, and river water from both pristine and polluted habitats. GABOSR, TOWOSR, and the upstream samples all had significantly higher ARG abundances compared to the average of the other environments tested here (Figure 4-5A). This high background level

of reads annotated as ARGs suggested that the Salinas Valley creek sediments are a natural reservoir for these genes. Furthermore, resistance genes to synthetic antibiotics such as florfenicol (*fexA* and *floR)* and ciprofloxacin (*qnrS)*, one of the most widely used antibiotics in humans worldwide, were absent or detected in very low abundance (less than 10 reads matching) in our datasets. Spurious matches to conserved gene regions can occur when analyzing short reads like the ones here, but the signal was not large enough to warrant further investigation using precise and targeted methods (e.g. ROCker). Overall, the absence of resistance genes to more recently introduced, synthetic antibiotics provides further evidence that the ARG signal observed in the Salinas Valley is likely autochthonous in origin. Future studies could involve deeper sequencing (higher community coverage) in order to recover long contigs and thus, determine the genomic background of the ARGs and if they are associated with mobile elements or plasmids for improved public health risk assessment. Still, our results highlight the importance of having a baseline or "pristine" sample to discern anthropogenic from naturally-occurring ARGs and have important implications for monitoring the spread of ARGs in the environment. For instance, without the upstream control samples, this study could have (speciously) concluded that GABOSR and TOWOSR are elevated in ARGs as a result of cattle ranching. However, the similar abundances found in the upstream samples indicated that the signal detected downstream could be inherent to this environment and that a more targeted analysis of specific ARGs was required to determine if the effect of cattle could be detected.

Tetracycline resistance genes have been shown to increase with and correlate to anthropogenic inputs along a river estuary system (Chen et al. 2013), suggesting that they can be useful indicators of anthropogenic pollution. However, tetracycline resistance genes

are also found in other pristine or natural environments (D'Costa et al. 2011; Allen et al. 2009; Cytryn 2013; Yang et al. 2013), and therefore can also be considered part of the autochthonous gene pool in some habitats. Here, we tested the hypothesis that if tetracycline resistance genes are naturally occurring, the production enzymes for tetracycline should also follow similar abundance patterns, as antibiotic resistance and biosynthesis genes are often encoded on the same operon to ensure antibiotic-producing species are resistant to the product they synthesize (Martín and Liras 1989). Thus, we expected to see a correlation between abundances of the tetracycline resistance gene, *tetM*, and its associated production genes (*oxyT*, *KSα*) if this system is not under heavy selection pressure of human-introduced antibiotics. The abundance of *tetM* in the Salinas Valley creek sediments was not correlated to *oxyT* and only moderately correlated to *KSα* (Figure 4-6). *OxyT* had very low abundance (less than 8 reads matching per sample), which suggested that the lack of correlation to *tetM* could be due to database limitations. That is, only a few reference *oxyT* genes are publicly available (13 sequences) and these likely do not capture the total diversity of this gene found in the environment. *KSα*, on the other hand, represents a broad class of synthesis genes for many different antibiotics with many more sequences in the reference databases and thus, a better estimate of antibiotic production potential was obtained based on these genes. Overall, these findings further supported that this ecosystem is a natural reservoir for ARGs, and the presence of tetracycline resistance is not likely to be solely caused by inputs from the cattle ranches. However, future investigations could involve additional antibiotic production gene references for more robust conclusions.

When compared to the other pristine or rural environmental metagenomes such as agricultural sediments and soils, permafrost, and river water, the abundances of reads annotated as human gut in the California sediments were not significantly different overall. However, the Ganges River (Agra) sample, collected from one of the most densely populated and highly polluted areas surrounding the river (Agra, Uttar Pradesh, India), was 1-2 orders of magnitude more abundant for human gut (open circle in Figure 4-5B), compared to the rest of the samples used in our study. Thus, a high human gut signal was expected for the Ganges River, consistent with previous results (S.-Y. Zhang et al. 2019) and served as a reference to assess relative levels of human fecal contamination. The rest of the samples included in our comparisons were from rural/agricultural or more remote areas, with lower population density, and consistently had lower signals of human fecal contamination than the Agra sample. Therefore, the low abundances of human gut sequences observed in Salinas Valley were consistent with the lower levels of human activity/density input relative to more human and animal populated sites, such as the Ganges River used for comparison here and indicated that our annotation and filtering methods were robust. Collectively, these results showed that metagenomics of river/creek sediments provide a reliable means for assessing the magnitude of the human presence/activity, consistent with recent studies of other riverine ecosystems (S.-Y. Zhang et al. 2019; Meziti et al. 2016).

Contrary to the results for human gut, the abundances of cow gut signal in the California samples were not consistent with our expectations. The TOWOSR and GABOSR sites are directly downstream of large cattle ranch operations and identical pathogen recovery from water and upstream cattle indicated the cattle ranches were the source of fecal

contamination (M. Cooley et al. 2007). As such, we expected to see a higher level of cow signal in the downstream metagenome samples, yet the abundance was not significantly different from the other environments or the upstream controls (Figure 4-5B&C). Notably, two of the samples from TOWOSR (T140210 and T140611) showed elevated signal for cow that was similar to the abundance observed in the highly polluted Ganges River reference metagenome (Figure 4-5C). These samples (especially T140210) had a higher abundance of the rumen enteric and cellulose degrading taxa (*Endomicrobia* and *Fibrobacteres*; Appendix B, Figure B 6), which supports the conclusion that these samples contained run-off from cattle, however the signal might be patchy or muted in the sediment and require more frequent sampling and/or larger sampling volumes than those used here to detect these signals.

Additionally, we were unable to detect any *E. coli* populations in any of the metagenomes, including samples that were positive for STEC via enrichment culture, indicating that it is not an abundant member of the sediment community (Table 4-3). This was consistent with imGLAD estimates that the sequencing effort applied to our metagenomes imposed a limit of detection for *E. coli*, and ddPCR results that showed abundance of STEC was low or absent in all samples. Overall, these results suggested that using shotgun metagenomics may not be sensitive (or economical) enough as a monitoring tool to detect a relatively low abundance microorganism in lotic sediments at the level of sequencing effort applied here, which was insufficient partly because of the extremely high community diversity (Appendix B, Figure B 1). More than the 2.5 to 5 Gbp/sample sequencing effort applied in this study would have been required to detect ~10 *E. coli* cells in a sample according to our estimates, which is not economical based on current standards and costs. More

specifically, obtaining the imGLAD minimum threshold of 0.12x coverage for an STEC genome (5 Mbp) in our metagenome libraries (average 4 Gbp), would require 0.6 Mbp of STEC reads, or 0.015% of the total metagenome, which translates to a relatively large number of cells *in situ*. For example, assuming 10^8 total cells/g of sediment, it would require ~10^4 STEC cells/g of sediment to robustly detect in the metagenomes (or 100 times more sequencing for detecting ~10 cells/g). Thus, the limit of detection of metagenomics, as applied here, was not low enough and should be combined with methods that offer lower detection limits and more precise counts (such as ddPCR).

Rivers are highly dynamic ecosystems and therefore subject to higher random variation and sampling artifacts that likely affect the dilution of the exogenous (human) input. Further, our samples represent relatively small volumes of sediment (~10 g) and the resulting metagenomic datasets did not saturate the sequence diversity in the DNA extracted from these samples (Appendix B, Figure B 1), which might introduce further experimental noise and stochasticity. Despite these technical limitations, our data consistently showed little evidence that agricultural or cattle ranching activities have a significant effect on the creek sediment microbial communities. The underlying reason for these results remains speculative but could include sediment absorption or dilution by the creek waters and should be the subject of future research in order to better understand the impact of these activities on the environment. Additionally, the functional and taxonomic diversity observed between our samples could not be attributed to the environmental and weather variables measured, especially for the TOWOSR samples that showed extensive sample heterogeneity (diversity). These results suggested that shorter intervals between sampling as well as more detailed *in-situ* geochemical data will be needed to elucidate the fine

scale processes driving the community composition within each location. Although the continued presence of STEC in Salinas watershed sediments is a public health risk, we did not find evidence that runoff from human activities has a substantial effect on the sediment microbial community when compared to more pristine sites. An imperative objective for public health is to assess how and where current agricultural practices impact the environment in order to determine best practices. Our study also provided important information on using metagenomics as a tool for public health risk studies of river water and sediment habitats, including what sampling volumes and frequencies to use, amount of sequencing to apply, and what bioinformatics analyses to perform on the resulting data for future public health risk studies of river water and sediment habitats. Finally, the ROCker models developed here for tetracycline resistance and production genes should be useful for robustly examining the prevalence of these genes in other samples and habitats.

4.6 Acknowledgements

This work was supported by the USDA (award 2030-42000-050-10), the US National Science Foundation (awards No 1511825 and 1831582 to KTK) and the US National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650044. The funding agencies had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

4.7 Materials and Methods

4.7.1 Sample collection and enrichment method for STEC

Sediment samples were collected from watersheds at public-access locations (Table 4-1). Weather information was downloaded from the California Irrigation Management Information System database (http://ipm.ucanr.edu/calludt.cgi) for the day of and five days prior to the sampling day from the closest monitoring station to the downstream sites (Table 4-2). Approximately 250mL of sediment was collected by dragging an open, sterile bottle attached to a 7.62m telescoping pole along the bottom of the stream in the upstream direction and in such a way that the majority of the sample was undisturbed sediment. Nevertheless, some mixing with the water column occurred. Sediment at GABOSR contained more sand than TOWOSR or any of the control locations. Nevertheless, even at GABOSR collection was selectively silt (with fine sand occupying less than 10% by volume). As such, an effort was made to collect comparable samples at different locations. Additionally, only the top 1-2 cm of sediment was collected. All samples were transported on ice and processed within 24 hours. Sediment was re-suspended in the lab just prior to sampling to ensure a uniform sub-sample. DNA from 10 g of the resuspended sediment/water mix was purified for sediment DNA using MoBio PowerSoil DNA extraction kit, following the manufacturer's protocol. A separate 100 mL of the sample was used for enrichment and isolation of STEC as previously described (M. B. Cooley et al. 2013).

4.7.2 PCR-based quantification method for STEC

Droplet digital PCR (ddPCR, BioRad) was performed on sediment DNA following the method of Cooley et al. (19). Each 20 µL reaction used 10 µL BioRad's Supermix for Probes, 2 µL primer (0.3µM final concentration) and probe (0.2µM), up to 1 µg DNA, 1.2 µL MgCl2 (1.5mM), and 0.2 µL HindIII (0.2 U/µL). Primer and probe sequences were as

previously published for STEC (M. B. Cooley, Carychao, and Gorski 2018). Droplets were created with Droplet Generation Oil for Probes in the QX-200 droplet generator (BioRad), and amplified for 5 min at 95°C, 45 cycles at 95°C for 30 s and 60°C for 90 s, then 5 min at 72°C and 5 min at 98°C. Droplets were processed with the QX-200 Droplet reader and template levels were predicted by QuantaSoft software version 1.7.4 (BioRad).

4.7.3 DNA sequencing and Bioinformatics sequence analysis

4.7.3.1 Metagenomic sequencing and community coverage estimates:

Shotgun metagenomic sequencing libraries were prepared using the Illumina Nextera XT library prep kit and HiSEQ 2500 instrument as described previously (Johnston et al. 2019). Short reads were passed through quality filtering and trimming as described previously (Rodriguez-R et al. 2015). In short, sequences were trimmed with a PHRED score cutoff of 20 and minimum length of 50bp. Only paired reads with both sisters longer than 50bp after trimming were used for further analysis. Average community coverage and diversity were estimated using Nonpareil 3.0 (Rodriguez-R and Konstantinidis 2014) with kmer kernel and default parameters. Sequences were assembled with IDBA (Peng et al. 2012) using kmer values ranging from 20 to 80**.**

4.7.3.2 Taxonomic analysis of rRNA gene-encoding sequences:

Metagenomic reads encoding short subunit (SSU) rRNA genes were extracted with Parallel-Meta v.2.4.1 using default parameters (Su et al. 2014). Closed reference OTU picking at 97% nucleotide identity with taxonomic assignment against the GreenGenes database (19) was performed using MacQiime v.1.9.1 (Caporaso et al. 2010) with the

reverse strand matching parameter enabled and the uclust clustering algorithm (Edgar 2010). Alpha diversity was calculated as the true diversity of order one (equivalent to the exponential of the Shannon index) and corrected for unobserved species using the Chao-Shen correction (Chao and Shen 2003) as implemented in the R package entropy (Hausser and Strimmer, n.d.). Richness was estimated using the Chao1 index (Chao 1984), and evenness was calculated from the estimated values of diversity divided by richness. Significant differences in taxonomic diversity, evenness, and richness were assessed using two-sided t-tests. Multiple rarefactions were performed on OTU tables as implemented in MacQiime v.1.9.1 (rarifying up to the minimum number of counts per sample: option -e 5,596).

4.7.3.3 Determination of the total community bacterial fraction:

In order to determine whether bacterial gene abundances need be corrected for relative bacterial fraction in the total metagenome libraries, the relative abundance of *Bacteria*, *Archaea*, and *Eukarya* was estimated in each dataset by searching a subset (~$1x10^5$ reads per sample) of randomly selected protein coding reads against the TrEMBL database ((UniProt Consortium et al. 2017); downloaded May 2018) using DIAMOND blastx v.0.9.22.123 (Buchfink, Xie, and Huson 2015) with the "--more sensitive" option and e-value cutoff of $1x10^{-5}$. The TrEMBL IDs for best hit matches were summarized at the domain level using custom scripts and the metadata files available at ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/ No significant difference in the relative abundance of *Bacteria* was found between the different samples, thus no correction for bacterial fraction was applied to gene abundance calculations.

4.7.3.4 Functional and ARG annotation of metagenomic sequences:

Protein prediction was performed using FragGeneScan adopting the Illumina 0.5% error model (Rho, Tang, and Ye 2010). Resulting amino acid sequences were searched against the Swiss-Prot (downloaded June 2017) (UniProt Consortium et al. 2017) and Comprehensive Antibiotic Resistance gene (CARD, downloaded May 2017; 26) databases using blastp (Camacho et al. 2009) for functional annotation. Best matches to the Swiss-Prot database with >80% query coverage, >40% identity and >35 amino acid alignment length were kept for further analyses. A more stringent cut off was used for best matches to the CARD (>40% identity over >90% of the read length) to minimize false positive matches.

4.7.3.5 Detection of cow and human gut microbiome associated sequences:

Searches for cow gut associated sequences were performed using our own collection of cow fecal metagenomes from six cow individuals collected in Georgia, USA. DNA extracted from cow fecal material underwent the same library prep, DNA sequencing and quality trimming and processing as described above. Short reads for both the cow gut and CA sediment metagenomes have been deposited to the SRA database (submission IDs: PRJNA545149 and PRJNA545542, respectively). Predicted genes (as nucleotides) from all six individual cows were pooled together and de-replicated at 95% identity using the CD-HIT algorithm **(**Options: -n 10, -d 0; (Fu et al. 2012)) resulting in 459,176 non-redundant cow gut metagenome "database" sequences. Human gut-associated sequences were assessed based on comparisons of short-reads against the Integrated Gene Catalog (IGC) of human gut microbiome genes (MetaHIT Consortium et al. 2014), heretofore

referred to as Human Gut Database (HG) for clarity. The abundance of cow and human gut signal in the short-read metagenomes was determined based on the number of reads from each dataset matching these reference sequences using blastn v2.2.29 with a filtering cut off of >95% identity and >90% query length coverage. Due to under-sampling of the total community diversity at our sequencing depth, these more comprehensive, whole gut microbiome databases were preferred over a specific suite of biomarkers for anthropogenic pollution, which are less likely to be detected in the metagenomes by chance, compared to the whole cow or human gut microbiome.

4.7.3.6 Abundance of specific antibiotic resistance (ARG) and production genes using ROCker:

Dynamic filtering cut-off models targeting a tetracycline resistance gene (*tetM*) and two antibiotic production genes (*oxyT* and *KSα)* were designed with ROCker v1.3.1, as previously described (Orellana, Rodriguez-R, and Konstantinidis 2017). Reference sequences for model building were manually selected from public databases and models were built for 150bp reads and default parameters. The reference sequences and ROCker models are available at http://enve-omics.ce.gatech.edu/rocker/models. Short reads were searched against the reference sequences used to build the model with blastx. The ROCker models were used to filter matches, which were subsequently divided by the median reference gene length in order to calculate sequencing coverage and were then normalized for genome equivalents as described below. Correlation between abundances of antibiotic production and resistance genes was determined using linear regression.

4.7.3.7 Quantification of genome equivalents (GE):

Average genome size and genome sequencing depth (i.e., the average sequencing depth of single copy genes) were determined for each sample using MicrobeCensus v1.0.6 with default parameters (Nayfach and Pollard 2015). The sequencing depth of reference genes with a given annotation was estimated for each dataset (in reads/bp), then divided by the corresponding average genome sequencing depth and summed to give the total GEs per sample.

4.7.3.8 Mash and multivariate analysis:

MASH v1.0.2 (Ondov et al. 2016) was used to assess overall whole-community similarity among metagenomes in a reference database-independent approach (Options: -s 100000). Functional gene and 16S rRNA gene-based OTU count matrices were median-normalized using the R package DESeq2 (v.1.16.1; (Anders and Huber 2010)). Pairwise Bray-Curtis and weighted UniFrac (16S only) dissimilarity indexes of the normalized counts were used for principal component analysis (PCA) and non-metric multidimensional scaling (NMDS) analysis in order to assess whole-community gene functional and taxonomic (16S rRNA gene OTUs) similarity. The significance of metadata parameters on the NMDS ordinations was performed using the ecodist and envfit functions of the R package vegan v2.4.4 (indices included: location, sampling time, ddPCR counts for STEC, same day precipitation, 5-day precipitation, solar radiation, air temp, soil temp, and humidity). The two west Salinas samples (WS1 and WS2) were excluded from this analysis in order to minimize confounding variation of temporal and spatial differences. In order to control for spatial variance, a more rigorous distance-based redundancy analysis (db-RDA; (Legendre and Anderson 1999)) was used to investigate the correlation to metadata using the capscale

function in the R package vegan (included same indices as above, but with Condition(location) constraint on ordinations).

4.7.3.9 *In-silico* detection of *E. coli* in sample metagenomes:

The presence of any *E. coli* in the metagenomes was determined using a blastn search of short reads against an STEC reference genome (accession NC_002695) that had been filtered to remove non-diagnostic (i.e. highly conserved among phyla) regions with MyTaxa (Luo, Rodriguez-R, and Konstantinidis 2014). Only matches with nucleotide identity >95% and alignment length >97% were used to calculate relative abundance of *E. coli* in the metagenomes. This level of sequence diversity (nucleotide identity >95%) encompasses well the diversity within the *E. coli-Shigella* spp. group; thus, any *E. coli* populations present in the metagenomes at high enough abundance would be detected at this filtering cutoff. The best hit output from blastn was also analyzed with imGLAD (Castro et al. 2018), a tool that can estimate the probability of presence and limit of detection of a reference/target genome in a metagenome.

4.7.3.10 Determination of DA taxa and gene functions:

Functional annotations of the recovered protein sequences were summarized into several hierarchical ranks including metabolic pathways and individual protein families based on the SEED classification system (Overbeek et al. 2005). The 16S rRNA gene OTUs were placed into taxonomic groups based on the lowest rank of taxonomic classification (genus, family etc.) shared by 90% or more of the sequences within the OTU using MacQiime v.1.9.1 (Caporaso et al. 2010). DA functional annotation terms (subsystems) or OTUs were identified in samples grouped by location (e.g., pairwise comparison of all 10 TOWOSR

vs. all 10 GABOSR and vs. all 7 upstream "pristine control" sites) using the negative binomial test and false discovery rate (P_{adj} <0.05) as implemented in DESeq2 v1.16.1 (Anders and Huber 2010). Subsystems with Log_2 fold change (L2FC) >1 or taxa with L2FC >2 were manually grouped into broader categories based on known functional or taxonomic similarities, respectively (Figures 3 & Appendix B, Figure B 6), which were then normalized by library size (per million read library). A larger L2FC cutoff was used for taxa to account for the larger dataset size and allow for inspection of the taxa contributing most to differential abundance between the locations. The taxonomic assignment of these DA taxa were confirmed against the SILVA database (downloaded October 2018; (Yilmaz et al. 2014)). Each subsystem or taxonomic category was then divided by its average sequencing depth across all samples to provide unbiased counts for presentation purposes.

4.7.3.11 Comparison of putative anthropogenic signals observed in California sediments to metagenomes from other environments:

Publicly available metagenomes from other studies were used to compare abundances of reads annotated as ARG, HG, and cow gut with the results obtained for the California sediment datasets reported here. These metagenomes included: three Montana River sediments (MT; (Gibbons et al. 2014)), two temperate agricultural soils from Illinois (Hav and Urb; (Orellana et al. 2018)), an Alaskan tundra soil (AK; (Johnston et al. 2016)), one sample from the Ganges River near Agra, Uttar Pradesh (Agra; (S.-Y. Zhang et al. 2019)), and one from the Kalamas River in Greece (Kal; (Meziti et al. 2016)). Short read metagenomes for MT samples were downloaded from MG-RAST ((Keegan, Glass, and Meyer 2016); MG-RAST IDs: 4481974.3, 4481983.3, 4481956.3). The remaining datasets

were obtained from the NCBI short read archive (SRA) database (Hav: ERR1939174, Urb: ERR1939274, AK: ERR1035437, Agra: SRR6337690, Kal: SRR3098772). Reads from these metagenomes were comparable to the ones from this study (100 – 150bp paired-end Illumina sequencing) and underwent the same trimming, annotation (against the CARD, HG, and cow gut databases only) and gene count normalization protocol as described above. The Kruskal-Wallis test in R was performed to determine significantly different mean abundances between groups . Alpha diversity and taxonomic comparisons were performed (for MT datasets only) based on metagenomic reads encoding fragments of the 16S rRNA gene, which were identified as described above.

CONCLUSIONS AND FUTURE PERSPECTIVES

Waterborne diseases resulting from fecal contamination in the environment are a significant public health issue worldwide. For example, diarrheal disease caused by waterborne infections are a main cause of death in children under five. Water scarcity is expected to rise along with global population and urbanization; as such, reclaimed wastewater will be necessary to meet growing water demands. Reclaimed wastewater or "brown" water is increasingly being used for agricultural irrigation to produce food and to replenish depleting groundwater. Therefore, reliable and consistent water quality monitoring will become even more important as these trends continue.

Current FIB and MST markers are becoming obsolete for many reasons (e.g., lack host specificity and sensitivity, poor correlation to pathogens, etc.) and are unlikely to serve as indicators for emerging fecal-related contaminants, such as antibiotic resistance genes (ARGs). Furthermore, qPCR only detects marker genes that match the specific primers and probe, so it cannot capture novel targets and may miss some of portions of the target host population that show nucleotide polymorphisms relative to the primer sequences. Metagenomic sequencing effectively captures both phylogenetic and functional diversity in a water sample simultaneously and is suitable for overcoming several of the limitations associated with culture-based and qPCR methods (described earlier in this thesis).

This thesis primarily focused on using meta-omics techniques to evaluate the "gold standard" FIB, *E. faecalis*, and discover novel targets from host specific gut sequences. We identified several metagenome-assembled genomes (MAGs) and functional genes as promising new targets for improved MST and for distinguishing fecal contamination from

human or livestock (cows and pigs) sources. Notably, the identified MAGs differ from the most commonly used FIB both taxonomically and functionally, which suggests that better biomarkers may be found among novel taxa that have not been previously considered for MST. Further, we were unable to effectively distinguish enteric *E. faecalis* from their naturalized counterparts based on our mesocosm incubations, an important limitation that is not applicable to the newly proposed MAGs because they are strict anaerobes and died off quickly under our simulated aerobic aquatic habitat. We have also not detected these MAGs in other (presumably) un-polluted aquatic samples. Overall, our results confirmed our overarching hypothesis that functional genes related to host-microbe interactions carried in strict anaerobes are likely the better targets for MST compared to facultative anaerobes like *E. faecalis* and *E. coli*.

This thesis also offered critical insights on the persistence of markers and their decay rates under oligotrophic, freshwater conditions. This decay information will be useful for determining the age of pollution events and integrating into more accurate quantitative microbial risk assessment (QMRA) models. Notably, although the MAGs we assembled were highly host-specific, they all showed similar decay characteristics and the majority of MAGs from all three hosts did not persist longer than four days (**Chapter 3**). Furthermore, our rRNA/rDNA analysis in **Chapter 2** showed that *E. faecalis* significantly reduced metabolic activity four days after being introduced into an aerobic, oligotrophic environment. These results were consistent with a recent QMRA study that predicted that the gastrointestinal infection risk from sewage contamination in surface waters is not significant (<3% chance of infection) after 3.3 days (Boehm et al. 2018) in accordance with the EPA risk threshold for bathing water (USEPA RWQC 2012). The prediction from

Boehm and colleagues was based on published decay constants for pathogens and FIB from more than 70 publications. The fact that different microbes (including the FIB *E. faecalis*) and pathogens from sewage, human, cow, and pig guts have apparently similar persistence of about 3-4 days in aerobic surface water environments suggests that this is likely a robust timeline on which to base public health risk assessments. In other words, fecal pollution that is more than 4 days old is unlikely to represent a significant risk to public health, however it is still unclear whether this is affected by the relative volume or concentration of pollution levels (e.g., if infection risk persists longer than four days in waters with very high concentrations of fecal pollution). Nevertheless, being able to accurately distinguish the sources of fecal pollution is still valuable for remediating chronically polluted waters. Thus, developing robust host-specific MST biomarkers is still critical.

We also demonstrated the advantages of metagenomic methods over traditional qPCR and culture-based tests such as qPCR assays targeting (allegedly) host-specific regions of the 16S rRNA gene in the *Bacteroides*. The most commonly used MST qPCR marker, HF183, performed quite poorly compared to the putative biomarker MAGs because it was not reliably detected in two of the three human stool samples used here, and thus, underestimated fecal contamination risk. In contrast, the other qPCR markers (ruminant specific *Bacteroides* 16S and human mtDNA) were detectable and persisted for 14 days, indicating they likely over-estimate the infection risk based on the QMRA predictions mentioned above. Although targeting an entire MAG or metagenome is currently not economically feasible for water quality monitoring assays, these results suggest that a single qPCR target is likely not sufficient to accurately determine public health risk. A suite

of markers that can account for imperfect host sensitivity and incorporate accurate decay information will likely give better risk estimates.

As cost of sequencing continues to decrease, it may be possible to monitor water quality directly with metagenomic techniques. Currently, metagenomics is not suitable for rapid water quality decision making because the total time needed from DNA extraction to sequencing takes ~39-55 hours with the Illumina platform **(**Hong et al. 2020), which is longer than even traditional culture-based tests (which take 24 hours). The turn-around time can potentially be reduced with new technologies such as the Oxford Nanopore sequencing platform, whose small, hand-held sequencers allow for metagenome sequencing in remote areas as well as standardization of the associated bioinformatics pipelines to process the sequence data (Quick et al. 2016). This feature of Oxford Nanopore may also make metagenomic techniques more accessible for developing nations that have limited infrastructure and resources (Roy et al. 2018), and often experience more serious outbreaks and prevalent issues with waterborne diseases as a result of poor water and sanitation systems.

In addition to wet lab and sequencing, there are also bioinformatics and data analysis challenges to address. A major bottleneck lies in establishing well-curated databases to facilitate the ability to make meaningful inferences from metagenomic data. Several recent studies have made considerable effort to sequence metagenomes and/or recover MAGs from cow rumen (Wilkinson et al. 2020, Almeida et al. 2020, Wang et. Al 2019, Stewart et al. 2019) as well as pig (Xiao et a. 2016, Wang et al. 2019) and chicken guts (Gilroy et al. 2020). However, this information has not yet been synthesized toward novel MST marker development. Future work should expand on the collection of host fecal samples

and integrate publicly available data to better assess the host specificity and sensitivity of markers across more broad geographical regions. A similar undertaking to catalogue the diversity of oil-associated microbes has already been successfully completed by other members of the Konstantinidis lab (Karthikeyan et al. 2020). The putative host-specific MAGs and functional genes identified in this thesis can be benchmarked against these more comprehensive databases in order to determine if they are applicable to more than just the local region of Georgia explored here. In addition to host specific databases for MST, curated collections of biological pathogens associated with fecal contamination (especially viruses and protozoa because less is known about them) would also be helpful for monitoring public health and predicting waterborne outbreaks since metagenomics can recover the pathogens, in addition to the FIB (or other biomarker) organisms. Ideally, MST and water quality monitoring markers should be quantitative and target abundance normalization methods be standardized in order for robust comparisons against regulatory standards. In this thesis we improved on the semi-quantitative capabilities of metagenomic methods using total cell counts from qPCR (**Chapter 3**) in order to estimate absolute abundances (cells/mL) and the limit of detection for target MST genomes in lake water mesocosm metagenomes. Therefore, the protocols presented in this thesis should be useful for applying metagenomics to water quality monitoring.

The work presented here applied cutting-edge, next-generation sequencing techniques for water quality monitoring. However, most municipalities are still primarily using traditional culture-based tests and official EPA guidelines have only recently begun to adopt molecular techniques like qPCR. For most local water quality monitoring agencies, even qPCR is not feasible because it requires specialized equipment and more

technical expertise compared to simple culturing tests. Therefore, in order for metagenomics to be viable for routine water monitoring, efforts should be made to automate not only the wet-lab components but also the bioinformatic analyses of the resulting data. This can be done by developing machine learning models to automatically classify different sources of fecal pollution with key genetic signatures from different hosts that are recovered in the metagenomic dataset. Ideally, the resulting models and databases would be wrapped into a single, centralized software that includes all quality filtering and trimming steps. Water managers could simply submit a water metagenome to the server, which reports results of the analysis as the probability of specific source pollution and public health risk.

This thesis also highlighted some of the limitations of using metagenomics for MST and public health surveillance such as issues related to the limits for detecting a target genome in an environmental metagenome. In high diversity environments such as sediments, detecting rare community members (e.g., pathogens) with metagenomics is likely not economically viable with brute force, high coverage sequencing. However, in more highly concentrated and less diverse microbial communities, such as sewage, public health surveillance with metagenomics has already shown promise based on the work conducted as part of this thesis as well as other parallel studies. For example, a large survey of sewage metagenomes from 60 countries found that antibiotic resistance gene (ARG) abundance was correlated to socioeconomic, health, and environmental factors and suggested that improving sanitation could limit global burden of antimicrobial resistance (Hendriksen et al. 2019). Viruses are also becoming increasingly important for MST and public health surveillance, especially in the wake of COVID19. The Konstantinidis and

Brown labs are already working on a project to monitor COVID19 cases and potential outbreaks on the Georgia Tech campus using raw sewage from the student dormitories. Viruses are the most abundant entities on the planet (orders of magnitude more abundant than bacteria). Bacteriophages, such as CrAssphage, are also promising areas to investigate for improved MST in environments where pollution is more dilute but still a significant risk to public health. As a result of cutting-edge metagenomic and viromic technologies, scientists have been able to characterize the immense microbial diversity in the human microbiome and use this information to help prevent and cure diseases. This thesis demonstrated that the application of these technologies for MST can similarly help to improve public health monitoring and risk assessment and ultimately help to reduce the incidence and burden of waterborne diseases.

APPENDIX A. SUPPLEMENTAL MATERIAL FOR CHAPTER 3

A.1 Supplemental figures and tables

Table A 1: Total community covered by our sequencing depth as determined by Nonpareil v3.0. Averages from three biological replicates and standard deviation are shown.

Day	Cow	Pig	Human	Neg Control
0	26.9 ± 2.4	49.5 ± 13.0	79.6 ± 3.4	46.4
1	27.4 ± 6.8	55.0 ± 3.0	80.9 ± 1.9	53.6
4	54.9 ± 4.0	59.2 ± 2.3	74.2 ± 10.9	70.8
7	79.6 ± 5.5	72.5 ± 9.5	77.7 ± 2.4	76.0

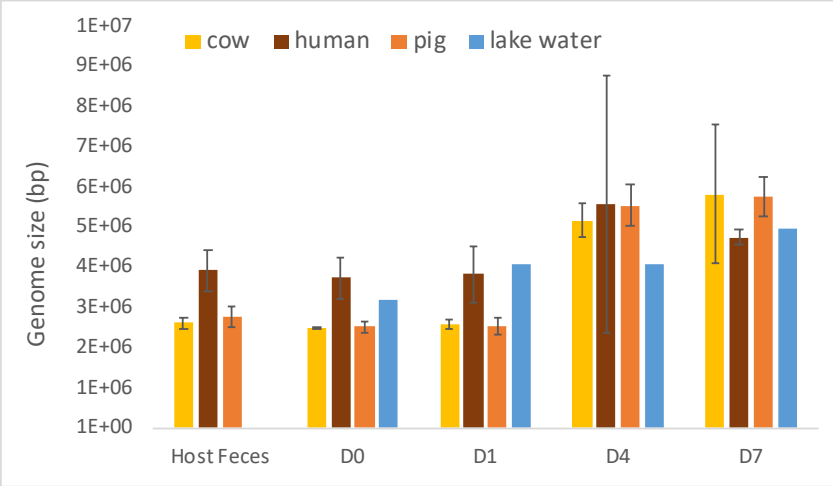


Figure A 2: Total community covered by our sequencing depth as determined by Nonpareil v3.0. Averages from three biological replicates and standard deviation are shown.

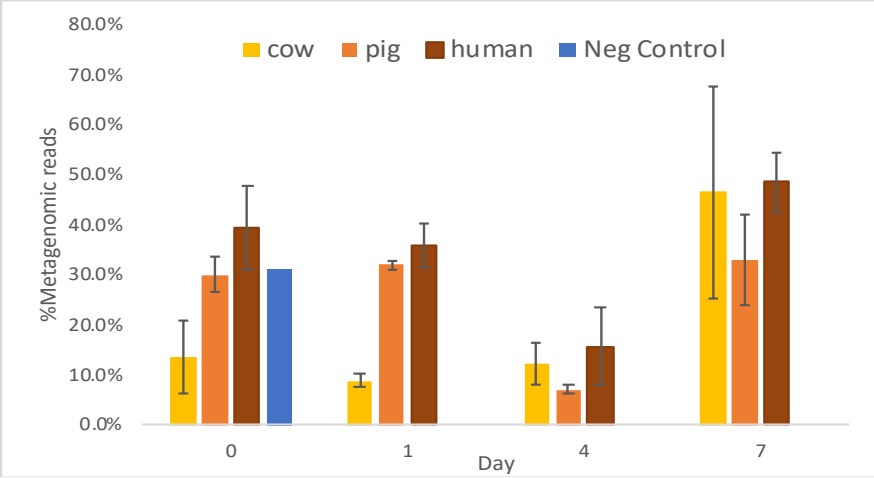


Figure A 2: Fraction of the microbial community represented by MAGs. The graph shows the total number of mesocosm metagenome reads that matched to any of the ~500 MAGs (fecal, D7 and Lake Lanier) that were used in this study.

Table A 2: Metagenome sample, trimming and nonpareil diversity information for dialysis bag mesocosm, lake water negative control and host fecal samples sequenced in this study. Results are reported for reads after quality trimming and removal of host DNA with bmtagger (for fecal samples only).

					Metagenomes			Nonpareil
Sample ID	Sample Type	Host Type	Day	Samplng Date	#paired reads (trimmed)	Avg. trimmed length (bp)	Sample size (Gbp)	Diversity
H1_D0	feces:lake water mix	human	0	180625	1.21E+07	131.27	1.59E+09	17.79
H2_D0	feces:lake water mix	human			8.64E+06	128.79	1.11E+09	18.17
H3_D0	feces:lake water mix	human			1.32E+07	125.77	1.67E+09	18.05
H1_D1	feces:lake water mix	human	1	180626	9.72E+06	130.66	1.27E+09	17.73
H2_D1	feces:lake water mix	human			1.25E+07	124.11	1.55E+09	18.11
H3_D1	feces:lake water mix	human			1.58E+07	124.72	1.97E+09	18.25
H1_D4	feces:lake water mix	human	4	180629	1.01E+07	131.89	1.33E+09	17.94
H2_D4	feces:lake water mix	human			1.73E+07	120.10	2.07E+09	18.11
H3_D4	feces:lake water mix	human			1.23E+07	122.47	1.50E+09	19.33
H1_D7	feces:lake water mix	human	7	180702	1.23E+07	126.89	1.56E+09	17.90
H2_D7	feces:lake water mix	human			8.35E+06	130.38	1.09E+09	17.74
H3_D7	feces:lake water mix	human			1.22E+07	123.73	1.52E+09	17.18
C7_D0	feces:lake water mix	cow	0	170928	8.19E+06	130.46	1.07E+09	21.24
C8_D0	feces:lake water mix	cow			8.64E+06	130.80	1.13E+09	21.01
C9_D0	feces:lake water mix	cow			8.78E+06	130.49	1.15E+09	21.14
C7_D1	feces:lake water mix	cow	1	170929	1.00E+07	121.65	1.22E+09	20.93
C8_D1	feces:lake water mix	cow			8.39E+06	131.85	1.11E+09	20.82
C9_D1	feces:lake water mix	cow			3.20E+06	121.48	3.89E+08	20.57
C7_D4	feces:lake water mix	cow	4	171002	7.82E+06	131.49	1.03E+09	19.52
C8_D4	feces:lake water mix	cow			9.76E+06	127.57	1.24E+09	18.74
C9_D4	feces:lake water mix	cow			7.19E+06	130.98	9.41E+08	19.28
C7_D7	feces:lake water mix	cow	7	171005	1.48E+07	125.75	1.86E+09	18.55
C8_D7	feces:lake water mix	cow			1.41E+07	124.22	1.75E+09	16.99
C9_D7	feces:lake water mix	cow			1.22E+07	123.14	1.50E+09	17.77
P7_D0	feces:lake water mix	pig	0	170928	1.07E+07	125.44	1.34E+09	19.94
P8_D0	feces:lake water mix	pig			8.99E+06	122.63	1.10E+09	20.53
P9_D0	feces:lake water mix	pig			1.14E+07	126.52	1.44E+09	19.11
P7_D1	feces:lake water mix	pig	1	170929	1.39E+07	128.11	1.79E+09	19.92
P8_D1	feces:lake water mix	pig			1.35E+07	126.77	1.72E+09	19.98
P9_D1	feces:lake water mix	pig			9.13E+06	131.38	1.20E+09	19.23
P7_D4	feces:lake water mix	pig	4	171002	1.07E+07	127.09	1.35E+09	19.22
P8_D4	feces:lake water mix	pig			1.05E+07	125.83	1.32E+09	19.29
P9_D4	feces:lake water mix	pig			6.19E+06	129.88	8.04E+08	19.05
P7_D7	feces:lake water mix	pig	7	171005	1.68E+07	123.94	2.08E+09	17.60
P8_D7	feces:lake water mix	pig			1.07E+07	124.64	1.33E+09	19.17
P9_D7	feces:lake water mix	pig			1.32E+07	125.95	1.66E+09	18.84
Human LLD0	Lake water	neg control	0	180625	1.11E+07	124.40	1.38E+09	20.33
Animal LLD0	Lake water	neg control	0	170928	1.01E+07	128.74	1.31E+09	20.73
Animal LLD1	Lake water	neg control	1	170929	3.54E+07	107.62	3.81E+09	20.94
Animal LLD4	Lake water	neg control	4	171002	3.39E+07	104.97	3.56E+09	19.52
Animal LLD7	Lake water	neg control	7	171005	3.64E+07	102.72	3.74E+09	19.00

Table A 2 continued

cow4	feces	cow	n/a	n/a	2.39E+07	123.46	2.96E+09	19.75
cow5	feces	cow	n/a	n/a	2.87E+07	118.37	3.39E+09	19.53
cow6	feces	cow	n/a	n/a	2.44E+07	121.66	2.97E+09	20.79
cow7	feces	cow	n/a	n/a	2.82E+07	125.81	3.55E+09	21.31
cow8	feces	cow	n/a	n/a	2.92E+07	125.42	3.67E+09	21.07
cow9	feces	cow	n/a	n/a	2.80E+07	121.80	3.42E+09	21.28
pig4	feces	pig	n/a	n/a	2.65E+07	124.30	3.29E+09	19.47
pig5	feces	pig	n/a	n/a	2.27E+07	124.20	2.82E+09	19.01
pig6	feces	pig	n/a	n/a	2.50E+07	128.33	3.21E+09	19.62
pig7	feces	pig	n/a	n/a	2.04E+07	123.17	2.51E+09	19.53
pig8	feces	pig	n/a	n/a	2.22E+07	124.54	2.76E+09	19.56
pig9	feces	pig	n/a	n/a	2.07E+07	123.44	2.55E+09	19.15
hum1	feces	human	n/a	n/a	7.46E+06	129.50	9.66E+08	17.83
hum2	feces	human	n/a	n/a	8.23E+06	130.07	1.07E+09	17.96
hum3	feces	human	n/a	n/a	9.39E+06	127.58	1.20E+09	17.14

Table A 3: qPCR assay reaction details and performance. (a) Primers and probes are listed in the following order: forward, reverse, hydrolysis probe. (b) All of the hydrolysis probes were labeled at the 5' end with the reporter dye FAM (6-carboxyfluorescein) and at the 3' end with a non-flourescent iowa black quencher (BHQ) with a minor groove binding moiety. Except the EPA1611 assay which used a TAMRA quencher dye as described in the EPA Method 1611. (c) Reported as the average for all plates ran EXCEPT for EPA1611, in which results for the composite curve used to calculate the average number of target sequences in calibrator cells are reported.

Assay ID	Target organism	Target gene	Primer/Probe name (a)	Primer/Probe sequence (5' to 3' direction) (b)	Primer/Probe Ref	Final Primer conc. (uM)	Standard curve intercept, slope, R2 (c)	%Efficiency
HF183	Human-specific *Bacteroides*	16S rRNA (V2 region)	HF183	ATCATGAGTTCACATGTCCG	Bernhard and Field 2000	0.25	39.52, -3.67, 1.00	87.90
			BFDRev	CGTAGGAGTTTGGACCGTGT	Converse 2009	0.25		
			BFDFAM	CTGAGAGGAAGGTCCCCCACATTGGA	Converse 2009	0.25		
RumBac	Ruminant-specific *Bacteroidetes*	16S rRNA	BacR_f	GCGTATCCAACCTTCCCG	Reischer et al. 2006	0.1	43.31, -3.64, 1.00	88.21
			BacR_r	CATCCCCATCCGTTACCG		0.1		
			BacR_p	CTTCCGAAAGGGAGATT		0.5		
HUMmt	Human mitochondrial genome	NADH dehydrogenase subunit 5	human forward	CAGCAGCCATTCAAGCAATGC	Caldwell et al. 2007	0.25	39.11, -3.58, 1.00	90.27
			human reverse	GGTGGAGACCTAATTGGGCTGATTAG		0.25		
			human probe	TATCGGCGATATCGGTTTCATCCTCG		0.25		
EF16S	*Enterococcus faecalis*	16S rRNA	E. faecalis forward	CGCTTCTTTCCTCCCGAGT	Santo-Domingo et al. 2003	0.25	39.59, -3.68, 1.00	87.08
			E. faecalis reverse	GCCATGCGGCATAAACTG		0.25		
			E. faecalis probe	CAATTGGAAAGAGGAGTGGCGGACG		0.25		
GenBac16S	Phylum Bacteria	Universal 16S rRNA	Bac1055YF	AATAAATCATAAACTCCTACGGGAGGCAGCAGT	Ritalahti et al. 2006	0.3	43.96, -3.87, 1.00	81.21
			Bac1392R	AATAAATCATAACCTAGCTATTACCGCGGCTGCT		0.3		
			Bac1115Probe	CGGCTAACTMCGTGCCAG		0.3		
EPA1611	*Enterococcus* spp.	23S rRNA	ECST748F	GAGAAATTCCAAACGAACTTG	EPA Method 1611	1	40.74, -3.42, 0.97	96.25
			ENC854R	CAGTGCTCTACCTCCATCATT		1		
			GPL813TQ	TGGTTCTCTCCGAAATAGCTTTAGGGCTA		0.08		

Table A 4: Assembly information for host fecal and Day 7 (D7) samples.

Sample ID	#contigs (>500bp)	Total length (bp)	N50	#Predicted Genes
cow4	34,981	9.35E+07	3169	115,590
cow5	37,498	8.18E+07	2283	102,542
cow6	40,433	8.50E+07	2170	108,942
cow7	43,338	9.01E+07	2113	109,649
cow8	51,152	1.01E+08	2007	125,567
cow9	45,384	8.65E+07	1913	105,379
pig4	41,805	1.22E+08	3964	143,550
pig5	32,798	8.84E+07	3277	107,477
pig6	48,532	1.33E+08	3354	158,585
pig7	36,439	9.41E+07	3009	113,188
pig8	40,874	1.05E+08	3008	125,345
pig9	40,308	1.06E+08	3062	127,158
hum1	15,219	5.70E+07	6111	55,972
hum2	16,772	6.35E+07	6464	65,157
hum3	13,127	4.71E+07	5913	49,717
H1_D7	16,680	4.90E+07	4149	52,059
H2_D7	10,259	4.17E+07	7402	42,521
H3_D7	7,784	2.46E+07	5105	26,536
C7_D7	20,697	7.74E+07	7095	79,699
C8_D7	7,523	3.05E+07	19034	31,461
C9_D7	19,243	7.03E+07	5130	69,004
P7_D7	28,659	8.33E+07	3763	90,883
P8_D7	30,318	7.65E+07	2744	83,548
P9_D7	26,478	8.84E+07	5184	93,057

Figure A 3: Heatmap of presence/absence for different phenotypes of all 13 human fecal MAGs as determined by Traitar.

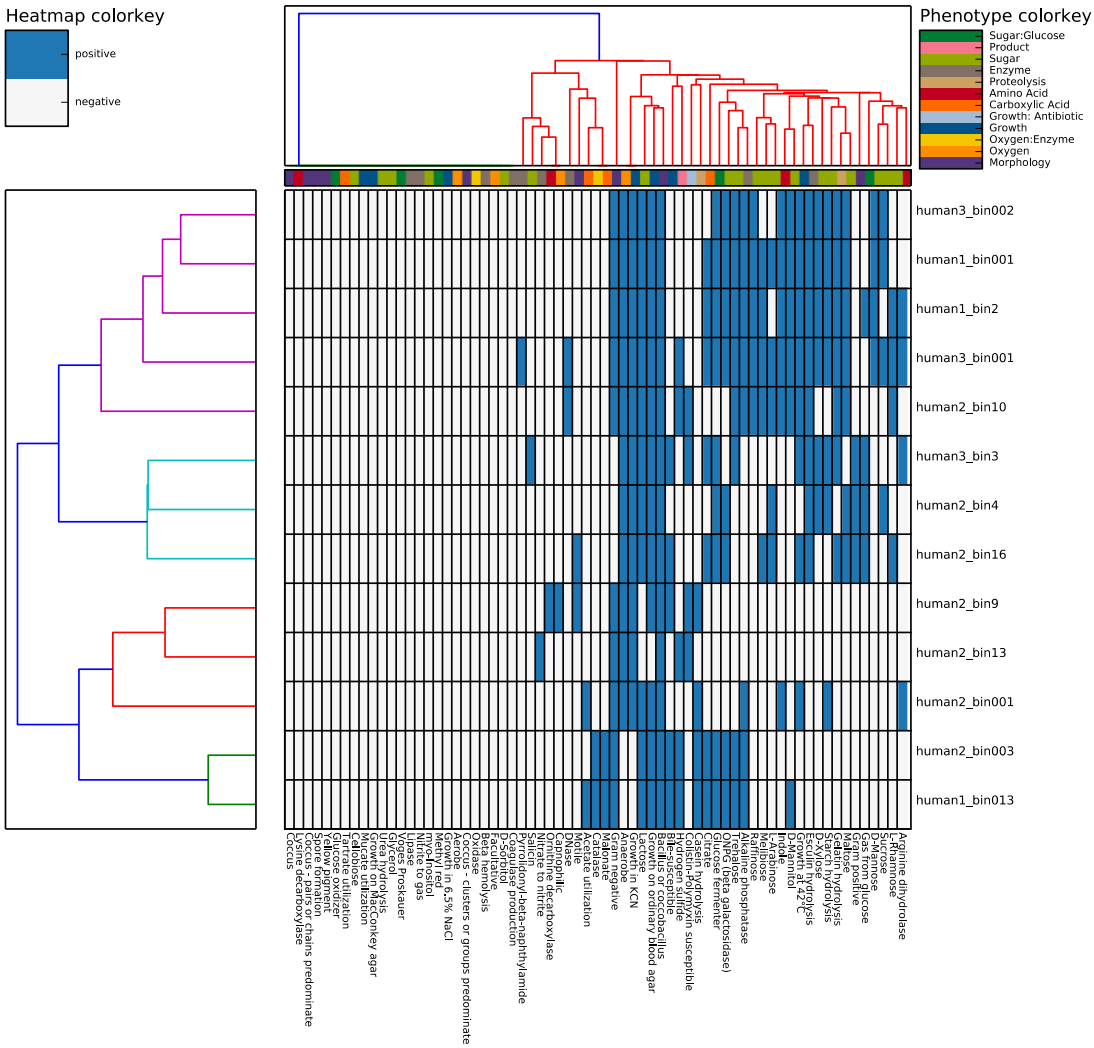


Figure A 4: Heatmap of presence/absence for different phenotypes of all 17 cow fecal MAGs as determined by Traitar.

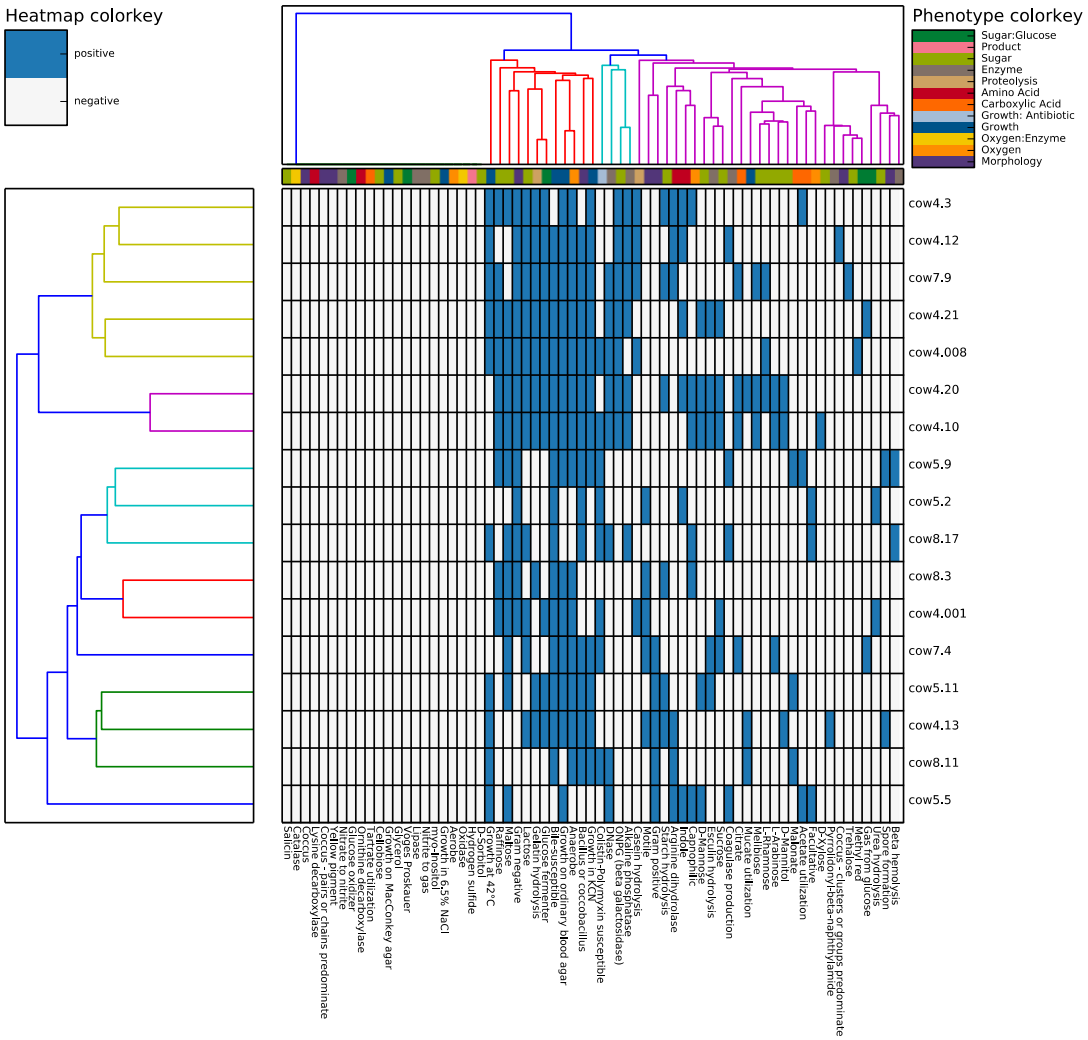


Figure A 5: Heatmap of presence/absence for different phenotypes of all 49 pig fecal MAGs as determined by Traitar.

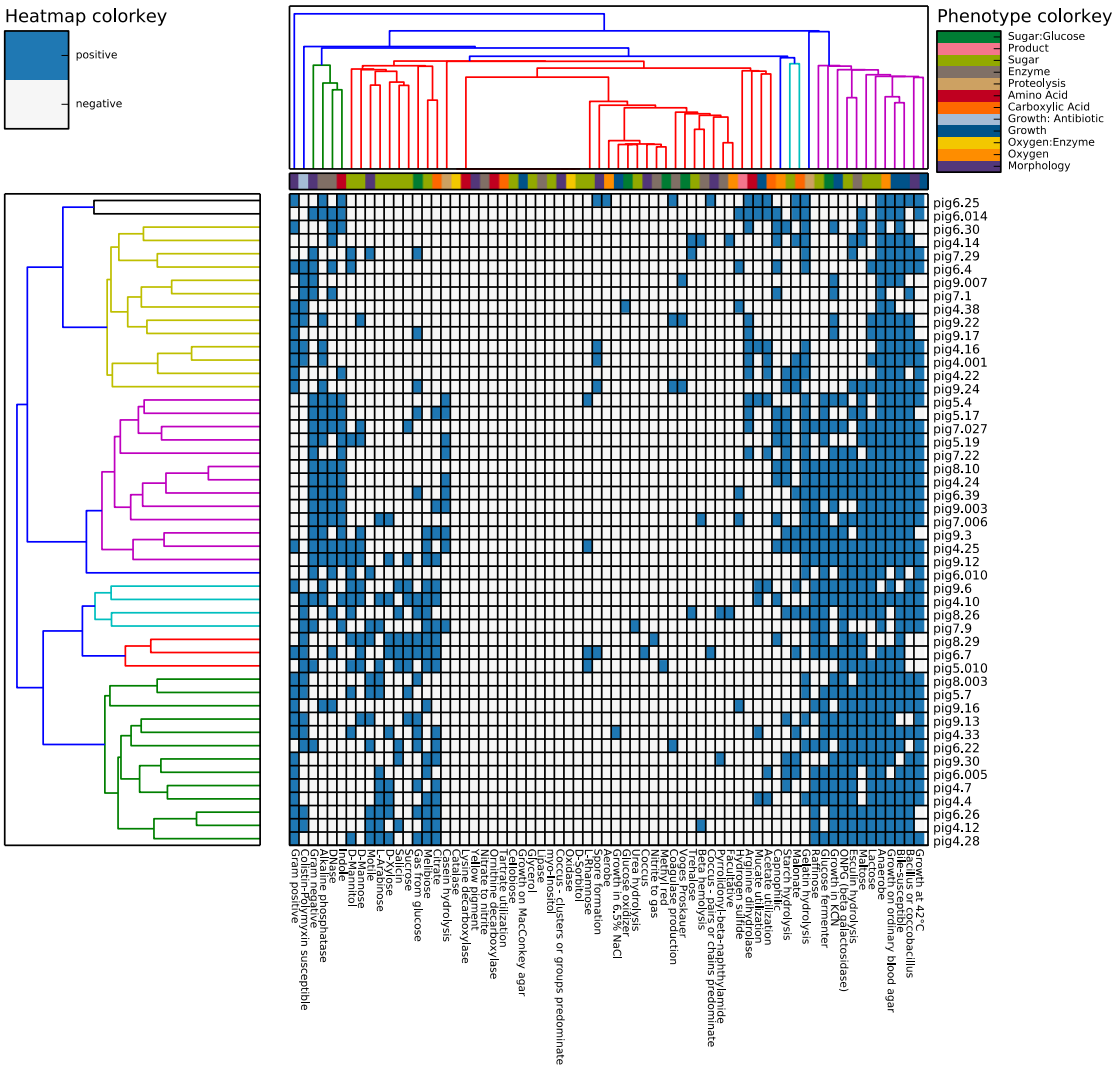


Figure A 6: KEGG modules present in host MAGs identified as potential biomarkers.

Only the MAGs identified as potential biomarkers are shown here for visualizaiton purposes.

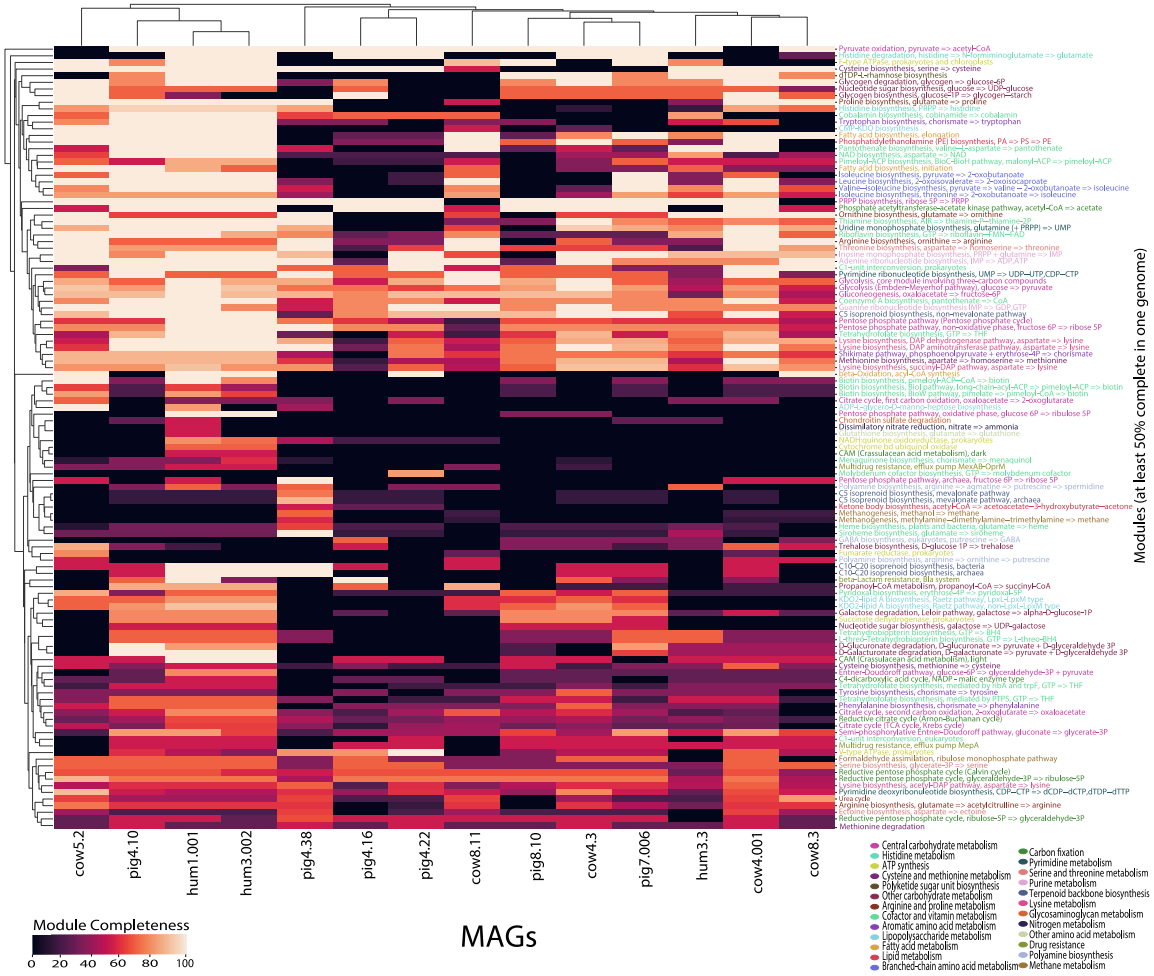


Figure A 7: Abundance kinetics of Day 7 (D7) MAGs in the fecal mesocosm samples over time. The collection of 17 D7 MAGs assembled from the D7 mesocosm metagenomes was searched against the time-series mesocosm metagenomes and their abundances (rows) are shown for all 17 MAGs that be detected in at least one metagenome. Each column is a mesocosm metagenome and only the D4 and D7 samples are shown because no D7 MAGs were detectable in any of the earlier time points. Naming style: number refers to the sample time in days while the H, C, or P refers the human, cow or pig biological replicate mesocosm. Abundance is expressed as % of total bacterial community (i.e. TAD80 divided by MicrobeCensus average genome sequencing depth).

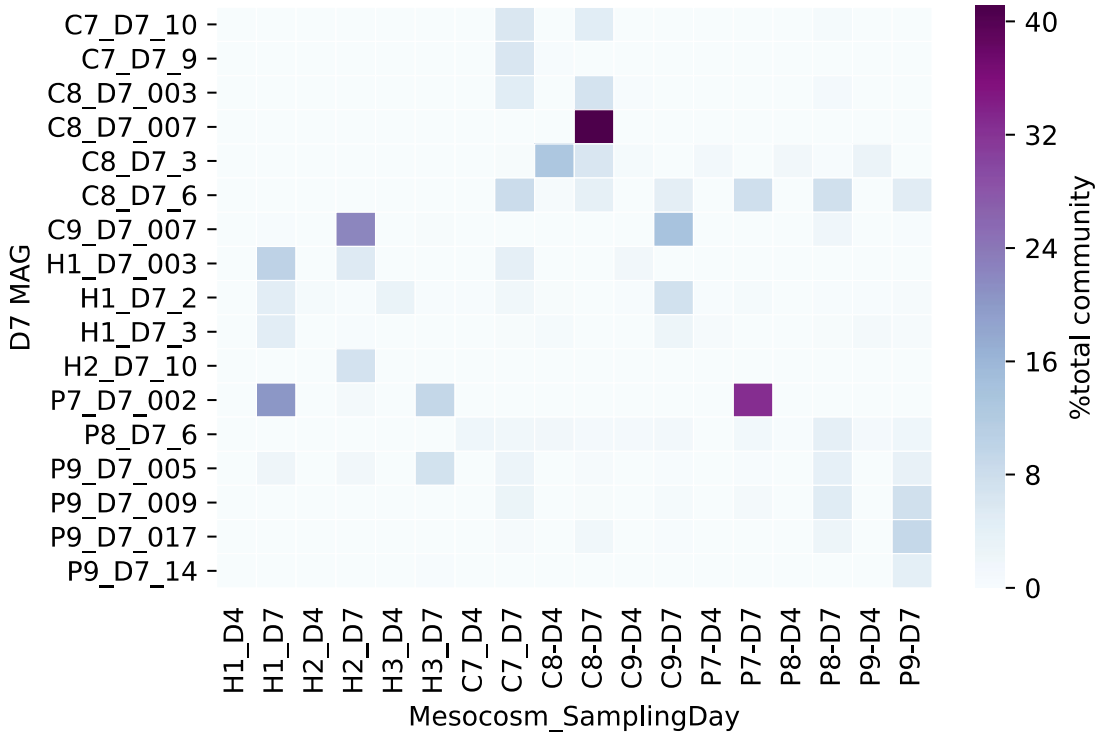
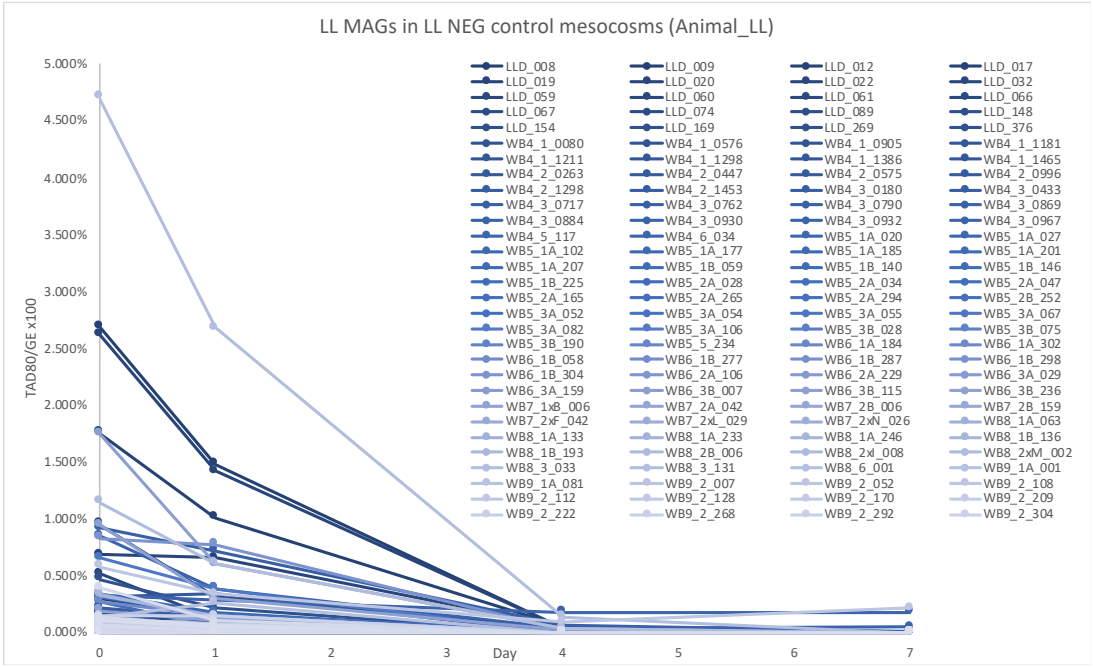


Figure A 8: Decay kinetics of MAGs in the uninoculated lake water negative control dialysis bags. Only a single set of negative control bags was used for the cow and pig experiments because they were carried out at the same time. Abundance is reported as the percent of the total bacterial community (i.e. TAD80 (>95%ID) divided by genome equivalents) **(Top)** Decay of the 139 LL MAGs that could be detected in any of the negative control mesocosms from the 477 LL MAG collection (Rodriguez-Rojas et al. 2019) (**Bottom**) Only 8 of the 17 high quality D7 MAGs could be detected in the D7 negative control sample. No D7 MAGs were detected in any of the earlier time points.



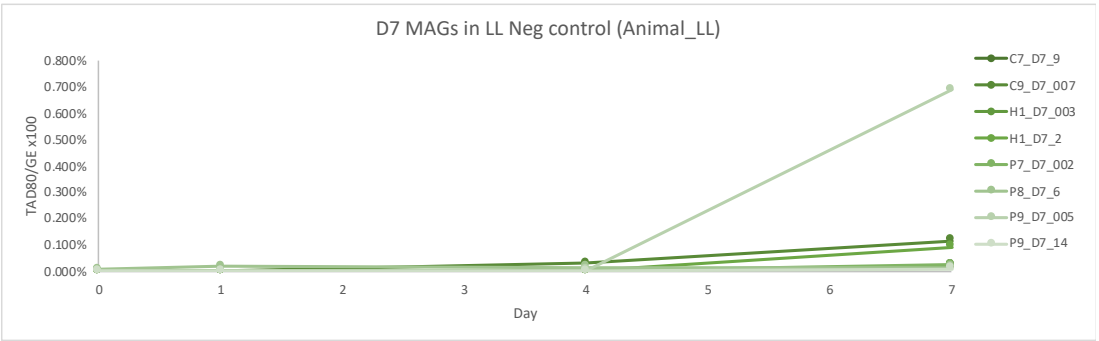


Figure A 9: Genetic relatedness among the host fecal and D7 MAG recovered by our study. Heatmap comparing average amino acid (%AAI) of the MAGs assembled from pig, cow, and human fecal and D7 mesocosm metagenomes.

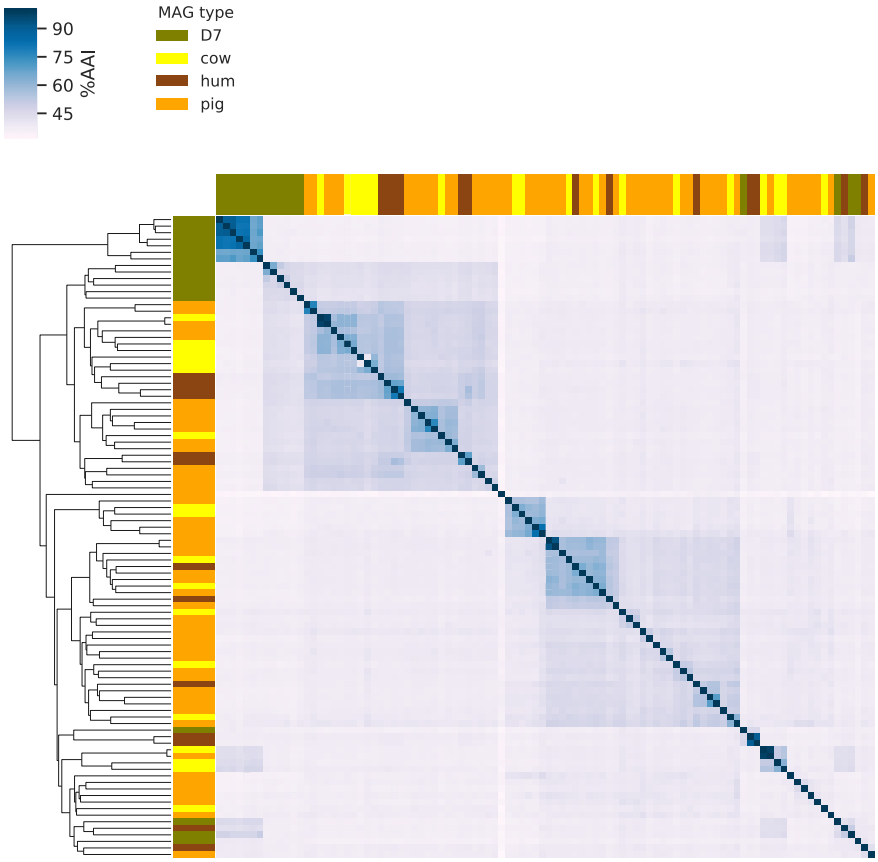


Figure A 10: Gene functions enriched between the host fecal and D7 mesocosm metagenomes. The heatmap shows the KEGG functions (rows) that were differentially abundant between the different host types (columns) with $P_{adj} < 0.05$ as determined by DESeq2 analysis. Color scale indicates the abundance relative to the average across all metagenome samples.

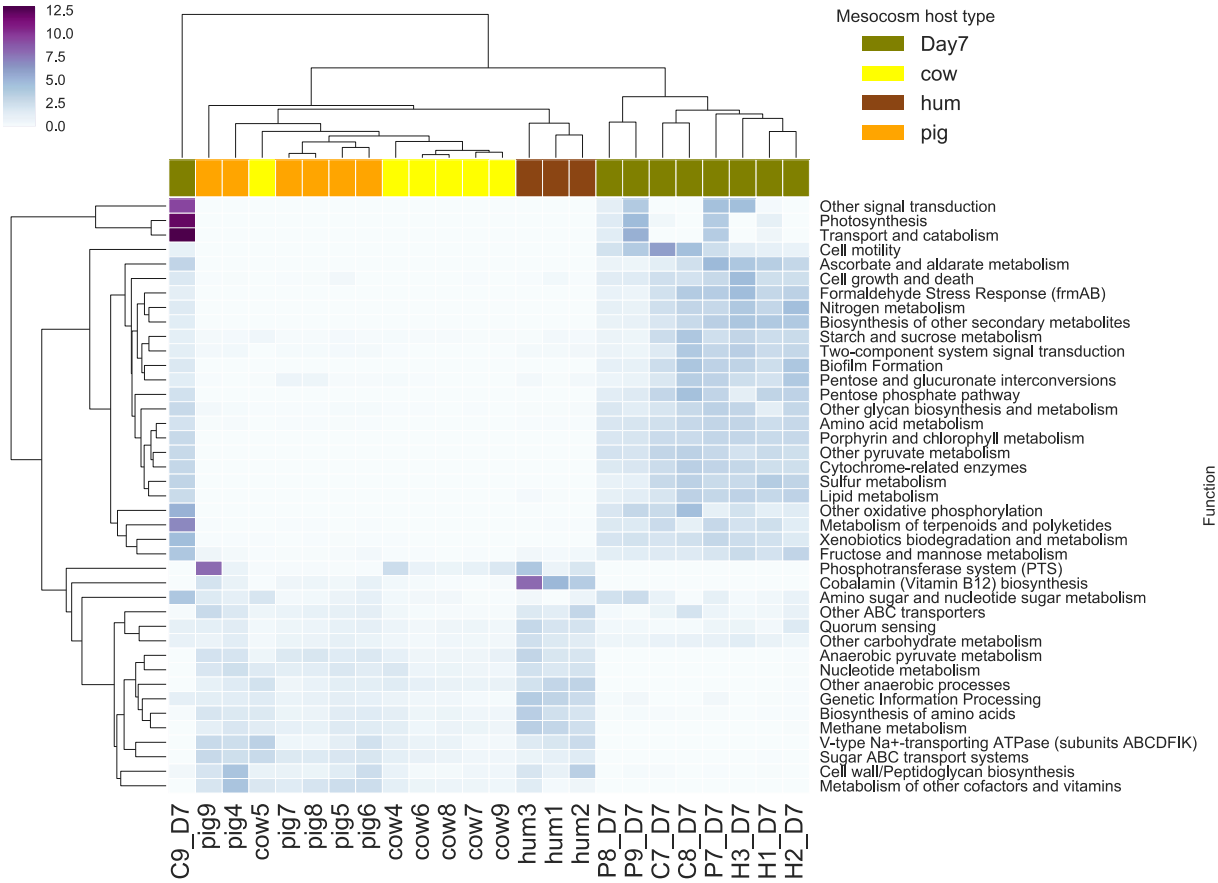
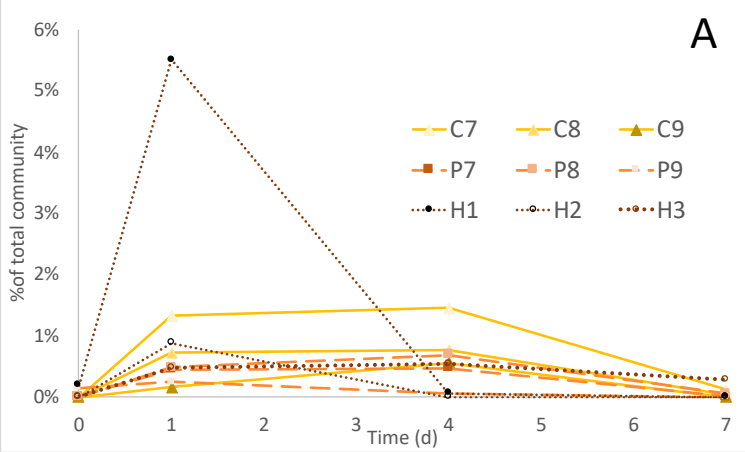
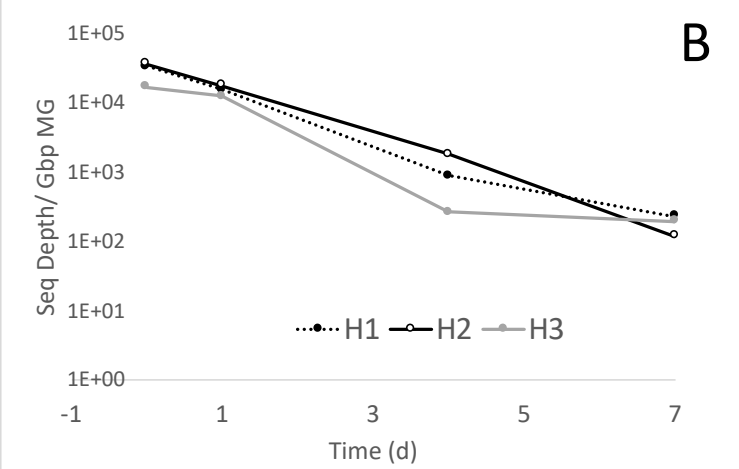


Figure A 11: Alignment of human fecal contigs to *B. dorei* 16S reference gene. Contigs from hum1 (top) and hum3 (bottom) fecal assemblies with best match (base on blastn search) to the 16S gene from the *B. dorei* reference genome (Table 3-1) with mismatches to the forward primer (highlighted in orange).

Figure A 12: Decay of refence genomes for common MST markers. The human mtGenome, *B. dorei* and CrAssphage were not detected in any of the cow or pig mesocosms or negative controls. **(A)** Abundance of the common enteric commensal strain, *Escherichia coli HS* (accession: NC_009800.1), in the fecal mesocosms over time. Abundance is reported as % of total bacterial community (i.e., TAD80 divided by genome equivalents; details in the main text). (**C**) Abundance of the human mitochondrial genome (accession: J01415.2) in the human fecal mesocosms expressed as sequencing depth (i.e., TAD80 at >95%ID) divided by MG library size in Gbp.





APPENDIX B. SUPPLEMENTAL MATERIAL FOR CHAPTER 4

B.1 Supplementary Figures and Tables

Table B 3: Trimming, assembly, and Nonpareil diversity information for all 27 metagenomes sequenced as part of this study.

Sample ID	Avg. read length (bp)	No. paired reads	Sample size (Gbp)	N50	No. contigs (>500bp)	Nonpareil %Coverage	Nonpareil Diversity
G130904	144.4	1.33E+07	3.85	769	58,546	24.10	22.43
G140116	142.5	1.53E+07	4.35	681	49,477	18.80	22.89
G140128	145.2	1.01E+07	2.93	725	14,356	16.60	22.58
G140210	137.1	1.48E+07	4.06	758	48,426	25.10	22.40
G140224	143.2	1.32E+07	3.79	805	60,908	27.00	22.27
G140301	134.4	1.73E+07	4.65	665	28,526	24.40	22.53
G140319	141.2	1.53E+07	4.31	706	55,117	24.20	22.49
G140402	132.6	1.87E+07	4.97	701	41,934	24.70	22.62
G140415	140.1	1.63E+07	4.56	674	41,527	22.00	22.69
G140611	135.2	1.79E+07	4.83	724	55,313	21.50	22.92
T130904	144.6	1.15E+07	3.24	648	24,568	12.80	23.39
T130918	144.5	1.18E+07	3.4	624	20,165	13.95	22.97
T131023	142.9	1.31E+07	3.75	637	27,470	19.32	22.85
T131230	144.1	1.45E+07	4.19	667	40,187	18.90	22.76
T140116	144.5	1.31E+07	3.78	676	26,568	23.25	*24.59*
T140128	143.9	1.14E+07	3.27	757	47,874	16.17	22.79
T140210	143.8	8.77E+06	2.52	656	22,643	12.12	22.95
T140224	140.9	1.50E+07	4.22	622	20,482	13.48	23.60
T140319	142.2	1.46E+07	4.16	662	38,677	23.33	22.93
T140611	131.9	2.01E+07	5.29	640	34,533	25.08	22.70
GC1	142.7	1.32E+07	3.68	604	8,409	9.39	23.45
GC2	141.5	1.26E+07	3.49	624	10,195	10.07	23.38
GC3	142.4	1.37E+07	3.78	645	21,043	13.01	23.15
TC1	145.7	1.51E+07	4.07	650	23,375	13.50	23.21
TC2	141.9	1.83E+07	4.71	745	23,752	17.90	23.03
WS1	141.7	1.51E+07	3.96	728	55,769	17.27	23.03
WS2	141.4	1.39E+07	3.62	716	53,809	14.49	23.28

Table B 2: Gene names for the top 30 most abundant ARGs listed in Appendix B, Figure B 7 as described in the ontology metadata files for the Comprehensive Antibiotic Resistance gene Database.

Gene ID	Gene Name or Description
bcrA	Bacitracin transport ATP-binding protein BcrA
macB	Macrolide export ATP-binding/permease protein MacB
APH(3'')-Ib	aminoglycoside phosphotransferase
sav1866	Putative multidrug export ATP-binding/permease protein SAV1866
PmrA	Response regulator for polymyxin resistance PmrA
arlR	Response regulator ArlR
mtrA	transcriptional activator of the MtrCDE multidrug efflux pump
dfrE	chromosome-encoded dihydrofolate reductase
novA	type III ABC transporter
otrC	tetracycline resistance efflux pump
vanRM	transcriptional activator
vanRF	Two-component response regulator
cpxR	Transcriptional regulatory protein CpxR
arlS	protein histidine kinase ArlS
lmrD	chromosomally-encoded efflux pump that confers resistance to lincosamides
baeR	Transcriptional regulatory protein BaeR
mdtC	Multidrug resistance protein MdtC
bacA	Undecaprenyl-diphosphatase
tlrC	Tylosin resistance ATP-binding protein TlrC
lmrC	chromosomally-encoded efflux pump that confers resistance to lincosamides
mexI	efflux pump membrane transporter MexI
carA	ABC transporter involved in macrolide resistance
dfrA3	integron-encoded dihydrofolate reductase
mexF	efflux pump membrane transporter MexF
oleB	ABC transporter in Streptomyces antibioticus and is involved in oleandomycin secretion
rosB	potassium antiporter rosB
msrC	chromosomal-encoded ABC-efflux pump
PmrF	Undecaprenyl-phosphate 4-deoxy-4-formamido-L-arabinose transferase
cpxA	Sensor histidine kinase CpxA
golS	Transcriptional regulatory protein GolS

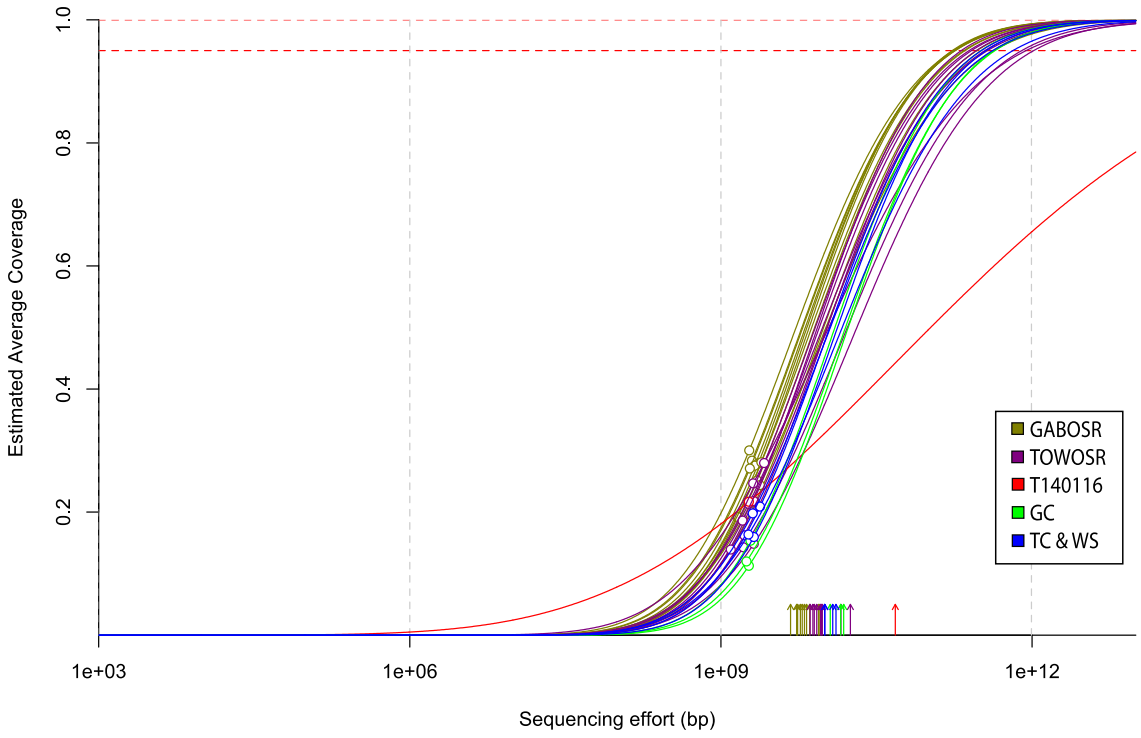


Figure B 1: Sequencing coverage of the microbial community datasets included in this study. Average community coverage was estimated (solid lines) using the Nonpareil algorithm. Empty circles represent the actual community coverage estimate at the sequencing effort applied and arrows are the Nonpareil diversity estimates. The lower horizontal red line indicates 95% average community coverage. The nonpareil diversity index and estimated average coverage for each sample can be found in Appendix B, Table B 2. Sample T140116 from TOWOSR was marked red due to its outlier diversity estimates.

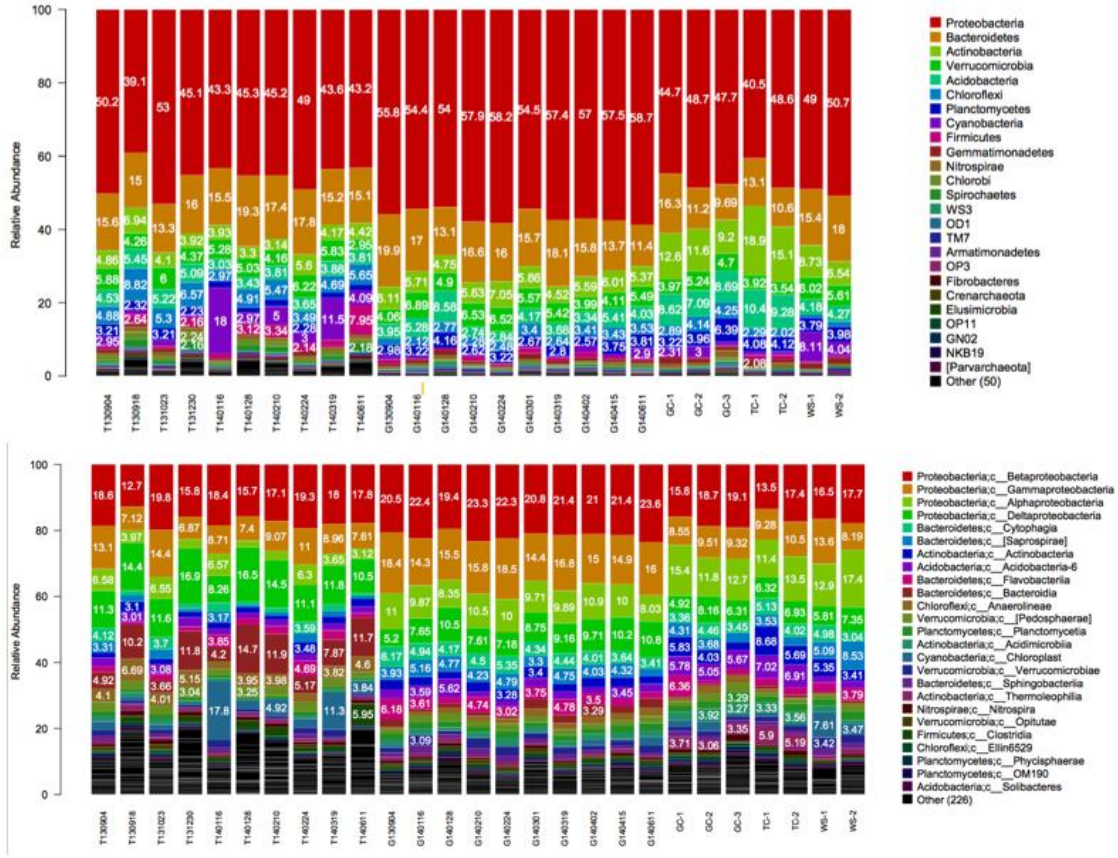


Figure B 2: Taxonomic composition of 16S rRNA gene OTUs. OTUs were formed by closed-reference OTU picking at 97% nucleotide identity level as implemented in MacQiime v1.9.1. **(A)** OTU abundances summarized at phylum level. **(B)** Class level OTUs with >3% relative abundance. Samples from TOWOSR and GABOSR are presented by location through time (indicated by letter T or G and the sample collection date).

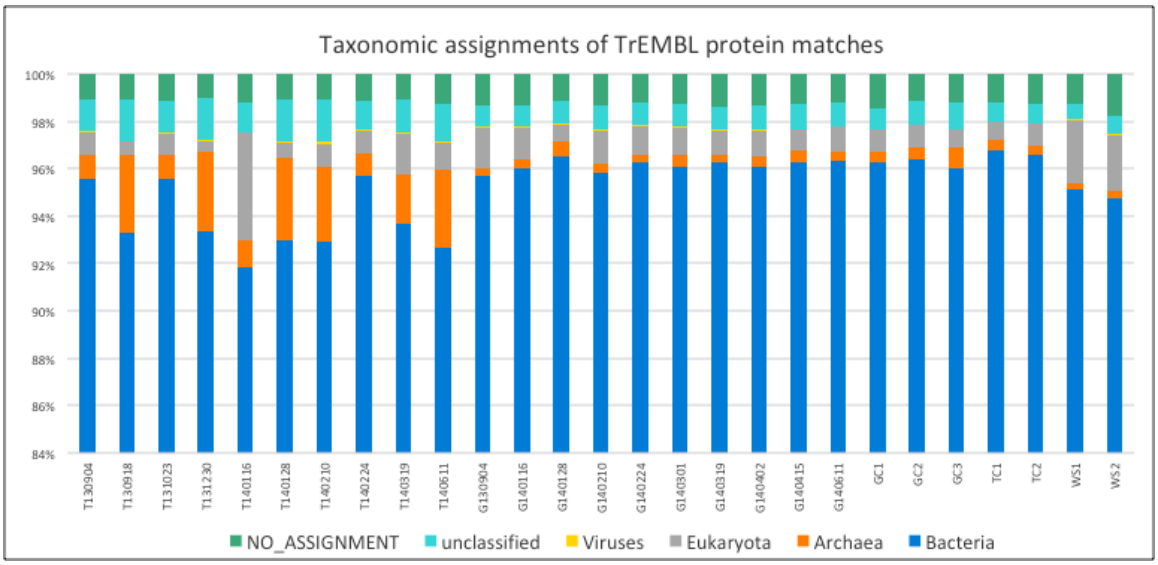


Figure B 3: Taxonomic assignment of functional gene-encoding reads at domain level based on TrEMBL annotations. Protein-coding short reads were annotated against the TrEMBL database and converted to the phylum level classification as described in the main text.

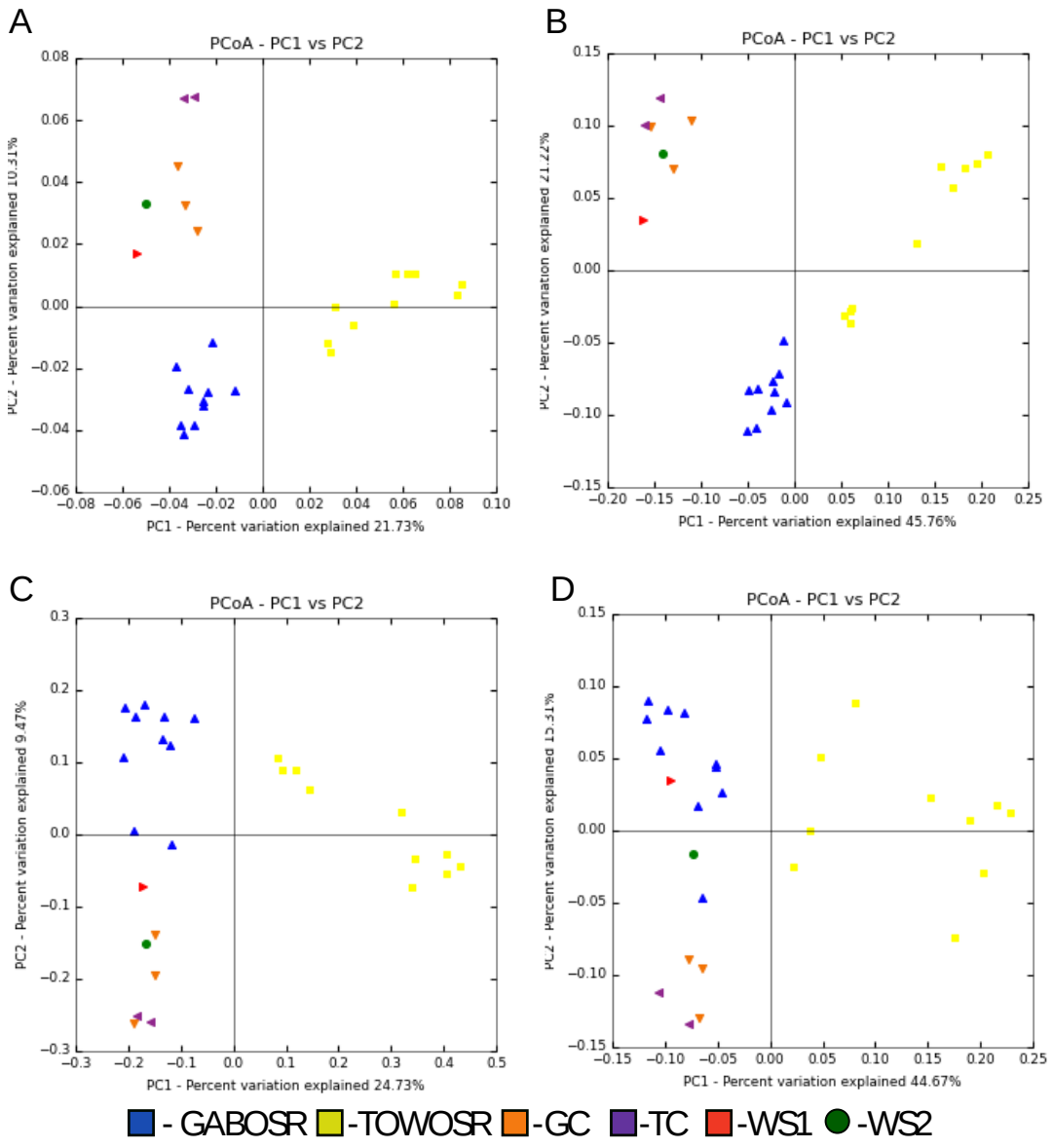


Figure B 4: Spatial separation is the strongest driver of functional and taxonomic diversity among Salinas Valley microbial communities. PCA ordination of (A) MASH (B) functional gene, (C) 16S rRNA gene OTU Bray-Curtis and (D) weighted UniFrac distances. Note that location (as shown by the different colors) was the only significant correlating parameter in all four dataset/ordinations (ADONIS: P=0.001, R2= 0.44, 0.67, 0.41, and 0.56, respectively).

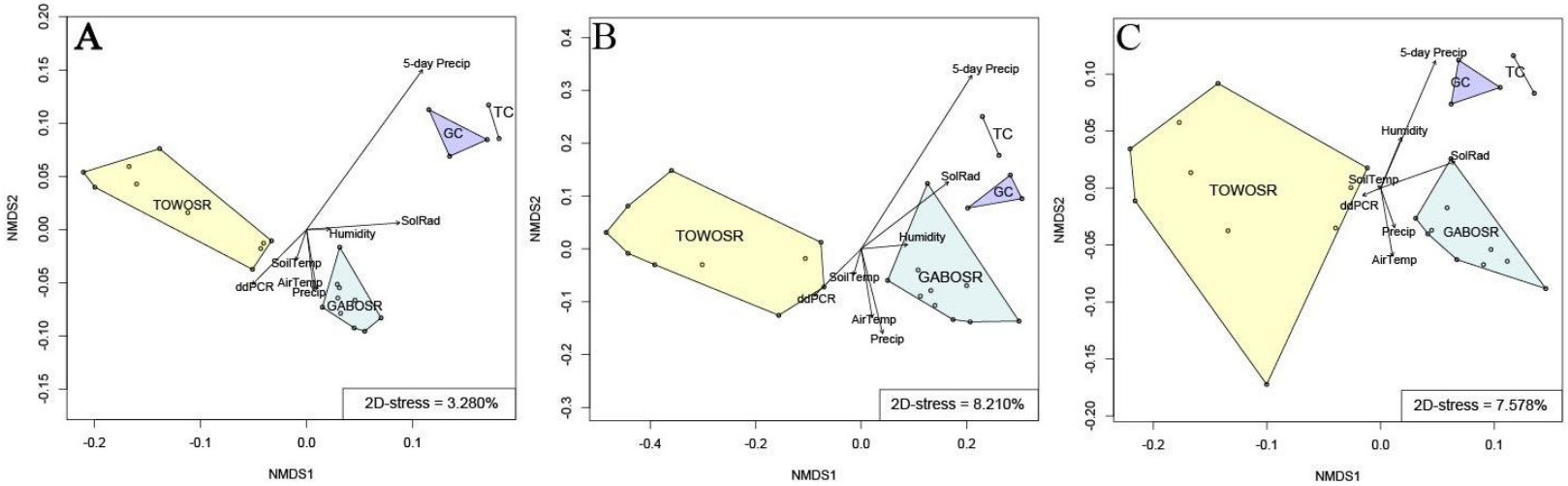


Figure B 5: Correlation of diversity patterns to measured environmental parameters. Non-metric multidimensional scaling using Bray-Curtis dissimilarities of (A) functional gene, (B) 16S rRNA gene OTU, and (C) Weighted Unifrac distances of OTU count data. West Salinas samples (WS1 and WS2) were omitted in order to minimize confounding variation of time and space differences (see main text for further details). Samples are grouped into locations as specified in Table 4-2, which was significantly correlated to all three datasets/ordinations (A: P=0.001, R^2=0.845; B: P=0.001, R^2=0.0.787; C: P=0.001, R^2=0.726). Arrow vectors show correlation with local weather data.

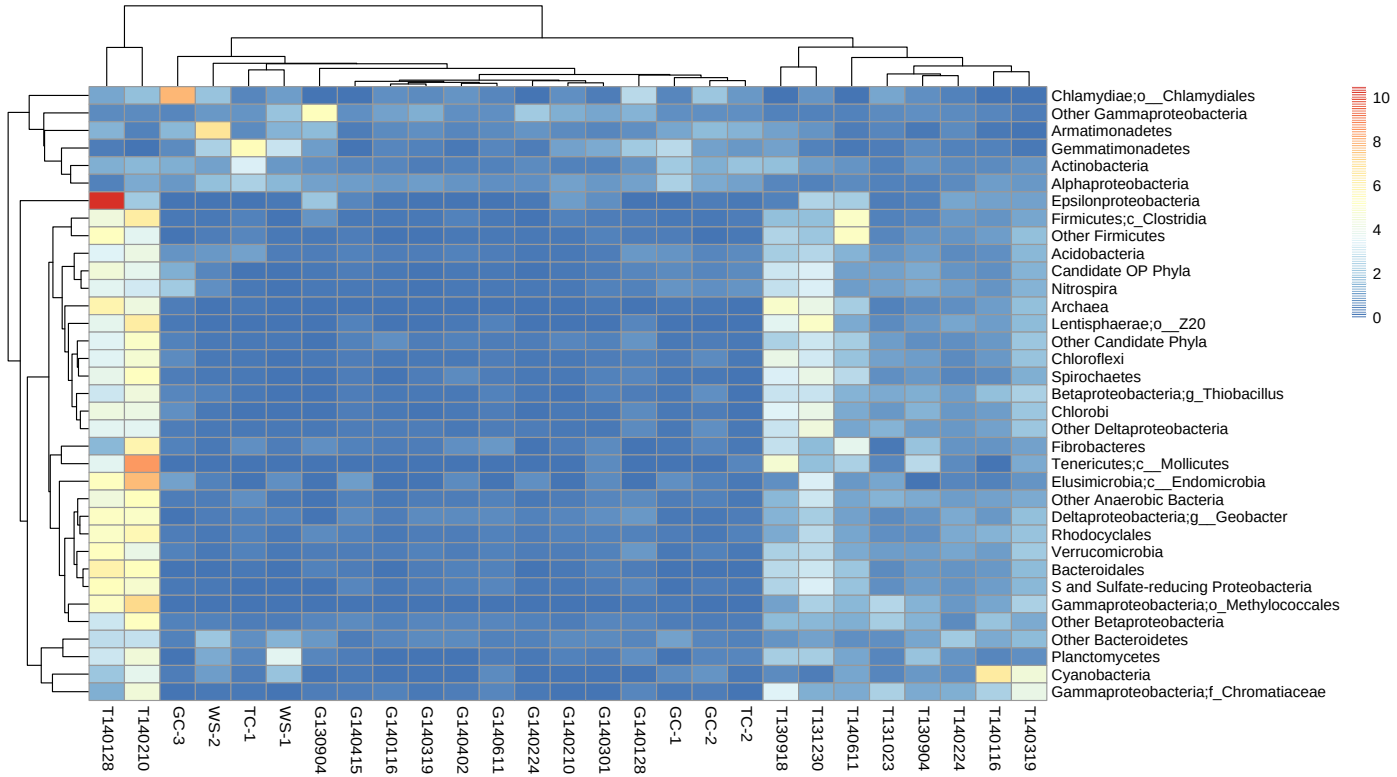


Figure B 6: Taxonomic profile of differentially abundant OTUs between the Salinas Valley creek sediment communities. Heatmap showing count data for 795 differentially abundant OTUs between locations (TOWOSR, GABOSR, and Upstream) with $\log_2$ fold change > 2 and $P_{adj} < 0.05$ that were summarized into 35 taxonomic groups as described in Chapter 4

Supplementary Data S2. Color scale indicates the abundance relative to the average of all samples (increasing from blue to red). Letters T or G and date in the column names represent the sample site (TOWOSR or GABOSR) and collection date, respectively. TC, GC and WS represent the upstream TOWOSR Control, GABOSR Control, and West Salinas, respectively.

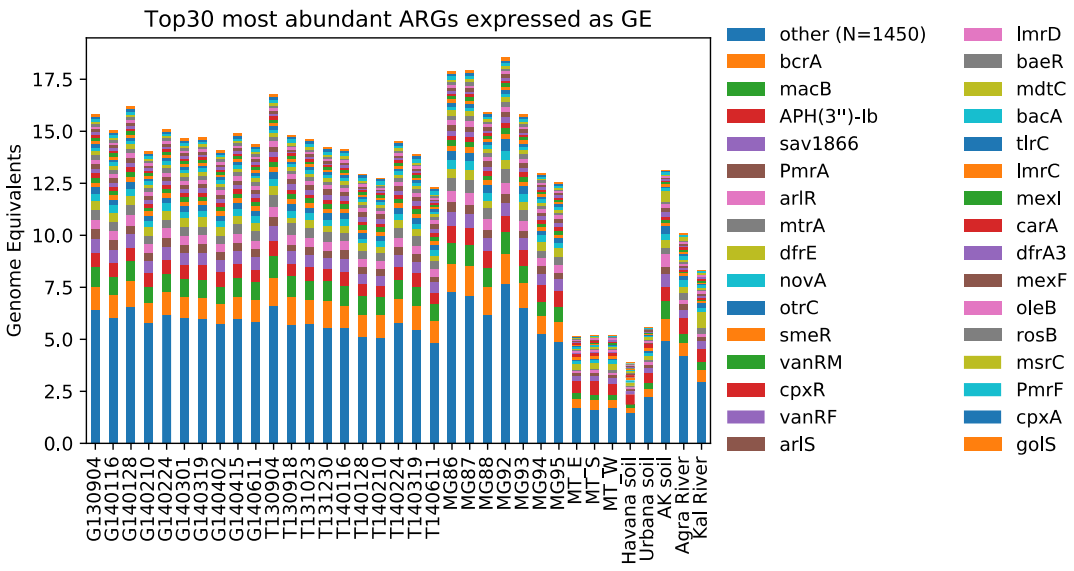


Figure B 7: Abundance of reads annotated as ARGs. Short reads were searched against the Comprehensive Antibiotic Resistance Database. The sequencing depth for each gene was divided by the normalization factor for each sample (average genome sequencing depth) and summed to get total genome equivalents per sample, (i.e., fraction of total genomes encoding the gene of interest, assuming a single-copy for each per genome). Full names or descriptions for the 30 genes listed here are provided in Appendix B, Table B 2.

REFERENCES

Abia, Akebe Luther King, Arghavan Alisoltani, Jitendra Keshri, and Eunice Ubomba-Jaswa. 2018. "Metagenomic Analysis of the Bacterial Communities and Their Functional Profiles in Water and Sediments of the Apies River, South Africa, as a Function of Land Use." *The Science of the Total Environment* 616–617 (March): 326–34. https://doi.org/10.1016/j.scitotenv.2017.10.322.

Ahmed, Warish, Bridie Hughes, and Valerie J. Harwood. 2016. "Current Status of Marker Genes of Bacteroides and Related Taxa for Identifying Sewage Pollution in Environmental Waters." *Water* 8 (6): 231.

Ahmed, Warish, Aldo Lobos, Jacob Senkbeil, Jayme Peraud, Javier Gallard, and Valerie J. Harwood. 2018. "Evaluation of the Novel CrAssphage Marker for Sewage Pollution Tracking in Storm Drain Outfalls in Tampa, Florida." *Water Research* 131 (March): 142–50. https://doi.org/10.1016/j.watres.2017.12.011.

Ahmed, Warish, Sudhi Payyappat, Michele Cassidy, Colin Besley, and Kaye Power. 2018. "Novel CrAssphage Marker Genes Ascertain Sewage Pollution in a Recreational Lake Receiving Urban Stormwater Runoff." *Water Research*, August. https://doi.org/10.1016/j.watres.2018.08.049.

Allen, Heather K., Luke A. Moe, Jitsupang Rodbumrer, Andra Gaarder, and Jo Handelsman. 2009. "Functional Metagenomics Reveals Diverse Beta-Lactamases in a Remote Alaskan Soil." *The ISME Journal* 3 (2): 243–51. https://doi.org/10.1038/ismej.2008.86.

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11 (10): R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Anderson, K. L., J. E. Whitlock, and V. J. Harwood. 2005. "Persistence and Differential Survival of Fecal Indicator Bacteria in Subtropical Waters and Sediments." *Applied and Environmental Microbiology* 71 (6): 3041–48. https://doi.org/10.1128/AEM.71.6.3041-3048.2005.

Arnold, Benjamin F., Timothy J. Wade, Jade Benjamin-Chung, Kenneth C. Schiff, John F. Griffith, Alfred P. Dufour, Stephen B. Weisberg, and John M. Colford. 2016. "Acute Gastroenteritis and Recreational Water: Highest Burden Among Young US Children." *American Journal of Public Health* 106 (9): 1690–97. https://doi.org/10.2105/AJPH.2016.303279.

Bae, S. W., and S. Wuertz. 2009. "Discrimination of Viable and Dead Fecal Bacteroidales Bacteria by Quantitative PCR with Propidium Monoazide." *Applied and*

Environmental Microbiology 75 (9): 2940–44. https://doi.org/10.1128/aem.01333-08.

Berendonk, Thomas U., Célia M. Manaia, Christophe Merlin, Despo Fatta-Kassinos, Eddie Cytryn, Fiona Walsh, Helmut Bürgmann, et al. 2015. "Tackling Antibiotic Resistance: The Environmental Framework." *Nature Reviews Microbiology* 13 (5): 310–17. https://doi.org/10.1038/nrmicro3439.

Boehm, A. B., L. C. Van De Werfhorst, J. F. Griffith, P. A. Holden, J. A. Jay, O. C. Shanks, D. Wang, and S. B. Weisberg. 2013. "Performance of Forty-One Microbial Source Tracking Methods: A Twenty-Seven Lab Evaluation Study." *Water Res* 47 (18): 6812–28. https://doi.org/10.1016/j.watres.2012.12.046.

Bowen. 2011. "Microbial Community Composition in Sediments Resists Perturbation by Nutrient Enrichment | The ISME Journal." 2011. https://www.nature.com/articles/ismej201122.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Byappanahalli, M. N., M. B. Nevers, A. Korajkic, Z. R. Staley, and V. J. Harwood. 2012. "Enterococci in the Environment." *Microbiol Mol Biol Rev* 76 (4): 685–706. https://doi.org/10.1128/MMBR.00023-12.

Byappanahalli, Muruleedhara N., Meredith B. Nevers, Richard L. Whitman, and Satoshi Ishii. 2015. "Application of a Microfluidic Quantitative Polymerase Chain Reaction Technique To Monitor Bacterial Pathogens in Beach Water and Complex Environmental Matrices." *Environmental Science & Technology Letters* 2 (12): 347–51. https://doi.org/10.1021/acs.estlett.5b00251.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (5): 335–36. https://doi.org/10.1038/nmeth.f.303.

Carducci, A., R. Battistini, E. Rovini, and M. Verani. 2009. "Viral Removal by Wastewater Treatment: Monitoring of Indicators and Pathogens." *Food and Environmental Virology* 1 (2): 85–91. https://doi.org/10.1007/s12560-009-9013-x.

Castro, Juan C., Luis M. Rodriguez-R, William T. Harvey, Michael R. Weigand, Janet K. Hatt, Michelle Q. Carter, and Konstantinos T. Konstantinidis. 2018. "ImGLAD:

Accurate Detection and Quantification of Target Organisms in Metagenomes." *PeerJ* 6 (November). https://doi.org/10.7717/peerj.5882.

Chao, Anne. 1984. "Nonparametric Estimation of the Number of Classes in a Population." *Scandinavian Journal of Statistics* 11 (4): 265–70.

Chao, Anne, and Tsung-Jen Shen. 2003. "Nonparametric Estimation of Shannon's Index of Diversity When There Are Unseen Species in Sample." *Environmental and Ecological Statistics* 10 (4): 429–43. https://doi.org/10.1023/A:1026096204727.

Chen, Baowei, Ximei Liang, Xiaoping Huang, Tong Zhang, and Xiangdong Li. 2013. "Differentiating Anthropogenic Impacts on ARGs in the Pearl River Estuary by Using Suitable Gene Indicators." *Water Research* 47 (8): 2811–20. https://doi.org/10.1016/j.watres.2013.02.042.

Cinek, Ondrej, Karla Mazankova, Lenka Kramna, Rasha Odeh, Abeer Alassaf, MaryAnn U. Ibekwe, Gunduz Ahmadov, et al. 2018. "Quantitative CrAssphage Real-Time PCR Assay Derived from Data of Multiple Geographically Distant Populations." *Journal of Medical Virology* 90 (4): 767–71. https://doi.org/10.1002/jmv.25012.

Cloutier, Danielle D., and Sandra L. McLellan. 2017. "Distribution and Differential Survival of Traditional and Alternative Indicators of Fecal Pollution at Freshwater Beaches." Edited by Shuang-Jiang Liu. *Applied and Environmental Microbiology* 83 (4). https://doi.org/10.1128/AEM.02881-16.

Cooley, Michael B., Diana Carychao, and Lisa Gorski. 2018. "Optimized Co-Extraction and Quantification of DNA From Enteric Pathogens in Surface Water Samples Near Produce Fields in California." *Frontiers in Microbiology* 9: 448. https://doi.org/10.3389/fmicb.2018.00448.

Cooley, Michael B., Michele Jay-Russell, Edward R. Atwill, Diana Carychao, Kimberly Nguyen, Beatriz Quiñones, Ronak Patel, et al. 2013. "Development of a Robust Method for Isolation of Shiga Toxin-Positive Escherichia Coli (STEC) from Fecal, Plant, Soil and Water Samples from a Leafy Greens Production Region in California." *PloS One* 8 (6): e65716. https://doi.org/10.1371/journal.pone.0065716.

Cooley, Michael B., Beatriz Quiñones, David Oryang, Robert E. Mandrell, and Lisa Gorski. 2014. "Prevalence of Shiga Toxin Producing Escherichia Coli, Salmonella Enterica, and Listeria Monocytogenes at Public Access Watershed Sites in a California Central Coast Agricultural Region." *Frontiers in Cellular and Infection Microbiology* 4: 30. https://doi.org/10.3389/fcimb.2014.00030.

Cooley, Michael, Diana Carychao, Leta Crawford-Miksza, Michele T. Jay, Carol Myers, Christopher Rose, Christine Keys, Jeff Farrar, and Robert E. Mandrell. 2007. "Incidence and Tracking of Escherichia Coli O157:H7 in a Major Produce Production Region in California." Edited by Debbie Fox. *PLoS ONE* 2 (11): e1159. https://doi.org/10.1371/journal.pone.0001159.

Costa, Patrícia S., Mariana P. Reis, Marcelo P. Ávila, Laura R. Leite, Flávio M. G. de Araújo, Anna C. M. Salim, Guilherme Oliveira, Francisco Barbosa, Edmar Chartone-Souza, and Andréa M. A. Nascimento. 2015. "Metagenome of a Microbial Community Inhabiting a Metal-Rich Tropical Stream Sediment." *PLoS ONE* 10 (3). https://doi.org/10.1371/journal.pone.0119465.

Cytryn, Eddie. 2013. "The Soil Resistome: The Anthropogenic, the Native, and the Unknown." *Soil Biology and Biochemistry* Complete (63): 18–23. https://doi.org/10.1016/j.soilbio.2013.03.017.

D'Costa, Vanessa M., Christine E. King, Lindsay Kalan, Mariya Morar, Wilson W. L. Sung, Carsten Schwarz, Duane Froese, et al. 2011. "Antibiotic Resistance Is Ancient." *Nature* 477 (7365): 457–61. https://doi.org/10.1038/nature10388.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. "Greengenes, a Chimera-Checked 16S RRNA Gene Database and Workbench Compatible with ARB." *Applied and Environmental Microbiology* 72 (7): 5069–72. https://doi.org/10.1128/AEM.03006-05.

Dhanji, Hiran, Niamh M. Murphy, Christine Akhigbe, Michel Doumith, Russell Hope, David M. Livermore, and Neil Woodford. 2011. "Isolation of Fluoroquinolone-Resistant O25b:H4-ST131 Escherichia Coli with CTX-M-14 Extended-Spectrum β-Lactamase from UK River Water." *Journal of Antimicrobial Chemotherapy* 66 (3): 512–16. https://doi.org/10.1093/jac/dkq472.

Dorner, Sarah M., William B. Anderson, Robin M. Slawson, Nicholas Kouwen, and Peter M. Huck. 2006. "Hydrologic Modeling of Pathogen Fate and Transport." *Environmental Science & Technology* 40 (15): 4746–53.

Durso, Lisa M, and Kimberly L Cook. 2014. "Impacts of Antibiotic Use in Agriculture: What Are the Benefits and Risks?" *Current Opinion in Microbiology* 19 (June): 37–44. https://doi.org/10.1016/j.mib.2014.05.019.

Dutilh, Bas E., Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, Lance Boling, Jeremy J. Barr, et al. 2014. "A Highly Abundant Bacteriophage Discovered in the Unknown Sequences of Human Faecal Metagenomes." *Nature Communications* 5 (1). https://doi.org/10.1038/ncomms5498.

Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster than BLAST." *Bioinformatics* 26 (19): 2460–61. https://doi.org/10.1093/bioinformatics/btq461.

Eisenberg, Joseph N.S., Jamie Bartram, and Timothy J. Wade. 2016. "The Water Quality in Rio Highlights the Global Public Health Concern Over Untreated Sewage." *Environmental Health Perspectives* 124 (10): A180–81. https://doi.org/10.1289/EHP662.

Eren, A. M., L. Maignien, W. J. Sul, L. G. Murphy, S. L. Grim, H. G. Morrison, and M. L. Sogin. 2013. "Oligotyping: Differentiating between Closely Related Microbial Taxa Using 16S RRNA Gene Data." *Methods Ecol Evol* 4 (12). https://doi.org/10.1111/2041-210X.12114.

Field, K. G., and M. Samadpour. 2007. "Fecal Source Tracking, the Indicator Paradigm, and Managing Water Quality." *Water Res* 41 (16): 3517–38. https://doi.org/10.1016/j.watres.2007.06.056.

Fisher, J. C., A. M. Eren, H. C. Green, O. C. Shanks, H. G. Morrison, J. H. Vineis, M. L. Sogin, and S. L. McLellan. 2015. "Comparison of Sewage and Animal Fecal Microbiomes by Using Oligotyping Reveals Potential Human Fecal Indicators in Multiple Taxonomic Groups." *Appl Environ Microbiol* 81 (20): 7023–33. https://doi.org/10.1128/AEM.01524-15.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data." *Bioinformatics* 28 (23): 3150–52. https://doi.org/10.1093/bioinformatics/bts565.

García-Aljaro, Cristina, Elisenda Ballesté, Maite Muniesa, and Juan Jofre. 2017. "Determination of CrAssphage in Water Samples and Applicability for Tracking Human Faecal Pollution." *Microbial Biotechnology* 10 (6): 1775–80. https://doi.org/10.1111/1751-7915.12841.

Gibbons, Sean M., Edwin Jones, Angelita Bearquiver, Frederick Blackwolf, Wayne Roundstone, Nicole Scott, Jeff Hooker, Robert Madsen, Maureen L. Coleman, and Jack A. Gilbert. 2014. "Human and Environmental Impacts on River Sediment Microbial Communities." Edited by A. Mark Ibekwe. *PLoS ONE* 9 (5): e97435. https://doi.org/10.1371/journal.pone.0097435.

Graves, Christopher J., Elizabeth J. Makrides, Victor T. Schmidt, Anne E. Giblin, Zoe G. Cardon, and David M. Rand. 2016. "Functional Responses of Salt Marsh Microbial Communities to Long-Term Nutrient Enrichment." Edited by G. Voordouw. *Applied and Environmental Microbiology* 82 (9): 2862–71. https://doi.org/10.1128/AEM.03990-15.

Green, Hyatt C., Orin C. Shanks, Mano Sivaganesan, Richard A. Haugland, and Katharine G. Field. 2011. "Differential Decay of Human Faecal Bacteroides in Marine and Freshwater." *Environmental Microbiology* 13 (12): 3235–49.

Harwood, V. J., C. Staley, B. D. Badgley, K. Borges, and A. Korajkic. 2014. "Microbial Source Tracking Markers for Detection of Fecal Contamination in Environmental Waters: Relationships between Pathogens and Human Health Outcomes." *FEMS Microbiol Rev* 38 (1): 1–40. https://doi.org/10.1111/1574-6976.12031.

Hausser, Jean, and Korbinian Strimmer. n.d. "Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks," 16.

Heim, Sabina, Maria Del Mar Lleo, Barbara Bonato, Carlos A. Guzman, and Pietro Canepari. 2002. “The Viable but Nonculturable State and Starvation Are Different Stress Responses of Enterococcus Faecalis, as Determined by Proteome Analysis.” *Journal of Bacteriology* 184 (23): 6739–45.

Huber, David H., Ifeoma R. Ugwuanyi, Sridhar A. Malkaram, Natalia A. Montenegro-Garcia, Vadesse Lhilhi Noundou, and Jesus E. Chavarria-Palma. 2018. “Metagenome Sequences of Sediment from a Recovering Industrialized Appalachian River in West Virginia.” *Genome Announcements* 6 (18). https://doi.org/10.1128/genomeA.00350-18.

Huttenhower, Curtis, The Human Microbiome Project Consortium, Dirk Gevers, Rob Knight, Sahar Abubucker, Jonathan H. Badger, Asif T. Chinwalla, et al. 2012. “Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature* 486 (7402): 207–14. https://doi.org/10.1038/nature11234.

Ishii, Satoshi, and Michael J. Sadowsky. 2008. “Escherichia Coli in the Environment: Implications for Water Quality and Human Health.” *Microbes and Environments* 23 (2): 101–8. https://doi.org/10.1264/jsme2.23.101.

Ishii, Satoshi, Tao Yan, Hung Vu, Dennis L. Hansen, Randall E. Hicks, and Michael J. Sadowsky. 2010. “Factors Controlling Long-Term Survival and Growth of Naturalized Escherichia Coli Populations in Temperate Field Soils.” *Microbes and Environments* 25 (1): 8–14. https://doi.org/10.1264/jsme2.ME09172.

Jang, J., H.-G. Hur, M. J. Sadowsky, M. N. Byappanahalli, T. Yan, and S. Ishii. 2017. “Environmental Escherichia Coli: Ecology and Public Health Implications-a Review.” *Journal of Applied Microbiology* 123 (3): 570–81. https://doi.org/10.1111/jam.13468.

Jang, Jeonghwan, Yae-Seul Suh, Doris Y. W. Di, Tatsuya Unno, Michael J. Sadowsky, and Hor-Gil Hur. 2013. “Pathogenic Escherichia Coli Strains Producing Extended-Spectrum β-Lactamases in the Yeongsan River Basin of South Korea.” *Environmental Science & Technology* 47 (2): 1128–36. https://doi.org/10.1021/es303577u.

Jay, Michele T., Michael Cooley, Diana Carychao, Gerald W. Wiscomb, Richard A. Sweitzer, Leta Crawford-Miksza, Jeff A. Farrar, et al. 2007. “Escherichia Coli O157:H7 in Feral Swine near Spinach Fields and Cattle, Central California Coast.” *Emerging Infectious Diseases* 13 (12): 1908–11. https://doi.org/10.3201/eid1312.070763.

Jechalke, Sven, Christoph Kopmann, Ingrid Rosendahl, Joost Groeneweg, Viola Weichelt, Ellen Krögerrecklenfort, Nikola Brandes, et al. 2013. “Increased Abundance and Transferability of Resistance Genes after Field Application of Manure from Sulfadiazine-Treated Pigs.” *Appl. Environ. Microbiol.* 79 (5): 1704–11. https://doi.org/10.1128/AEM.03172-12.

Johnston, Eric R., Minjae Kim, Janet K. Hatt, Jana R. Phillips, Qiuming Yao, Yang Song, Terry C. Hazen, Melanie A. Mayes, and Konstantinos T. Konstantinidis. 2019. "Phosphate Addition Increases Tropical Forest Soil Respiration Primarily by Deconstraining Microbial Population Growth." *Soil Biology and Biochemistry* 130 (March): 43–54. https://doi.org/10.1016/j.soilbio.2018.11.026.

Johnston, Eric R., Luis M. Rodriguez-R, Chengwei Luo, Mengting M. Yuan, Liyou Wu, Zhili He, Edward A. G. Schuur, et al. 2016. "Metagenomics Reveals Pervasive Bacterial Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem." *Frontiers in Microbiology* 7 (April). https://doi.org/10.3389/fmicb.2016.00579.

Karkman, Antti, Katariina Pärnänen, and D.G. Joakim Larsson. 2018. "Fecal Pollution Explains Antibiotic Resistance Gene Abundances in Anthropogenically Impacted Environments." https://doi.org/10.1101/341487.

Keegan, Kevin P., Elizabeth M. Glass, and Folker Meyer. 2016. "MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function." *Methods in Molecular Biology (Clifton, N.J.)* 1399: 207–33. https://doi.org/10.1007/978-1-4939-3369-3_13.

Landers, Timothy F., Bevin Cohen, Thomas E. Wittum, and Elaine L. Larson. 2012. "A Review of Antibiotic Use in Food Animals: Perspective, Policy, and Potential." *Public Health Reports (1974-)* 127 (1): 4–22.

Legendre, Pierre, and Marti J. Anderson. 1999. "DISTANCE-BASED REDUNDANCY ANALYSIS: TESTING MULTISPECIES RESPONSES IN MULTIFACTORIAL ECOLOGICAL EXPERIMENTS." *Ecological Monographs* 69 (1): 1–24. https://doi.org/10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2.

Liang, Yuying, Xin Jin, Yuan Huang, and Shuiping Chen. 2018. "Development and Application of a Real-Time Polymerase Chain Reaction Assay for Detection of a Novel Gut Bacteriophage (CrAssphage)." *Journal of Medical Virology* 90 (3): 464–68. https://doi.org/10.1002/jmv.24974.

Luo, Chengwei, Luis M. Rodriguez-R, and Konstantinos T. Konstantinidis. 2014. "MyTaxa: An Advanced Taxonomic Classifier for Genomic and Metagenomic Sequences." *Nucleic Acids Research* 42 (8): e73. https://doi.org/10.1093/nar/gku169.

Luo, Chengwei, Seth T. Walk, David M. Gordon, Michael Feldgarden, James M. Tiedje, and Konstantinos T. Konstantinidis. 2011. "Genome Sequencing of Environmental Escherichia Coli Expands Understanding of the Ecology and Speciation of the Model Bacterial Species." *Proceedings of the National Academy of Sciences* 108 (17): 7200–7205. https://doi.org/10.1073/pnas.1015622108.

Maal-Bared, Rasha, Karen H. Bartlett, William R. Bowie, and Eric R. Hall. 2013. "Phenotypic Antibiotic Resistance of Escherichia Coli and E. Coli O157 Isolated

from Water, Sediment and Biofilms in an Agricultural Watershed in British Columbia." *The Science of the Total Environment* 443 (January): 315–23. https://doi.org/10.1016/j.scitotenv.2012.10.106.

Mantha, Sirisha, Angela Anderson, Saraswati Poudel Acharya, Valerie J. Harwood, and Jennifer Weidhaas. 2017. "Transport and Attenuation of Salmonella Enterica, Fecal Indicator Bacteria and a Poultry Litter Marker Gene Are Correlated in Soil Columns." *The Science of the Total Environment* 598 (November): 204–12. https://doi.org/10.1016/j.scitotenv.2017.04.020.

Mar Lleò, Maria del, Maria Carla Tafi, and Pietro Canepari. 1998. "Nonculturable Enterococcus Faecalis Cells Are Metabolically Active and Capable of Resuming Active Growth." *Systematic and Applied Microbiology* 21 (3): 333–39.

Martín, M. F., and P. Liras. 1989. "Organization and Expression of Genes Involved in the Biosynthesis of Antibiotics and Other Secondary Metabolites." *Annual Review of Microbiology* 43: 173–206. https://doi.org/10.1146/annurev.mi.43.100189.001133.

McArthur, Andrew G., Nicholas Waglechner, Fazmin Nizam, Austin Yan, Marisa A. Azad, Alison J. Baylay, Kirandeep Bhullar, et al. 2013. "The Comprehensive Antibiotic Resistance Database." *Antimicrobial Agents and Chemotherapy* 57 (7): 3348–57. https://doi.org/10.1128/AAC.00419-13.

MetaHIT Consortium, Junhua Li, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, et al. 2014. "An Integrated Catalog of Reference Genes in the Human Gut Microbiome." *Nature Biotechnology* 32 (8): 834–41. https://doi.org/10.1038/nbt.2942.

Meziti, Alexandra, Despina Tsementzi, Konstantinos Ar. Kormas, Hera Karayanni, and Konstantinos T. Konstantinidis. 2016. "Anthropogenic Effects on Bacterial Diversity and Function along a River-to-Estuary Gradient in Northwest Greece Revealed by Metagenomics: Diversity Patterns along a River-to-Estuary Gradient." *Environmental Microbiology* 18 (12): 4640–52. https://doi.org/10.1111/1462-2920.13303.

Morlon, Hélène, Timothy K. O'Connor, Jessica A. Bryant, Louise K. Charkoudian, Kathryn M. Docherty, Evan Jones, Steven W. Kembel, Jessica L. Green, and Brendan J. M. Bohannan. 2015. "The Biogeography of Putative Microbial Antibiotic Production." Edited by Hauke Smidt. *PLOS ONE* 10 (6): e0130659. https://doi.org/10.1371/journal.pone.0130659.

Nayfach, Stephen, and Katherine S. Pollard. 2015. "Average Genome Size Estimation Improves Comparative Metagenomics and Sheds Light on the Functional Ecology of the Human Microbiome." *Genome Biology* 16 (1): 51. https://doi.org/10.1186/s13059-015-0611-7.

Negi, Vivek, and Rup Lal. 2017. “Metagenomic Analysis of a Complex Community Present in Pond Sediment.” *Journal of Genomics* 5 (March): 36–47. https://doi.org/10.7150/jgen.16685.

Newton, R. J., S. L. McLellan, D. K. Dila, J. H. Vineis, H. G. Morrison, A. M. Eren, and M. L. Sogin. 2015. “Sewage Reflects the Microbiomes of Human Populations.” *MBio* 6 (2): e02574. https://doi.org/10.1128/mBio.02574-14.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. “Mash: Fast Genome and Metagenome Distance Estimation Using MinHash.” *Genome Biology* 17 (1): 132. https://doi.org/10.1186/s13059-016-0997-x.

Orellana, Luis H., Joanne C. Chee-Sanford, Robert A. Sanford, Frank E. Löffler, and Konstantinos T. Konstantinidis. 2018. “Year-Round Shotgun Metagenomes Reveal Stable Microbial Communities in Agricultural Soils and Novel Ammonia Oxidizers Responding to Fertilization.” *Applied and Environmental Microbiology* 84 (2). https://doi.org/10.1128/AEM.01646-17.

Orellana, Luis H., Luis M. Rodriguez-R, and Konstantinos T. Konstantinidis. 2017. “ROCker: Accurate Detection and Quantification of Target Genes in Short-Read Metagenomic Data Sets by Modeling Sliding-Window Bitscores.” *Nucleic Acids Research* 45 (3): e14–e14. https://doi.org/10.1093/nar/gkw900.

Overbeek, Ross, Tadhg Begley, Ralph M. Butler, Jomuna V. Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, et al. 2005. “The Subsystems Approach to Genome Annotation and Its Use in the Project to Annotate 1000 Genomes.” *Nucleic Acids Research* 33 (17): 5691–5702. https://doi.org/10.1093/nar/gki866.

Peng, Yu, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. 2012. “IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth.” *Bioinformatics (Oxford, England)* 28 (11): 1420–28. https://doi.org/10.1093/bioinformatics/bts174.

Petit, Fabienne, Olivier Clermont, Sabine Delannoy, Pierre Servais, Michèle Gourmelon, Patrick Fach, Kenny Oberlé, Matthieu Fournier, Erick Denamur, and Thierry Berthe. 2017. “Change in the Structure of Escherichia Coli Population and the Pattern of Virulence Genes along a Rural Aquatic Continuum.” *Frontiers in Microbiology* 8: 609. https://doi.org/10.3389/fmicb.2017.00609.

Probert, William S., Glen M. Miller, and Katya E. Ledin. 2017. “Contaminated Stream Water as Source for Escherichia Coli O157 Illness in Children.” *Emerging Infectious Diseases* 23 (7): 1216–18. https://doi.org/10.3201/eid2307.170226.

Reischer, Georg H., James E. Ebdon, Johanna M. Bauer, Nathalie Schuster, Warish Ahmed, Johan Åström, Anicet R. Blanch, et al. 2013. “Performance Characteristics of QPCR Assays Targeting Human- and Ruminant-Associated Bacteroidetes for

Microbial Source Tracking across Sixteen Countries on Six Continents." *Environmental Science & Technology* 47 (15): 8548–56. https://doi.org/10.1021/es304367t.

Rho, Mina, Haixu Tang, and Yuzhen Ye. 2010. "FragGeneScan: Predicting Genes in Short and Error-Prone Reads." *Nucleic Acids Research* 38 (20): e191. https://doi.org/10.1093/nar/gkq747.

Roberts, Marilyn C. 2005. "Update on Acquired Tetracycline Resistance Genes." *FEMS Microbiology Letters* 245 (2): 195–203. https://doi.org/10.1016/j.femsle.2005.02.034.

Rodriguez-R, Luis M., and Konstantinos T. Konstantinidis. 2014. "Nonpareil: A Redundancy-Based Approach to Assess the Level of Coverage in Metagenomic Datasets." *Bioinformatics* 30 (5): 629–35. https://doi.org/10.1093/bioinformatics/btt584.

Rodriguez-R, Luis M, Will A Overholt, Christopher Hagan, Markus Huettel, Joel E Kostka, and Konstantinos T Konstantinidis. 2015. "Microbial Community Successional Patterns in Beach Sands Impacted by the Deepwater Horizon Oil Spill." *The ISME Journal* 9 (9): 1928–40. https://doi.org/10.1038/ismej.2015.5.

Salyers, A A, N B Shoemaker, A M Stevens, and L Y Li. 1995. "Conjugative Transposons: An Unusual and Diverse Set of Integrated Gene Transfer Elements." *Microbiological Reviews* 59 (4): 579–90.

Schmitz, Bradley W., Masaaki Kitajima, Maria E. Campillo, Charles P. Gerba, and Ian L. Pepper. 2016. "Virus Reduction during Advanced Bardenpho and Conventional Wastewater Treatment Processes." *Environmental Science & Technology* 50 (17): 9524–32. https://doi.org/10.1021/acs.est.6b01384.

Sinclair, R.G., E.L. Jones, and C.P. Gerba. 2009. "Viruses in Recreational Water-Borne Disease Outbreaks: A Review." *Journal of Applied Microbiology* 107 (6): 1769–80. https://doi.org/10.1111/j.1365-2672.2009.04367.x.

Soller, Jeffrey A., Mary E. Schoen, Timothy Bartrand, John E. Ravenscroft, and Nicholas J. Ashbolt. 2010. "Estimated Human Health Risks from Exposure to Recreational Waters Impacted by Human and Non-Human Sources of Faecal Contamination." *Water Research* 44 (16): 4674–91. https://doi.org/10.1016/j.watres.2010.06.049.

Stachler, Elyse, Benay Akyon, Nathalia Aquino de Carvalho, Christian Ference, and Kyle Bibby. 2018. "Correlation of CrAssphage QPCR Markers with Culturable and Molecular Indicators of Human Fecal Pollution in an Impacted Urban Watershed." *Environmental Science & Technology* 52 (13): 7505–12. https://doi.org/10.1021/acs.est.8b00638.

Stachler, Elyse, Catherine Kelty, Mano Sivaganesan, Xiang Li, Kyle Bibby, and Orin C. Shanks. 2017. "Quantitative CrAssphage PCR Assays for Human Fecal Pollution

Measurement." *Environmental Science & Technology* 51 (16): 9146–54. https://doi.org/10.1021/acs.est.7b02703.

Su, Xiaoquan, Weihua Pan, Baoxing Song, Jian Xu, and Kang Ning. 2014. "Parallel-META 2.0: Enhanced Metagenomic Data Analysis with Functional Annotation, High Performance Computing and Advanced Visualization." *PLOS ONE* 9 (3): e89323. https://doi.org/10.1371/journal.pone.0089323.

UniProt Consortium, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, Mark Bingley, et al. 2017. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 45 (D1): D158–69. https://doi.org/10.1093/nar/gkw1099.

Unno, Tatsuya, Christopher Staley, Clairessa M. Brown, Dukki Han, Michael J. Sadowsky, and Hor-Gil Hur. 2018. "Fecal Pollution: New Trends and Challenges in Microbial Source Tracking Using next-Generation Sequencing." *Environmental Microbiology* 0 (0). https://doi.org/10.1111/1462-2920.14281.

USEPA. 2009. "Review Of Published Studies To Characterize Relative Risks From Different Sources Of Fecal Contamination In Recreational Water." EPA822-R-09–001.

USEPA. 2012. "2012 Recreational Water Quality Criteria." EPA820-F-12–061. Washington, DC: Office of Water Regulations and Standards.

US-FDA. 2015. "Antimicrobials Sold or Distributed for Use in Food-Producing Animals." Food and Drug Administration: Department of Health and Human Services.

Van Rossum, Thea, Michael A. Peabody, Miguel I. Uyaguari-Diaz, Kirby I. Cronin, Michael Chan, Jared R. Slobodan, Matthew J. Nesbitt, et al. 2015. "Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality." *Frontiers in Microbiology* 6. https://doi.org/10.3389/fmicb.2015.01405.

Vital, Marius, Benli Chai, Bjørn Østman, James Cole, Konstantinos T. Konstantinidis, and James M. Tiedje. 2015. "Gene Expression Analysis of E. Coli Strains Provides Insights into the Role of Gene Regulation in Diversification." *The ISME Journal* 9 (5): 1130–40. https://doi.org/10.1038/ismej.2014.204.

Wade, Timothy J., Nitika Pai, J. Eisenberg, and J. M. Colford. 2003. "Do US EPA Water Quality Guidelines for Recreational Waters Prevent Gastrointestinal Illness? A Systematic Review and Metaanalysis." *Environ. Health Perspect* 111 (8).

Walsh, Timothy R, Janis Weeks, David M Livermore, and Mark A Toleman. 2011. "Dissemination of NDM-1 Positive Bacteria in the New Delhi Environment and Its Implications for Human Health: An Environmental Point Prevalence Study." *The Lancet Infectious Diseases* 11 (5): 355–62. https://doi.org/10.1016/S1473-3099(11)70059-7.

Weidhaas, Jennifer, Sirisha Mantha, Elliott Hair, Bina Nayak, and Valerie J. Harwood. 2015. “Evidence for Extraintestinal Growth of Bacteroidales Originating from Poultry Litter.” *Applied and Environmental Microbiology* 81 (1): 196–202.

Weigand, M. R., N. J. Ashbolt, K. T. Konstantinidis, and J. W. Santo Domingo. 2014. “Genome Sequencing Reveals the Environmental Origin of Enterococci and Potential Biomarkers for Water Quality Monitoring.” *Environ Sci Technol* 48 (7): 3707–14. https://doi.org/10.1021/es4054835.

WHO. 2003. “Guidelines for Safe Recreational Water Environments.” Geneva: World Health Organization.

WHO. 2004. *Waterborne Zoonoses: Identification, Causes, and Control*. Emerging Issues in Water and Infectious Disease Series. London: IWA Publ.

WHO. 2014. “Antimicrobial Resistance: An Emerging Water, Sanitation and Hygiene Issue.” http://apps.who.int/iris/bitstream/handle/10665/204948/WHO_FWC_WSH_14.7_eng.pdf?sequence=1.

Xu, Meiying, Qin Zhang, Chunyu Xia, Yuming Zhong, Guoping Sun, Jun Guo, Tong Yuan, Jizhong Zhou, and Zhili He. 2014. “Elevated Nitrate Enriches Microbial Functional Genes for Potential Bioremediation of Complexly Contaminated Sediments.” *The ISME Journal* 8 (9): 1932–44. https://doi.org/10.1038/ismej.2014.42.

Yang, Jing, Chao Wang, Chang Shu, Li Liu, Jianing Geng, Songnian Hu, and Jie Feng. 2013. “Marine Sediment Bacteria Harbor Antibiotic Resistance Genes Highly Similar to Those Found in Human Pathogens.” *Microbial Ecology* 65 (4): 975–81. https://doi.org/10.1007/s00248-013-0187-2.

Yilmaz, Pelin, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner. 2014. “The SILVA and ‘All-Species Living Tree Project (LTP)’ Taxonomic Frameworks.” *Nucleic Acids Research* 42 (D1): D643–48. https://doi.org/10.1093/nar/gkt1209.

Zhang, Qian, Jessica J. Eichmiller, Christopher Staley, Michael J. Sadowsky, and Satoshi Ishii. 2016. “Correlations between Pathogen Concentration and Fecal Indicator Marker Genes in Beach Environments.” *Science of The Total Environment* 573 (December): 826–30. https://doi.org/10.1016/j.scitotenv.2016.08.122.

Zhang, Si-Yu, Despina Tsementzi, Janet K. Hatt, Aaron Bivins, Nikunj Khelurkar, Joe Brown, Sachchida Nand Tripathi, and Konstantinos T. Konstantinidis. 2019. “Intensive Allochthonous Inputs along the Ganges River and Their Effect on Microbial Community Composition and Dynamics.” *Environmental Microbiology* 21 (1): 182–96. https://doi.org/10.1111/1462-2920.14439.

Zhu, Yong-Guan, Timothy A. Johnson, Jian-Qiang Su, Min Qiao, Guang-Xia Guo, Robert D. Stedtfeld, Syed A. Hashsham, and James M. Tiedje. 2013. “Diverse and Abundant Antibiotic Resistance Genes in Chinese Swine Farms.” *Proceedings of the National Academy of Sciences* 110 (9): 3435–40. https://doi.org/10.1073/pnas.1222743110.